

Quantitative Methods for Health Research

Quantitative Methods for Health Research

A Practical Interactive Guide to Epidemiology and Statistics

Second Edition

Nigel Bruce

*The University of Liverpool
UK*

Daniel Pope

*The University of Liverpool
UK*

Debbi Stanistreet

*The University of Liverpool
UK*

WILEY

This edition first published 2018
© 2018 John Wiley & Sons Ltd

Edition History

John Wiley & Sons (1e, 2009)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Nigel Bruce, Daniel Pope and Debbi Stanistreet to be identified as the authors of this work has been asserted in accordance with law.

Registered Office(s)

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting scientific method, diagnosis, or treatment by physicians for any particular patient. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Bruce, Nigel, 1955– author. | Pope, Daniel, 1969– author. | Stanistreet, Debbi, 1963–
Title: Quantitative methods for health research : a practical interactive guide to epidemiology and statistics / by Nigel Bruce,
Daniel Pope, Debbi Stanistreet.
Description: Second edition. | Hoboken, NJ : Wiley, 2018. | Includes index. |
Identifiers: LCCN 2017030435 (print) | LCCN 2017031313 (ebook) | ISBN 9781118665268 (pdf) | ISBN 9781118665404 (epub) |
ISBN 9781118665411 (pbk.)
Subjects: | MESH: Epidemiologic Methods | Biometry—methods | Biomedical Research author.—methods
Classification: LCC RA652.4 (ebook) | LCC RA652.4 (print) | NLM WA 950 | DDC 614.4072—dc23
LC record available at <https://lccn.loc.gov/2017030435>

Cover Design: Wiley

Cover Images: (Crowd) © samxmeg/Gettyimages; (Map) © Arthimedes/Shutterstock

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface	xv
About the Companion Website	xxi
1 Philosophy of Science and Introduction to Epidemiology	1
Introduction and Learning Objectives	1
1.1 Approaches to Scientific Research	2
1.1.1 History and Nature of Scientific Research	2
1.1.2 What is Epidemiology?	6
1.1.3 What are Statistics?	7
1.1.4 Approach to Learning	8
1.2 Formulating a Research Question	8
1.2.1 Importance of a Well-Defined Research Question	8
1.2.2 Development of Research Ideas	10
1.3 Rates: Incidence and Prevalence	11
1.3.1 Why Do We Need Rates?	11
1.3.2 Measures of Disease Frequency	12
1.3.3 Prevalence Rate	12
1.3.4 Incidence Rate	12
1.3.5 Relationship Between Incidence, Duration, and Prevalence	15
1.4 Concepts of Prevention	16
1.4.1 Introduction	16
1.4.2 Primary, Secondary, and Tertiary Prevention	17
1.5 Answers to Self-Assessment Exercises	18
2 Routine Data Sources and Descriptive Epidemiology	25
Introduction and Learning Objectives	25
2.1 Routine Collection of Health Information	26
2.1.1 Deaths (Mortality)	26
2.1.2 Compiling Mortality Statistics: The Example of England and Wales	28
2.1.3 Suicide Among Men	29
2.1.4 Suicide Among Young Women	31
2.1.5 Variations in Deaths of Very Young Children	31
2.2 Descriptive Epidemiology	33
2.2.1 What is Descriptive Epidemiology?	33
2.2.2 International Variations in Rates of Lung Cancer	33
2.2.3 Illness (Morbidity)	34
2.2.4 Sources of Information on Morbidity	35

2.2.5	Notification of Infectious Disease	35
2.2.6	Illness Seen in General Practice	38
2.3	Information on the Environment	39
2.3.1	Air Pollution and Health	39
2.3.2	Routinely Available Data on Air Pollution	39
2.4	Displaying, Describing, and Presenting Data	41
2.4.1	Displaying the Data	41
2.4.2	Calculating the Frequency Distribution	42
2.4.3	Describing the Frequency Distribution	44
2.4.4	The Relative Frequency Distribution	57
2.4.5	Scatterplots, Linear Relationships and Correlation	60
2.5	Routinely Available Health Data	69
2.5.1	Introduction	69
2.5.2	Classification of Routine Health Information Sources	69
2.5.3	Demographic Data	71
2.5.4	Health Event Data	73
2.5.5	Population-Based Health Information	78
2.5.6	Deprivation Indices	79
2.5.7	Routine Data Sources for Countries Other Than the UK	80
2.6	Descriptive Epidemiology in Action	80
2.6.1	The London Smogs of the 1950s	80
2.6.2	Ecological Studies	82
2.7	Overview of Epidemiological Study Designs	84
2.8	Answers to Self-Assessment Exercises	86
3	Standardisation	101
	Introduction and Learning Objectives	101
3.1	Health Inequalities in Merseyside	101
3.1.1	Socio-Economic Conditions and Health	101
3.1.2	Comparison of Crude Death Rates	102
3.1.3	Usefulness of a Summary Measure	104
3.2	Indirect Standardisation: Calculation of the Standardised Mortality Ratio (SMR)	105
3.2.1	Mortality in Liverpool	105
3.2.2	Interpretation of the SMR	107
3.2.3	Dealing With Random Variation: The 95 per cent Confidence Interval	107
3.2.4	Increasing Precision of the SMR Estimate	108
3.2.5	Mortality in Sefton	108
3.2.6	Comparison of SMRs	110
3.2.7	Indirectly Standardised Mortality Rates	110
3.3	Direct Standardisation	110
3.3.1	Introduction	110
3.3.2	An Example: Changes in Deaths From Stroke Over Time	111
3.3.3	Using the European Standard Population	112
3.3.4	Direct or Indirect: Which Method is Best?	113
3.4	Standardisation for Factors Other Than Age	114
3.5	Answers to Self-Assessment Exercises	115
4	Surveys	123
	Introduction and Learning Objectives	123
	Resource Papers	124

4.1	Purpose and Context	124
4.1.1	Defining the Research Question	124
4.1.2	Political Context of Research	126
4.2	Sampling Methods	127
4.2.1	Introduction	127
4.2.2	Sampling	127
4.2.3	Probability	129
4.2.4	Simple Random Sampling	130
4.2.5	Stratified Sampling	131
4.2.6	Cluster Random Sampling	132
4.2.7	Multistage Random Sampling	133
4.2.8	Systematic Sampling	133
4.2.9	Convenience Sampling	133
4.2.10	Sampling People Who are Difficult to Contact	133
4.2.11	Quota Sampling	134
4.2.12	Sampling in Natsal-3	135
4.3	The Sampling Frame	137
4.3.1	Why Do We Need a Sampling Frame?	137
4.3.2	Losses in Sampling	137
4.4	Sampling Error, Confidence Intervals, and Sample Size	139
4.4.1	Sampling Distributions and the Standard Error	139
4.4.2	The Standard Error	140
4.4.3	Key Properties of the Normal Distribution	145
4.4.4	Confidence Interval (CI) for the Sample Mean	146
4.4.5	Estimating Sample Size	149
4.4.6	Sample Size for Estimating a Population Mean	149
4.4.7	Standard Error and 95 per cent CI for a Population Proportion	150
4.4.8	Sample Size to Estimate a Population Proportion	151
4.5	Response	153
4.5.1	Determining the Response Rate	153
4.5.2	Assessing Whether the Sample is Representative	154
4.5.3	Maximising the Response Rate	154
4.6	Measurement	157
4.6.1	Introduction: The Importance of Good Measurement	157
4.6.2	Interview or Self-Completed Questionnaire?	157
4.6.3	Principles of Good Questionnaire Design	158
4.6.4	Development of a Questionnaire	161
4.6.5	Checking How Well the Interviews and Questionnaires Have Worked	161
4.6.6	Assessing Measurement Quality	165
4.6.7	Overview of Sources of Error	169
4.7	Data Types and Presentation	171
4.7.1	Introduction	171
4.7.2	Types of Data	172
4.7.3	Displaying and Summarising the Data	173
4.8	Answers to Self-Assessment Exercises	176
5	Cohort Studies	185
	Introduction and Learning Objectives	185
	Resource Papers	186

5.1	Why Do a Cohort Study?	186
5.1.1	Objectives of the Study	186
5.1.2	Study Structure	188
5.2	Obtaining the Sample	188
5.2.1	Introduction	188
5.2.2	Sample Size	190
5.3	Measurement	190
5.3.1	Importance of Good Measurement	190
5.3.2	Identifying and Avoiding Measurement Error	190
5.3.3	The Measurement of Blood Pressure	191
5.3.4	Case Definition	192
5.4	Follow-Up	193
5.4.1	Nature of the Task	193
5.4.2	Deaths (Mortality)	193
5.4.3	Non-Fatal Cases (Morbidity)	194
5.4.4	Challenges Faced with Follow-Up of a Cohort in a Different Setting	194
5.4.5	Assessment of Changes During Follow-Up Period	196
5.5	Basic Presentation and Analysis of Results	198
5.5.1	Initial Presentation of Findings	198
5.5.2	Relative Risk	199
5.5.3	Hypothesis Test for Categorical Data: The Chi-Squared Test	201
5.5.4	Hypothesis Tests for Continuous Data: The z -Test and the t -Test	209
5.6	How Large Should a Cohort Study Be?	214
5.6.1	Perils of Inadequate Sample Size	214
5.6.2	Sample Size for a Cohort Study	215
5.6.3	Example of Output from Sample Size Calculation	216
5.7	Assessing Whether an Association is Causal	218
5.7.1	The Hill Viewpoints	218
5.7.2	Confounding: What Is It and How Can It Be Addressed?	220
5.7.3	Does Smoking Cause Heart Disease?	222
5.7.4	Confounding in the Physical Activity and Cancer Study	222
5.7.5	Methods for Dealing with Confounding	224
5.8	Simple Linear Regression	224
5.8.1	Approaches to Describing Associations	224
5.8.2	Finding the Best Fit for a Straight Line	226
5.8.3	Interpreting the Regression Line	227
5.8.4	Using the Regression Line	228
5.8.5	Hypothesis Test of the Association Between the Explanatory and Outcome Variables	228
5.8.6	How Good is the Regression Model?	229
5.8.7	Interpreting SPSS Output for Simple Linear Regression Analysis	231
5.8.8	First Table: Variables Entered/Removed	232
5.9	Introduction to Multiple Linear Regression	235
5.9.1	Principles of Multiple Regression	235
5.9.2	Using Multivariable Linear Regression to Study Independent Associations	235
5.9.3	Investigation of the Effect of Work Stress on Bodyweight	235
5.9.4	Multiple Regression in the Cancer Study	239
5.9.5	Overview of Regression Methods for Different Types of Outcome	240
5.10	Answers to Self-Assessment Exercises	242

6	Case–Control Studies	251
	Introduction and Learning Objectives	251
	Resource Papers	252
6.1	Why do a Case–Control Study?	253
6.1.1	Study Objectives	253
6.1.2	Study Structure	254
6.1.3	Approach to Analysis	255
6.1.4	Retrospective Data Collection	257
6.1.5	Applications of the Case–Control Design	258
6.2	Key Elements of Study Design	259
6.2.1	Selecting the Cases	259
6.2.2	The Controls	260
6.2.3	Exposure Assessment	262
6.2.4	Bias in Exposure Assessment	263
6.3	Basic Unmatched and Matched Analysis	265
6.3.1	The Odds Ratio (OR)	265
6.3.2	Calculation of the OR–Simple Matched Analysis	269
6.3.3	Hypothesis Tests for Case–Control Studies	271
6.4	Sample Size for a Case–Control Study	273
6.4.1	Introduction	273
6.4.2	What Information is Required?	273
6.4.3	An Example of Sample Size Calculation Using OpenEpi	274
6.5	Confounding and Logistic Regression	276
6.5.1	Introduction	276
6.5.2	Stratification	277
6.5.3	Logistic Regression	278
6.5.4	Example: Multivariable Logistic Regression	281
6.5.5	Matched Studies – Conditional Logistic Regression	287
6.5.6	Interpretation of Adjusted Results from the New Zealand Study	287
6.6	Answers to Self-Assessment Exercises	289
7	Intervention Studies	297
	Introduction and Learning Objectives	297
	Typology of Intervention Study Designs Described in This Chapter	297
	Terminology	298
	Resource Papers	299
7.1	Why Do an Intervention Study?	299
7.1.1	Study Objectives	299
7.1.2	Structure of a Randomised, Controlled Intervention Study	300
7.2	Key Elements of Intervention Study Design	303
7.2.1	Defining Who Should be Included and Excluded	303
7.2.2	Intervention and Control	304
7.2.3	Randomisation	306
7.2.4	Outcome Assessment	307
7.2.5	Blinding	308
7.2.6	Ethical Issues for Intervention Studies	308
7.3	The Analysis of Intervention Studies	309
7.3.1	Review of Variables at Baseline	310
7.3.2	Loss to Follow-Up	311
7.3.3	Compliance with the Treatment Allocation	311

7.3.4	Analysis by Intention-to-Treat	312
7.3.5	Analysis per Protocol	313
7.3.6	What is the Effect of the Intervention?	313
7.3.7	Drawing Conclusions	315
7.3.8	Adjustment for Variables Known to Influence the Outcome	315
7.3.9	Paired Comparisons	315
7.3.10	The Crossover Trial	317
7.4	Testing More-Complex Interventions	318
7.4.1	Introduction	318
7.4.2	Randomised Trial of Individuals for a Complex Intervention	319
7.4.3	Factorial Design	322
7.4.4	Analysis and Interpretation	323
7.4.5	Departure from the Ideal Blinded RCT Design	327
7.4.6	The Cluster Randomised Trial	328
7.4.7	The Community (Cluster) Randomised Trial	330
7.4.8	Non-Randomised Intervention Designs	332
7.4.9	The Natural Experiment	333
7.5	Analysis of Intervention Studies Using a Cluster Design	334
7.5.1	Why Does the Use of Clusters Make a Difference?	334
7.5.2	Summarising Clustering Effects: The Intra-Class Correlation Coefficient	334
7.5.3	Multi-Level Modelling	335
7.5.4	Analysis of the Cluster RCT of Physical Activity	335
7.6	How Big Should the Intervention Study Be?	337
7.6.1	Introduction	337
7.6.2	Sample Size for a Trial with Categorical Data Outcomes	337
7.6.3	One-Sided and Two-Sided Tests	339
7.6.4	Sample Size for a Trial with Continuous Data Outcomes	339
7.6.5	Sample Size for an Intervention Study Using Cluster Design	340
7.6.6	Estimation of Sample Size is not a Precise Science	341
7.7	Intervention Study Registration, Management, and Reporting	341
7.7.1	Introduction	341
7.7.2	Registration	342
7.7.3	Trial Management	342
7.7.4	Reporting Standards (CONSORT)	343
7.8	Answers to Self-Assessment Exercises	344
8	Life Tables, Survival Analysis, and Cox Regression	355
	Introduction and Learning Objectives	355
	Resource Papers	356
8.1	Survival Analysis	356
8.1.1	Introduction	356
8.1.2	Why Do We Need Survival Analysis?	356
8.1.3	Censoring	357
8.1.4	Kaplan–Meier Survival Curves	359
8.1.5	Kaplan–Meier Survival Curves	361
8.1.6	The Log-Rank Test	362
8.1.7	Interpretation of the Kaplan–Meier Survival Curve	365
8.2	Cox Regression	371
8.2.1	Introduction	371

8.2.2	The Hazard Function	371
8.2.3	Assumption of Proportional Hazards	372
8.2.4	The Cox Regression Model	372
8.2.5	Checking the Assumption of Proportional Hazards	372
8.2.6	Interpreting the Cox Regression Model	373
8.2.7	Prediction	374
8.2.8	Application of Cox Regression	375
8.3	Current Life Tables	377
8.3.1	Introduction	377
8.3.2	Current Life Tables and Life Expectancy at Birth	377
8.3.3	Life Expectancy at Other Ages	379
8.3.4	Healthy or Disability-Free Life Expectancy	379
8.3.5	Abridged Life Tables	380
8.3.6	Summary	381
8.4	Answers to Self-Assessment Exercises	381
9	Systematic Reviews and Meta-Analysis	385
	Introduction and Learning Objectives	385
	Increasing Power by Combining Studies	386
	Resource Papers	387
9.1	The Why and How of Systematic Reviews	387
9.1.1	Why is it Important that Reviews be Systematic?	387
9.1.2	Method of Systematic Review – Overview and Developing a Protocol	388
9.1.3	Deciding on the Research Question and Objectives for the Review	389
9.1.4	Defining Criteria for Inclusion and Exclusion of Studies	390
9.1.5	Identifying Relevant Studies	391
9.1.6	Assessment of Methodological Quality	396
9.1.7	Extracting Data	399
9.1.8	Describing the Results	399
9.2	The Methodology of Meta-Analysis	402
9.2.1	Method of Meta-Analysis – Overview	402
9.2.2	Assessment of Publication Bias – the Funnel Plot	403
9.2.3	Heterogeneity	405
9.2.4	Calculating the Pooled Estimate	407
9.2.5	Presentation of Results: Forest Plot	408
9.2.6	Sensitivity Analysis	409
9.2.7	Statistical Software for the Conduct of Meta-Analysis	410
9.2.8	Another Example of the Value of Meta-Analysis – Identifying a Dangerous Treatment	411
9.3	Systematic Reviews and Meta-Analyses of Observational Studies	414
9.3.1	Introduction	414
9.3.2	Why Conduct a Systematic Review of Observational Studies?	414
9.3.3	Approach to Meta-Analysis of Observational Studies	415
9.3.4	Method of Systematic Review of Observational Studies	416
9.3.5	Method of Meta-Analysis of Observational Studies	416
9.4	Reporting and Publishing Systematic Reviews and Meta-Analyses	418
9.5	The Cochrane Collaboration	419
9.5.1	Introduction	419
9.5.2	Cochrane Collaboration Logo	422

9.5.3	Collaborative Review Groups	422
9.5.4	Cochrane Library	422
9.6	Answers to Self-Assessment Exercises	423
10	Prevention Strategies and Evaluation of Screening	429
	Introduction and Learning Objectives	429
	Resource Papers	430
10.1	Concepts of Risk	430
10.1.1	Relative and Attributable Risk	430
10.1.2	Calculation of AR	431
10.1.3	Attributable Fraction (AF) for a Dichotomous Exposure	432
10.1.4	Attributable Fraction for Continuous and Multiple Category Exposures	434
10.1.5	Years of Life Lost (YLL) and Years Lived with Disability (YLD)	434
10.1.6	Disability-Adjusted Life Years (DALYs)	436
10.1.7	Burden Attributable to Specific Risk Factors	438
10.2	Strategies of Prevention	440
10.2.1	The Distribution of Risk in Populations	440
10.2.2	High-Risk and Population Approaches to Prevention	443
10.2.3	Safety and the Population Strategy	446
10.2.4	The High-Risk and Population Strategies Revisited	447
10.2.5	Implications of Genomic Research for Disease Prevention	448
10.3	Evaluation of Screening Programmes	450
10.3.1	Purpose of Screening	451
10.3.2	Criteria for Programme Evaluation	451
10.3.3	Assessing Validity of a Screening Test	452
10.3.4	Methodological Issues in Studies of Screening Programme Effectiveness	460
10.3.5	Are the Wilson–Jungner Criteria Relevant Today?	461
10.4	Cohort and Period Effects	463
10.4.1	Analysis of Change in Risk Over Time	463
10.4.2	Example: Suicide Trends in UK Men and Women	464
10.5	Answers to Self-Assessment Exercises	468
11	Probability Distributions, Hypothesis Testing, and Bayesian Methods	477
	Introduction and Learning Objectives	477
	Resource Papers	478
11.1	Probability Distributions	478
11.1.1	Probability – A Brief Review	478
11.1.2	Introduction to Probability Distributions	479
11.1.3	Types of Probability Distribution	481
11.1.4	Probability Distributions: Implications for Statistical Methods	487
11.2	Data That Do Not Fit a Probability Distribution	488
11.2.1	Robustness of an Hypothesis Test	488
11.2.2	Transforming the Data	488
11.2.3	Principles of Non-Parametric Hypothesis Testing	492
11.3	Hypothesis Testing: Summary of Common Parametric and Non-Parametric Methods	493
11.3.1	Introduction	493
11.3.2	Review of Hypothesis Tests	494

11.3.3	Fundamentals of Hypothesis Testing	494
11.3.4	Summary: Stages of Hypothesis Testing	495
11.3.5	Comparing Two Independent Groups	496
11.3.6	Comparing Two Paired (or Matched) Groups	500
11.3.7	Testing for Association Between Two Groups	506
11.3.8	Comparing More Than Two Groups	508
11.3.9	Association Between Categorical Variables	513
11.4	Choosing an Appropriate Hypothesis Test	517
11.4.1	Introduction	517
11.4.2	Using a Guide Table for Selecting a Hypothesis Test	517
11.4.3	The Problem of Multiple Significance Testing	520
11.5	Bayesian Methods	520
11.5.1	Introduction: A Different Approach to Inference	520
11.5.2	Bayes' Theorem and Formula	521
11.5.3	Application and Relevance	522
11.6	Answers to Self-Assessment Exercises	525
	Bibliography	529
	Index	533

Preface

Introduction

Welcome to *Quantitative Methods for Health Research*, a study programme designed to introduce you to the knowledge and skills required to make sense of published health research, and to begin designing and carrying out studies of your own.

The book is based closely on materials developed and tested over more than 15 years with the campus-based and online Master of Public Health (MPH) programmes at the University of Liverpool, UK. A key theme of our approach to teaching and learning is to ensure a reasonable level of theoretical knowledge (as this helps to provide a solid basis to understanding), while placing at least as much emphasis on the application of theory to practice (to demonstrate what actually happens when theoretical ‘ideals’ come up against reality). For these reasons, the learning materials have been designed around a number of published research studies and information sources that address a variety of topics from around the world, including both developed and developing countries. The many aspects of study design and analysis illustrated by these studies provide examples which are used to help you understand the fundamental principles of good research, and to practise these techniques yourself.

The MPH programme on which this book is based consists of two postgraduate taught epidemiology and statistics modules, one **Introductory** and the other **Advanced**, each of which requires 150 hours of study (including assessments), and provides 15 postgraduate credits (1 unit). As students and tutors using the book may find it convenient to follow a similar module-based approach, the content of chapters has been organised to make this as simple as possible. The table summarising the content of each chapter on pages xvii to xix indicates which sections (together with page numbers) relate to the introductory programme, and which to the advanced programme.

The use of computer software for data analysis is a fundamental area of knowledge and skills for the application of epidemiological and statistical methods. A complementary study programme in data analysis using IBM SPSS software has been prepared; this relates closely to the structure and content of the book. Full details of this study programme, including the data sets used for data analysis exercises, are available on the companion website for this book www.wiley.com/go/bruce/quantitative-health-research.

The book also has a number of other features designed to enhance learning effectiveness, summarised in the following sections.

Learning Objectives

Specific, detailed learning objectives are provided at the start of each chapter. These set out the nature and level of knowledge, understanding, and skills required to achieve a good standard at the master's level, and can be used as one point of reference for assessing progress.

Resource Papers and Information Sources

All sections of published studies that are required, in order to follow the text and answer the self-assessment exercises, are reproduced as excerpts in the book. However, we strongly recommend that all resource papers be obtained and read fully, as indicated in the text. This will greatly enhance the understanding of how the methods and ideas discussed in the book are applied in practice, and how research papers are prepared. All papers are fully referenced and available through open-access, in journals that are easily available through higher education establishments.

Key Terms

In order to help identify which concepts and terms are most important, those regarded as core knowledge appear in ***bold italic*** font. These can be used as another form of self-assessment, as a good grasp of the material covered in this book will only have been achieved if all these key terms are familiar and understood.

Sample Size Calculations

In several chapters, sample size calculations are explained and used as a basis for self-assessment exercises. We use OpenEpi webtools for this purpose, which can be found at <http://www.openepi.com>.

SPSS Dataset Used for Illustrating Examples of Statistical Analysis

A reference dataset is used in the book to illustrate analytical output from regression analyses (Chapters 5 and 6) using SPSS. It is also used for other data manipulation and analysis exercises for workbooks located on the companion website for the book (www.wiley.com/go/bruce/quantitative-health-research). This reference dataset relates to how aspects of work in manual occupational settings are associated with the outcome of low back pain, and has the following features:

- The aim of the study was to see what features of the occupational environment were associated with **low back pain**.
- The dataset relates to information collected on 775 employees selected randomly from manual occupational settings in the North West of England.
- The dataset includes information on **demography** (age, sex, height, weight and social class), **physical working environment** (working postures, manual handling activities and repetitive upper limb movements – the duration of these activities was recorded for 60 minutes

of one shift), **psychosocial working environment** (psychological demands of work) and **psychological distress** (a score based on responses to a psychological questionnaire – a higher score indicates a higher level of psychological distress).

Self-Assessment Exercises

Each chapter includes self-assessment exercises, which are an integral part of the study programme. These have been designed to assess and consolidate the understanding of theoretical concepts and competency in practical techniques. The exercises have also been designed to be worked through as they are encountered, as many of the answers expand on issues that are introduced in the main text. The answers and discussion for these exercises are provided at the end of each chapter.

Mathematical Aspects of Statistics

It has been our experience that many students interested in health research, while motivated and very capable, nevertheless do find that the mathematical aspects of statistical methods, such as formulae and mathematical notation, are quite daunting. This is an area of study that does require some persistence, as it is valuable to gain at least a basic mathematical understanding of the most commonly used statistical concepts and methods. We recognise, however, that creating the expectation of much more in-depth knowledge for all readers would be very demanding, and arguably unnecessary. For the most part, therefore, this book avoids detailed mathematical explanation and formulae.

Readers with more affinity for and knowledge of mathematics may be interested to know more, and such understanding is very important for more advanced research work and data analysis. In order to meet these objectives, all basic concepts, simple mathematical formulae, etc., the understanding of which can be seen as core knowledge, are included in the main text. More detailed explanations, including some more complex formulae and examples, are included in statistical reference sections [RS], marked with a start and finish as indicated below.



RS – Reference Section on Statistical Methods

Text, formulae, and examples.

We hope that you will enjoy this study programme, and find that it meets your expectations and needs.

Organisation of Subject Matter by Chapter

The following table summarises the subject content for each chapter, indicating which sections are introductory and which are advanced.

Chapter content and level			
Chapter	Level	Pages	Topics covered
1. Philosophy of science and introduction to epidemiology	Introductory	1–24	<ul style="list-style-type: none"> Approaches to scientific research What is epidemiology? What is statistics? Formulating a research question Rates, incidence, and prevalence Concepts of prevention
2. Routine data sources and descriptive epidemiology	Introductory	25–100	<ul style="list-style-type: none"> Routine collection of health information Descriptive epidemiology Information on the environment Displaying, describing, and presenting data Association and correlation Summary of routinely available data relevant to health Descriptive epidemiology in action, ecological studies, and the ecological fallacy Overview of epidemiological study designs
3. Standardisation	Introductory	101–122	<ul style="list-style-type: none"> Rationale for standardisation Indirect standardisation Direct standardisation
4. Surveys	Introductory	123–184	<ul style="list-style-type: none"> Rationale for survey methods Sampling methods The sampling frame Sampling error, sample size, and confidence intervals Response rates Measurement, questionnaire design, and validity Data types and presentation: categorical and continuous
5. Cohort studies	Introductory	185–250	<ul style="list-style-type: none"> Rationale for cohort study methods Obtaining a sample Measurement and measurement error Follow-up for mortality and morbidity Basic analysis – relative risk, hypothesis testing (the <i>t</i>-test and the chi-squared test) Introduction to the problem of confounding
	Advanced		<ul style="list-style-type: none"> Sample size for cohort studies Simple linear regression Multiple linear regression: dealing with confounding factors
6. Case–control studies	Introductory	251–296	<ul style="list-style-type: none"> Rationale for case–control study methods Selecting cases and controls Matching – to match or not? The problem of bias Basic analysis – the odds ratio for unmatched and matched designs
	Advanced		<ul style="list-style-type: none"> Sample size for case control studies Matching with more than one control Multiple logistic regression

Chapter content and level			
Chapter	Level	Pages	Topics covered
7. Intervention studies	Introductory	297–354	<ul style="list-style-type: none"> • Rationale for experimental study methods • The randomised controlled trial (RCT) • Randomisation • Blinding, controls, and ethical considerations • Analysis of trial outcomes: analysis by intention-to-treat and per-protocol • Paired data and cross-over trials
	Advanced		<ul style="list-style-type: none"> • Adjustment when confounding factors are not balanced by randomisation • Sample size for experimental studies • Testing more complex interventions; cluster and non-randomised experimental designs • Factorial design • Multilevel analysis • Trial management and reporting
8. Life tables, survival analysis, and Cox regression	Advanced	355–384	<ul style="list-style-type: none"> • Nature of survival data • Kaplan–Meier survival curves • Cox proportional hazards regression • Introduction to life tables
9. Systematic reviews and meta-analysis	Advanced	385–428	<ul style="list-style-type: none"> • Purpose of systematic reviews • Method of systematic review • Method of meta-analysis • Special considerations in systematic reviews and meta-analysis of observational studies • The Cochrane Collaboration
10. Prevention strategies and evaluation of screening	Advanced	429–476	<ul style="list-style-type: none"> • Relative and attributable risk, population attributable risk, and attributable fraction • High-risk and population approaches to prevention • Measures and techniques used in the evaluation of screening programmes, including sensitivity, specificity, predictive value, and receiver operator characteristic (ROC) curves • Methodological issues and bias in studies of screening programme effectiveness • Cohort and period effects
11. Probability distributions, hypothesis testing, and Bayesian methods	Advanced	477–528	<ul style="list-style-type: none"> • Theoretical probability distributions • Steps in hypothesis testing • Transformation of data • Paired t-test • One-way analysis of variance • Non-parametric tests for paired data, two or more independent groups, and for more than two groups • Spearman's rank correlation • Fisher's exact test • Guide to choosing an appropriate test • Multiple significance testing • Introduction to Bayesian methods

Acknowledgements

The preparation of this book has involved the efforts of a number of people whose support we wish to acknowledge: Francine Watkins for preparing the first section of Chapter 1 'Approaches to Scientific Research'; Jo Reeve for preparing Section 5 of Chapter 2; James Higgerson for providing valuable advice for the Current Life Tables section in Chapter 8; Paul Blackburn for assistance with graphics and obtaining permission for the reproduction of the resource papers; Chris West for his advice on statistical methods, and data management and analysis; Nancy Cleave and Gill Lancaster for their invaluable contributions to early versions of the statistical components of the materials; our students and programme tutors who have provided much valued, constructive feedback that has helped to guide the development of the materials upon which this book is based; the staff at Wiley for their encouragement and support; and Chris and Ian at Ty Cam for providing a tranquil refuge in a beautiful part of Denbighshire.

About the Companion Website

Quantitative Methods for Health Research – A Practical Interactive Guide to Epidemiology and Statistics is accompanied by a companion website:

www.wiley.com/go/bruce/quantitative-health-research

The website includes:

- SPSS workbooks and datasets

1

Philosophy of Science and Introduction to Epidemiology

Introduction and Learning Objectives

In this chapter, we will begin by looking at different approaches to *scientific research*, how these have arisen, and the importance of recognising that there is no single, right way to carry out investigations in the health field. Rather, we will see that different perspectives can be complementary in providing a more complete understanding of any given issue. We will then go on to explore the *research task*, discuss what is meant by *epidemiology* and *statistics*, and look at how these two disciplines are introduced and developed in the book. The next section introduces the concept of *rates* for measuring the frequency of disease or characteristics we are interested in, and in particular the terms *incidence* and *prevalence*. These definitions and uses of rates are fundamental ideas with which you should be familiar before we look in more detail at research methods and study design. In the final section, we will look at key concepts in disease prevention, including the commonly used terms *primary, secondary, and tertiary* prevention.

The reason for starting with a brief exploration of the nature of scientific methods is to see how historical and social factors have influenced the biomedical and social research traditions that we take for granted today. This will help you understand your own perceptions of, and assumptions about, health research, based on the knowledge and experience you have gained to date. It will also help you understand the scientific approach being taken in this book and how this both complements and differs from that developed in books and courses on qualitative research methods – as and when you may choose to study these. Being able to draw on a range of research traditions and their associated methods is especially important for the discipline of public health, but it is also important for many other aspects of health and health care.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Briefly describe the key differences between the main approaches to research that are used in the health field.
- Describe what is meant by epidemiology, and list the main uses to which epidemiological methods and thought can be put.
- Describe what is meant by statistics, and list the main uses to which statistical methods and thought can be put.
- Define and calculate rates, prevalence, and incidence, and give examples of their use.
- Define primary, secondary, and tertiary prevention and give examples of each.

1.1 Approaches to Scientific Research

1.1.1 History and Nature of Scientific Research

Scientific research in health has a long history going back at least to the classical period. There are threads of continuity, as well as new developments in thinking and techniques, that can be traced from the ancient Greeks and through the fall of the Roman Empire, the Dark Ages, and the Renaissance to the present time. At each stage, science has influenced, and has been influenced by, the culture and philosophy of the time. Modern scientific methods reflect these varied historical and social influences. So it is useful to begin this brief exploration of scientific health research by reflecting on our own perceptions of science and how our own views of the world fit with the various ways research can be approached. As you read this chapter you might like to think about the following questions:

- What do you understand by the terms *science* and *scientific research*, especially in relation to health?
- How has your understanding of research developed?
- What type of research philosophy best fits your view of the world and the health issues you are most interested in?

Thinking about the answers to these questions will help you understand what we are trying to achieve in this section and how this can best support the research interests that you have and are likely to develop in the years to come. The history and philosophy of science is of course a whole subject in its own right, and this is of necessity a very brief introduction.

Scientific Reasoning and Epidemiology

Health research involves many different scientific disciplines, many of which you will be familiar with from previous training and experience. Here we are focusing principally on epidemiology, which is concerned with the study of the distribution and determinants of disease within and between populations. In epidemiology, as we shall see, there is an emphasis on *empiricism*, that is, the study of observable phenomena by scientific methods, detailed observation, and accurate measurement. The scientific approach to epidemiological investigation has been described as

- **Systematic** – There is an agreed system for performing observations and measurement.
- **Rigorous** – The agreed system is followed exactly as prescribed.
- **Reproducible** – All the techniques, apparatus, and materials used in making the observations and measurements are written down in enough detail to allow another scientist to reproduce the same process.
- **Repeatable** – Scientists often repeat their own observations and measurements several times in order to increase the reliability of the data. If similar results are obtained each time, the researcher can be more confident the phenomena have been accurately recorded.

These are characteristics of most epidemiological study designs and are an important part of the planning and implementation of the research. However, this approach is often taken for granted by many investigators in the health field (including epidemiologists) as the only way to conduct research. Later we will look at some of the criticisms of this approach to scientific research, but first we need to look in more detail at the reasoning behind this perspective.

Positivism

The view of science and knowledge known as *positivism* is the dominant philosophy underlying contemporary epidemiology. The evolution of positivism has been extensively documented elsewhere (see Guba and Lincoln, 1994; Halfpenny, 1982; and Feigl, 1969), its early development being attributed mainly to August Comte during the early 19th century. However, its roots can be traced back to the 17th century, to a time when scientists stopped relying on religion, conjecture, and faith to explain phenomena, and instead began to use reason and rational thought. This period saw the emergence of the view that it is only by using scientific thinking and practices that we can reveal the truth about the world.

Positivism assumes a stable observable reality that can be measured and observed. So, for positivists, scientific knowledge is proven knowledge, and theories are therefore derived in a systematic, rigorous way from observation and experiment. This approach to studying human life is the same approach that scientists take to study the natural world. Human beings are believed by positivists to exist in causal relationships that can be empirically observed, tested, and measured (Bilton *et al.*, 2002) and to behave in accordance with various laws. Because this reality exists whether we look for it or not, it is the role of scientists to reveal its existence but not to attempt to understand the inner meanings of these laws or express personal opinions about these laws. One of the primary characteristics of a positivist approach is that the researcher takes an objective distance from the phenomena so that the description of the investigation can be detached and undistorted by emotion or personal bias (Davey, 1994). This means that within epidemiology, various study designs and techniques have been developed to increase objectivity; you will learn more about these in later chapters.

More recently, some of the earlier tenets of positivism have been challenged through the work of Karl Popper and other scholars such as Bronowski (Bronowski, 1956; Popper, 1959), and as a result, a post-positivistic approach has emerged since the mid-20th century or so, and this approach now underpins much contemporary empirical research activity (Philips, 1990). Post-positivism still advocates that there is an objective reality, but it suggests that reality can only be measured imperfectly due to the limitations of the scientific approach. It also asserts a realist perspective, stating that there are phenomena that can't be observed but nevertheless do exist, so science is not limited to only those phenomena that can be measured. A more-detailed discussion of post-positivism is outside of the scope of this book. For the purposes of this chapter, when we refer to positivism, this can be assumed to refer to both positivist and post-positivist approaches within epidemiology.

A second aspect of scientific thinking, which also evolved over this period, derives from the work of Thomas Kuhn (1922–1996), who challenged the concept of absolute evidence. In *The Structure of Scientific Revolutions*, (Kuhn, 1970), Kuhn argued that one scientific paradigm – one 'conceptual worldview' – may be dominant at a particular period in history. Over time, however, this paradigm is challenged, and eventually it is replaced by another view (paradigm), which then becomes accepted as the most important and influential. He termed these revolutions in science 'paradigm shifts'. Although questioned by other writers, this perspective suggests that scientific methods we may take for granted as being the only or best way to investigate health and disease are to an extent the product of historical and social factors, and they can be expected to evolve – and maybe change substantively – over time.

Induction and Deduction

There are two main forms of scientific reasoning: *induction* and *deduction*. Both have been important in the development of scientific knowledge, and it is useful to appreciate the difference between the two in order to understand the approach taken in epidemiology.

Induction

With inductive reasoning, researchers make repeated observations and use this evidence to generate theories to explain what they have observed. For example, if a researcher made a number of observations in different settings of women cooking dinner for their partners, they might then inductively derive a general theory:

All women cook dinner for their partners.

Deduction

Deduction works in the opposite way to induction, starting with a theory (known as an *hypothesis*) and then testing it by observation. Thus, a very important part of deductive reasoning is the formulation of the hypothesis – that is, the provisional assumption researchers make about the population or phenomena they wish to study before starting with observations. A good hypothesis must enable the researcher to test it through a series of *empirical observations*. So, in deductive reasoning, the hypothesis would be

All women will cook dinner for their partners.

Observations would then be made in order to test the validity of this statement. This would allow researchers to check the consistency of the hypothesis against their observations, and if necessary, the hypothesis can be discarded or refined to accommodate the observed data. So, if they found even one woman not cooking for her partner, the hypothesis would have to be re-examined and modified. This characterises the approach taken in epidemiology and by positivists generally.

Karl Popper (1902–1994) argued that hypotheses can never be proved true for all time, and scientists should aim to refute their own hypotheses even if this goes against what they believe (Popper, 1959). He called this the *hypothetico-deductive method*, and in practice this means that an hypothesis should be capable of being falsified and then modified. Thus, to be able to claim the hypothesis is true would mean that all routes of investigation have been carried out. In practice, this is impossible, so research following this method does not set out with the intention of proving that an hypothesis is true. In due course we will see how important this approach is for epidemiology and in the statistical methods used for testing hypotheses.

Alternative Approaches to Research

It is important to be aware that positivism is only one of many different approaches to scientific research. Many social scientists, for example, believe that these approaches are not relevant for the study of human behaviour. From this perspective, they believe that human beings do not act in accordance with observable rules or laws. This makes humans different from phenomena in the natural world, and so they need to be studied in a different way. Positivism (and post-positivism) have also been criticised because they cannot explain how people interpret or make sense of the world. As Green and Thorogood (2004, p. 12) argue,

Unlike atoms (or plants or planets), human beings make sense of their place in the world, have views on researchers who are studying them, and behave in ways that are not determined in law-like ways. They are complex, unpredictable, and reflect on their behaviour. Therefore, the methods and aims of the natural sciences are unlikely to be useful for studying people and social behaviour: instead of explaining people and society, research should aim to understand human behaviour.

Many social scientists therefore hold different beliefs about how we should carry out research into human behaviour. Consequently, they are more likely to take an *inductive* approach to research because they argue that they do not want to make assumptions about the social world until they have observed it in and for itself. They therefore do not want to formulate hypotheses because they believe these are inappropriate for making sense of human action. Rather, they believe that human action cannot be explained but must be understood.

Whereas positivists would be concerned mainly with observing patterns of human behaviour, other researchers principally wish to understand that behaviour. This latter group requires a different starting point that will encompass their view of the world, or different *theoretical positions* to make sense of the world. It turns out that there are many different positions that can be adopted, and while we cannot go into them all here, we briefly consider one of the most important of these, known as an *interpretative* approach.

An interpretative approach assumes an interest in the meanings underpinning human action, and the role of the researcher is therefore to unearth that meaning. The researcher would not look to measure the reality of the world but would seek to understand how people interpret the world around them (Green and Thorogood, 2004).

Let's look at an example of positivist and interpretivist approaches in respect of a common health problem with multiple physiological, social, and behavioural aspects, namely, asthma. A *positivist* approach to researching this condition may be to obtain a series of objective measurements of symptoms and lung function using a standard procedure on a particular sample of people over a specified period of time. An *interpretative* approach might involve talking in-depth to a small number of participants with asthma to try to understand how they view the impact of their symptoms on their lives. Obviously, in order to do this, these two types of approaches require the use of different research methods. Those planning interpretative research would use *qualitative methods* (e.g. interviews, focus groups, and ethnographic methods), whereas positivists (e.g. epidemiologists) would choose *quantitative methods* (e.g. surveys and cohort studies involving lung-function measurements and highly structured questionnaires). These two different approaches would draw on different *research paradigms* and would therefore produce different types of findings.

Those drawing on an interpretative perspective would also differ from positivists in respect of the view that researchers can have an objective, unimpaired, and unprejudiced stance in the research that allows them to make value-free statements. Interpretative research accepts that researchers are human beings and therefore cannot stand objectively apart from the research. In a sense, they are part of the research process, as their presence can influence the nature and outcome of the investigation.

With the asthma example, we can see how complementary the findings of these two different approaches to research could be. The positivist approach can help determine whether a new medication provides any benefit in terms of control of symptoms or lung function, for example. On the other hand, the interpretative approach can help us understand why the activities of some people may be more affected by their condition than others, for example.

Researchers working with a post-positivist framework have to some extent narrowed the divide between quantitative and qualitative approaches to research, since post-positivism does not reject approaches that focus on the meanings people give to their actions as seen in interpretative approaches to research. Mixed-methods approaches (using both qualitative and quantitative methods) to research can therefore help build up a more-complete picture of effective health care and support that allows a better understanding of many aspects of a person's life and health experience. Further discussion of mixed methods is outside of the scope of this book; for a useful guide, see Cresswell and Plano Clark (2011).

Exercise for Reflection

1. Make brief notes on the type of scientific knowledge and research with which you are most familiar.
2. Is this predominantly positivistic (hypothetico-deductive) or interpretative in nature, or is it more of a mixture?

There are no answers provided for this exercise, as it is intended for personal reflection.

1.1.2 What is Epidemiology?

The term *epidemiology* is derived from the following three Greek words:

Epi – among
Demos – the people
Logos – study of

We can translate this in more modern terms into *the study of the distribution and determinants of disease frequency in human populations*. The following exercise will help you to think about the uses to which the discipline of epidemiology is put.

**Self-Assessment Exercise 1.1.1**

Make a list of some of the *applications* of epidemiological methods and thought that you can think of. In answering this, avoid listing types of epidemiological study that you may already know. Try instead to think in general terms about the practical applications of these methods.

Answers in Section 1.5

This exercise shows the very wide application of epidemiological methods and thought. It is useful to distinguish between two broad functions of epidemiology, one very practical, the other more philosophical:

- The range of epidemiological research methods provides a toolbox for obtaining the best scientific information in a given situation (assuming, that is, you have established that a positivist approach is most appropriate for the topic under study!).
- Epidemiology helps us use knowledge about the population determinants of health and disease to inform the full range of investigative work, from the choice of research methods, through analysis and interpretation, to the application of findings to policy. With experience, this becomes a way of thinking about health issues over and above the mere application of good methodology.

You will find that your understanding of this second point grows as you learn about epidemiological methods and their application. This is because epidemiology provides the means of describing the characteristics of populations, comparing them, and analysing and interpreting the differences, as well as the many social, economic, environmental, behavioural, ecological, and genetic factors that determine those differences.

1.1.3 What are Statistics?

A statistic is a numerical fact. Your height and weight and the average daily rainfall in Liverpool are examples of statistics. The academic discipline of statistics is concerned with the collection, presentation, analysis, and interpretation of numerical information (also called *quantitative* information).

Statistics are Everywhere!

We are surrounded by, and constantly bombarded with, information from many sources: the cereal box, unemployment figures, football results, opinion polls, and articles in scientific journals. The science of statistics allows us to make sense of this information and is thus a fundamental tool for investigation in many disciplines, including health, education, economics, agriculture, and politics, to name but a few. The next exercise encourages you to explore how statistics are used in everyday life.



Self-Assessment Exercise 1.1.2

Look at some general information sources such as newspapers or websites and find up to five items in which statistics are used. List the ways numerical information is presented.

Examples in Section 1.5

The scientific term for pieces of information is *data*. The singular is *datum*, meaning a single piece of information, such as, for example, one person's weight. A set of data may consist of many items, such as the heights, weights, blood pressures, smoking habits, and exercise levels of several hundred people. In its raw state, this mass of figures tells us little. There are two ways we use statistics to help us interpret data:

- To *describe* the group about which the data have been collected. This may be a group of people or a group of hospitals or a group of laboratory specimens. We describe the group by summarising the information into a few meaningful numbers and pictures. We discuss this further in Chapter 2.
- To *infer* something about the population of which the group we are studying is a part. We often want to know something about a population, such as everyone older than 65 years in Liverpool, but, practically, can only collect information about a subset of that population. This subset is called a sample, and it is explored in Chapter 3 on surveys. With inference, we want to know what generalisations to the population can be made from the sample and with what degree of certainty.



Self-Assessment Exercise 1.1.3

Can you find one example of *description* and one example of *inference* in your newspaper or Web search? If you have found an example of making an inference, to which population does it apply?

Examples in Section 1.5

1.1.4 Approach to Learning

We will explore the use and interpretation of statistical techniques through a number of published studies. Whether or not you go on to carry out research and use statistical methods yourself, you are certain to be a consumer of statistics through published research. We will emphasise both the use of appropriate techniques and the critical interpretation of published results. You will also be learning about epidemiology and statistics in an integrated way. This approach recognises that the two disciplines embody many closely related concepts and techniques. There are also certain very distinct qualities, which you will find are emphasised through the more theoretical discussion relating to one or another discipline. Your learning of these research methods will be based primarily on practical examples of data and published studies in order to help you to see how epidemiology and statistics are used in practice, and not just in theoretical or ideal circumstances.

Summary

- There is no single, right philosophy of health research. The approach taken is determined primarily by the nature of the problem being investigated as well as by a range of other factors, including historical and social influences.
- As an individual, your education, training, and experience strongly influence the scientific paradigm with which you are familiar and comfortable.
- A variety of approaches to, and methods for, research are both appropriate and necessary in the health field.
- Epidemiology provides us with a range of research tools, which can be used to obtain the information required for preventing health problems, providing services, and evaluating health care. One of the most important contributions of epidemiology is the insight gained about the factors that determine the health of populations.
- Statistics is concerned with the collection, presentation, analysis, and interpretation of numerical information. We may use statistical techniques to describe a group (of people, hospitals, etc.) and to make inferences about the population to which the group belongs.

1.2 Formulating a Research Question

1.2.1 Importance of a Well-Defined Research Question

This is arguably the most important section of the whole book. The reason for our saying this is that the methods we use, and ultimately the results that we obtain, must be determined by the question we are seeking to answer.

So how do we go about formulating that question? This does not (usually) happen instantaneously, and there is a good deal of debate about how the question ought to be formulated and how it is formulated in practice. For example, Karl Popper argued that research ideas can come from all kinds of sources. But the idea is not enough on its own, and it usually needs working on before it is a clearly formulated *research question*. Here are some of the factors we have to take into account in fashioning a clear research question:

- What does all the other work on the topic tell us about the state of knowledge, and what aspects need to be addressed next? Our idea might actually arise from a review such as this and therefore be already fairly well defined, but more often than not, the idea or need

arises before we have had a chance to fully evaluate the existing body of knowledge. This also depends on how far we already are into researching a particular subject.

- Different types of problem and topic areas demand, and/or have been traditionally associated with, different research traditions. Does our idea require a positivist approach with an hypothesis that can be falsified? If so, the research question needs to be phrased in a way that allows this. Alternatively, we might be trying to understand how a certain group of people view a disease and the services provided for them. This question must also be precisely defined, but not in the same way: There is nothing here to be falsified; rather, we wish to gain as full an understanding as possible of people's experience and perspectives to guide the development of services.
- If our idea is ambitious, the necessary research may be too demanding, expensive, or complex to answer the question(s) in one go. Perhaps it needs to be done in separate studies or in stages.

In practice, defining the research question does not usually happen cleanly and quickly but is a process that gradually results in a more and more sharply defined question as the existing knowledge, research options, and other practical considerations are explored and debated. There may appear to be exceptions to this – for instance, a trial of a new drug. On the face of it, the question seems simple enough: The drug is now available, so is it better than existing alternatives or not? However, as we will discover later, the context in which the drug would be used can raise a lot of issues that play a part in defining the research question.

Although knowledge of appropriate research methods is important in helping you to formulate a clear research question, it is nevertheless useful to start the *process* of developing your awareness of, and skills in, this all-important aspect of research. The following exercise provides an opportunity for you to try this.



Self-Assessment Exercise 1.2.1

A Research Idea ...

Your work in an urban area involves aspects of the care and management of people with asthma. You are well aware of the current concern about the effect of air pollution on asthma and the view that whereas pollution (e.g. ozone, nitrogen oxides) almost certainly exacerbates asthma, this may not be the cause of the underlying asthmatic tendency.

In recent years, you have noticed that asthmatics (especially children) living in the poorest parts of the city seem to suffer more severe and frequent asthma episodes than those living in better-off parts.

Although you recognise that pollution from traffic and industry might be worse in the poorer areas, you have been wondering whether other factors, such as diet (e.g. highly processed foods) or housing conditions such as dampness and associated moulds, might be the real cause of the difference.

You have reviewed the literature on this topic and have found a few studies that are somewhat conflicting and that do not seem to have distinguished very well among the pollution, diet, and housing factors you are interested in.

Have a go at converting the idea described above into a well-formulated research question appropriate for epidemiological enquiry. Note that there is no single, right research question here. You do not need to describe the study methods you might go on to use.

Specimen Answer in Section 1.5

1.2.2 Development of Research Ideas

We have seen that research is a process that evolves over a period of time. It is influenced by many factors, including other work in the field, the prevalent scientific view, political factors, finance, and so on. Research is not a socially isolated activity with a discrete (inspirational) beginning, (perfect) middle, and (always happy) ending (Figure 1.2.1).

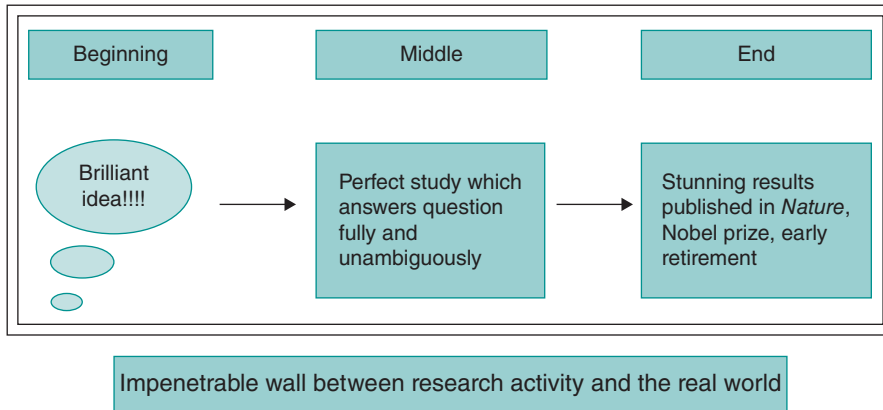


Figure 1.2.1 Research fantasy time.

A more realistic way to describe the process of research development is cyclical, as illustrated in Figure 1.2.2. A well-defined and realistic question, which is (as we have seen) influenced by many factors, leads to a study that, it is hoped, provides much of the information required.

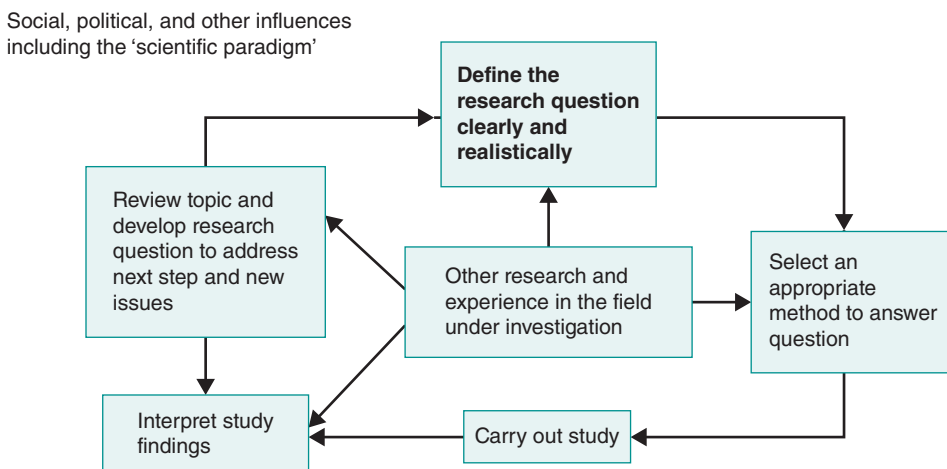


Figure 1.2.2 Research is a process that can usefully be thought of as being cyclical in nature and subject to many influences from both inside and outside the scientific community.

These findings, together with developments in the scientific field as well as social, political, or other significant influences, will lead to further development of the research question.

Summary

- Defining a clear research question is a fundamental step in research.
- Well-defined research questions do not (usually) appear instantly (although the underlying idea might arise in a moment of inspiration)!
- Research is a process, which can usefully be thought of as cyclical, albeit subject to many external influences along the way.

1.3 Rates: Incidence and Prevalence

1.3.1 Why Do We Need Rates?

Incidence and prevalence are terms you will have encountered commonly in everyday language as well as in research. In quantitative health research, they have clear and specific definitions, and they represent fundamental concepts of how the frequencies of diseases or characteristics among groups of people are described. Why then are they so important?

One way to approach this question is by considering the problem that arises when we try to interpret a change in the number of events (which could be deaths, hospital admissions, etc.) occurring during, say, a period of one year in a given setting. Exercise 1.3.1 is an example of this type of problem and is concerned with an increase in numbers of hospital admissions.

**Self-Assessment Exercise 1.3.1**

Over a period of 12 months, the accident and emergency department of a city hospital noted that the number of acute medical admissions for people over 65 had increased by 30 per cent. In the previous 5 years, there had been a steady increase of only about 5 per cent per year.

1. List the possible reasons for the 30 per cent increase in hospital accident and emergency admissions.
2. What other information could help us to interpret the reasons for this sudden increase in admissions?

Answers in Section 1.5

In this exercise we have seen the importance of interpreting changes in numbers of events in the light of knowledge about the *population* from which those events arose. This is why we need *rates*. A rate has a *numerator* and a *denominator* and must be determined over a specified *period of time*. It can be defined as follows:

$$\text{RATE} = \frac{\text{Number of events arising from defined population in a given period}}{\text{Number in defined population, in same period}}$$

“Numerator”

“Denominator”

1.3.2 Measures of Disease Frequency

We can view rates as measures of the *frequency* of disease or of characteristics in the population. Two of the most important ways of presenting this information are provided by *prevalence* and *incidence* rates.

1.3.3 Prevalence Rate

The prevalence rate tells us how many cases of a disease (or people with a characteristic, such as smoking) there are in a given population at a specified time. The *numerator* is the number of cases, and the *denominator* is the population we are interested in.

$$\text{Prevalence} = \frac{\text{Number of cases at a given time}}{\text{Number in population at that time}}$$

This can be expressed as a percentage, or per 1,000 population, or per 10,000, etc., as convenient. The following are therefore examples of prevalence:

- In a local government area with a population of 400,000, there are 100,000 smokers. The prevalence is therefore 25 per cent, or 250 per 1,000.
- In the same area, there are known to be 5,000 people with diagnosed schizophrenia. The prevalence is therefore 1.25 per cent, or 12.5 per 1,000.

Note that these two examples represent snapshots of the situation at a given time. We do not have to ask about people starting or giving up smoking or about people becoming ill with (or recovering from) schizophrenia. It is a matter of asking, ‘In this population, how many are there now?’ This snapshot approach to measuring prevalence is known as *point prevalence*, since it refers to one point in time, and it is the usual way the term *prevalence* is used. If, on the other hand, we assess prevalence over a period of time, it is necessary to think about cases that exist at the start of the period and new cases that develop during the period. This measure is known as *period prevalence*, and this, together with point prevalence, is illustrated in Figure 1.3.1 and Exercise 1.3.2.

Point prevalence (Figure 1.3.1a) is assessed at one point in time (time A), whereas period prevalence (Figure 1.3.1b) is assessed over a period (period B). The horizontal bars represent patients becoming ill and recovering after varying periods of time – the start and end of one episode is marked in Figure 1.3.1a. Period prevalence includes everyone who has experienced the disease at some time during this period.



Self-Assessment Exercise 1.3.2

1. In the examples in Figure 1.3.1, calculate the point prevalence and period prevalence.
2. Why do the point prevalence and period prevalence differ?

Answers in Section 1.5

1.3.4 Incidence Rate

Whereas the *prevalence* rate gives us a measure of how many cases there are in the population at a given time (or period when period prevalence is used), the *incidence rate* tells us the rate at which *new* cases are appearing in the population over a specified time period. The time period

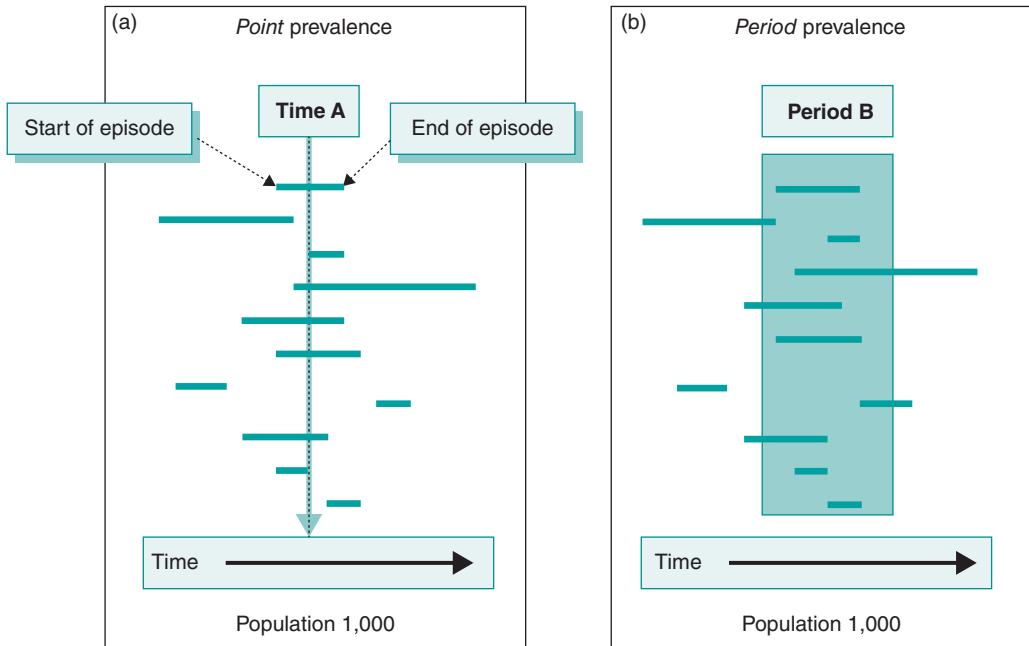


Figure 1.3.1 Period and point prevalence.

must always be specified for an incidence rate. To determine the incidence rate, we need to know the number of new cases appearing over a specified period of time and the number of people in the population who could become cases over that same period of time (the at-risk population). This is called the **cumulative incidence rate** and is calculated as follows:

Incidence rate (also termed ‘cumulative incidence rate’)

$$\text{Incidence} = \frac{\text{Number of new cases arising from a defined population in a specified time period}}{\text{Number in defined at-risk population over the same time period}}$$

This rate can be expressed per 1,000 per year (or other convenient time period) or per 10,000 per year, etc., as appropriate, over the specified time period. Thus, if there were 20 new cases of a disease in an at-risk population of 2,500 over a period of one year, the cumulative incidence rate, expressed per 1,000 per year, is as follows:

$$\text{Incidence} = \frac{20}{2,500} \times 1,000$$

This gives a rate of 8 cases per 1,000 per year. The denominator – the defined population – must be exclusively the population that could become cases (termed **at risk**). For example, if we are considering the rate of hysterectomy (removal of the uterus) among UK women aged 50 years and older, the denominator population must be women 50 and older but excluding those who have had a hysterectomy.

Quite often, however, when a large group of the population is being studied – for example, all men in a city such as Liverpool – the exact number at risk throughout the year in this defined population will not be readily available, and an estimate has to be made. The corresponding midyear population estimate is often used as the denominator population, since it is usually available from published official statistics.

Using a source of information on the population such as this means that existing cases who are not at risk of becoming new cases of the disease (because they are already affected) will be included in the denominator of the equation (unless data on the numbers of cases of the disease in question are available to allow these to be removed). Where the number of existing cases is not known, the resulting incidence measure will not be much affected so long as the number of cases is relatively small in comparison to the total population, but you should be aware that the true incidence rate of the disease will be slightly underestimated. The following exercise will help consolidate your understanding of the cumulative incidence rate.



Self-Assessment Exercise 1.3.3

1. In Figure 1.3.1b, how many new cases arose during period B?
2. Surveillance data for episodes of food poisoning in a given population showed that among children aged 0–14 years, there had been 78 new cases among boys and 76 new cases among girls over a period of 1 year. The population breakdown for the area is as shown in the following table.

Age group (years)	Female	Male
0–14	37,100	41,000
15–24	59,610	60,100
25–44	62,050	57,300
45–64	42,450	39,610
65+	28,790	21,990
Total	230,000	220,000

Calculate the annual cumulative incidence rates per 10,000 for boys aged 0–14 years and for girls aged 0–14 years. Comment on what you find.

Answers in Section 1.5

Person-Time

The cumulative incidence rate assumes that the entire population at risk at the beginning of the study has been followed up for the same amount of time. It therefore measures the proportion of unaffected individuals who, on average, will contract the disease over the specified time period. However, in some study designs, such as cohort studies (described in Chapter 5), people may be entered into the study at different times and then be followed up to a specific end-of-study date. In addition, some might withdraw from the study or might die before the end of the study. Study participants therefore have differing lengths of follow-up.

To account for these varying times of follow-up, a denominator measure known as *person-time* is used. This is defined as the sum of each individual's time at risk while remaining free of disease. When *person-time* is used, incidence is calculated slightly differently and is known as the *incidence density*, which is the average person-time incidence rate. Note, we do not include 'rate' in the term 'incidence density', as this is implied. Incidence density is calculated as follows:

$$\text{Incidence density} = \frac{\text{Number of new cases arising from a defined population}}{\text{Total at risk person-time of observation}}$$

Since person-time can be counted in various units such as days, months, or years, it is important to specify the time units used. For example, if six new cases of a disease are observed over a period of 30 person-years, then the incidence would be $6/30 = 0.2$ per person-year or, equivalently, 20 per 100 person-years or 200 per 1,000 person-years. If people are lost to follow-up or withdraw from the study prematurely, their time at risk is taken to be the time they were under observation in the study. This next exercise will help with understanding incidence density.



Self-Assessment Exercise 1.3.4

Time of follow-up in study, or until disease develops, for 30 subjects:

Subject number	Years of follow-up	Disease (Y or N)	Subject number	Years of follow-up	Disease (Y or N)
1	19.6	N	16	0.6	Y
2	10.8	Y	17	2.1	Y
3	14.1	Y	18	0.8	Y
4	3.5	Y	19	8.9	N
5	4.8	N	20	11.6	Y
6	4.6	Y	21	1.3	Y
7	12.2	N	22	3.4	N
8	14.0	Y	23	15.3	N
9	3.8	Y	24	8.5	Y
10	12.6	N	25	21.5	Y
11	12.8	Y	26	8.3	N
12	12.1	Y	27	0.4	Y
13	4.7	Y	28	36.5	N
14	3.2	N	29	1.1	Y
15	7.3	Y	30	1.5	Y

1. Assuming that all subjects in the table enter the study at the same time and are followed up until they leave the study (end of follow-up) or develop the disease, find the total observation time for the 30 subjects and estimate the incidence density. Give your answer per 1,000 (10^3) person-years.
2. It is more usual for follow-up studies to be of limited duration, where not all the subjects will develop the disease during the study period. Calculate the incidence density for the same 30 subjects if they were observed for only the first 5 years, and compare this with the rate obtained in question 1.

Answers in Section 1.5

1.3.5 Relationship Between Incidence, Duration, and Prevalence

There is an important relationship between the incidence rate, illness duration, and prevalence rate. Consider the following two examples:

- Urinary tract infections among women are seen very commonly in general practice, reflecting the fact that the incidence rate is high. The duration of these infections (with treatment) is

usually quite short (a few days), so at any one time there are not as many women with an infection as one might imagine given the high incidence. Thus, the (point) prevalence is not particularly high.

- Schizophrenia is a chronic psychiatric illness from which the majority of sufferers do not recover. Although the incidence is quite low, it is such a long-lasting condition that at any one time the prevalence is relatively high, between 0.5 per cent and 1 per cent, or 5 to 10 per 1,000 in the UK.

Thus, for any given incidence, a condition with a longer duration will have a higher prevalence. Mathematically, we can say that the prevalence rate is proportional to the product of the incidence rate and the average duration of the disease. In particular, when the prevalence rate is low (less than about 10 per cent), the relationship can be expressed as follows:

$$\text{Prevalence} = \text{incidence} \times \text{duration}$$

This formula holds so long as the units concerned are consistent, the prevalence rate is low, and the duration is constant (or an average can be taken). Thus, if the incidence rate is 15 per 1,000 per year, and the duration is, on average, 26 weeks (0.5 years), then the prevalence rate will be (15×0.5) per 1,000 = 7.5 per 1,000.

Summary

- Rates are a very important concept in epidemiology, and they allow comparison of information on health and disease from different populations.
- A rate requires a numerator (cases) and a denominator (population), each relating to the same specified time period.
- The prevalence rate is the number of cases in a defined population at a particular point in time (point prevalence) or during a specified period (period prevalence).
- The incidence rate is the number of new cases that occur during a specified time period in a defined at-risk population (cumulative incidence).
- Incidence density is a more precise measure of incidence and uses person-time of observation as the denominator.
- Without using rates, comparison of numbers of cases in different populations may be very misleading.
- The relationship between incidence and prevalence is determined by the duration of the condition under consideration.

1.4 Concepts of Prevention

1.4.1 Introduction

In this section, we look at the ways we can describe *approaches to prevention*. This is a well-established framework that provides important background to many of the studies we will examine as we learn about research methods, as well as for services such as screening.

The following examples illustrate three different approaches to disease prevention. Please read through these, and complete Exercise 1.4.1. We will then look at the formal definitions of these approaches.

Example 1: Road Accidents Among Children

In 2012, accidents were the most common cause of death among male children and young adults aged 5–19 years in the UK, and the majority of these accidents occurred on the roads. Lower speed limits, linked to stricter enforcement, offer one way of reducing the number of these deaths arising from road accidents.

Example 2: Breast Cancer

Breast cancer is one of the most common cancers among women. Despite this, we know little for certain about the causes beyond genetic, hormonal, and some dietary factors. For a number of years, mammography, a radiographic (X-ray) examination of the breast, has been routinely offered to women 50–64 years of age. Abnormalities suggestive of cancer are investigated by biopsy (removal of a small piece of tissue for microscopic examination), and if the biopsy is positive, the cancer is treated.

Example 3: Diabetes and the Prevention of Foot Problems

Diabetes, a disorder of blood glucose (blood sugar) metabolism, is generally a progressive condition. The actual underlying problem does not usually resolve, and control of blood glucose has to be achieved through attention to diet, and usually also with medication, which may be in tablet form or as injected insulin. Associated with this disordered glucose metabolism are a range of chronic degenerative problems, including atherosclerosis (which leads to heart attacks and to poor blood supply to the lower legs and feet), loss of sensation in the feet due to nerve damage, and eye problems. Many of these degenerative processes can be slowed down, and associated problems prevented, by careful management of the diabetes. One important example is care of the foot when blood supply and nerves are affected. This involves educating the diabetic patient about the problem and about how to care for the foot, and providing the necessary treatment and support.

**Self-Assessment Exercise 1.4.1**

For each of the above examples, describe in everyday language (that is, avoiding technical terms and jargon) how prevention is being achieved. In answering this question, think about the way the prevention measure acts on the development and progression of the disease or health problem concerned.

Answers in Section 1.5

1.4.2 Primary, Secondary, and Tertiary Prevention

The three examples of prevention that we have just discussed are (respectively) illustrations of *primary*, *secondary*, and *tertiary* prevention. These terms can be defined as shown in Table 1.4.1.

Table 1.4.1 Primary, secondary, and tertiary prevention.

Term	Definition	Example studied
<i>Primary</i>	Preventing an infection, injury, or other disease process from occurring	Limiting vehicle speeds reduces the likelihood of a young person or child being involved in a road accident; if an accident does occur, the risk of serious injury or death from this cause is reduced.
<i>Secondary</i>	Early detection of a disease process at a stage where the course of the disease can be stopped or reversed	Offering mammography to the population of women aged 50–64 years allows earlier detection of breast cancer and a better chance of successful treatment.
<i>Tertiary</i>	Management of a condition that is already established in such a way as to minimise consequences such as disability and other complications	The diabetic patient generally requires lifelong treatment, and the underlying condition cannot be cured. The complications and disability arising from foot problems, for example, can be avoided or ameliorated by an active approach to prevention.



Self-Assessment Exercise 1.4.2

For each of the following activities, state whether this is primary, secondary, or tertiary prevention, giving brief reasons for your answer:

1. Measles immunisation
2. Smear tests for cervical cancer every 5 years
3. A well-managed programme of terminal care for a patient with cancer
4. Use of bed nets impregnated with insecticide in malaria-endemic areas
5. Smoking-cessation programme in middle-aged men recovering from a heart attack

Answers in Section 1.5

1.5 Answers to Self-Assessment Exercises

Section 1.1

Exercise 1.1.1

The uses of epidemiological methods and thought: This list is not necessarily exhaustive, but it covers the most important applications:

- By studying populations rather than those already in the health-care system, one can gain a more-representative and complete picture of the distribution of disease. Studies of populations and those receiving care can identify the factors that determine who does, and does not, take up health care, and why.
- Describing the frequency of a disease, health problem, or risk factor; who is affected, where, and when. This may be used for *planning*, as in epidemic control, service provision, etc.
- Understanding the *natural history* of health problems; that is, what happens if there is no treatment or other intervention.
- Understanding the *causes* of disease, thus laying the basis for prevention.

- Determining the **effectiveness** of health interventions, whether drugs, surgical operations, or health promotion through, for example, raising awareness or establishing public policy.
- Through an understanding of the determinants of the health of populations, epidemiology contributes to the development of **prevention policy**.

Thus, while basic research may add to our biologic understanding of why an exposure causes or prevents disease, only epidemiology allows the quantification of the magnitude of the exposure–disease relationship in humans and offers the possibility of altering the risk through intervention. Indeed, epidemiologic research has often provided information that has formed the basis for public health decisions long before the basic mechanism of a particular disease was understood.

Hennekens and Buring, 1987, p. 13

Study of groups that are particularly healthy or vulnerable is often the beginning of the search for causes, and so of prevention.

Morris, 1957, p. 263

Exercise 1.1.2

These are just a few examples of the use of statistics found on the Internet. Note the different ways that the information is presented.

1. Cricket scores with examples of counts and averages (BBC Sport website)

The screenshot shows the BBC Sport website's 'English One-Day Rankings' page. The page title is 'English One-Day Rankings' with a dropdown menu set to 'English One-Day' and an 'UPDATE' button. Below the title is a table titled 'Royal London One PCA Rankings'. The table has 10 rows and 10 columns: Player, County, Batting, Bowling, Fielding, Captain, Wins, Played, Points, and Average. The data is as follows:

Player	County	Batting	Bowling	Fielding	Captain	Wins	Played	Points	Average	
1	Collingwood	Durham	78.43	67.94	1.0	0.0	7.0	11	154.37	14.03
2	Rudolph	Glamorgan	132.09	3.41	1.0	0.0	4.0	8	140.5	17.56
3	Stokes	Durham	85.28	43.76	5.0	0.0	5.0	7	139.04	19.86
4	Rashid	Yorkshire	29.66	82.72	1.0	0.0	6.0	9	119.38	13.26
5	Patel	Wanwickshire	7.91	104.1	1.0	0.0	6.0	10	119.01	11.9
6	Chopra	Wanwickshire	97.34	0.0	8.0	6.0	6.0	9	117.34	13.04
7	Clarke	Wanwickshire	23.79	77.05	5.0	0.0	6.0	10	111.94	11.18
8	Stoneman	Durham	93.14	0.0	3.0	7.0	7.0	11	110.14	10.01
9	Westley	Essex	64.57	36.99	0.0	0.0	5.0	8	106.56	13.32
10	Taylor	Nottinghamshire	90.81	0.0	4.0	5.0	5.0	7	104.81	14.97

At the bottom right of the table, it says 'Last updated: Tuesday, 14 October 2014 10:12'. Two callout boxes are present: one pointing to the 'Played' column with the text 'Counts: number of matches played in', and another pointing to the 'Average' column with the text 'Averages: the average number of points per match (calculated as the total points divided by the number of matches played)'.

- Voting intentions from opinion polls for the 2014 European elections, showing examples of percentages (Source: YouGov)

European Elections

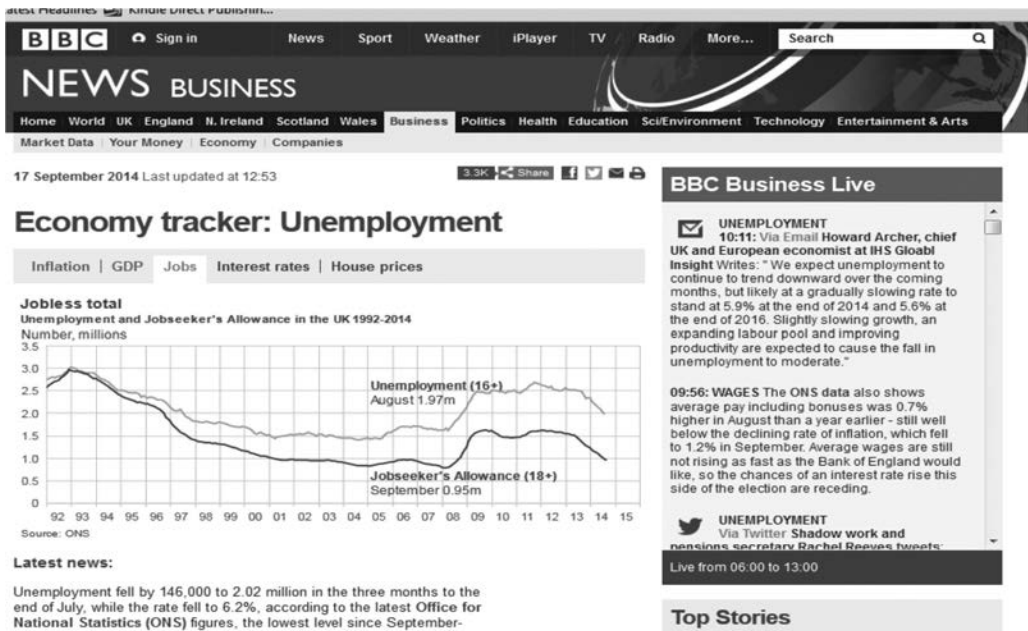
2014 European Voting Intention

Survey End Date	CON	LAB	LDEM	UKIP	GRN	BNP
YouGov/Sun/Times	22	26	9	27	10	1
Opinium	21	25	6	32	6	1
Survation/Mirror	23	27	9	32	4	1
YouGov/Sun	23	27	10	27	8	1
TNS	21	28	7	31	?	?
YouGov/Sun	21	28	10	24	12	1
ComRes/ITV (O)	20	27	7	33	6	1
YouGov/Sunday Times	23	27	9	26	9	1
Opinium	20	29	5	31	5	3
ICM/Sunday Telegraph (O)	26	29	7	25	6	?
ComRes/Independent on Sunday (O)	20	24	6	35	7	2
YouGov/Sun	22	28	10	25	10	1
Opinium	22	28	7	30	5	2
ICM/Guardian	27	24	7	26	10	0
ComRes/Coalition for Marriage	22	24	8	34	5	1

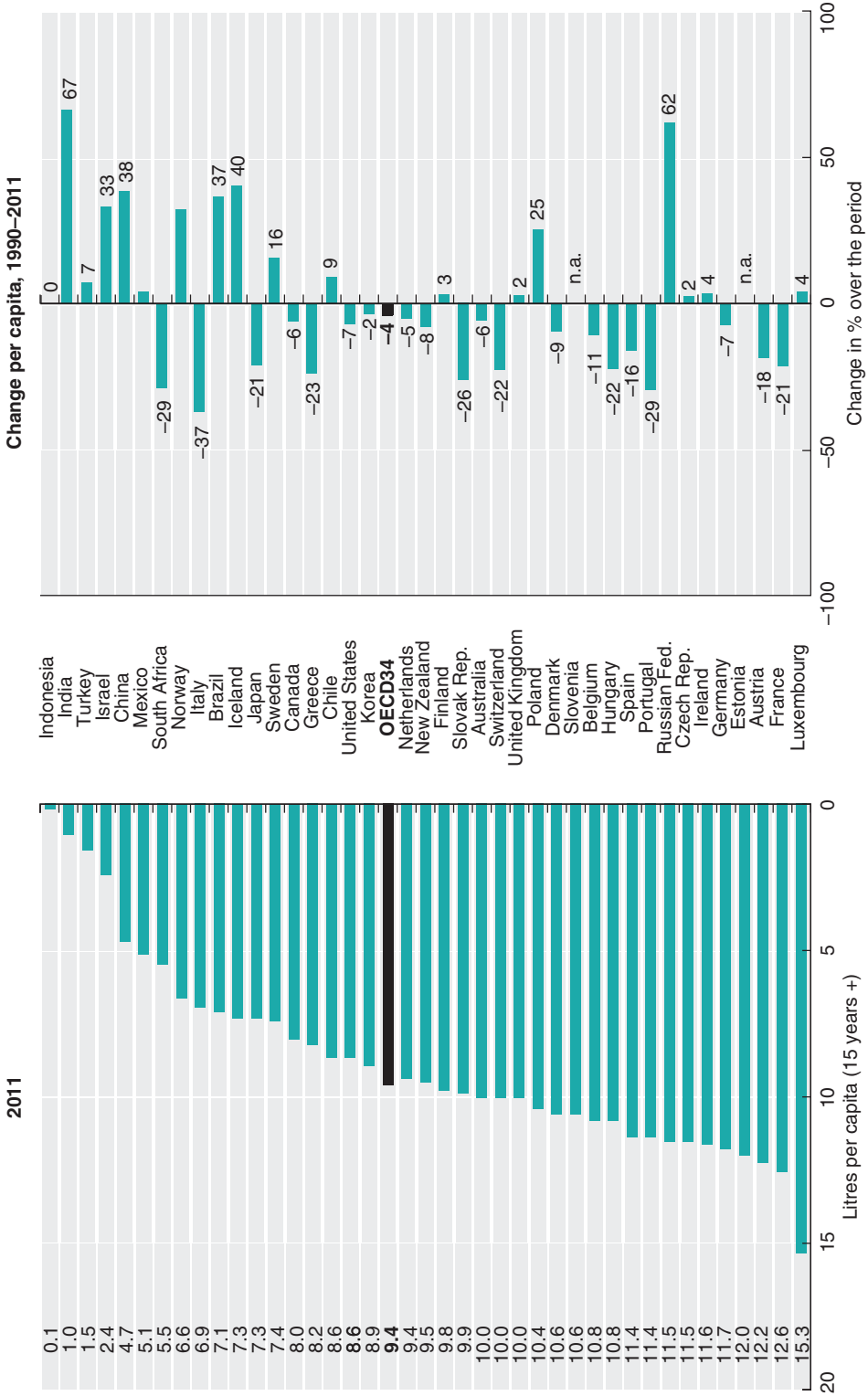
SWINGOMETER
BASIC GB SWINGOMETER
PROVISIONAL BOUNDARIES
SWINGOMETER
GRAPHICAL SWINGOMETER
ADVANCED SWINGOMETER

AMAZON ADS
Affluence, Austerity and ...
Paul Whiteley, Ha...
£17.51
The Conservative
Tim Bale (Paperb...
£14.99

- Unemployment data for the UK 1992–2014, showing percentage unemployment and those claiming the job-seekers’ allowance, presented as line graphs



- Alcohol consumption (litres per capita) among adults by country in 2011 and change over the period 1990–2011 (as a percentage), presented as bar charts (Source: Organisation for Economic Co-operation and Development)



Section 1.2

Exercise 1.2.1

This is a complex problem, so don't worry if you found it challenging. Many different research questions could arise from this, depending on the type of research your experience and interests relate to. For example, at one end of the spectrum, a laboratory-based scientist might wish to carry out some physiologically based animal experiments involving chemicals from traffic pollution, food additives, and fungal spores (from damp housing). Laboratory-based experimentation with humans is out of the question (apart from specific volunteer studies), so we will concentrate on epidemiological investigation. The following two research questions represent different levels of investigation. The first aims to start the research process off by describing the situation (from which associations between asthma and other factors can be investigated), and the second takes a more analytical approach. Which one is appropriate would depend on exactly what has already been studied, resources available, and so on. You will note that both questions define the population as children aged 5–15 years, in order to help focus the study.

Research Question 1

How are asthma, levels of air pollution, damp housing, consumption of [specified] processed foods, and socioeconomic circumstances distributed among children aged 5–15 years living in [named city]?

Research Question 2

Among children aged 5–15 years living in [named city], are the presence and/or frequency of asthma associated with [specified] processed foods or damp housing, after taking account of levels of air pollution?

Section 1.3

Exercise 1.3.1

The reasons for this very large increase could be as follows:

- A **chance** (random) variation (although this is unlikely given the large numbers involved). The role of chance variation is a very important concept in epidemiology and statistical methods, and we will begin to examine this in Chapter 2.
- An **artefact** of the system, such as a change, or error, in the system for recording admissions. This is also unlikely to cause such a dramatic increase, but it needs to be considered. The quality of information and how it is defined and handled are very important issues in research, and we will begin to look at these in Chapter 2.
- A **real** increase, which could result from changes in referral procedures by GPs (although this is again unlikely to cause such a large increase in one year, especially given the more gradual increase seen in previous years). A more likely explanation is a sudden increase in the population that the accident and emergency department is serving. Natural increase in population is unlikely, unless there had been an event such as a rapid influx of refugees for example. Closure of another (smaller) accident and emergency department in the city is the most likely reason for the large increase in the population being served.

The key point here is that in order to make some judgment about this increase, we need to know the size of the population from which the admissions are coming. If this population has increased by 30 per cent, then, all other things being equal, we would expect the number of admissions to increase by 30 per cent. Changes in numbers of events (whether deaths, cases,

admissions, etc.) cannot be interpreted usefully without information on changes in the population from which these events arose and the time period concerned.

Exercise 1.3.2

1. Calculation of point and period prevalence. **Point prevalence:** In this diagram, 7 cases were 'active' at time A. With a population of 1,000, the point prevalence is $7 \div 1,000 = 0.007$. This is a rather untidy way of expressing the prevalence, so we can state it as 0.7 per cent or 7 per 1,000. **Period prevalence:** Here we have a longer period to consider, during which some cases resolve and new ones occur. We include the cases that got better, because they were cases at some time during period B. We must also include the new cases that start during period B. A total of 10 cases were active during period B, so the period prevalence is $10 \div 1,000 = 0.01$. This can be presented as 1 per cent or 10 per 1,000 over the period concerned (e.g. 1 year).
2. The point and period prevalence rates differ because, for period prevalence, we included some cases that had not yet recovered and some new cases that appeared during period B.

Exercise 1.3.3

1. During period B, a total of seven new cases arose.
2. The incidence rate for boys aged 0–14 years is $78 \div 41,000 \times 10,000$ per year = 19.0 per 10,000 per year. The incidence rate for girls aged 0–14 years is $76 \div 37,100 \times 10,000$ per year = 20.5 per 10,000 per year. So although there were more cases among boys, the incidence rate was actually higher among girls. The reason for this is that the population of girls was smaller. This again emphasises why rates are so important.

Exercise 1.3.4

1. The total observation time for the 30 subjects is 261.9 person-years, during which time $n = 20$ subjects developed the disease. The incidence rate is $20/261.9 = 0.0764$ per person-year or 76.4 per 1,000 (10^3) person-years.
2. In this case the total observation time is:

$$5 + 5 + 5 + 3.5 + 4.8 + 4.6 + \dots 5 + 1.1 + 1.5 = 115.8 \text{ person-years}$$

The total number of cases is now 11 (as we only include cases occurring within the first 5 years), so the incidence rate is $11/115.8 = 0.095$ per person-year or 95.0 per 1,000 person-years. This rate is somewhat higher than that for the longer follow-up.

Section 1.4

Exercise 1.4.1

Road Accident Prevention

Reducing vehicle speeds reduces the likelihood of a road accident, and it also reduces the chance of serious injury or death if a collision (with a person, vehicle, or other object) does occur. The prevention process here is principally through preventing the accident (injury or death) happening in the first place, but if it does happen, the severity of injury and the likelihood of death are reduced.

Breast Cancer Prevention

From the information we have, it is apparent that we do not yet know enough to prevent most cases of breast cancer from occurring in the first place. What we can do is detect the disease at

a stage where, if treated promptly, it is possible to cure the disease in a substantial proportion of affected women. Thus, in contrast to the road accident example, the prevention begins after the disease process has started, but it is at a stage where it is still possible to cure the disease.

Prevention of Foot Problems in the Diabetic Patient

In this last example, the disease process is well established, and it cannot be removed or cured. That does not mean that the concept of prevention has to be abandoned, however. In this situation, preventative action (education, support, treatment, etc.) is being used to prevent damage to the skin of the feet, with the infection and ulceration that can follow. Prevention activities are carried out, even though the underlying disease remains present.

Exercise 1.4.2

Example	Prevention approach	Reasons and explanation
Measles immunisation	Primary	Immunisation raises immunity of the recipient and prevents infection.
Smear tests	Secondary	The smear test is a screening procedure designed to detect the disease process at an early stage. Note that the test on its own is not prevention, as it does not alter the course of the disease; it must be followed up (if positive) by biopsy and treatment as necessary.
Terminal care	Tertiary	The patient is dying (e.g. from cancer), and nothing can be done to alter that. That does not mean the preventative approach is abandoned. Good terminal care can prevent a lot of pain and emotional distress in both the patient and the patient's family and friends. This can have lasting benefits.
Bed nets	Primary	Impregnated bed nets prevent contact between the feeding mosquito and humans, especially at night, when most biting occurs. This prevents introduction of the malaria parasite into the body, thus preventing the occurrence of the disease.
Smoking cessation	Tertiary?	This one does not fit definitions easily. Smoking cessation in a healthy younger person could properly be regarded as primary prevention of heart disease. In our example, the men had already had a heart attack. This is not secondary prevention (detection at early stage, etc.). It is probably best seen as tertiary, especially bearing in mind that this smoking cessation should generally be part of a broader rehabilitation package helping the man to get mentally and physically better, reduce his risk of another heart attack, and get back to work or other activity.

2

Routine Data Sources and Descriptive Epidemiology

Introduction and Learning Objectives

In this chapter, we examine a range of information sources of value to health research, and learn how to present, analyse, and interpret this information. Exercises will help you understand how the information is collected and its strengths and weaknesses, and they will introduce you to methods for presenting and comparing the data. We will finish by looking at how these descriptive epidemiological research methods fit into an overview of study designs, and consider the nature of the research evidence that we have obtained from the methods studied.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the key sources of routinely available data relevant to research in public health and health care, including the census, and examples of registrations (deaths, cancer, etc.), notifiable diseases, service utilisation, surveys, and environmental monitoring.
- Examine the uses, strengths, and weaknesses of selected examples of routine data.
- Describe the usefulness of studying variations in health, and health determinants, by time, place, and person, with examples.
- Calculate and present a frequency distribution for a set of continuous data.
- Display continuous data in a histogram.
- Describe the shape of a distribution of continuous data.
- Summarise distributions of continuous data: calculate mean, median, mode, standard deviation, and interquartile range, and use these appropriately.
- Interpret a scatterplot, describe the relationship between two continuous variables, and interpret a (Pearson) correlation coefficient.
- Describe what is meant by an ecological study and how the ecological fallacy can arise.
- Present an overview of the different types of epidemiological study design, and the nature of descriptive studies in relation to other study designs.

Most of the data sources used as examples in this chapter are drawn from England and Wales, in part because these are among the most comprehensive available anywhere. Many of the principles relating to obtaining and interpreting routine data highlighted in these examples also apply more generally. Some international examples are also used for comparison.

2.1 Routine Collection of Health Information

2.1.1 Deaths (Mortality)

We begin by describing the collection and recording of death data (termed *mortality data*), drawing on the system used in England and Wales. This includes how the cause of death is decided upon and recorded, and how other information about the person who died (such as their occupation) is added. The following two (fictional) stories are about deaths occurring in very different circumstances. Read through these, and as you do so, note down the pathways by which information about the person dying, and the mode of death, is obtained and recorded. Note the roles of relatives, medical staff and other officials, and other significant people such as witnesses.

Story A – An Expected Death

On a cold January afternoon, tired and laden with shopping, Joan Williams walked slowly up to her front door. Hearing the phone, she put down the bags and fumbled with the key in her anxiety about the call. Perhaps it was the hospital with news about her mother.

'Hello, Joan Williams', she said quietly.

'Oh Hello, Mrs Williams, it's Sister Johnson here. I'm afraid your mother has had a relapse. She is conscious, but I think you should come over to the hospital as soon as you can.'

'Yes, of course I'll be there', Joan replied.

Joan had been expecting this call for a week or more, but that did not prevent the fear she now felt about what lay ahead. Mechanically, she put away the shopping, as her thoughts drifted back to the summer holidays. Her 79-year-old mother had joined the family for their holiday in Cornwall, and despite her age she was active and helpful with the children, who were very fond of her. Just before they were due to go home, her mother developed a cough and fever, and was found to have pneumonia. She improved on antibiotics, but then she fell ill again with a recurrence, this time with blood-stained phlegm. She was found to have lung cancer, and for much of the next 6 months was in and out of hospital for palliative radiotherapy and treatment of chest infections. She had deteriorated a lot in the last few weeks, and Joan knew that her mother was close to death.

When Joan arrived at the infirmary, her mother was unconscious. A few minutes before, she had suffered a heart attack, and apart from monitoring her condition, no other active treatment was to be given. For a while after her mother died, Joan sat beside the bed thinking of all that had happened in the last 6 months, and how she would tell the children. She had not had the responsibility of dealing with the death of a family member before, since her father had been killed on active military service when she was very young. Now she would have to get the death certificate, make all the arrangements for the funeral, and tell the rest of the family. She just wanted to be left alone with her memories of her mother.

The staff at the hospital were very kind, and they helped as much as they could. 'If you come back tomorrow morning, the death certificate will be ready', Sister Johnson told her. The doctor who filled in the death certificate had to decide on the cause of death, which she put down as myocardial infarction, with the underlying cause as carcinoma of the bronchus. Since a clear diagnosis of the cancer had already been established, and Mrs Williams' mother had been seen both by her GP and at the hospital regularly over the previous 6 months, all the information the doctor needed was in the medical records. The next day, Joan collected the death certificate and took it to the local registry office, where she had to register the death. The clerk asked her a few more questions about her mother, thanked her as he handed her the disposal order for the undertaker, and said that she could now go ahead with the funeral arrangements.

Walking out into the bright January sunshine, Joan was grateful that at least this bit of the proceedings had been quite straightforward, though painful nonetheless. She thought of her sister Mary, whose husband had died suddenly at home last year: Mary had suffered terribly with all the delay while the coroner's office spoke to the doctors and a postmortem examination was carried out, all the time wondering whether they would order an inquest. At least her mother's death had been expected, and the officials could fill in the forms without all kinds of delays and investigations.



Self-Assessment Exercise 2.1.1

1. Who completed the death certificate for Joan Williams's mother?
 - a. Sister Johnson
 - b. the coroner
 - c. the hospital doctor
 - d. Joan Williams
 - e. the registrar of births and deaths.
2. We are interested in how accurately mortality statistics report the true cause of death.
 - a. How accurate do you think the information on the certificate was for Mrs Williams' mother?
 - b. Do you think a post-mortem examination (autopsy) would have improved the quality of the information?

Answers in Section 2.8

Story B – A Sudden, Unexpected Death

John Evans had been driving trains for over 20 years. Late one Sunday afternoon he experienced the nightmare that all drivers fear more than anything else.

After leaving York on the way south to London, he brought his Intercity train up to full speed as he ran down the beautiful coastal track of East Yorkshire. Rounding a gentle bend at over 120 miles per hour, he noted someone on a bridge about half a mile ahead. That was a common enough sight on this stretch of track, but something about the person's movements held his attention. His heart missed a beat as he realised a man was climbing the parapet, and he instantly applied the brakes, knowing only too well that stopping the train was impossible. Unable to look away, he caught sight of the blank face of a young man as he fell past the windscreen. He barely heard the thud as the man's body went under the train, and there was little left for the forensic scientists. Dental records were enough though, and the victim was eventually identified as a 24-year-old homeless man who had been living in a hostel in Hull. He had no record of psychiatric illness, had left no suicide note, and had not told anyone of his intentions.

At the coroner's inquest, only his mother was available to give evidence; his father had left home when the victim was very young and had died some 5 years before from heart disease. She had not seen her son for 6 months before he died, and on the occasion of their last meeting he had not appeared unduly upset. The train driver also gave his evidence, and he felt certain this was a deliberate act of suicide. However, without evidence of definite suicidal intent such as a note or a record of several previous attempts, the coroner was obliged to return a verdict of 'death uncertain as to whether deliberately or accidentally caused', known as an open verdict.



Self-Assessment Exercise 2.1.2

1. Why did the coroner not return a verdict of suicide on the homeless man, who seemingly deliberately threw himself under a high-speed train?
2. What proportion of all true suicides do you think receive a verdict of suicide in England, and hence appear in the mortality statistics as such?

Answers in Section 2.8

2.1.2 Compiling Mortality Statistics: The Example of England and Wales

These examples have given you part of the picture of how information about deaths in the UK is obtained and compiled into mortality statistics. Figure 2.1.1 summarises all of the stages for a death (like that of Joan Williams' mother) not requiring notification to the coroner. The coroner is an independent judicial officer of the Crown who has a statutory duty to investigate the circumstances of certain categories of death for the protection of the public.

The circumstances in which the coroner needs to be involved are summarised in the box below. It is important for you to be aware of these circumstances for three reasons:

1. The coroner's inquest will result in more complete information on the cause of death than would otherwise be available.
2. There is usually a delay of several months, and sometimes more than a year, before the inquest is completed, and the information about the cause of death may not enter the mortality statistics until the inquest is closed.
3. The coroner's records are a potentially useful source of information for local research (subject to permission).

Circumstances in Which the Coroner Must be Informed of a Death

A death must be reported to the coroner where

- no doctor saw the deceased during his or her last illness;
- although a doctor attended the deceased during the last illness, the doctor is not able or available, for any reason, to certify the death;
- the cause of death is unknown;
- the death occurred during an operation or before recovery from the effects of an anaesthetic;
- the death occurred at work or was due to industrial disease or poisoning;
- the death was sudden and unexplained;
- the death was unnatural;
- the death was due to violence or neglect;
- the death was in other suspicious circumstances; or the death occurred in prison, police custody, or another type of state detention.

Source: Guide to Coroners Services. Ministry of Justice. Accessed 2015.

In 2014, just under half (45%) of all deaths were reported to coroners in England and Wales. Coroners carried out post-mortems on 40% of all cases and inquests on 12% of the deaths reported to them. (Source: *Coroners Statistics 2014 England and Wales. Ministry of Justice Statistics Bulletin. London. Ministry of Justice. Accessed 2015*)

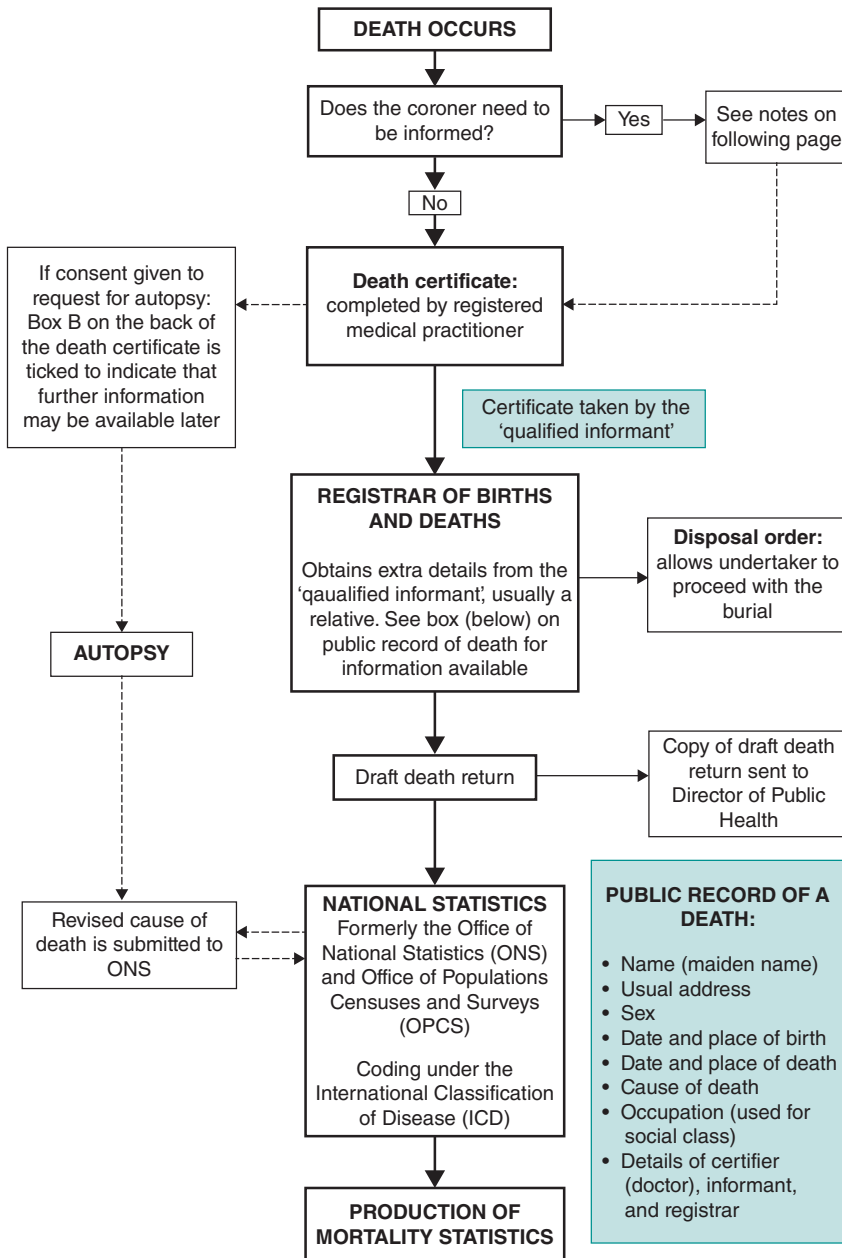


Figure 2.1.1 Process of recording deaths in England and Wales.

2.1.3 Suicide Among Men

Having looked at how information about deaths is collected, we now examine some actual data on deaths (known as *mortality data*) to find out what we can learn from simply presenting the information in a graph and making some comparisons among different years. In this example we look at mortality rates for suicide among men in England and Wales for the years 1980 to 2011, as shown in Table 2.1.1.

Table 2.1.1 Death rates per million population per year for 'suicide and self-inflicted injury' for all men, and selected age groups for men and women, England and Wales 1980 to 2011 (ICD codes E950-E959 for mortality statistics based on ICD-9, 1980–2000 and X60-X84 for mortality statistics based on ICD-10, 2001–2011)*.

Year	All men	Men 25–34	Men 65–74	Women 25–34
1980	110	130	182	42
1981	114	145	165	47
1982	115	145	173	46
1983	116	140	175	34
1984	118	140	170	38
1985	121	153	169	44
1986	116	146	176	39
1987	113	146	158	38
1988	121	155	162	38
1989	108	147	133	35
1990	117	160	136	38
1991	116	172	108	32
1992	118	159	129	37
1993	108	145	106	32
1994	113	181	109	33
1995	109	168	110	27
1996	103	163	98	37
1997	103	157	91	32
1998	109	186	90	34
1999	108	162	92	35
2000	100	150	99	33
2001	100	157	95	28
2002	98	148	82	35
2003	97	138	85	33
2004	102	142	93	33
2005	93	116	86	32
2006	97	122	90	32
2007	93	125	81	30
2008	99	131	84	35
2009	100	126	87	29
2010	96	107	81	26
2011	104	111	92	27

*See further explanation of ICD codes in the text.

Source: Office for National Statistics: Cause (Series DH2 up to 2006 and Series DR 2006 onwards available online) <http://www.ons.gov.uk>

International Classification of Disease (ICD) codes are a standardized set of codes used for classifying diseases. The tenth revision of the *International Statistical Classification of Diseases and Related Health Problems* (ICD-10) was introduced on 1 January 2001. ICD-10 replaced ICD-9, which had been used between 1979 and 2000. As a result, ICD-10 mortality data are not directly comparable with ICD-9. However, overall suicide data are not affected by the change (although codes do vary at the third- and fourth-digit level). If you wish to read about the effect of the introduction of ICD-10 in more detail, see Rooney *et al.* (2002).



Self-Assessment Exercise 2.1.3

1. Plot the data in Table 2.1.1 for all men and for each of the two male age groups separately for each year using computer software, e.g. MS Excel, or graph paper if you are not yet familiar with the software. We will examine the data for women shortly.
2. Describe what you see for the three groups (all men, 25 to 34 years, and 65 to 74 years), over the years 1980 to 2011.

Answers in Section 2.8

This exercise illustrates variation in an important measure of health over time, and it starts to raise questions about the reasons for changes that we can see have occurred since 1990, especially since these are so different for younger and older men. Do not spend any more time at this stage on why there are differences between the two age groups, as we will return to the interpretation of these changes shortly.

2.1.4 Suicide Among Young Women

We now look at the same information on suicide rates for young women.



Self-Assessment Exercise 2.1.4

1. Using data for women aged 25 to 34 years (Table 2.1.1), add the rates to your male mortality graph.
2. What do you observe about the trends for women aged 25 to 34 years and men in the same age group?

Answers in Section 2.8

2.1.5 Variations in Deaths of Very Young Children

Here is another striking example of how an important health indicator, in this case the *infant mortality rate* (IMR), varies among different groups in society. The IMR for a given area (e.g. country, region, or district) is defined as follows:

$$\text{IMR} = \frac{\text{The number of deaths per year occurring within the first year of life}}{\text{Total number of live births in the year}} \times 1000$$

The IMR is one of the best indicators we have of the health status of a population, and it generally relates quite closely to the level of socioeconomic development. This is true for both developed and developing countries, and for this reason it is widely used as an indicator of health and development, either on its own or combined with other information.

Table 2.1.2 is taken from Child Mortality Statistics: Childhood, Infant and Perinatal in England and Wales 2013. It shows IMRs for children as defined by the Standard Occupational Classification 2010 based on the highest (most socioeconomically advantaged) occupational level of either parent. The data in this table include parents who are married, in civil partnerships, or jointly registered as living at the same address. This information is recorded by the registrar of births and deaths when the parents go to register the death (Figure 2.1.1).

Table 2.1.2 Infant mortality rates (IMR) for live-born children less than 1 year of age, England and Wales 2013.

Socioeconomic Classification*	IMR per 1,000 live births
1.1	1.6
1.2	2.2
2	2.2
3	2.4
4	4.3
5	4.7
6	4.5
7 and 8	7.9
Not classified	5.2

*Classification of categories is derived from Standard Occupational Classification 2010 as follows: 1. Higher Managerial, administrative, and professional occupations; 1.1 Large employers and higher managerial and administrative occupations; 1.2 Higher Professional occupations; 2 Lower managerial, administrative, and professional occupations; 3 Intermediate occupations; 4 Small employers and own account workers; 5 Lower supervisory and technical occupations; 6 Semi-routine occupations; 7 Routine occupations; 8 Never worked and long-term unemployed. 'Non classified' comprises a mixed group including full-time students, occupations not stated or inadequately described, and employment not possible to classify.

Source: Child Mortality Statistics: Childhood, Infant and Perinatal, 2013. Last accessed 2015.

The category classifications are explained in the footnotes underneath the table.



Self-Assessment Exercise 2.1.5

1. What is the ratio of the IMR for children in group 7 and 8 compared to group 1.1?
2. Briefly list the reasons that you think could explain the striking variations in IMR across these socioeconomic classification groups in England and Wales.

Answers in Section 2.8

Summary

- The accuracy of mortality statistics (including the cause of death and information about the deceased person) depends on the accuracy of death certificate completion.
- Autopsies and coroner's inquests provide additional information for those deaths on which they are performed.
- Variations in levels of mortality, e.g. by time, or between groups defined by socioeconomic circumstances, can provide valuable insight into possible causes of disease and death.

2.2 Descriptive Epidemiology

2.2.1 What is Descriptive Epidemiology?

In the last section we looked at some dramatic differences between population groups defined by sex and socioeconomic circumstances in trends in suicide rates over time and in trends in infant mortality. These two examples show how a relatively simple study of data over time, and between the sexes (suicide) or socioeconomic groups (IMR), which we can term differences between groups of persons, can begin to throw light on very important issues that might determine health. This is the essence of *descriptive epidemiology*, in which we can begin to learn about the determinants of population health by exploring the patterns of variation in measures of health (in this case the suicide and infant mortality rates) and factors that tell us something about what is causing these variations. So far, we have looked at variations by time and person, and we will shortly add a third dimension to this: variation by place.

Descriptive epidemiology is the study of variations in measures of population health by *time, person, and place*.

So what seems at first sight to be a very simple investigative technique nevertheless sets us thinking about fundamental social, economic, and environmental issues that are among the most important influences on health and the incidence of diseases.

2.2.2 International Variations in Rates of Lung Cancer

In this final part of our discussion of routine mortality statistics, we look at variations in death rates for lung cancer in two European countries. Table 2.2.1 shows the numbers of deaths and the death rates per 100,000 population from lung cancer among women aged 75 years and

Table 2.2.1 Numbers of deaths and death rates per 100,000 per year from lung cancer for women aged 75 years and older.

Country (year)	Number of deaths	Rate per 100,000/year
Austria (2012)	475	108.3
UK (2012)	7,587	254.5

Source: World Health Organisation Mortality Database (2016). Last accessed February 2016.

older for Austria and for the UK for 2012 taken from the World Health Organisation (WHO) Mortality Database.



Self-Assessment Exercise 2.2.1

1. Describe the differences between the data for the two countries.
2. What reasons can you think of that might explain these differences? In answering this question, it is useful to think about whether observed differences relate to chance, are an artefact, or are real.

Answers in Section 2.8

In the two examples of suicide data we have been considering, the data were collected by the system outlined in Figure 2.1.1 and were for England and Wales. The systems used in different countries vary in a number of respects, and this should be borne in mind when comparing data across countries. Table 2.2.2 summarises key information on preparation of mortality statistics in the UK (England and Wales) and Austria, taken from the 2005 WHO Survey on mortality data.

Table 2.2.2 Preparation of mortality statistics in the UK and Austria.

Procedural aspects	UK	Austria
Percentage of deaths certified by medical doctor or coroner	100%	100%
Percentage of deaths occurring in hospital	70%–79%	30%–39%
Percentage of deaths for which autopsy performed	20%–29%	20%–29%
Follow-up enquiries to certifier	Yes 5%	Yes 2%
Coding procedure	Centrally coded	Centrally coded

Source: WHO Survey on mortality data (2005). Last accessed February 2016.



Self-Assessment Exercise 2.2.2

1. What differences are there between the two systems for collecting mortality data?
2. Do you think this could have influenced the reported lung cancer data? If so, in what way?
3. In the light of this new information, would you alter your views about the likely cause(s) of the large difference in lung cancer death rates for women aged 75 years and older in the two countries?

Answers in Section 2.8

2.2.3 Illness (Morbidity)

So far in this chapter, we have looked at data on deaths, which are termed *mortality* statistics. Although death is an important and useful measure of disease, especially where accurate information is available on the cause, we also want to know about episodes of disease not resulting in death. The term used for episodes of illness is *morbidity*.

Good information on mortality is more readily available than information on morbidity. This is because in most countries, it is a legal requirement that all deaths be certified by a doctor and officially registered. Thus, in a high-income country such as the UK, although there are some inconsistencies and inaccuracies in the way certificates are completed, virtually all deaths are recorded, and the information in the mortality statistics is of a reasonably high standard of accuracy (although this does vary by age and cause of death). By contrast, not all episodes of illness are recorded with anything like the level of attention that attends the certification of a death, and the recording that is done is determined primarily by the contact that the ill person has with the health system:

- Many illness episodes are not brought to the attention of the health-care system, and so are not recorded. This is a very important perspective to bear in mind, and it is one of the reasons that studies of the *prevalence* of illness often require population surveys (Chapter 4).
- Episodes of illness brought to the attention of primary care are often recorded inconsistently, although it is increasingly the case that in England, general practitioners record episodes of certain conditions through the Quality Outcomes Framework for a wide range of conditions (see Section 2.5), and the majority of consultations are recorded electronically. Although some of this information may be made available with permission, there may nevertheless be inconsistencies in methods of recording and between different systems.
- More-serious conditions requiring investigation and treatment in hospital have more-detailed records (attendances at accident and emergency, as outpatients, and as inpatients), but it is only the data transferred onto *patient information systems* that are potentially available for analysis.

2.2.4 Sources of Information on Morbidity

For these reasons, information on the majority of illness episodes in the population is not available for routine analysis in the way that all deaths are. Having said that, there are some sources that do provide this type of information, such as, for instance, on illness episodes seen in general practice. Up until 1992, the *National Morbidity Studies* were the main source of GP morbidity data for England and Wales. These surveys commenced in the 1950s and obtained information approximately every 10 years in selected practices. This was superseded by the General Practice Research Database, now renamed the Clinical Practice Research Datalink; you can read more about this in Section 2.5. There are also systems for the routine recording of some key disease types, two of the most important examples being the *notification of infectious diseases*, and the *registration of cancers*. We now examine some examples of infectious disease notification.

2.2.5 Notification of Infectious Disease

Despite the decline in infectious disease associated with improvements in living conditions in more-developed countries, serious diseases such as meningitis, tuberculosis and, food poisoning remind us of the continuing importance of infections and the need for vigilance and control. Timely information about cases of infectious diseases such as these is vital to the work of those responsible for control measures. In the UK, a system of notification for certain specified infectious diseases has been in operation for many years. The responsibility for notification rests with the doctor who makes the diagnosis, and this includes cases where the diagnosis is made by laboratory testing of blood or other samples. All diseases notifiable under the Health Protection (Notification) Regulations 2010 are listed in the accompanying box.

Notifiable Diseases		
Acute encephalitis	Infectious bloody diarrhoea	Severe acute respiratory syndrome (SARS)
Acute infectious hepatitis	Invasive group A streptococcal disease	Scarlet fever
Acute meningitis	Legionnaires' disease	Smallpox
Acute poliomyelitis	Leprosy	Tetanus
Anthrax	Malaria	Tuberculosis
Botulism	Measles	Typhus
Brucellosis	Meningococcal septicaemia	Viral haemorrhagic fever
Cholera	Mumps	Viral hepatitis
Diphtheria	Plague	Whooping cough
Enteric fever	Rabies	Yellow fever
Food poisoning	Rubella	
Haemolytic uraemic syndrome		


Source: Public Health England (2010) Notifiable diseases as causative organisms, how to report. Last accessed February 2016.

Not surprisingly, doctors do not always get around to completing the notification forms, especially for the less-serious conditions. This means that the statistics tend not to be complete (not all cases are notified), and this is termed *underreporting*. Despite this, levels of underreporting remain fairly constant over time, so it is possible to look at trends with a fair degree of confidence. We now look at three examples: food poisoning, meningitis, and mumps.

Communicable Diseases Currently Notifiable in the UK

Time Trends in Food Poisoning

Table 2.2.3 shows the numbers of notified cases of food poisoning for England and Wales for each of the years 1984 to 2009, together with the population numbers. Notifications include notified cases and also cases otherwise ascertained. Cases otherwise ascertained were no longer collected after week 35 of 2010, so it is not possible to compare trends for 2010 onwards with previous years.



Self-Assessment Exercise 2.2.3

1. Calculate the crude incidence rates for food poisoning for 2001 onward expressed per 1,000 population per year, and enter these in the last column.
2. Plot the data (using computer software or on graph paper).
3. Comment briefly on any observed trend. What might be the explanations for this trend?

Answers in Section 2.8

Seasonal and Age Patterns in Communicable Disease

As well as looking for changes in disease patterns over time, it is also common to monitor incidence of disease according to the season and by age and sex. For example, incidence rates for meningitis demonstrate that age is a very strong determinant of susceptibility to meningitis (rates are significantly higher in children younger than one year and gradually decrease until adulthood); being male also appears to carry some additional risk in young children. Further, meningitis shows seasonal trends, peaking in the winter months, falling to its lowest level during the summer, and rising again towards the end of the year, probably due to people being

Table 2.2.3 Cases of food poisoning, and population (thousands), England and Wales, 1984 to 2009.

Year	Number of cases ¹	Population (1000s) ²	Crude rate/1,000/year
1984	20,702	49,713.1	0.42
1985	19,242	49,860.7	0.39
1986	23,948	49,998.6	0.48
1987	29,331	50,123.0	0.58
1988	39,713	50,253.6	0.79
1989	52,557	50,407.8	1.04
1990	52,145	50,560.6	1.03
1991	52,543	50,748.0	1.04
1992	63,347	50,875.6	1.25
1993	68,587	50,985.9	1.35
1994	81,833	51,116.2	1.60
1995	82,041	51,272.0	1.60
1996	83,233	51,410.4	1.62
1997	93,901	51,559.6	1.82
1998	93,932	51,720.1	1.82
1999	86,316	51,933.5	1.66
2000	86,528	52,140.2	1.66
2001	85,468	52,360.0	
2002	72,649	52,567.2	
2003	70,895	52,792.2	
2004	70,311	53,053.2	
2005	70,407	53,416.3	
2006	70,603	53,725.8	
2007	72,382	54,082.3	
2008	68,962	54,454.7	
2009	74,974	54,809.1	

Source: ¹Public Health England (2014) Statutory Notifications of Infectious Diseases (NOIDS). Last accessed February 2016.

² Mid-year population estimates available from Office for National Statistics. Last accessed February 2016.

indoors more in the colder weather and concurrent respiratory infections increasing the risk of transmission. Fortunately, more recently, the incidence of meningitis has fallen significantly. Data collated by England's Health Protection Agency demonstrate that following the introduction of the *Haemophilus influenzae B* (Hib), pneumococcal, and meningitis C vaccine in 2013, the numbers of annual cases of meningitis had fallen to around half of those seen 25 years previously.

A further example of a communicable disease with a seasonal trend is influenza. The Health Protection Agency monitors influenza (and other related respiratory illness) trends on a weekly basis so that if community transmission increases, targeted appropriate action can be taken, including promoting vaccination among vulnerable people, providing advice to the public on reducing transmission and action to be taken in the event of symptoms, and monitoring impact on providing health services.

We have now looked at some examples of variation by time in the incidence of communicable disease. In doing so, we have learned about how the diseases behave in the population, and we have started to think about factors determining these variations. This emphasises again the value of simple *descriptive epidemiology*, and how it can be a useful first step in any investigation.

2.2.6 Illness Seen in General Practice

Although much illness (morbidity) is never brought to the attention of the health system, general practice (in the UK) represents a very comprehensive first point of contact between episodes of illness in the population and the opportunity to make and record a diagnosis. Thus, although it can never provide true population *incidence* and *prevalence* (as not all cases of disease are presented to GPs), information on contacts with general practice is an important resource for morbidity data.

Case Study of Asthma

The UK, along with other developed countries, has seen an increase in the incidence (and prevalence) of asthma since the 1980s. It might be thought that routine statistics should provide a simple enough explanation for this increase, but this is not the case. One problem with following trends in health data over a number of years is that the way a given condition is described can change. Changes in definitions can result from increasing knowledge about the condition and its relationship to other similar problems, or it can result from a change in fashion about what is the proper term for a condition.

Changes such as these are certainly true for asthma. Since the 1980s, the treatment of asthma has improved greatly. In the past, fear of this disease and the lack of adequate treatment led to unwillingness to discuss the diagnosis, especially for children, for whom the term ‘wheezy bronchitis’ was often used.

As knowledge of asthma has increased, doctors have become more aware of the need to make a diagnosis earlier, institute effective treatment, and help the child and his or her parents to understand how to manage the condition in a way that minimises the adverse effects on the child’s life and education. The term *diagnostic fashion* is used to describe how these various influences can change the way a disease or condition is described in medical practice over time or among different places. Although the National Morbidity Studies were discontinued many years ago, they provide a good opportunity to study historical trends in asthma in conjunction with the other conditions with which it may have been confused over the years. Table 2.2.4, reproduced from data in the 1981 report, shows annual consultation rates per 1,000 persons at risk of acute bronchitis, asthma, and hay fever.

Table 2.2.4 Annual consultation rates per 1,000 persons at risk.

Condition	1955–56	1971–72	1981–82
Acute bronchitis	48.9*	59.1	58.2
Asthma	—**	9.6	17.8
Hay fever	—**	11.0	19.7

*Includes bronchitis unspecified.

**Data not available separately for these two conditions.

Source: Third National Morbidity Study (1981), OPCS.

Even this source of information does not provide a full picture. For one thing, in the first (1955–56) report, data for asthma and hay fever were not available separately. Nevertheless, there is some important information here, which is examined through this next exercise.



Self-Assessment Exercise 2.2.4

1. Make brief notes on the trends in consultation rates for the three conditions.
2. Does this help answer the question as to whether the change in asthma rates has resulted from a change in diagnostic fashion? That is, some cases of what used to be called acute (wheezy) bronchitis in the past are now (more properly) termed asthma?
3. In the light of your observations, what do these data suggest happened to the true rates of asthma over the period studied?

Answers in Section 2.8

This example has illustrated that for a given problem, any one routine data source can provide some answers, but some key issues remain uncertain and incomplete. Valuable though these routine data sources are, there is still often a need to examine the research question more specifically and precisely through the various study designs examined in subsequent chapters. However, simple descriptive epidemiological analysis can be very powerful and can lead to policy changes. Section 2.6, on the health effects of the smogs (air pollution) in London during the 1950s, provides a good example of this.

Summary

- The key to descriptive analysis is the presentation and interpretation of variations in measures of health and health determinants by time, place, and person.
- When trying to understand the reasons for variations, it is useful to think whether the observed patterns are due to chance, artefact, or real effects.
- Good morbidity data are generally harder to come by than mortality data, but very useful information is nevertheless available from sources such as infectious disease notifications, cancer registration, and GP databases such as The Clinical Practice Research Datalink. Some of the important survey-based sources in England are considered in Section 2.5.

2.3 Information on the Environment

2.3.1 Air Pollution and Health

The influence of the environment on health is becoming an increasingly important social and political issue. Air pollution is one such environmental concern, due to increased risk of respiratory and cardiovascular illness and death.

2.3.2 Routinely Available Data on Air Pollution

Local government environmental health departments in the UK have carried out air-quality monitoring for many years, and in Section 2.6 we look at some of the historical information

recorded at the time of the serious smogs (mixtures of smoke from coal fires and fog) that occurred in London in the early 1950s. The UK government developed the National Air Quality Strategy, one component of which is the development of accessible air-quality data on key pollutants derived from national monitoring sites. These data are available on the Internet at <http://uk-air.defra.gov.uk> (Last accessed 1 July 2017), where information is updated hourly (Table 2.3.1).

Table 2.3.1 Concentrations of PM₁₀ in µg/m³ from selected UK automatic monitoring sites at midday on a typical working day.

Location	Concentration (µg/m ³)	Location	Concentration (µg/m ³)
Rural particulates (PM₁₀)			
Northern Ireland	1.0	Rochester	22.0
Urban particulates (PM₁₀)			
Edinburgh centre	0.4	Birmingham east	15.5
Glasgow centre	7.0	Birmingham centre	15.7
Newcastle centre	–	Thurrock	14.6
Belfast centre	14.3	London Bloomsbury	29.4
Middlesbrough	12.2	London Bexley	23.2
Leeds centre	18.0	London Hillingdon	30.7
Hull centre	15.6	London Brent	25.0
Stockport	–	Sutton roadside	24.8
Bury roadside	38.0	London Eltham	–
Manchester Piccadilly	16.6	London Kensington	24.9
Bolton	–	Haringey roadside	–
Liverpool centre	15.2	Camden kerbside	32.8
Sheffield centre	–	Swansea centre	23.0
Nottingham centre	13.0	Cardiff centre	15.0
Leicester centre	18.5	Port Talbot	20.1
Wolverhampton centre	16.2	Bristol centre	13.4
Leamington Spa	–	Southampton centre	30.0

The concentrations of a number of different pollutants are measured at various monitoring sites, which cover towns and cities in the UK. The pollutant we will look at is termed PM₁₀.

This stands for particulate matter of aerodynamic diameter 10 micrometres (µm) or less (a micrometre is one millionth of a metre). This pollutant, one of the most harmful components of smoke, is emitted by vehicle engines, burning of coal and other solid fuels in power stations, and other industrial processes. The importance of the size is that the smaller the particle, the farther it can penetrate into the lungs, and very small particles can cross from the lungs to the circulation and thereby have systemic effects on the body.

The data in Table 2.3.1 are concentrations of PM₁₀ recorded at each site at a particular point in time. The concentrations are measured in micrograms per cubic metre (one-millionth of a gram per cubic metre of air), written as µg/m³. The locations are divided into rural and urban. There are only two rural locations, so we do not have much information about air pollution in rural areas. But what can the rest of the data tell us about levels of PM₁₀ in urban areas? How do concentrations vary across the country? What is a ‘typical’ value? How can we compare PM₁₀

concentrations from one time, or day, with another? We can gain some idea of the situation by looking at the individual data values listed, known as the *raw data*, explored in the next exercise.



Self-Assessment Exercise 2.3.1

1. Find the locations with the lowest and highest concentrations of PM_{10} .
2. Identify which locations did not record a level of PM_{10} at this time. What percentage of the urban locations is this?

Answers in Section 2.8

We can answer some of our other questions by picturing and summarising the data. This is worked through in the next section, where we will look at these air-pollution data in more detail.

2.4 Displaying, Describing, and Presenting Data

2.4.1 Displaying the Data

In this section we will use the air-pollution data presented in Table 2.3.1 to introduce some basic statistical methods that are fundamental to the understanding and interpretation of information.

Taking the air pollution example, although we need the individual data values, also called *observations* or *observed values* of PM_{10} , to see the concentration of PM_{10} for any particular location, it is difficult to get the overall picture from a list of numbers. The first step in analysing the information in a dataset is usually to picture the data in some way. Figure 2.4.1 shows one way of doing this.

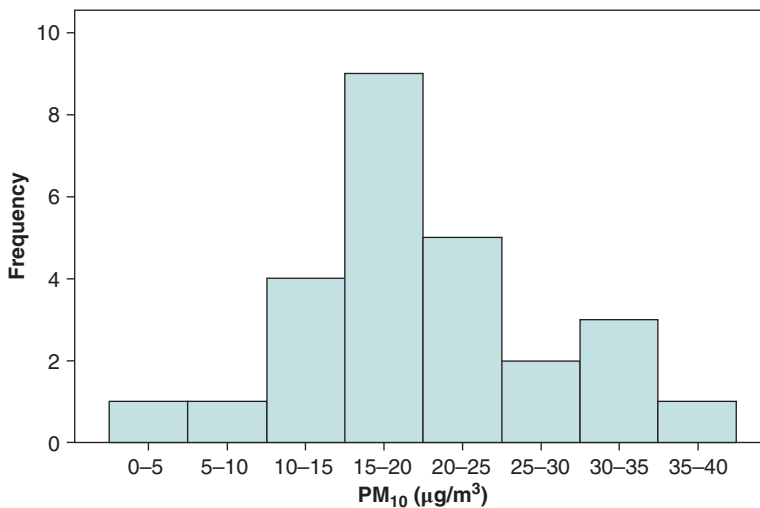


Figure 2.4.1 Histogram showing frequency distribution of PM_{10} .

This illustration is known as a **histogram** and presents the **frequency distribution** of the data. In this example, the histogram shows the number (or **frequency**) of air-pollution monitoring sites with PM_{10} concentrations in the intervals 0–5, 5–10, 10–15, and so on. Notice that there are no gaps between the bars representing frequency in the histogram. This is because PM_{10} concentration is a **continuous** measurement: Theoretically, the concentration can be any value greater than or equal to 0, such as 0.5 or 23.65.

2.4.2 Calculating the Frequency Distribution

The frequency distribution is a table of numbers that shows how many of the data values are within each of a number of **intervals**. The frequency distribution of PM_{10} concentrations that we displayed in the histogram is the number of measurements in each of the intervals 0–5, 5–10, and so on. The frequency distribution of a set of data is not unique, because the frequencies for each interval depend on how we choose the intervals. So to find the frequency distribution of a set of data, we first need to decide how to divide the data into intervals. Computer software can do this for you automatically, though you can also specify the intervals. To calculate the frequency distribution of the PM_{10} data by hand, the following procedure can be used.

Divide the PM_{10} Scale into Intervals

Between 5 and 20 intervals is a reasonable number to use. The smaller the number of observations, the fewer intervals we want (generally, not more than 10 intervals if we have fewer than 50 observations). The data values for the urban monitoring sites vary from 0.4 to 38.0 – a range of 37.6. To include all the data, we could start the first interval from 0.4, and, to achieve 10 intervals, each interval would have a width of $37.6/10 = 3.76$, or 4 to the nearest whole number. The intervals would then be

0.4 – 4.4 4.4 – 8.4 8.4 – 12.4 ... and so on

These are very messy numbers! We do not need to start at the smallest value, as long as the first interval starts at a smaller value. And dividing the data into any particular number of intervals, such as 10, is less important than having intervals of an ‘easy’ size, such as 5 or 10. So let’s start again, with the first interval starting at 0 and each interval having width 5:

0 – 5 5 – 10 10 – 15 15 – 20 20 – 25 25 – 30 ... etc.

Count How Many Data Values are within Each Interval

This is straightforward until we get to London Brent, which had a PM_{10} concentration of $25.0 \mu\text{g}/\text{m}^3$. Does this value fall in the interval 20–25 or 25–30? We can only count each value once – it cannot be in two intervals, so the intervals must not overlap. It is usual to define each interval to contain its lower boundary, but not its upper. So the interval from 20 to 25 includes 20.0, but not 25.0. Since the data are recorded to one decimal place, the largest value that can be included in 20–25 is $24.9 \mu\text{g}/\text{m}^3$. To emphasise this convention, the intervals may be written in one of the following alternative ways:

Alternative 1	0–	5–	10–	... etc.
Alternative 2	0–4.9	5–9.9	10–14.9	
Alternative 3	0–5–	5–10–	10–15–	

Seven urban locations do not have a value recorded, so we cannot include them in any of the intervals, but they should be noted. Now we can write down the complete frequency distribution as shown in Table 2.4.1. A frequency distribution is often just called a distribution.

Table 2.4.1 Frequency distribution of PM_{10} concentrations in 34 urban locations.

PM_{10} $\mu\text{g}/\text{m}^3$	Frequency
0–	1
5–	1
10–	5
15–	9
20–	5
25–	2
30–	3
35–	1
Total*	27

*Seven locations did not record PM_{10} at this time.

Summary: Calculating a Frequency Distribution for Continuous Data

- Divide the range of the data (generally) into 5 to 20 convenient intervals of equal width.
- Avoid overlaps in terms of how the interval ranges are written (i.e., 0–4.9, 5–9.9, etc.).
- Count how many observations fall within each interval.



Self-Assessment Exercise 2.4.1

1. Calculate the frequency distribution of the urban-centre PM_{10} data using the intervals 0–4, 4–8, etc.
2. Draw a histogram of this distribution and compare it with the histogram with wider intervals (Figure 2.4.1).

Answers in Section 2.8

Using too few intervals results in a histogram with few peaks and troughs, which does not tell us very much about how the data values vary. Using too many intervals means that several of them may be empty or might contain only one observation. Although more detail can be seen, the histogram ceases to be a summary, and we do not get a very useful overall impression of the data. Choosing an appropriate number of intervals for the frequency may seem to be something of an art rather than a science. Experience and following the general rule of having between 5 and 20 intervals should result in a useful histogram, and the number of intervals can always be adjusted if the first attempt is not suitable.

2.4.3 Describing the Frequency Distribution

We have seen that a list of raw data values is not very useful by itself. We generally want to summarise the data in a short description and a few numbers so that we can present the important features of whatever we are measuring or so we can compare these data with another dataset. We can summarise the information in a set of data by describing the *shape, location, and spread* of the distribution.

Shape

We can see the *shape* from the histogram. We are interested in the overall shape, not the detail that results, for example, from having only a (relatively) small number of observations – we have only 34 monitoring sites.

You can imagine enclosing the histogram in a smooth curve, smoothing out the small ups and downs, to see the general shape. In Figure 2.4.2, this has been done (in SPSS) by fitting the normal distribution curve.

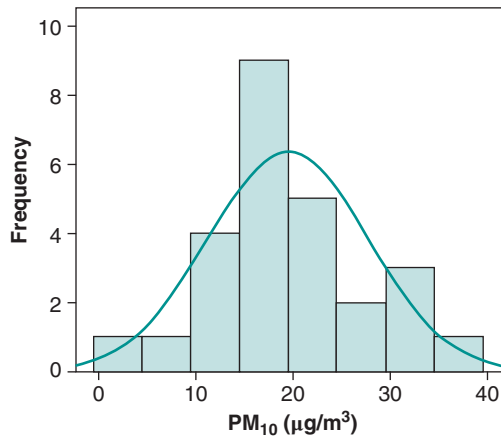


Figure 2.4.2 Histogram for PM₁₀ showing a symmetric frequency distribution.

We describe the overall shape of the distribution in terms of how many peaks it has, whether it is approximately *symmetric* or *skewed* in one direction, and whether there are any marked deviations from the overall shape.

Peaks are called *modes*. A distribution with one distinct peak is called *unimodal*; with two distinct peaks, it is called *bimodal*. The distribution of the PM₁₀ values we are studying here is unimodal (Figure 2.4.2). The fact that the interval 30–35 has one more observation than the interval 25–30 is not important. It is sometimes difficult to decide whether a distribution is truly bimodal. It may help you to decide by knowing that most distributions of data are in fact unimodal, and that if a distribution really is bimodal, there is probably an explanation for this. For example, the histogram in Figure 2.4.3 illustrates the weights of a group of 92 students.

This distribution appears to have two distinct peaks. With the additional information that these are the weights of a mixed group of students (57 males and 35 females), we can be confident that this is a truly bimodal distribution: We would expect male students to be heavier than female students on average. In fact we have two overlapping (unimodal) distributions, one for female students and the other for male students.

The second feature of the distribution in which we are interested is whether it is symmetric or skewed (Figure 2.4.4). A distribution is *symmetric* if we can mentally divide it into two halves

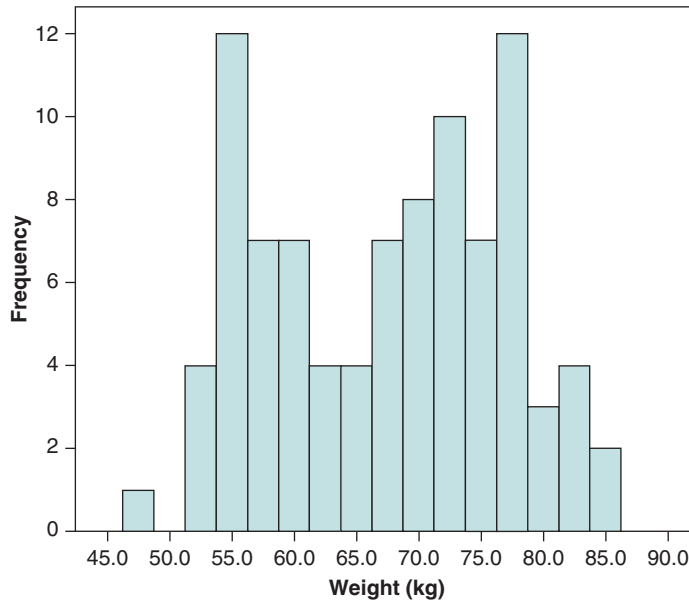


Figure 2.4.3 Histogram showing the distributions of student weights; in fact, there are two overlapping distributions, one for male students and one for female students – see text for explanation.

that are approximate mirror images. It is *skewed* to the right (also called *positively skewed*) if the right tail (that is, the higher values) is much longer than the left tail. Again, we do not insist on exact symmetry in giving a general description of the shape of a distribution.

Figure 2.4.4 shows two frequency distributions. The distribution of length of hospital stay (Figure 2.4.4 (a)) is right skewed: The larger values are more spread out than the smaller values, giving a long right tail. The distribution of height (Figure 2.4.4 (b)) is approximately symmetric. In both distributions, there are gaps; these gaps are small and not important in describing the overall shape.

Most distributions of data are symmetric or right skewed. If the left tail is longer than the right tail, the distribution is *left skewed* (also called *negatively skewed*). Left and right refer to the sides of a histogram.

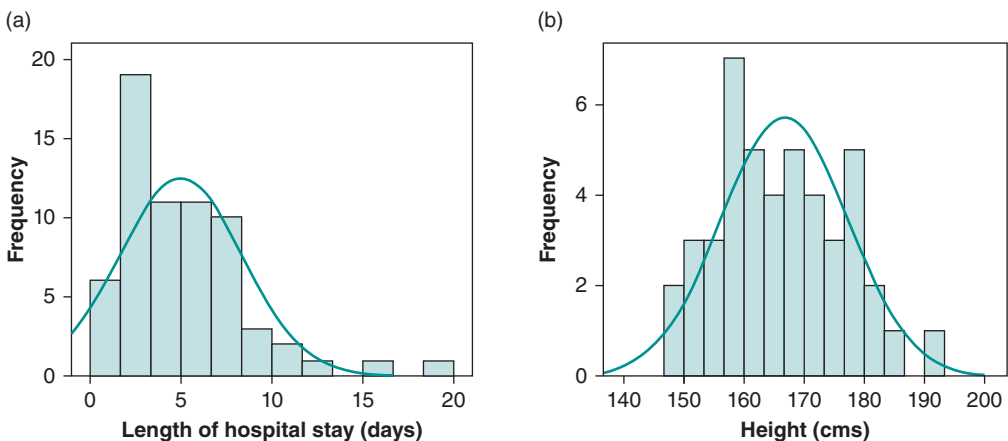


Figure 2.4.4 (a) A positively (right) skewed distribution; (b) a symmetric distribution.

Lastly, in describing the shape of the distribution, we look for any marked deviations from the overall shape. Is there a large gap? Is there an isolated value (or values) far from the majority of data values? An individual observation that falls well outside the overall pattern of the data is called an **outlier**. An outlier can arise in a number of ways. The value may have been measured or recorded incorrectly. Alternatively, it could be a correct value that is simply unusual, in which case it may be giving us important information. Apparent outliers should never be dismissed (see the story about data suggesting a hole in the ozone layer, below). If possible, we should find out how the value arose.

In 1985, British scientists reported a hole in the ozone layer of the Earth's atmosphere over the South Pole. This was disturbing, because ozone protects us from cancer-causing ultraviolet radiation. The report was at first disregarded, because it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down showed nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used to analyse the data had automatically set these values aside as suspicious outliers, and not included them in the analysis. Readings back to 1979 were reanalysed and showed a large and growing hole in the ozone layer. Fortunately, although the outliers were omitted from the initial analysis, they were not discarded. It was therefore possible to reanalyse the data.

Reported in the New York Times [date N/A]

In Figure 2.4.5 are some more data reporting student weights; these illustrate another example of an outlier and how this should be dealt with. Most of the data have an approximately symmetric distribution, but one value (in the interval 107.5–112.5 kg) is far from the rest of the data. Is this the weight of a particularly heavy student, or has the value been recorded or entered into the computer incorrectly? This value should be checked if possible, and any error corrected. In fact, a weight of around 110 kg is by no means impossible. Although it is much heavier than the other weights, this is not exceptionally heavy. If we cannot check the value, we should accept it as being correct. Outliers should never be dismissed or ignored simply because they spoil the look of the data!

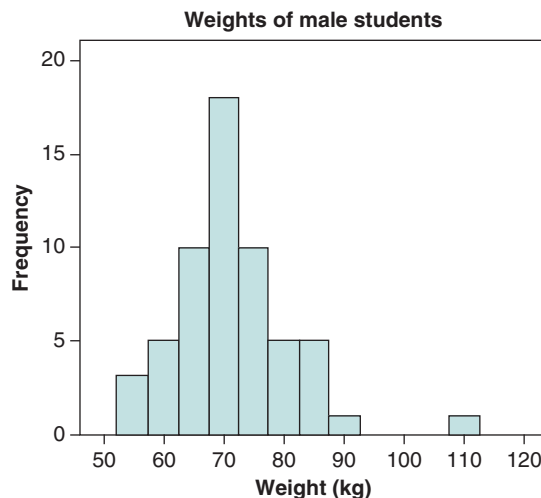


Figure 2.4.5 An outlier?

Summary: Questions in Describing the Shape of a Distribution

- Is the distribution unimodal or bimodal?
- Is it symmetric or skewed?
- Are there any large gaps or outliers? If there are outliers, are they likely to be erroneous, or correct but unusual?

**Self-Assessment Exercise 2.4.2**

Look at Figure 2.4.1 (or the version in Figure 2.4.2) and describe the shape of the PM_{10} distribution.

Answers in Section 2.8

Location

We call the value around which a distribution is centred its location. We can roughly estimate this value from a histogram; this is easiest for an approximately symmetrical distribution. The central value of the distribution of PM_{10} values (Figure 2.4.1) is around $20 \mu\text{g}/\text{m}^3$. The central value of the distribution of height (Figure 2.4.4(b)) is around 167 cm. Locating the centre of the distribution by eye is rather imprecise. To be more precise, we need to define what we mean by 'centre'. There is in fact more than one definition, resulting in different measures of location. These are the *mode, median, and mean*.

Mode

We have already met one measure of location in describing the shape of the distribution: The *mode* is the value that occurs most frequently. We might say it is typical of the sorts of values that occur. When continuous data are grouped in a frequency distribution, the mode is the interval with the highest frequency, sometimes called the modal group, and this is the peak of the histogram. We can see from the histogram of the PM_{10} data (Figures 2.4.1 and 2.4.2) that most values lie between 15 and $20 \mu\text{g}/\text{m}^3$. The mode (or modal group) is $15\text{--}20 \mu\text{g}/\text{m}^3$. Note that the modal group depends on how the intervals of the frequency distribution are chosen. The number of modes is important in describing the shape of a distribution, but the mode itself is not generally used to describe the location of a distribution of continuous values.

Summary: Mode

- The mode of a distribution is the value, or group of values, that occur most often.
- The mode corresponds to the highest peak of a histogram.

Median

The simplest measure of location is literally the central, or middle, value of the distribution. This is called the *median*, and half the observations fall above it and half below it. The median is often denoted by M . To find the median, we need to sort the data into ascending order. For example, the median of the values 12, 7, 6, 9, and 14 is 9.

Median of five values:

6 7 9 12 14

We have 27 observations of urban PM_{10} concentration (Table 2.4.2). The middle value of 27 is the 14th value (this is the $[(27 + 1)/2]$ th value). Arranging the observations in ascending order and counting 14 from the smallest, we find the median concentration is $16.6 \mu\text{g}/\text{m}^3$, for Manchester Piccadilly.

Table 2.4.2 Finding the median value for PM_{10} data.

Location	PM_{10}	Order	
Edinburgh centre	0.4	1	
Glasgow centre	7.0	2	
Middlesbrough	12.2	3	
Nottingham centre	13.0	4	
Bristol centre	13.4	5	
Belfast centre	14.3	6	
Thurrock	14.6	7	
Cardiff centre	15.0	8	
Liverpool centre	15.2	9	
Birmingham cast	15.5	10	
Hull centre	15.6	11	
Birmingham centre	15.7	12	
Wolverhampton centre	16.2	13	
Manchester Piccadilly	16.6	14	← middle value is 16.6
Leeds centre	18.0	15	
Leicester centre	18.5	16	
Port Talbot	20.1	17	
Swansea centre	23.0	18	
London Bexley	23.2	19	
Sutton Roadside	24.8	20	
London Kensington	24.9	21	
London Brent	25.0	22	
London Bloomsbury	29.4	23	
Southampton centre	30.0	24	
London Hillingdon	30.7	25	
Camden kerbside	32.8	26	
Bury roadside	38.0	27	

If we have an odd number of observations, the median is the middle value. If we have an even number of observations, the median is defined to be halfway between the *two* middle values. For example, the median of the values 12, 7, 6, 9, 14, and 18 is halfway between 9 and 12; that is, 10.5.

6	7	9	12	14	18
$\text{Median} = \frac{9 + 12}{2} = 10.5$					

Summary: Median

The median M of a distribution is the middle value. To find the median:

1. Arrange the observations in order of size from smallest to largest.
2. If the number n of observations is odd, the median is the middle observation. This is the $(n + 1)/2$ th observation, counting from the smallest observation.
3. If the number n of observations is even, the median is halfway between the middle two observations. This is the average of the $n/2$ th and $(n/2 + 1)$ th observations.

The median is usually stated to the accuracy of the data, or to one more decimal place.

Although the *median* is the simplest measure of location, it is not the most commonly used. The most commonly used is the *mean*.

Mean

The *mean* is the arithmetic average of the observations, and is calculated by adding together all the observations and dividing by the number of observations. The mean of the urban PM_{10} observations (Table 2.3.1) therefore is

$$\frac{0.4 + 7.0 + \cdots + 30.0}{27} = \frac{523.1}{27} = 19.37 \text{ } \mu\text{g}/\text{m}^3$$

The mean is denoted by \bar{x} (pronounced ‘x bar’), and the individual observations are denoted by x_1, x_2, \dots and so on. For example, the first observation in the urban PM_{10} dataset is 0.4, so $x_1 = 0.4$.

Summary: Mean

If we denote n observations by x_1, x_2, \dots, x_n , then the mean of the observations is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

This can also be written $\bar{x} = \frac{1}{n} \sum x_i$

The mean is usually stated to one more significant figure than the data.

The Greek letter Σ (capital sigma) is shorthand for ‘add up all the values’.

**Self-Assessment Exercise 2.4.3**

1. Use Table 2.3.1 to list the PM_{10} concentrations for the London locations. These include Sutton roadside, Haringey roadside, and Camden kerbside.
2. Find the mean and median of these values.

Answers in Section 2.8**The Median and Mean Compared**

The median is just one observation (or the average of two observations) from the distribution. The other observations do not have a direct effect on the median; only their relative values are important. The mean, however, is calculated from *all* the observations: They all contribute equally to the value of the mean. The largest observed value of PM_{10} concentration was

38.0 $\mu\text{g}/\text{m}^3$ at Bury roadside. Suppose this value was not 38.0 $\mu\text{g}/\text{m}^3$, but 138.0 $\mu\text{g}/\text{m}^3$ – an outlier (which we will assume is not erroneous). The median of the new distribution is still 16.6 $\mu\text{g}/\text{m}^3$ – the middle value has not changed. However, the mean changes from 19.37 $\mu\text{g}/\text{m}^3$ to 23.08 $\mu\text{g}/\text{m}^3$, a substantial increase of almost 20 per cent. All the observations are included in the calculation of the mean, so it is **sensitive** to a few extreme observations.

In this example, we may consider excluding the outlier of 138.0 $\mu\text{g}/\text{m}^3$ as atypical (the mean of the remaining 26 observations is 18.66 $\mu\text{g}/\text{m}^3$; the median is 16.4 $\mu\text{g}/\text{m}^3$). However, a skewed distribution with no outliers also pulls the mean towards its long tail. So, for a right-skewed distribution, the mean is greater than the median. If the distribution is left skewed, the median is greater than the mean. Because the mean is sensitive to extreme observations – that is, it cannot resist their influence – we say that it is not a **resistant measure** of location. The median is a resistant measure. A resistant measure is not strongly influenced by outliers or by changes in a few observations, no matter how large the changes are.

Median or Mean?

We have defined three measures of the location of a distribution: the mode, the median, and the mean. We usually choose between the mean and the median. These are calculated in different ways, and as we have seen for the PM_{10} data, they do not generally have the same value. They also have different properties: The median is resistant, whereas the mean is not. So which should we use to describe the central value of a distribution? The mean and the median both have the same value for symmetric distributions. In this situation and when simply describing the location of the distribution, it does not really matter which we use. However, the mean is generally used in preference to the median for several reasons:

- It contains more information than the median (it is calculated from all the data values).
- The data do not need to be ordered to find the mean.
- The mean has a number of mathematical properties that are very useful in further describing and comparing distributions.

For a skewed distribution, the mean and median differ, and the stronger the skew, the greater the difference. In this case, the median is often used to describe the location of the distribution. This is because it is resistant to the extreme values, and it is thought to represent better the centre of the distribution. However, although the median is a useful summary measure, we shall see that when we want to make comparisons between different distributions, we often use the mean even for skewed distributions because of its mathematical properties.

Summary: Measuring the Location of a Distribution

1. Location can be measured by the mean or median.
2. The mean is generally preferred to the median because of its mathematical properties.
3. The median may be a better measure of location for skewed distributions: It is resistant.



Self-Assessment Exercise 2.4.4

1. Which of the following are correct statements? The shape of a distribution can be described by
 - a. a histogram
 - b. the mean
 - c. the mode
 - d. the number of modes.

2. Which **two** of the following are correct statements?
- In a right-skewed distribution, the median is greater than the mean.
 - A skewed distribution is unimodal.
 - In a right-skewed distribution, most observations are less than the mean.
 - In a left-skewed distribution, the left tail is shorter than the right tail.
 - In a positively skewed distribution, the left tail is shorter than the right tail.
3. Figure 2.4.6 shows two distributions. In each case, at which point (a, b, or c) does the mean lie, and at which point does the median lie?

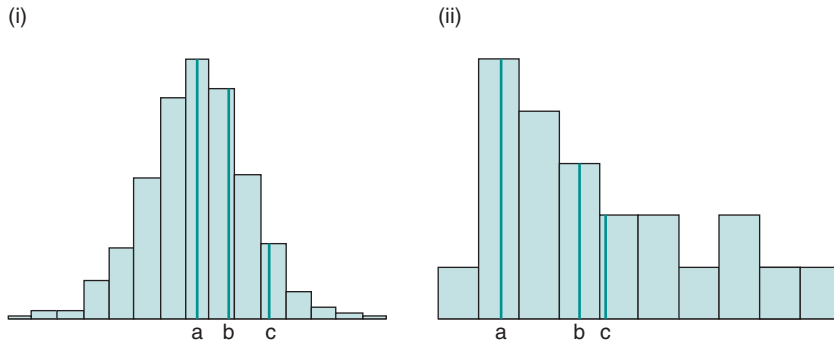


Figure 2.4.6 Mean and median.

Answers in Section 2.8

Spread

Lastly, we summarise the information in a set of data by a measure of how variable the data are, or in other words, how spread out they are. As well as the location of the distribution, it is important to know whether the values are bunched together or well spread out (Figure 2.4.7).

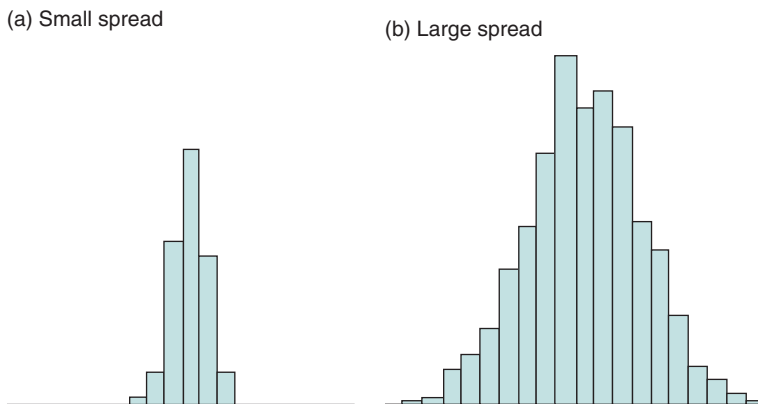


Figure 2.4.7 A distribution bunched together (a) or well spread out (b).

Again, there are several ways we can measure the distribution. The simplest measure is the **range**, which is simply the difference between the smallest and largest values. However, this

is not a very useful measure. Because it is calculated from the two extreme values, it is not a resistant measure, and it does not tell us how spread out the majority of the data values are.

Interquartile Range

An alternative measure that is resistant is the *interquartile range* (IQR). This is the range covered by the middle half of the data (Table 2.4.3).

Table 2.4.3 Finding the interquartile range for PM₁₀ data from urban monitoring sites.

Location	PM ₁₀	Order	
Edinburgh centre	0.4	1	
Glasgow centre	7.0	2	
Middlesbrough	12.2	3	
Nottingham centre	13.0	4	
Bristol centre	13.4	5	
Belfast centre	14.3	6	
Thurrock	14.6	7	←first quartile is 14.6
Cardiff centre	15.0	8	
Liverpool centre	15.2	9	
Birmingham east	15.5	10	
Hull centre	15.6	11	
Birmingham centre	15.7	12	
Wolverhampton centre	16.2	13	
Manchester Piccadilly	16.6	14	←middle value is 16.6
Leeds centre	18.0	15	
Leicester centre	18.5	16	
Port Talbot	20.1	17	
Swansea centre	23.0	18	
London Bexley	23.2	19	
Sutton roadside	24.8	20	
London Kensington	24.9	21	←third quartile is 24.9
London Brent	25.0	22	
London Bloomsbury	29.4	23	
Southampton centre	30.0	24	
London Hillingdon	30.7	25	
Camden kerbside	32.8	26	
Bury roadside	38.0	27	

One method for calculating the IQR is now described. The value below which 25 per cent, or one quarter, of the data values fall is called the *first (or lower) quartile*. The value below which 75 per cent of the data values fall (that is, one quarter of the data are above this value) is called the *third (or upper) quartile*. The second quartile is the median. The first and third quartiles

are usually denoted by Q_1 and Q_3 , and the interquartile range IQR is the difference between the first and third quartiles, $IQR = Q_3 - Q_1$.

To find the quartiles, we need to sort the data into ascending order, as we did to find the median. Remember that the median is the middle value of the distribution; that is, 50 per cent of the values fall below it and 50 per cent are above it. So the first quartile is halfway between the smallest value and the median, and the third quartile is halfway between the median and the largest value. Here are the ordered PM_{10} data for urban monitoring sites again:

There are 27 values, so the median is the 14th value. There are 13 values below the median and 13 values above, so the first quartile is the 7th value (the middle of 13), counting from the smallest, and the third quartile is the 7th value, counting back from the largest; that is, the 21st value. Using the ordered data given above, the first quartile is 14.6 and the third quartile is 24.9. The interquartile range is therefore $IQR = 24.9 - 14.6 = 10.3 \mu\text{g}/\text{m}^3$.

Summary: Interquartile Range

- The interquartile range is the difference between the first and third quartiles.
- The first and third quartiles (Q_1 and Q_3) are the values such that 25 per cent and 75 per cent, respectively, of the observations fall below them.
- To find the quartiles, first locate the median of the distribution. The first quartile is then the median of the values below the median, and the third quartile is the median of the values above the median.
- The interquartile range is $IQR = Q_3 - Q_1$.

Variance and Standard Deviation

The most commonly used measure of the spread of a distribution is the **standard deviation**, or its square, the **variance**. This is not a resistant measure. It is a measure of how spread out the data are around the mean, so we only use the standard deviation as the measure of spread in conjunction with the mean as the measure of location. The variance is calculated by adding up the squared deviations from the mean (which are the differences between each observation and the mean value), and dividing by the number of observations less one. This gives the average squared deviation. For the PM_{10} data, the mean is $19.37 \mu\text{g}/\text{m}^3$, and the deviations from the mean are $(0.4 - 19.37)$, $(7.0 - 19.37)$, $(14.3 - 19.37)$, and so on. The sum of squared deviations is therefore

$$(0.4 - 19.37)^2 + (7.0 - 19.37)^2 + (14.3 - 19.37)^2 + \dots + (38.0 - 19.37)^2 = 1778.0119$$

and the variance is $1778.0119/(27 - 1) = 68.38507308 \approx 68.39$.

The sum of squared deviations is shortened to **sum of squares**, and the number of observations less one ($n - 1$) is called the **degrees of freedom** (see box below). So we have that

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

If the observations are widely spread about the mean, the deviations are large, and the variance is large. If the observations are all close to the mean, the variance is small. To calculate the variance, we square the deviations from the mean, so the variance does not have the same units of measurement as the data.

Degrees of Freedom

The term **degrees of freedom** (df) is a measure of the number of independent pieces of information on which the precision of an estimate is based. The degrees of freedom for an estimate equals the number of observations minus the number of additional parameters estimated for that calculation.

Standard Deviation

Rather than measuring spread by the variance, we generally use the square root of the variance, called the **standard deviation**. The standard deviation measures spread about the mean in the original units.

The standard deviation of the urban PM₁₀ concentrations is 42.17. The variance and standard deviation are usually stated to one more significant figure than the data. However, when calculating them, full accuracy should be retained for intermediate calculations, and only the final answer should be rounded. The variance is denoted by s^2 and the standard deviation by s .



RS – Reference Section on Statistical Methods

This section summarises the mathematical formula for variance and the relationship between variance and standard deviation. It also provides an easier alternative formula for standard deviation, for use with calculators. The variance of n observations x_1, x_2, \dots, x_n , is

$$s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]$$

This can also be written

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The standard deviation s is the square root of the variance. The variance and standard deviation are usually stated to one more significant figure than the data. Many calculators calculate variance and standard deviation directly from the data. If you need to calculate them by hand, it is easier to use an alternative, but equivalent, formula, for the variance:

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 \right]$$

This formula uses the sum of all the data values ($\sum x_i$) and the sum of all the squared data values x . To calculate the variance of the PM₁₀ data using this formula, we need the sum of all the values, which is 523.1 (we found this when we calculated the mean in Section 2.4.3), and the sum of all the squared values. This is

$$0.4^2 + 7.0^2 + \dots + 38.0^2 = 11912.59$$

So the variance is

$$\frac{1}{26} \left[11912.59 - \frac{1}{27} (523.1)^2 \right] = 68.38507123 \approx 68.39 \mu\text{g}/\text{m}^3$$

as we found before. The slight difference in the unrounded value is because, when we first calculated the variance, we used the value of the mean rounded to two decimal places, not the exact value.

Measures of Spread for Mean and Median

We have seen that the interquartile range is related to the median (the quartiles are the medians of the two halves of the data, below and above the median), and that the standard deviation is related to the mean (it is a measure of variation about the mean). These measures of location and spread are always used in these combinations: ***mean and standard deviation*** or ***median and interquartile range***. It does not make sense to mix them.

Summary: Measuring the Spread of a Distribution

1. Spread can be measured by the standard deviation or the interquartile range.
2. The standard deviation is generally preferred.
3. The interquartile range may be more useful for skewed distributions, and is resistant.
4. The standard deviation is used when the mean is used as the measure of location.
5. The interquartile range is used when the median is used as the measure of location.



Self-Assessment Exercise 2.4.5

1. Find the interquartile range of the PM_{10} concentrations for the London locations that you listed in Exercise 2.4.3.
2. Use a calculator (if you wish to try calculating this yourself) or computer to find the standard deviation.

Answers in Section 2.8

Summary: Describing the Distribution of a Set of Data

- We describe a distribution by its shape, location, and spread.
- The shape indicates whether the distribution is unimodal or bimodal, symmetric or skewed in one direction, and whether there are any gaps or outliers.
- The location is the value around which the distribution is centred. This is measured by the mean or median. The median is resistant to extreme values and outliers; the mean is not resistant to such values.
- The spread is how variable the values are. This is measured by the standard deviation or the interquartile range. The interquartile range is a resistant measure.
- We generally use the mean together with the standard deviation to summarise the location and spread of an approximately symmetric distribution.
- The median and the interquartile range together are also useful for summarising the location and spread of a distribution, particularly where this is skewed, due to the resistant nature of these measures.
- The mean and standard deviation have mathematical properties that make them especially useful in further description and comparison of distributions, to which we will return in later chapters.

An Example of Summarised Data

This lengthy explanation of how we describe a distribution may lead you to think that summarising and describing data is a lengthy task. We should, however, be able to summarise the important features of a set of data quite briefly; otherwise it is not a summary! Here is a summary of the data on PM_{10} concentrations drawn from the text and exercises.

The concentration of PM_{10} in the air was recorded at 27 urban locations at a particular point in time. No data were available for a further seven locations with monitoring equipment.

The histogram (Figure 2.4.8, originally shown as Figure 2.4.1) shows the frequency distribution of the data. The distribution is unimodal and slightly right skewed, with no outliers. The mean PM_{10} concentration is $19.37 \mu\text{g}/\text{m}^3$, and the standard deviation $8.27 \mu\text{g}/\text{m}^3$.

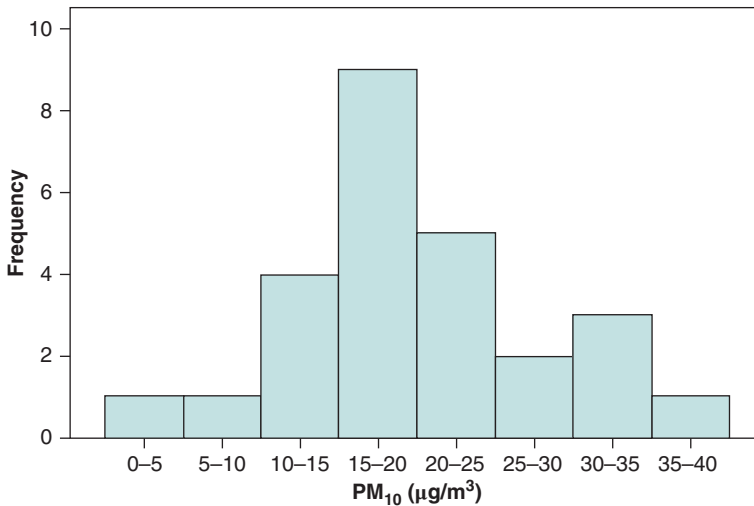


Figure 2.4.8 Histogram showing frequency distribution of PM_{10} .

Remember that throughout this section we have been dealing with *continuous* data: measurements that can take any value within some range (such as any value greater than zero). Later we will meet other types of data, which can take only particular values, such as number of children of a partnership (0, 1, 2, etc.) and blood group (one of four groups). We will use different techniques to describe and summarise this type of data. For now, in the next section, we will look at another important and useful way of displaying continuous data: the *relative frequency distribution*.



Self-Assessment Exercise 2.4.6

1. Using the data in Table 2.4.4, calculate the frequency distribution in intervals of width $5 \mu\text{g}/\text{m}^3$, and present the distribution in a table.
2. Prepare a histogram (using computer software or on graph paper by hand) of the frequency distribution and describe its shape.
3. Use a calculator (if you wish to try calculating this yourself) or computer software to find the mean and standard deviation of the data.
4. Write two or three sentences comparing the distribution of PM_{10} concentrations at 5 PM with that at midday (see Section 2.3.4 for an example of how to summarise data).

Table 2.4.4 Concentrations of urban PM₁₀ from UK automatic monitoring sites 5 hours after concentrations previously recorded.

Location	Concentration ($\mu\text{g}/\text{m}^3$)	Location	Concentration ($\mu\text{g}/\text{m}^3$)
Edinburgh centre	29.0	Birmingham east	31.0
Glasgow centre	43.0	Birmingham centre	31.0
Newcastle centre	33.0	Thurrock	28.0
Belfast centre	49.0	London Bloomsbury	34.0
Middlesbrough	–	London Bexley	25.0
Leeds centre	164.0	London Hillingdon	25.0
Hull centre	39.0	London Brent	28.0
Stockport	28.0	Sutton roadside	32.6
Bury roadside	–	London Eltham	23.6
Manchester Piccadilly	37.0	London Kensington	27.2
Bolton	–	Haringey roadside	33.5
Liverpool centre	34.0	Camden kerbside	46.3
Sheffield centre	45.0	Swansea centre	29.0
Nottingham centre	51.0	Cardiff centre	42.0
Leicester centre	33.0	Port Talbot	2.0
Wolverhampton centre	38.0	Bristol centre	42.0
Leamington Spa	32.0	Southampton centre	25.0

Answers in Section 2.8**2.4.4 The Relative Frequency Distribution**

In this section we will look at how histograms can also be used to display the *relative frequency distribution*, rather than the straightforward frequency distribution we have just discussed (Table 2.4.5). We will also look at situations where relative frequency distributions are the more useful approach.

Table 2.4.5 Relative frequency distribution of PM₁₀ concentrations in 34 urban locations at a particular point in time.

PM ₁₀ ($\mu\text{g}/\text{m}^3$)	Frequency	Relative Frequency (%)
0–5	1	3.7
5–10	1	3.7
10–15	5	18.5
15–20	9	33.3
20–25	5	18.5
25–30	2	7.4
30–35	3	11.1
35–40	1	3.7
Total*	27	99.9

*Seven locations did not record PM₁₀ at this time.

The relative frequencies are the percentages or proportions of the total frequency that are in each interval. For example, for the PM_{10} data (Table 2.4.1), there is one observation (Edinburgh $0.4 \mu\text{g}/\text{m}^3$) in the interval $0-5 \mu\text{g}/\text{m}^3$. This is one out of 27 observations; that is, 3.7 per cent of the observations. So, the relative frequency of the $0-5 \mu\text{g}/\text{m}^3$ interval is 3.7 per cent (Table 2.4.2). Similarly, the relative frequency of PM_{10} values in the interval $15-20 \mu\text{g}/\text{m}^3$ is $(9/27) \times 100\% = 33.3\%$:

Figure 2.4.9 is identical to our original histogram of the frequency distribution in Figure 2.4.1 apart from having a different scale on the vertical axis (percentage, instead of frequency/number).

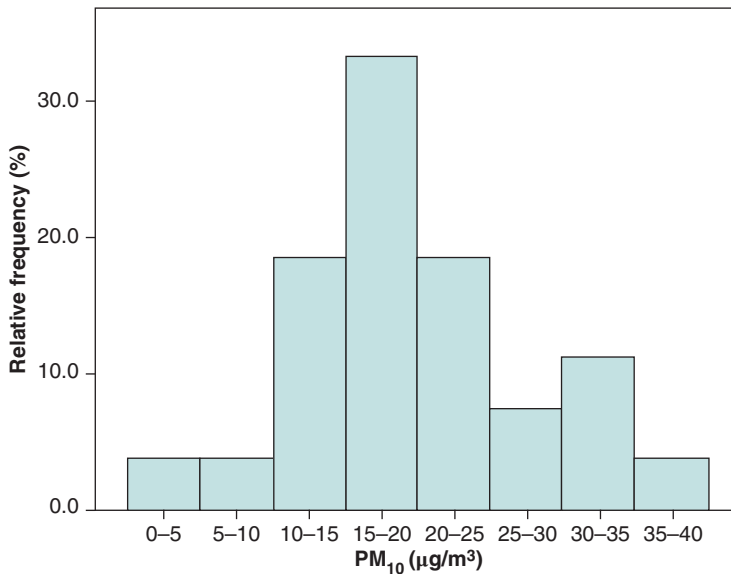


Figure 2.4.9 Relative frequency distribution of PM_{10} data.

This new histogram has the same shape as the original version, but the relative frequency is on the y -axis rather than actual frequency (numbers of observations). So what use is it? Relative-frequency histograms are very useful if we want to compare two or more distributions with different numbers of observations.

For example, if we want to compare the age distribution of Liverpool with that of England and Wales, the frequency histogram for the latter towers over the one for Liverpool because the population numbers are so much greater, and the bars of the histogram for Liverpool cannot be seen (Figure 2.4.10(a)).

By contrast, using relative frequency histograms (Figure 2.4.10(b)) in which the vertical scale shows the percentage of the population in each interval, we can easily compare the shapes of the two distributions. We can see that although the age distributions of Liverpool and of England and Wales were quite similar in 2001, Liverpool had a larger proportion of young adults, especially in the 20–24 age group.

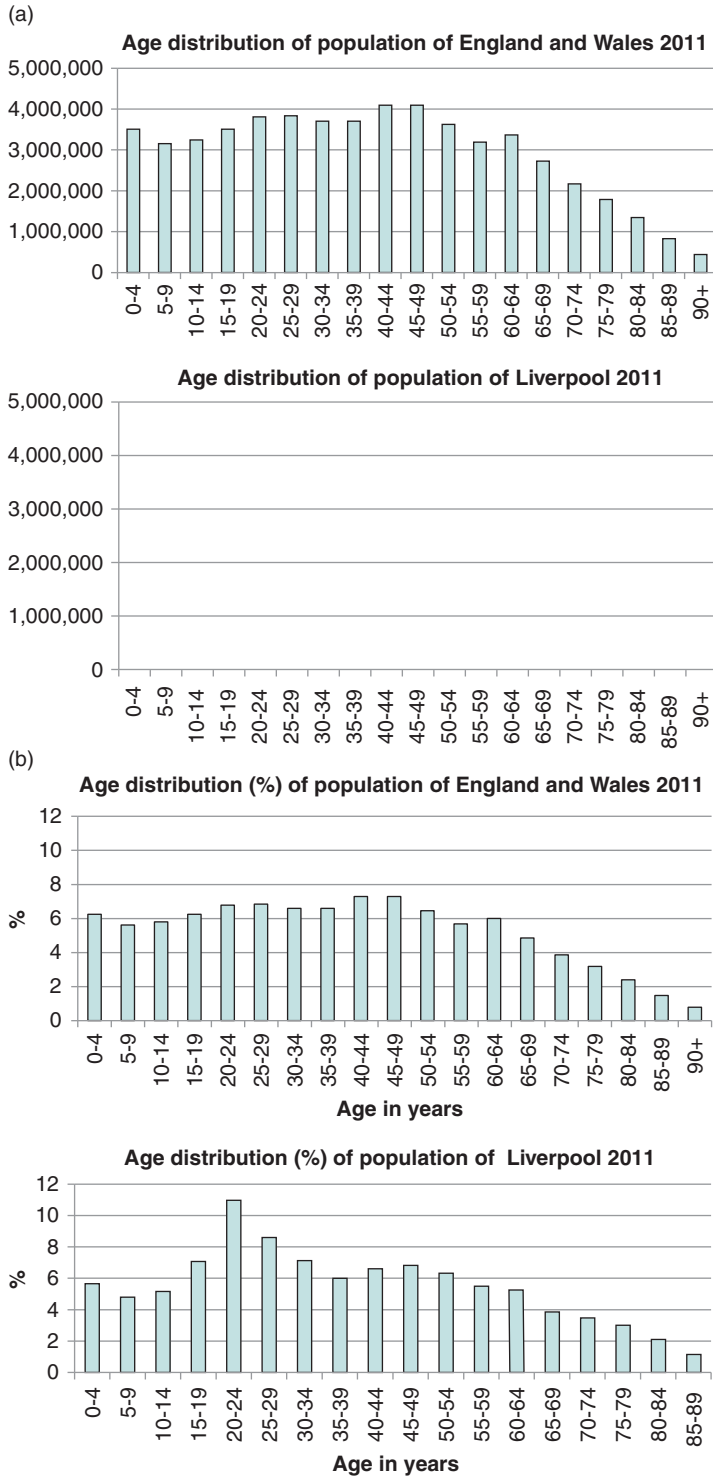


Figure 2.4.10 (a) Frequency histograms for England and Wales (upper panel) and Liverpool (b) Relative frequency histograms for England and Wales (upper panel) and Liverpool (lower panel).

2.4.5 Scatterplots, Linear Relationships and Correlation

We have been studying measurements that vary from one location or time to another. A quantity that varies is called a *variable*. Thus, PM₁₀ concentration at a particular time is a variable (it varies from location to location); blood pressure is another variable (it varies from person to person). These are both examples of *continuous variables*: the values they can take form a continuous range.

Two variables are said to be related, or associated, if knowing the value of one variable provides some information about the value of the other variable. For example, knowing a person's height tells us something about what their weight might be: weight and height are related. It is not a perfect relationship of course: Knowing someone's height does not tell us exactly how much they weigh. In this section, we look at how to display and summarise the relationship between two continuous variables, each measured for the same individuals (people, locations, countries, or whatever).

For our first example, we look at the gestational ages and birthweights for 24 babies, shown in Table 2.4.6.

Table 2.4.6 Gestational age and birthweight for 24 babies.

Gestational age (weeks)	Birthweight (kg)	Gestational age (weeks)	Birthweight (kg)
40	2.968	40	3.317
38	2.795	36	2.729
40	3.163	40	2.935
35	2.925	38	2.754
36	2.625	42	3.210
37	2.847	39	2.817
41	3.292	40	3.126
40	3.473	37	2.539
37	2.628	36	2.412
38	3.176	38	2.991
40	3.421	39	2.875
38	2.975	40	3.231

Scatterplots

We have a set of gestational ages and a set of birthweights for the same babies. The two values in each half-row are linked by the fact that they correspond to the same baby. We would like to display and summarise the data, just as we did with a set of values of one variable.

As before, it is important to begin by picturing the information we have. We could construct one histogram of the babies' ages and one of their birthweights, but we would then lose the vital information that the measurements are linked. The appropriate picture is a *scatterplot* (Figure 2.4.11).

Each point on the scatterplot corresponds to one baby and shows the gestational age and birthweight for each. Note that the scales on the two axes do not start at zero; it is not necessary

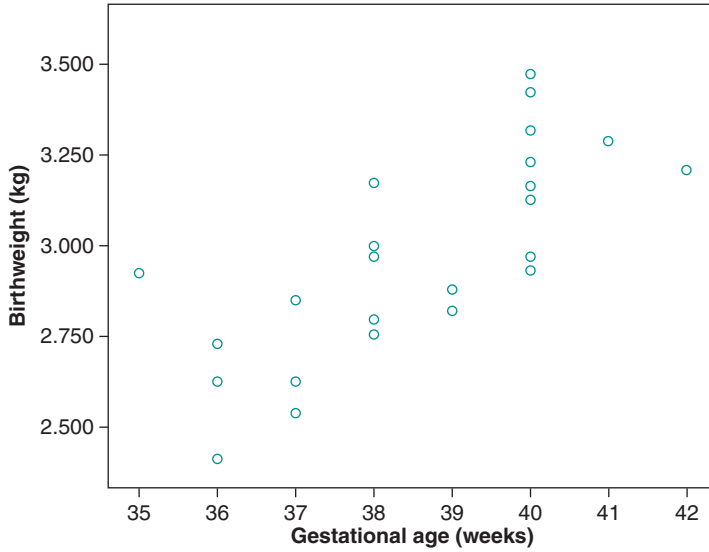


Figure 2.4.11 Scatterplot comparing gestational age (35–42 weeks) with birthweight.

for scales to start at zero, but they should always be clearly labelled and should show the units of measurement as well as the values. If we started at zero, the plot would include a lot of empty space, and the data, the vital information that we are interested in, would be squashed up and less clear (Figure 2.4.12).

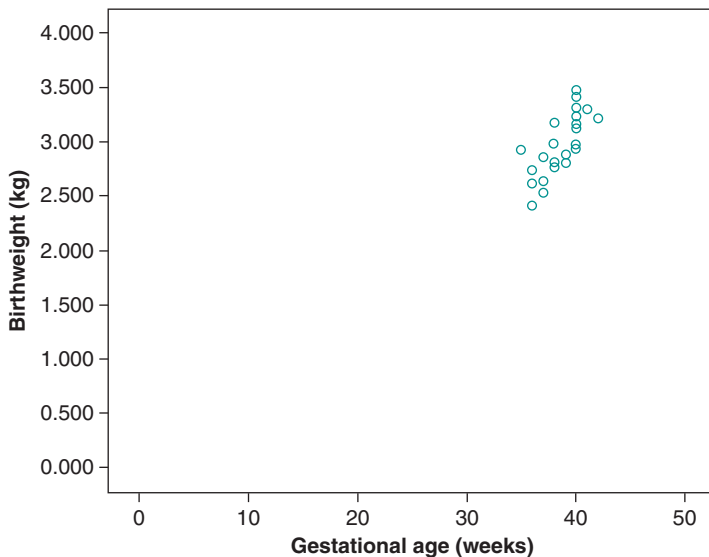


Figure 2.4.12 Scatterplot comparing gestational age (0–50 weeks) with birthweight.

This scatterplot shows that the babies' gestational ages and birthweights are clearly related, despite a lot of variation in birthweight for any particular gestational age. Overall, high values of one are associated with high values of the other. We describe the relationship between the variables in terms of the *form* of the relationship, the *strength* of the relationship, and whether it is *positive* or *negative*.

Linear and Non-Linear Relationships

The simplest form of relationship between two variables is a *linear* relationship. This means that the overall pattern of the data can be described by a straight line. It does not mean that the points in the scatterplot lie exactly on a straight line; this would be a perfect linear relationship, which is rarely observed in practice. In our example, we do not expect gestational age and birthweight to have a perfect relationship, because gestational age is only one of a number of factors that can affect birthweight. Even when a relationship is governed by an exact physical law, such as the relationship between voltage and current for fixed resistance, we would not expect to get a straight line plot from measurements of voltage and current. This is because there would be errors in our measurements.

What we are looking for is whether the points on the scatterplot are grouped around a straight line, without showing any obvious non-linear pattern such as is seen in Figure 2.4.13, an example of road casualties that show a cyclical pattern.

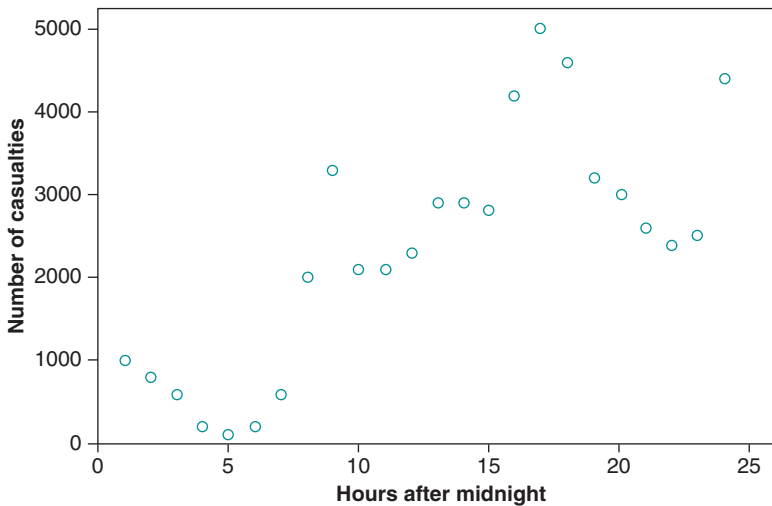
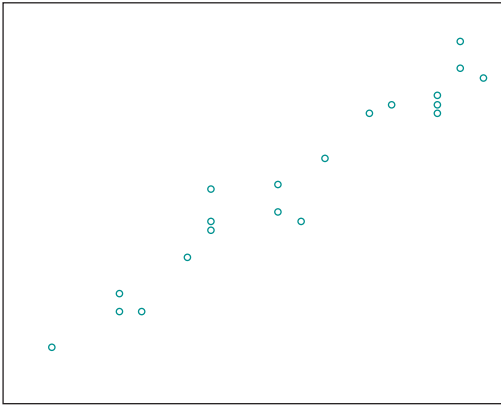


Figure 2.4.13 Example of a non-linear relationship: Casualties to road users on Friday.

We can describe the relationship between gestational age and birthweight as approximately linear: there is no obvious non-linear pattern. If the points in a scatterplot lie close to a straight line, we say there is a strong linear relationship. Conversely, if they are widely spread, the relationship is weak (Figure 2.4.14).

(a) A strong linear relationship



(b) A weak linear relationship

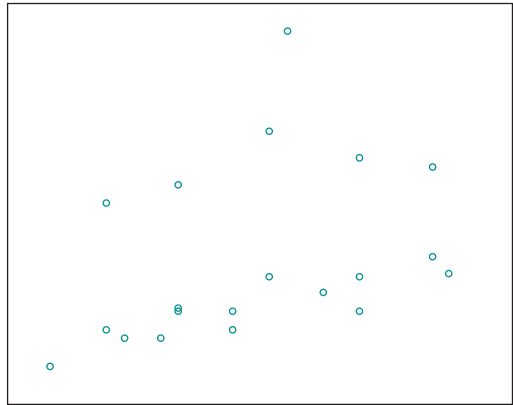
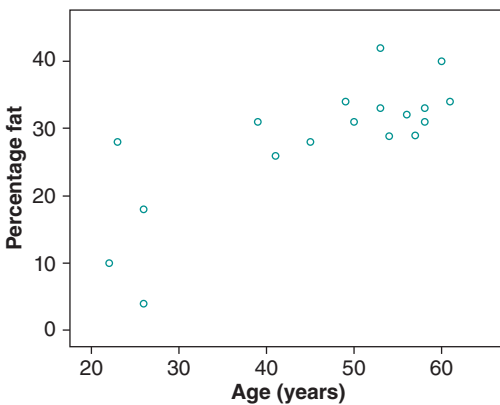


Figure 2.4.14 Examples of (a) strong and (b) weak linear relationships. (a) A positive linear relationship between age and the percentage of body fat; (b) a negative linear relationship between exam scores for maths and English.

Finally, we need to say whether the relationship is positive or negative. In a positive relationship, high values of one variable are associated with high values of the other (and low values of the two variables are associated). In a negative relationship, high values of one variable are associated with low values of the other (Figure 2.4.15). We can therefore describe the relationship between gestational age and birthweight for the babies in Table 2.4.4 as a fairly strong, positive, linear relationship.

(a)



(b)

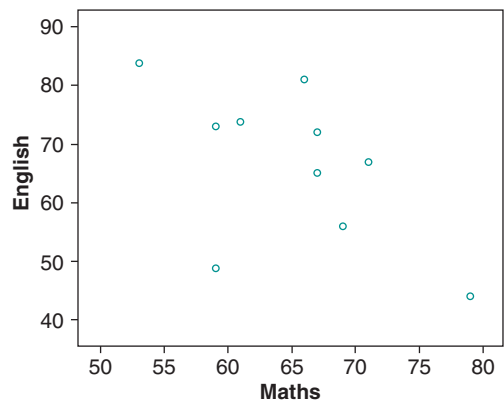
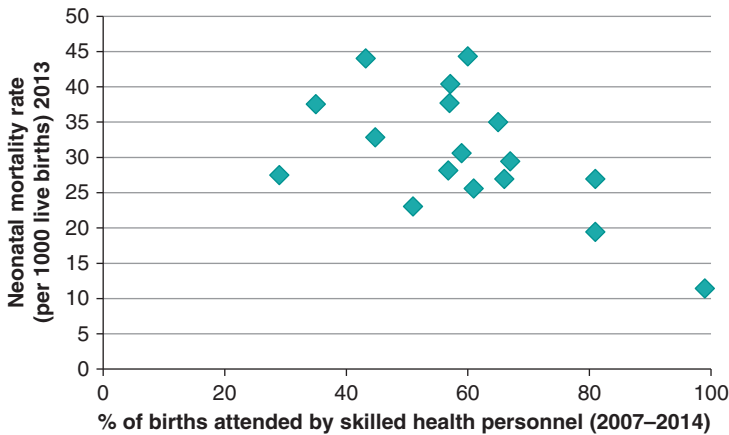


Figure 2.4.15 Examples of positive (a) and negative (b) linear relationships. (a) A positive linear relationship between age and the percentage of body fat; (b) a negative linear relationship between exam scores for maths and English. *Source:* World Health Statistics 2015. Last accessed December 2015.



Self-Assessment Exercise 2.4.7

The scatterplot below shows the percentage of births attended by skilled health personnel (2007–2014) and the neonatal mortality rate (2013) for 17 countries in West Africa. Describe the relationship between neonatal mortality and the percentage of births attended.



Source: World Health Statistics 2015. Last accessed December 2015.

Answers in Section 2.8

The Correlation Coefficient

A scatterplot shows the direction, form and strength of any association, and it is an important first step in investigating the data. However, the interpretation of a scatterplot by eye is subjective. For example, changing the scale or the amount of white space on the plot affects our perception of a linear, or any other, pattern. We can summarise the strength of a linear relationship with a single number, the **correlation coefficient**. For the type of data we have been looking at, which is the association between two continuous variables, we use the Pearson correlation coefficient (also known as product–moment correlation). A second type used for ranked or nonparametric data is called the Spearman correlation coefficient; this is described in Chapter 11.

The correlation coefficient is a measure calculated from the data, and it has an objective interpretation. The value of the correlation coefficient always lies between -1 and $+1$, with -1 corresponding to a perfect negative relationship (the points lie on a straight line sloping from top left to bottom right), and $+1$ corresponding to a perfect positive relationship (the points lie on a straight line sloping from bottom left to top right). For any relationship that is not exactly linear, the correlation coefficient lies somewhere between -1 and $+1$. A positive relationship has a positive correlation coefficient, and a negative relationship has a negative coefficient. A value of zero means that there is no *linear* association between the two variables (but note that it is possible for variables with a correlation coefficient of zero to be nonlinearly related).

The Pearson correlation coefficient calculated from a set of data is usually labelled r . Some sets of data with their correlation coefficients are shown in Figure 2.4.16.

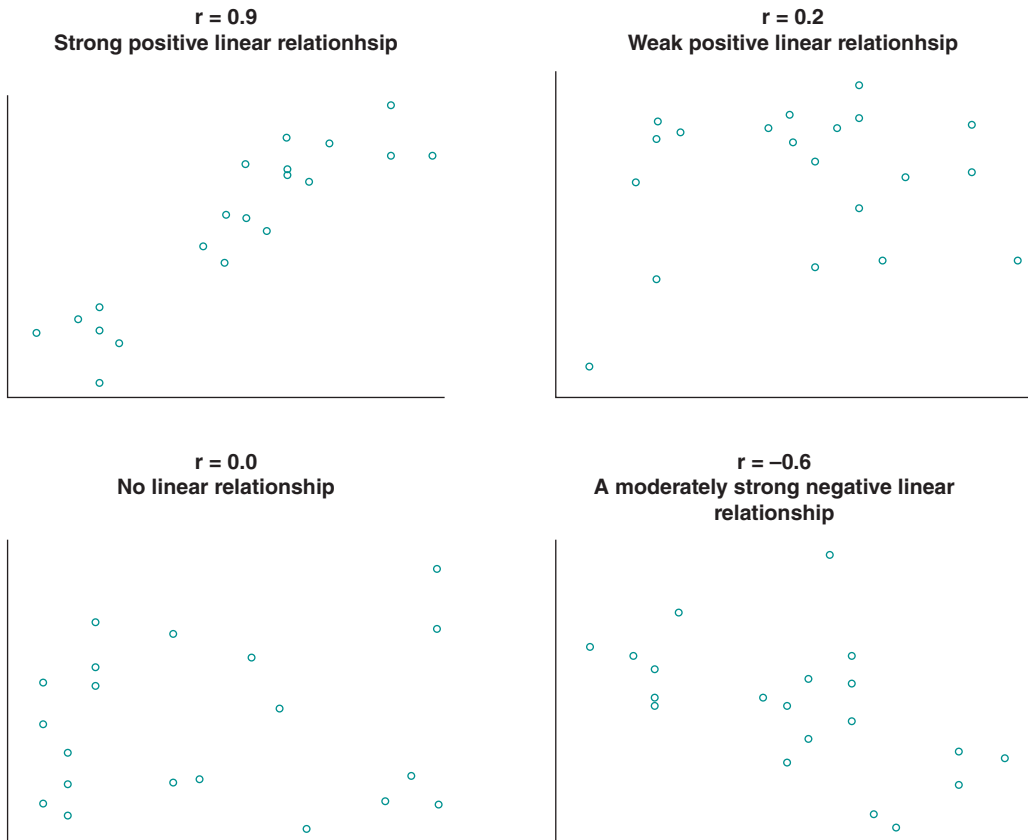


Figure 2.4.16 Examples of scatterplots for different distributions with correlation coefficients.



RS – Reference Section on Statistical Methods

Calculation of the Pearson Correlation Coefficient

To show how to calculate the correlation coefficient, with an example, we will label the values of one variable x and the other y . In our example, we may call the gestational ages of the babies x and the birthweights y . If we have n pairs of values (x, y) , ($n = 24$ for the babies), then the correlation coefficient is defined as follows:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

where \bar{x} and \bar{y} are the means of the x and y values. The terms $(x - \bar{x})$ and $(y - \bar{y})$ are deviations from the means of the two sets of values. Many calculators will find the correlation coefficient directly, without the need to use the above formula. Correlation coefficients are also calculated by many computer spreadsheets and statistical packages. To find the correlation coefficient with a calculator

that does not have statistics functions, it is easier to use an alternative, but equivalent, version of the formula:

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}$$

The expression $\sum xy$ means 'multiply each x value by the corresponding y value and then add them all together'. We can set out the calculation in a table. To calculate the correlation coefficient between gestational age and birthweight (Table 2.4.6), we need the following values:

Gestational age (x)	Birthweight (y)	x^2	y^2	xy	
40	2.968	1,600	8.809024	118.720	
38	2.795	1,444	7.812025	106.210	
40	3.163	1,600	10.004569	126.520	
35	2.925	1,225	8.555625	102.375	
36	2.625	1,296	6.890625	94.500	
37	2.847	1,369	8.105409	105.339	
41	3.292	1,681	10.837264	134.972	
40	3.473	1,600	12.061729	138.920	
37	2.628	1,369	6.906384	97.236	
38	3.176	1,444	10.086976	120.688	
40	3.421	1,600	11.703241	136.840	
38	2.975	1,444	8.850625	113.050	
40	3.317	1,600	11.002489	132.680	
36	2.729	1,296	7.447441	98.244	
40	2.935	1,600	8.614225	117.400	
38	2.754	1,444	7.584516	104.652	
42	3.210	1,764	10.304100	134.820	
39	2.817	1,521	7.935489	109.863	
40	3.126	1,600	9.771876	125.040	
37	2.539	1,369	6.446521	93.943	
36	2.412	1,296	5.817744	86.832	
38	2.991	1,444	8.946081	113.658	
39	2.875	1,521	8.265625	112.125	
40	3.231	1,600	10.439361	129.240	
Total	925	71,224	35,727	213,198,964	2,753,867

So the correlation coefficient is

$$r = \frac{2753.867 - 925 \times 71.224/24}{\sqrt{(35727 - 925)^2/24)(213.198964 - 71.224^2/24)}} = 0.74$$

(Remember, in any mathematical calculation, always complete complete multiplication and division before addition and subtraction).

The correlation coefficient is positive, showing that, on the whole, birthweight increases as gestational age increases. The value of 0.74 is closer to 1 than to 0 and indicates a fairly strong relationship between birthweight and gestational age. The correlation coefficient does not, of itself, tell us whether or not the relationship is linear, but we arrived at that conclusion from the scatterplot.



Self-Assessment Exercise 2.4.8

The following data are population and total daily water consumption for 10 regions of Scotland in 1995.

Region	Population (thousands)	Water consumption (megalitres/day)
Borders	105	32
Central	273	218
Dumfries and Galloway	148	76
Fife	351	146
Grampian	528	166
Highland	207	95
Lothian	754	284
Orkney and Shetland	43	22
Tayside	395	123
Western Isles	29	12

1. Draw a scatterplot of these data with water consumption on the vertical (y) axis and describe the relationship between water consumption and population.
2. Using the information and formulae provided in the reference section (above) or a computer, we find the correlation coefficient between water consumption and population to be 0.89 (if you wish to try calculating the coefficient with the above formulae, the answer and explanation are provided in Section 2.8). Interpret this result as fully as you can.

Answers in Section 2.8

Remember that the correlation tells us only about the strength of a *linear* association. Variables with a small correlation coefficient may have a strong non-linear relationship (Figure 2.4.17). This re-emphasises the importance of *picturing* the data on a scatterplot before calculating a correlation coefficient.

For the examples in Figure 2.4.17, it is clearly inappropriate to calculate a measure of the strength of linear association. There are statistical methods for determining the strength of nonlinear associations, but these are beyond the scope of this book.

Coefficient of Determination

A useful interpretation of the numerical value of r is provided by the value r^2 , termed the *coefficient of determination*. It can be thought of as the amount of the total variation in one variable that can be explained by the variation in the other variable.

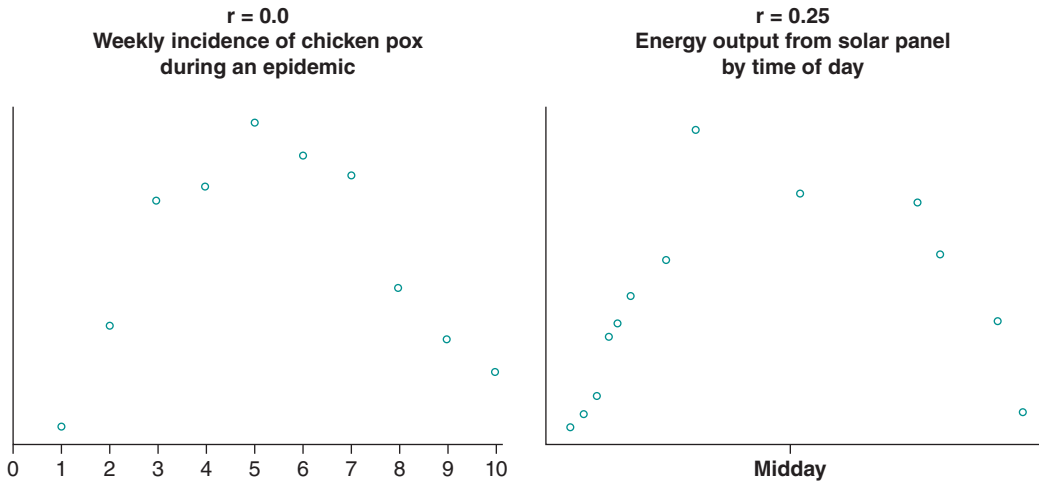


Figure 2.4.17 Scatterplots showing examples of non-linear associations.

In our example of birthweight and gestational age, it makes sense to try to explain birthweight in terms of gestational age, but not vice versa. We have already seen in the scatterplot how the birthweights varied for different gestational ages, and how the correlation coefficient $r = 0.74$ implied quite a strong positive linear relationship between the two variables. If we now calculate $r^2 = 0.74 \times 0.74 = 0.55$, we get an idea of how much of the total variation in birthweight can actually be explained by the variation in the ages. In fact, 55 per cent of the total variation in birthweight is explained by the variation in gestational age; the remaining 45 per cent of the variation is left unexplained and must be due to other factors that have not been considered in the analysis, or that are unknown.

Summary: The Relationship Between Two Continuous Variables

- The values of two continuous variables, each observed for the same individuals, can be displayed in a scatterplot.
- The scatterplot shows the form, strength, and direction of any relationship between the variables; this is a very important first step in describing the relationship.
- The strength and direction of a linear association between continuous variables can be summarised by the Pearson (product-moment) correlation coefficient.
- The correlation coefficient has a value between -1 and $+1$. A value of -1 indicates a perfect negative relationship, $+1$ indicates a perfect positive relationship, and 0 indicates that there is no linear relationship between the variables.
- The correlation coefficient is commonly stated to two decimal places, but there is no general rule.
- Useful interpretation of the strength of a relationship, in terms of how much variation in one variable is explained by the other, is provided by the coefficient of determination, r^2 . Note, however, that use of the coefficient of determination should be consistent with the direction and plausibility of the relationship.

2.5 Routinely Available Health Data

2.5.1 Introduction

The examples of routinely available data we have looked at so far cover death statistics, infectious disease, illness episodes seen in general practice, and environmental pollution data. These sources cover a wide range of data, including legally required registration and notifications, health service records, and measurements made outside the health-care system. This emphasises the great variety of information routinely available about health and disease, particularly in a developed country such as the UK, and also the many factors that influence public health. These data sources are a valuable resource, but it is important to understand potential limitations when using them, and to ask yourself the following questions:

- What can I find out about how this information was collected and prepared?
- What is the nature and extent of any error or bias that has occurred in the data collection and preparation?
- If there are errors and bias, how much do these matter for the purposes that I intend to use these data?

The next section reviews the nature, uses, and limitations of some of the key routine data sources, mainly from the UK, that are commonly used in health research.

2.5.2 Classification of Routine Health Information Sources

Routine data can be classified into three types:

- **Demographic data:** These data describe populations and therefore the number or characteristics of people with risk of ill health or mortality. Such information is required to interpret morbidity and mortality data.
- **Health (disease) events data:** These data describe health-related or disease events recorded through contact with health services.
- **Population-based health-relevant information:** These data include information on lifestyle and other aspects of health status that does not rely on contact with health services. This kind of information is often one of the most useful in health-related research, but it is less easily identified through routine sources.

Routine data sources share a number of strengths and weaknesses from an epidemiological perspective, and these are listed in Table 2.5.1. Table 2.5.2 outlines the different routine data sources covered in this section.

The remainder of this section explores the most commonly used health data sources in the UK, as laid out in Table 2.5.2. To get the most out of the section, it is recommended that you also explore the websites for the different sources and take a little time to familiarize yourself with the different datasets available. It is also worth noting that at the time of this book going to press, some of the changes to routinely available data sources were still under development, so the relevant websites will be the most up-to-date source of information available.

Table 2.5.1 Strengths and weaknesses of routine data sources.

Strengths	Weaknesses
Readily available data; regularly updated.	Often incomplete with a risk of imprecision and bias.
Repeated analysis over time allows assessment of trends.	Limited details on some determinants of health, such as ethnicity.
Useful for baseline description of expected levels of disease and to identify hypotheses.	Accessible, but may be presented in a form that is difficult to interpret.
	Many health services data are primarily intended for health services management rather than investigation of health status.

Table 2.5.2 Routinely available data sources covered in this section.

Types of data	Examples
2.5.3 Demographic data	The National Population Census Population estimates and projections Vital registrations; births and deaths
2.5.4 Health event data	Routinely available health service data <ul style="list-style-type: none"> ● Patient level health service data ● Secondary Uses Service ● Hospital Episode Statistics ● Confidential Inquiry datasets ● Mental Health and Learning Disability datasets ● Other datasets ● National Cancer Intelligence Network ● Clinical Practice Research Datalink ● National Drug Treatment and Monitoring System ● Disease Surveillance ● Mortality datasets Aggregate Level Health Service Data <ul style="list-style-type: none"> ● Public Health England Data and Knowledge Gateway ● Outcomes frameworks ● The Quality and Outcomes Frameworks and Disease Registers ● Benchmark tools
2.5.5 Population-based information	The Health Survey for England The ONS Longitudinal Study
2.5.6 Deprivation indices	Index for Multiple Deprivation The Carstairs Index
2.5.7 Routine data sources for countries other than the UK	

2.5.3 Demographic Data

Demography is the study of the size and structure of populations. Population size is affected by fertility, death, and migration. Population structure includes consideration of age, sex, ethnicity, socioeconomic status, and geographical location. Key routine sources of data include the national population census, vital statistics (births and deaths), records of inward and outward migration, and population estimates and projections.

The National Population Census

The national population census is a survey carried out at household level every ten years. The census provides extensive demographic and socioeconomic data collected across the whole country at one point in time, data that can also be compared with previous censuses. Completion of the census is a legal requirement in the UK (as it is in most countries), so coverage of households (but not travelling populations) is good, with a 94% per cent response rate being achieved in 2001 and 2011. Census data are important as they provide a reference-standard population estimate for analysis of population trends and essential statistical information for epidemiology. This is important for planning and funding of public services, as well as supporting research and business. Questions on the census form are piloted to minimise errors due to ambiguity in phrasing, and internal audit mechanisms aim to maximise the quality of data collation and analysis. Important gaps in data collection have also been addressed over time. For example, questions on ethnicity were added in 1991.

However, the census is expensive, time-consuming, and an immense administrative task. In the UK (as in many countries), the census is carried out once every 10 years and is used as a base to calculate sub-national population estimates in the intervening years. In the run-up to the next census date, the accuracy of mid-year population estimates decreases. This is because whilst estimates can adjust for births and deaths, they cannot accurately account for internal or international migration (see below). There is also consistent undercounting of certain subgroups of the population, including young men, armed forces personnel, the homeless, and travelling communities.

Data collection in the census depends on self-report, and this can lead to inaccuracy and possible bias, for example, if people were not willing to disclose certain information. Moreover, in the UK, when forms are submitted, they are electronically read, so inaccurate completion of the form can lead to coding errors. In terms of analysis of population trends, the introduction of the poll tax (a local tax based on individuals eligible to vote) negatively affected coverage rates in 1991 (and possibly 2001), with substantial undercounts among certain subgroups of people. Some changes in the wording of questions over time may also affect the interpretation of data over more than one census period.

Population Estimates and Projections

Estimates of population size and distribution between census dates are produced using data on births, deaths, and migration. Whereas registration of births and deaths is a legal requirement, migration is more complex. External migration (in and out of the UK) is based on the International Passenger Survey. In the UK during the 2001 to 2011 period, there was significant underestimation of immigrants arriving, and methods are now being put into place to overcome this problem. In addition, the effects of illegal immigration are hard to quantify. Internal migration, that is, migration within the UK, is estimated based on proxy sources including changes to the NHS Central Register (a register of all patients registered with a GP in England, Wales, or the Isle of Man) and the patient registration data service and so is open to underestimation, particularly among certain groups. In fact, the majority of public health statistics are affected by the

accuracy of mid-year population estimates; for example, when mortality rates are calculated at a small-area level, the accuracy of the data is affected by the accuracy of internal migration estimates applied to census data.

The UK Office of National Statistics (ONS) also produces national and sub-national population projections. Some short-term population projections can be quite accurate: for example, the numbers of people aged 65 to 74 years projected over a period of 20 years, based on the number of people currently aged 45 to 54 years and expected rates of mortality. However, other projections are more difficult, for example, the number of schoolchildren 20 years into the future is harder to predict, because fertility rates might change. Population projections therefore often include a *sensitivity analysis*, giving a range of estimates for different assumptions, for example, fertility rates.

Vital Registration: Births and Deaths

Registration of births and deaths contributes to updating estimates of population size between census dates. As we have already seen, registration in the UK is a legal requirement, and a network of superintendent and local registrars oversee the process of reporting these events to the ONS. The qualified informant (usually the nearest available relative) is responsible for notification.

Births are legally required to be notified within 42 days. Information recorded includes the name, sex, and date and place of birth of the child, mother, and father, together with the parents' occupations. In the UK, it is not a requirement to register the father's name if the mother and father are not married.

The process of death registration has already been described at the beginning of this chapter, and as discussed previously, mortality data are considered an extremely important routinely available data source due to the completeness of registration.

For the compilation of death statistics, if information on cause is not available at the time of death, the registrar can indicate that it will be provided at a later date. This information is added to the statistical system when available, but the public record (i.e., the death certificate) is not revised. As we saw, there are clear guidelines on reporting deaths to the coroner for further investigation if the cause of death is unclear. The coroner is responsible for following this up by postmortem examination (autopsy) and inquest, if necessary. Cause of death is coded according to an international classification system, the International Classification of Disease (ICD) (introduced in Section 2.1.3), allowing statistical comparisons by cause of death.

Inaccuracies can be introduced during the process of collating mortality data, including clinical diagnosis of cause of death, completion of the death certificate, and interpretation of statistics. Careful death certification is not always a high priority for busy clinicians, yet coding relies entirely on data obtained from the certificate. In addition, death certification is becoming increasingly complex with multifactorial causes of death; as a result, accuracy of diagnosis varies between subgroups of the population. For example, cause of death in children is more likely to be investigated and correctly diagnosed than in the elderly, in whom multiple disease conditions and diagnoses may be present. We have also seen how diagnostic fashion and less-socially acceptable causes of death, such as suicide or alcohol-related disease, also influence what is entered on the certificate.

ICD coding undergoes periodic revision, which has implications for analysis of disease trends, and *bridging tables* are produced to allow any necessary adjustments to be made. Finally, some of the data categories on a death certificate may be recorded less reliably than others. For example, occupational mortality data rely on the accuracy of information provided by informants about the type of employment and position held, which is not always reported accurately

(there is a recognized tendency for relatives to ‘promote’ the deceased person’s employment status) or systematically.

2.5.4 Health Event Data

Routinely Available Health Service Data

This section describes some of the main types of health event data available in the UK that we have not already covered earlier in this chapter and how these can be used in descriptive epidemiology.

The introduction of the Health and Social Care Act in the UK in 2012 brought about a number of changes to the way public health and health service data are organised. The following section provides a description of the main types of routinely available health event data in the UK at the time of writing, but many of the changes to the collection and access to these data were ongoing at this time. As of 2015, most data sources were available through the Health and Social Care Information Centre (HSCIC) (resulting from the 2012 Health and Social Care Act), which was responsible for collecting, analysing, and presenting national health and social care data.

The following section is split into patient-level health service data and aggregate-level health service data.

Patient-Level Health Service Data

Although mortality data provide an indication of the most important disease conditions in a population, they do not provide an adequate assessment of the extent and nature of health need within the population. Morbidity (illness) data provide a better measure of overall health and health care need, but collection of these data brings particular challenges in terms of completeness and accuracy. Currently, most morbidity data are collected through records of contact with health services, although many factors other than the nature of a disease can influence whether or not these contacts take place – whether through primary care or hospital use.

No single exclusive data source exists for morbidity data. A system of routinely collecting data within secondary care was originally established to assist service management rather than to provide data on morbidity for epidemiological purposes and health-care planning. This is evident from the way that the data are collected, which, to some extent, limits its utility from a population health perspective. For example, data are collected according to hospital episode rather than by individual, so repeat admissions for a disease such as asthma in the same person are not easily distinguished from admissions among different people. However, recent changes to the way data are collected and linked should lead to a more complete picture of morbidity both regionally and nationally in the future (known as the Care.Data programme). In practice, the HSCIC collects morbidity data from a number of sources and makes these available to service providers and commissioners via the Secondary Uses Service (SUS), described next.

Secondary Uses Service (SUS)

The Secondary Uses Service (SUS) is the single, comprehensive data warehouse for secondary health-care data in England, allowing reporting and analyses to support the delivery of health-care services. Individual-level patient data are recorded and can be identifiable or anonymised, depending on requirements. The data are available to both providers and commissioners of services, subject to strict governance conditions for secondary use; in other words, for purposes other than primary clinical care. SUS also provides a range of services for analysing and presenting these data.

Hospital Episode Statistics (HES)

Hospital Episode Statistics are derived from SUS data and are based on finished consultant episodes. A record is kept of each individual distinct episode of care (including day case interventions), defined as a period of treatment under a particular consultant. In the National Health Service (NHS) system, data collected include details of

- the hospital and GP
- geographical information
- the patient (age, sex, ethnicity)
- the process of arrival and discharge
- the process of care, including number of consultant episodes and treatment received
- outcomes including disease diagnosis (by ICD code) and final outcome (death, discharged, etc.).

HES data are collected routinely during a patient's time at or in hospital via the Patient Administration System, and they are then used to ensure that hospitals receive payment based on the care provided. HES data cover three areas of care: accident and emergency, admissions, and outpatients. Although the primary focus of the data is service management, ICD codes are available allowing detailed information on morbidity. Although this is useful for analysing the data by clinical diagnosis, the accuracy of diagnostic coding has been questioned because coding is based on the finished consultant episode, not on admission and discharge. For example, a patient admitted to hospital from the accident and emergency department is registered under a particular consultant. Subsequently, their care may be transferred to another consultant – for example, a specialist in cardiology. These are recorded as two separate episodes, because episodes are defined as periods of care under one consultant. As medical teams in hospitals become increasingly specialised, this may happen more often and add further complexity to interpretation of the data.

Community Information Data Set

The Community Information Data Set (CIDS) is also a person-based, SUS dataset providing person-based information on patients who are in contact with community services. The data are collected in a consistent way across the country so patient data can be compared and used to inform commissioning decisions and service management. The minimum dataset required is under review at the time of writing, and a minimum standard dataset will be mandated in due course.

Mental Health Services Data Set

The Mental Health Services Data Set (MHSDS) (which has superseded the Mental Health and Learning Disabilities dataset) contains person-based data about the care of children, young people, and adults using mental health services. It also includes access to psychological therapies and elements of the Learning Disabilities Census. As a SUS dataset, it seeks reuse clinical and operational data for purposes not directly related to patient care.

Other Patient Datasets

There are a number of other person-based datasets that use clinical and operational data for purposes other than direct patient care. These include a maternity dataset, a children and young people's health services dataset, and a child and adolescent mental health services dataset. The dataset on cancer outcomes and services is described further next. Information about these datasets can be found on the Health and Social Care Information Centre website.

National Cancer Intelligence Network

Now part of Public Health England (an executive agency of the Department of Health responsible for protecting and improving health and reducing health inequalities), the National Cancer Intelligence Network (NCIN) produces the Cancer Outcomes and Services Dataset (COSD), replacing the National Cancer Dataset. The dataset is compiled from individual patient data collected from the UK cancer registries. The NCIN also produces a number of different publications reporting on incidence of and mortality and survival from different types of cancers.

Clinical Practice Research Datalink (CPRD)

The CPRD is jointly funded by the National Institute for Health Research (NIHR) and the Medicine and Healthcare Products Regulatory Agency. CPRD was previously known as the General Practice Research Database (GPRD). The name change reflects the intention to develop this database to improve understanding of the entire journey through clinical care by maximizing the linkage of anonymised data, and hence it is not reliant on a single database. Key linked data now covers primary care, secondary care via HES plus GP medication data, demographic data, and central mortality data. It is envisaged that in the future, hospital medication data and many of the available audit datasets covering a huge variety of topics will also be linked to GPRD and that many of these will provide national coverage.

In addition to meeting service information requirements, the CPRD is increasingly being used for epidemiological research because it provides the most comprehensive routinely available morbidity data source. There are many examples in the literature of case–control studies that have been conducted in both clinical and public health fields using CPRD data.

The National Drug Treatment and Monitoring System (NDTMS)

Data are collected from providers of treatment services on the number of individuals receiving treatment for drug and alcohol misuse. The NDTMS produces monthly, quarterly, and annual reports summarizing the number of people receiving treatment by area and the type of treatment being given. These data are available on the system website.

Disease Surveillance

Monitoring is necessary for disease prevention and control, and surveillance falls under the remit of Public Health England (PHE). Notification systems for surveillance require the reporting of a particular disease or event to an official authority. For infectious diseases, the Notification of Infectious Disease System (NOIDS) in England is the surveillance body. NOIDS is subject to health protection legislation, which was updated in 2010 to have an all-hazards approach, including both infectious diseases and non-infectious hazards that could present a risk to human health. NOIDS notifications are provided by medical practitioners (such as general practitioners or hospital doctors), who have a duty to notify the Health Protection Agency of any relevant disease, infection, or contamination that occurs to any patient. Notifiable diseases were detailed in Section 2.2.5, but other diseases that present significant risk to human health (e.g. a new emerging pathogen such as Severe Acute Respiratory Syndrome (SARS) or a new chemical or physical hazard, such as polonium used for assassination) should also be reported. Public health laboratories are also legally required to pass on information of any notifiable diseases that come to their attention.

The Second Generation Surveillance System (SGSS) is a system for collating disease surveillance data. It includes a real-time surveillance system based on GP consultation data, and it also includes enhanced surveillance systems for certain diseases, such as that for tuberculosis, which collect a wide range of additional information including outcomes and risk factors. In addition, it includes mandatory surveillance systems for health care–associated infections such

as *Escherichia coli* (*E. coli*) and *methicillin-resistant Staphylococcus aureus* (MRSA). All these systems are managed by PHE. NOIDS returns are updated weekly and are available to public health teams at a local level.

There are also many other surveillance systems for the reporting of specific disease outcomes, including sexually transmitted diseases, and rarer diseases such as Creutzfeld–Jakob disease (CJD).

Mortality Datasets

The Primary Care Mortality Database The Primary Care Mortality Database is managed by the Health and Social Care Information Centre and links mortality data to an individual patient's GP practice where they were registered, as well as the ward/Lower Super Output Area¹ they lived in, and the location in which they died (e.g. hospital, care home, at their own residence). At the time of writing it holds monthly data based on Local Authority (LA) and Clinical Commissioning Group (CCG) structures, which are available for extraction by LA analysts, via a secure login. Counts of patients by GP practice are maintained quarterly and are made available to public health departments in Local Authorities and NHS organisations.

Summary Hospital-Level Mortality Indicator (SHMI) The Summary Hospital-Level Mortality Indicator (SHMI) dataset reports on mortality at hospital-trust level across the NHS in England. The indicator itself is the ratio of the number of patients who die following hospitalisation at a trust and the number who would be expected to die based on average England figures, taking into account the characteristics of the patients treated there. These figures are published quarterly by the HSCIC.

Mortality Data for Specific Causes: Confidential Inquiries Deaths from certain causes where preventative measures or better clinical management could reduce death rates and/or case fatality undergo additional routine investigation. These Confidential Inquiries now come under the umbrella of the Healthcare Quality Improvement Partnership, which aims to promote quality in health care, and in particular to increase the impact that clinical audit has on health-care quality in England and Wales. These include the following:

- The National Confidential Inquiry into Suicide and Homicide (NCISH) examines descriptive epidemiology (i.e., cases by time, place, and person, etc.) and risk factors for suicide and homicide involving people in contact with mental health services, along with cases of sudden unexplained death amongst psychiatric inpatients.
- The Maternal, Infant and Perinatal programme includes a series of reports in three areas: Child Death Review, Maternal Deaths, and Perinatal Mortality. These reports are now completed via MBRRACE-UK (Mothers and Babies – Reducing Risk through Audits and Confidential Enquiries across the UK). MBRRACE-UK is a recently formed collaboration that has run the national Maternal, Newborn and Infant Clinical Outcomes Review Programme since 2012 and has taken over responsibility for the Confidential Enquiry into Maternal Deaths. These reports are available on the Healthcare Quality Improvement Partnership (HQIP) website.
- The National Confidential Enquiry into Patient Outcome and Death (CEPOD) examines deaths following medical or surgical intervention and makes recommendations to improve the quality of the delivery of care for the benefit of the public. The reports are available on the

¹ Lower Layer Super Output areas (LSOAs) were developed to improve small-area statistics reporting and define areas containing populations between 1,000 and 3,000 residents.

HQIP website and tend to focus on individual sub-categories each year, limiting the scope for analysis of trends.

- A number of other clinical audits and reviews regularly take place on specific conditions, one example being the National Review into Asthma Deaths.

Aggregate-Level Health Service Data

Public Health England Data and Knowledge Gateway

The PHE Data and Analysis tools offer direct access to a number of tools and data sources from a single point. The tools were developed by a number of organisations that are now part of PHE, including the Public Health Observatories, the Health Protection Agency, cancer registries, UK screening programmes, and the National Treatment Agency for Substance Misuse. Thus far, data are collated on

- specific health conditions such as cancer, mental health, cardiovascular disease;
- lifestyle risk factors such as smoking, alcohol, and obesity;
- wider determinants of health such as environment, housing, and deprivation;
- health protection, and differences between population groups, including adults, older people, and children.

These data can be accessed via the Public Health England website and include such topics as cancer, general health profiles, end of life care, health impact assessment, and screening.

Outcomes Frameworks

Outcomes frameworks provide indicators for measuring outcomes in health and social care. The three main outcomes frameworks include the National Health Service Outcomes Framework, the Public Health Outcomes Framework, and the Adult Social Care Outcomes Framework.

The National Health Service (NHS) Outcomes Framework sets out the conditions that are used to hold the NHS Commissioning Board to account for improvements in health outcomes. The rationale is to improve quality through emphasis on health outcomes as opposed to process. The NHS Outcomes Framework operates over five domains: preventing people from dying prematurely, enhancing quality of life for people with long-term conditions, helping people recover from episodes of ill health or injury, ensuring people have a positive experience of care and treatment, and caring for people in a safe environment and protecting them from avoidable harm.

The Public Health Outcomes Framework has two main outcomes: increased healthy life expectancy and reduced differences in life expectancy and healthy life expectancy between communities. The framework focuses on improving the wider determinants of health, improving health, protecting health, and healthcare, public health, and preventing premature mortality.

The Adult Social Care Outcomes Framework is the main tool for setting direction and measuring adult social care. Its domains include enhancing quality of life for people with care and support needs, delaying and reducing the need for care and support, ensuring that people have a positive experience of care, and safeguarding adults whose circumstances make them vulnerable and protecting from avoidable harm.

These three frameworks are supported by education outcomes. Education outcomes measure progress in education, training, and workforce development across the whole system.

The Quality and Outcomes Framework and Disease Registers

The Quality and Outcomes Framework is a reward programme for GP practices achieving particular results. It is a voluntary scheme, and the indicators change each year. General practice disease registers form part of the Quality and Outcomes Framework (QOF) dataset.

Disease registers collect information on patients with certain conditions such as diabetes, coronary heart disease, and epilepsy registered with each practice; this information is used to promote delivery of standardised and high-quality health care. In addition, at the national level the registers provide useful information for health-service planning and research. Such registers are potentially resource intensive and require identification of adequate resources from the outset and skilled organisation to set up and maintain. Nevertheless, they provide an important source of data for audit and research, can help with service delivery and patient care, and are important in planning services through describing the occurrence of disease.

Aggregated files showing the prevalence for all the disease registers are provided by the HSCIC under the QOF datasets for each Clinical Commissioning Group (CCG) in the UK. Using this framework, it is possible to examine one specific GP practice and compare its performance to the CCG as a whole or compare it with national performance.

Potential problems with disease registers include selective registration and ascertainment bias (the most serious cases or patients who attend the GP more often may be registered), administrative problems of completeness and duplication, maintenance of confidentiality, and variation in data collection and quality across regions.

Benchmark Tools

The data sources described above are also used to produce benchmarks.

NHS comparators are designed to help NHS organisations improve the quality of care delivered by benchmarking and comparing activity and costs on a local, regional and national level. This benchmark mainly uses HES and focuses on secondary care.

Better Care, Better Value indicators have been developed to identify potential areas for improvement in efficiency. They can be used locally to help inform planning, and to inform views on the scale of potential efficiency savings in different aspects of care.

The spending and outcomes tool (SPOT tool) helps commissioners to link health outcomes and expenditure. NHS England commissioned Public Health England to develop this tool, which is now available to CCGs and uses programme budgeting, a well-established technique for assessing investment in programmes for specific disease areas such as cancer or mental health. The tool enables CCGs to identify areas that require attention, where shifts in investment will lead to improved local health gains.

2.5.5 Population-Based Health Information

In this section so far, we have looked at sources of information that rely on individuals with illness coming into contact with the health system. For some outcomes, the condition is unequivocal (e.g. maternal mortality) or serious enough (most cancers) that all – or virtually all – cases do reach the health-care system. For most illness, however, only a proportion does. Many different factors determine who does and does not come into contact with the health system, including age, sex, ethnicity, language, illness severity, and service access and quality, among others. Hence, data on disease frequency and determinants in the whole population are very important in filling the gap, but they can only be obtained by going to the community and carrying out a survey or by establishing data monitoring that captures all (or a sample, if this can be representative) of events such as road traffic accidents.

Population-based health and related socioeconomic information includes routine surveys as well as the ONS longitudinal study (described later). Information from a range of agencies, including local government and the police, on transport, road traffic accidents, the environment and education is also available. Two of the most commonly used sources are described briefly below, but this is not an exhaustive list, and there are many more examples, such as the British Household Panel Survey and the Labour Force Survey.

The Health Survey for England

The Health Survey for England (HSE) has run annually since 1994 and is now commissioned by the Health and Social Care Information Centre. The survey is carried out on a random sample of approximately 10,000 people in private households, and since 1995 it has included children aged 2–15 years. Data are collected on household, socioeconomic, health, social care, and lifestyle factors by questionnaire. In addition, physical measurements of height, weight, waist-to-hip ratio, blood pressure, urine, and blood samples are taken on a sub-sample. Key topics are repeated annually, allowing comparison of trends. Specific issues, such as coronary heart disease, physical activity, or accidents, are also covered in greater depth at periodic intervals.

Office of National Statistics Longitudinal Study

Established in 1971, the primary purpose of the Office of National Statistics (ONS) Longitudinal Study is to provide more-accurate information on occupational mortality. It follows up a small sample of individuals born in England and Wales on census day 1971, and immigrants with this birthday. It relies on linkage of records for these individuals with other ONS and NHS Central Register information, the latter containing details of all NHS-registered patients. Events recorded include deaths of a study member or their spouse, births to women in the cohort, infant deaths, cancer, immigration, and emigration. Data are not released directly to the public but are available in the form of reports describing how the topics included in the study vary by occupation. A particular advantage of the ONS Longitudinal Study in the investigation of the relationships between work and health is that health status can be determined in advance of changes in employment status, and vice versa, and the timescale is also known. It is therefore possible to untangle a problem that beset many studies of work and health; that is, knowing whether poor health led to loss (or change in type) of employment, or whether the type of work led to the change in health status. This temporal advantage of longitudinal (prospective) studies in studying causal relationships is explored in more detail in Chapter 5 on cohort studies.

2.5.6 Deprivation Indices

Socioeconomic deprivation and its contributing factors are recognised to be very important determinants of health. Measures of deprivation are now well established, and a number are in common use. All of these indices assess the proportion of individuals or households in a socially or geographically defined area that have poor living conditions. It should be noted, however, that not all people living in an area with a high deprivation score are deprived, and vice versa. This issue, a feature of group-based (or ‘ecological’) analyses, is discussed further in Section 2.6.2. Deprivation indices provide a summary measure of key social, economic, and environmental factors, which together have a very substantial impact on health. The current most commonly used indices include the Indices of Multiple Deprivation and the Carstairs Index.

The Indices of Multiple Deprivation

The English Indices of Deprivation 2015 provide a relative measure of deprivation at the small-area level across England. Areas are ranked from least deprived to most deprived on seven

different dimensions of deprivation and on an overall composite measure of multiple deprivation. Most of the data underlying the 2015 Indices are based on data from 2012–2013. The domains used are income deprivation, employment deprivation, health deprivation and disability, education deprivation, crime deprivation, barriers to housing and services deprivation, and living environment deprivation. Each of these domains has its own scores and ranks, allowing users to focus on specific aspects of deprivation. In addition, two supplementary indices measure income deprivation amongst specific age cohorts: the Income Deprivation Affecting Children Index (IDACI) and the Income Deprivation Affecting Older People Index (IDAOPI).

The Carstairs Index

The Carstairs Index is an index of deprivation used in Scotland. It was originally developed by Carstairs and Morris in the 1980s and is based on four indicators derived from the Small Area Statistics Census Tables: low social class, lack of car ownership, overcrowding, and male unemployment. The index has since been updated based on subsequent census data from 1991, 2001, and 2011.

2.5.7 Routine Data Sources for Countries Other Than the UK

This section (Section 2.5) has focused on the UK as an example of a country with well-developed national systems for collecting and collating health-related data. Other countries have their own national data, and systems for recording person-level health data may differ according to the approach taken. Denmark, for example, has very advanced systems for digital exchange of health data, with secure access to hospital and GP records for patients and health care staff. The system is organized based on a unique personal identifier that all Danish citizens receive at birth, along with a secure Web identifier.

There are also European-level and global data sources that provide a wealth of useful data. You met an example of national comparative data in Section 2.2.2 when we compared mortality rates for lung cancer and systems for collecting mortality data between Ireland and Austria. The World Health Organisation Statistical Information System (WHOSIS) and Eurostat are two of the most widely used sources of international health information, and if you are interested in European-level statistical data, you may find it useful to familiarize yourself with the databases that are available via their websites.

2.6 Descriptive Epidemiology in Action

2.6.1 The London Smogs of the 1950s

In this example we will see how investigation of severe episodes of air pollution in London during the 1950s using relatively simple techniques of descriptive epidemiology contributed to the introduction of new legislation to control the burning of solid fuels in urban areas.

Contemporary descriptions of these smogs, a mixture of smoke and fog, tells of the streets being so dark that vehicles had to use lights in daytime, and people were barely able to see a few feet in front of them. This is how serious the air pollution had become in London during the early 1950s.

The smoke arose mainly from household use of coal for heating and cooking, so the pollution was worse in the coldest months of the year. The fog was associated with temperature inversions: cold, still air trapped in the bowl of the Thames Valley. Serious pollution episodes caused by coal burning no longer occur in London due to regulatory control of coal use (only 'smokeless' types can be used) and the widespread availability of clean and more convenient fuels such as electricity and natural gas. Urban air pollution in the UK still presents a risk to health, but the main source now is motor vehicles.

We start by looking at health and air-pollution data for this now infamous episode, using descriptive epidemiological methods, and consider what can be discovered from interpretation of variations by time, place, and person. Table 2.6.1 shows data for Greater London on deaths (all causes), air temperature, and air pollution for each day over the period 1–15 December 1952.

Table 2.6.1 Deaths, air temperature, and air pollution in London 1–15 December 1952.

Variable	Date								
	Period 1–8 December								
	1st	2nd	3rd	4th	5th	6th	7th	8th	
Deaths									
– central London	112	140	143	120	196	294	513	518	
– outer London	147	161	178	168	210	287	331	392	
Air temperature:									
– daily mean (°F) at Kew ¹	36.9	34.2	39.0	36.5	29.5	28.9	28.9	31.5	
Air pollution:									
– smoke (µg/m ³) Kew ¹	340	340	190	420	1470	1750	870	1190	
– smoke (µg/m ³) County Hall ²	380	490	610	490	2640	3450	4460	4460	
– Sulphur dioxide (ppm)	0.09	0.16	0.22	0.14	0.75	0.86	1.34	1.34	
				Period 9–15 December					
		9th	10th	11th	12th	13th	14th	15th	
Deaths									
– central London		436	274	255	236	256	222	213	
– outer London		362	269	273	248	245	227	212	
Air temperature:									
– daily mean (°F) at Kew ¹		36.6	43.3	45.1	40.1	37.2	35.2	32.0	
Air pollution:									
– smoke (µg/m ³) Kew ¹		470	170	190	240	320	290	180	
– smoke (µg/m ³) County Hall ²		1,220	1,220	320	290	500	320	320	
– Sulphur dioxide (ppm)		0.47	0.47	0.22	0.23	0.26	0.16	0.16	

¹Kew is southwest London, about 6 miles from the centre of the city (Westminster).

²County Hall is on the south bank of the Thames, opposite the Houses of Parliament (Westminster).



Self-Assessment Exercise 2.6.1

1. Plot the data for all variables in a way that allows you to compare the variations of each across the period under study (1–15 December 1952).
2. Describe the findings for the deaths occurring in central London and outer London.
3. Describe the findings for smoke pollution (Kew and County Hall) and for sulphur dioxide (only County Hall data available).
4. Describe the findings for temperature (Kew). Do you think that these temperature data are representative of the whole city?
5. Interpret the information you now have on variations in these atmospheric variables and the deaths, paying careful attention to the timing of the variations.
6. Do we have information by person, time, and place (all three)? What additional information would you like to have had for this investigation?

Answers in Section 2.8

The Clean Air Act

Clean Air Act was passed in 1956, in part because of this and other episodes of severe air pollution that resulted in substantial loss of life. The Act established smokeless zones and controlled domestic smoke emissions for the first time. It resulted from the action of the National Society for Smoke Abatement, a pressure group consisting of MPs and some Medical Officers of Health, and it led to a marked decline in smoke emissions, which were reduced by around 65 per cent between 1954 and 1971. Sulphur dioxide emissions continue to fall.

In this section we have looked at an example of how descriptive epidemiology has contributed to national public health policy. The focus on environmental data also illustrates the value of information sources from outside the health services.

2.6.2 Ecological Studies

Introduction

Ecological studies are a type of descriptive epidemiological study design that we have already encountered. An *ecological study* or analysis is essentially one that examines associations between units of grouped (or *aggregated*) data, such as electoral wards, regions, or even whole countries.

There are many examples of this approach: for instance, a scatterplot of socioeconomic condition by electoral ward, versus the percentage of smokers in each ward. The point is that the data are for the group (in this case a ward), and this is in contrast to surveys and cohort and case–control studies, where we look at information on exposures and outcomes for every individual. Figure 2.6.1 shows the ecological association between a measure of fat consumption in a number of countries, and the age-adjusted death rate in each country.



Self-Assessment Exercise 2.6.2

This ecological study demonstrates a very clear association, and it is tempting to think this is *causal* – that is, higher fat consumption increases the risk of mortality. Can you think of any reasons why this type of analysis could be misleading?

Answers in Section 2.8

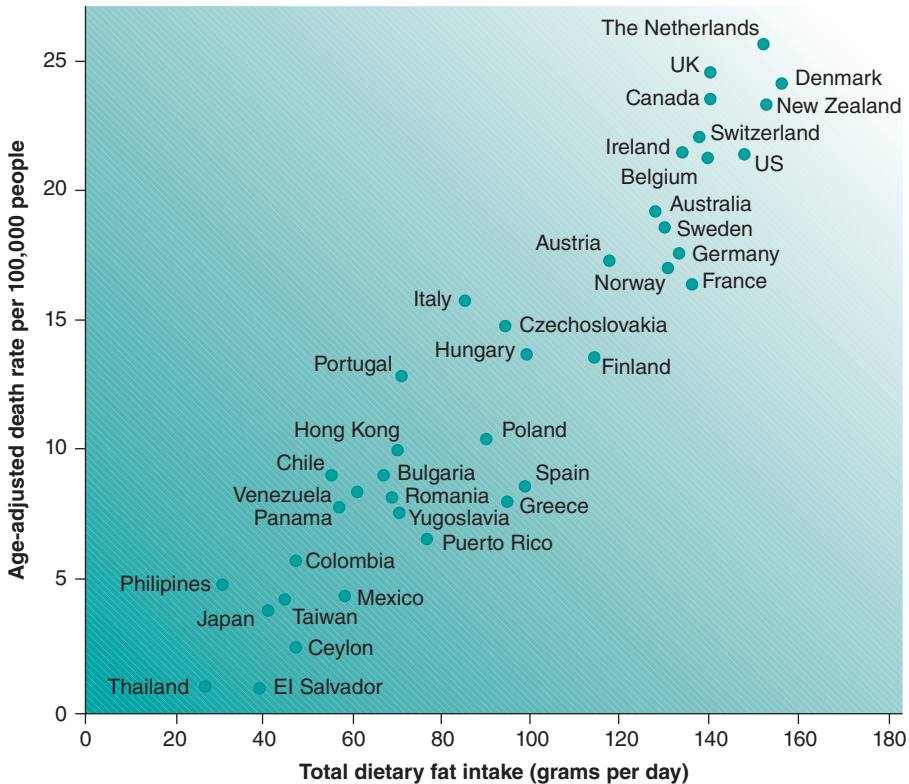


Figure 2.6.1 Total dietary intake and age-adjusted death rate per 100,000 people. *Source:* Bingham 2004. Reproduced with permission of Nature Publishing Group.

The Ecological Fallacy

As we have seen, aggregate measures tell you nothing for certain about individuals. This gives rise to the possibility of an *ecological fallacy*. This occurs when we ascribe to members of a group one or more characteristics they do not possess as individuals. The consequence of an ecological fallacy can be that an association seen between aggregated data does not exist at the individual level and is in fact explained by other associated factors. If the association is seen at the ecological level but not at the individual level, it cannot be causal.

The risk of encountering the ecological fallacy does not mean that such analyses are unhelpful but that they need to be used with caution. As an initial step in studying disease causation, this is a useful means of exploring associations worthy of further investigation at the individual level.

Summary

- Ecological studies are a useful initial stage in the epidemiological investigation of association and causation.
- Ecological studies are carried out at an aggregated level.
- There is an implicit assumption that the characteristics of the aggregated group apply to the individuals that make up the group. If this is wrong, an ecological fallacy will arise.

2.7 Overview of Epidemiological Study Designs

In this chapter on descriptive epidemiology, based as it has been mainly on routine data sources, we have seen how relatively simple analyses by *person, time, and place* have actually been remarkably informative. At least some of our examples, however, have raised more questions than were answered, leaving us with questions such as these:

- People with greater socioeconomic deprivation have higher mortality, but why? We do not have the detailed and specific kind of information about these people's lifestyles, views, experiences, and so on, that we need in order to explain the findings or to advise what should be done to alleviate the problems.
- How can we be sure that the factors we found to be associated with death rates actually cause the disease? Some of the apparent associations could easily be caused by other factors about which we have little or no information.
- If we are planning to publish our research findings and develop public health and health-care policies based on them, surely it is not going to be sufficient to say, for example, that air pollution seems to be associated with asthma. We need to be able to show what levels of specific pollutants cause how much disease, and what benefit can be expected from feasible reductions in these pollutants. We need to be able to measure these effects and to state what the margins of error are.

These and other questions, which we will discuss throughout the book, return again and again in research. They are, if you like, the crux of why we need more-sophisticated research methods, and they give an idea of what these methods have to offer. Descriptive epidemiology has been a good starting point to set the context, to open up ideas about explanations and causes, and to help us in handling and presenting information. What we need now, however, are more-powerful and more-focused research designs: methods that can answer questions in ways that can really advance our understanding of causation or that can measure the effects of health care and prevention in ways that are useful to those who carry out the work, as well as to those who have to manage ever more restricted budgets.

Table 2.7.1 presents a summary of the main types of health studies, including case studies (which we will not consider further) and epidemiological studies (which are the main focus of the book). We have dealt with the first of these epidemiological designs, descriptive studies, in this part of the book. We look at the next design, surveys, in Chapter 4. The main purpose of surveys is to collect information, from individuals, that was not available from routine sources. Here we are setting out to explore the first of a series of increasingly focused and rigorous scientific research methods that can provide the tools to address our unanswered questions.

Some Reassurance

Do not worry if you find it difficult to understand all the terms and concepts included in Table 2.7.1. These will become clearer as we progress through each study design in the following chapters, and at this point you should just try to understand the general ideas. As we work through these study designs, you will see how each successive method gains in quality. Stronger research design means that questions can be answered with greater certainty, especially when we are dealing with questions of causation, and when we are trying to measure (quantify) the effects of risk factors, health-care interventions, and so on. This greater strength generally comes at a price, though. The more rigorous the study design, the more difficult, time-consuming, and expensive it usually is to carry out. All designs have their place and value, however, and you will appreciate this better as you become more familiar with the strengths and limitations of each. You may find it useful to refer back to this table as we start to look at each new study design.

Table 2.7.1 An overview of research study designs.

Study type	Nature of investigation	Comment on research design
(a) Case studies		
<ul style="list-style-type: none"> • Case reports: descriptions of one or more cases (of disease), together with circumstances of interest. • Case series: as above, but a consecutive series of cases. 	A case study is good way of getting ideas about what might be causing, or predisposing to, a health problem. We can call this hypothesis generation.	As a research method, this approach is very weak. For example, we have no way of knowing for sure whether people without the disease are any different from those with it.
(b) Observational epidemiological study designs		
Descriptive epidemiology		
Studies of patterns of disease in populations, and variations by time, place, and person.	With this design, we are beginning to test ideas on cause and effect, but they are still largely dependent on routinely available data.	Stronger than case studies, as formal comparisons between population groups can be made.
Surveys		
Cross-sectional studies, often using samples, designed to measure prevalence, and more particularly to study associations between health status and various (possibly causal) factors.	In contrast to working with routinely available data, surveys are designed to collect information exactly as we specify, from the individuals we are especially interested in.	They are more focused than descriptive studies, and as a result are a more-powerful means of investigating associations. The inclusion of information on a range of factors means that some allowance can be made for these in studying causal links.
Case-control studies		
In these studies, people with a disease ('cases') and people without ('controls') are investigated, essentially to find out whether exposure to a factor of interest differs between cases and controls.	This is a focused research design that allows measurement of the strength of the association between a disease and its possible cause, or factors (such as health interventions) that can afford protection. It is the most appropriate and practical study design for less-common diseases.	This is a valuable and widely used study design but one that is subject to bias. Careful design is vital, but even so there is often room for a lot of debate about the interpretation. Compared to cohort studies and trials (see below), this design is relatively quick and cheap to carry out.
Cohort studies		
By first surveying, and then following up, a sample of people (a 'cohort'), it is possible to measure by how much the incidence of disease differs between those who were exposed to a factor of interest (at the time of the survey) and those who were not exposed, or were exposed less.	This design also provides a good measure of the strength of association, and it is less open to bias than the case-control design. This major advantage is, however, gained at the expense of time, cost, and complexity. In contrast to case-control studies, cohort designs are not suitable for uncommon conditions due to the time it would take for enough cases to arise.	This is a valuable and widely used study design. Cohort studies, as a result of their design, provide true estimates of incidence and are less subject to bias than case-control studies due to the opportunity to measure exposure before the disease occurs. The information about temporal relationships (that is, that exposure preceded the disease) helps in assessing whether the association seen between exposure and outcome is causal.

(continued)

Table 2.7.1 (Continued)

Study type	Nature of investigation	Comment on research design
(c) Intervention study design		
By taking a sample of people, and intervening (e.g. with a medication, operation, or prevention measure) among some (the 'intervention' group), but not others (the 'control' group), we can assess how much the level of health or incidence of disease has been altered by the intervention.	In many respects, this is the ultimate research design because of the control we have over who does (and does not) have exposure to the factor we wish to test. Where the allocation to intervention and control is random (as with a randomised control trial, RCT), all other factors that can influence the outcome are equally distributed across the two groups.	Evidence from intervention studies is powerful, especially if the study is randomised, controlled, and blinded (that is, neither investigators nor subjects know who was in intervention or control groups). It can be appreciated, however, that there are many situations where allocating some people to receive a health-care intervention or be exposed to a risk factor, and others not, would be impractical and/or unethical.

Natural Experiments

One final design, which in some respects is like an intervention study (as there is a marked change in exposure) but is in fact an observational design (as the researchers observe the effects of change that happened to take place), is the *natural experiment*. This might be used, opportunistically, when some substantial natural or human-made event or policy change takes place. A good example of such an event was the nuclear reactor explosion at Chernobyl, Ukraine, which provided the opportunity (and duty) to study the effects of radiation exposure. Such studies may provide unique opportunities to study high-risk exposures, such as to radiation or toxic chemicals. We will not look further at natural experiments as a specific study design in this book, but many aspects of other designs and epidemiological methods in general would be relevant in designing and carrying out such a study.

2.8 Answers to Self-Assessment Exercises

Section 2.1

Exercise 2.1.1

1. Certification

The correct answer is the hospital doctor. Death certificates must be completed by a registered medical practitioner. As explained in the story, there was no need to involve the coroner's office in this death, because the death was not sudden and Mrs Williams' mother was under the care of the hospital. The circumstances in which the coroner should be involved are discussed in more detail shortly.

2. Quality of information

- In this case the information is likely to be accurate, because a diagnosis of lung cancer had already been confirmed by investigation (chest radiography, bronchoscopy, and biopsy).
- The autopsy probably would not make a difference in this case. If the lung cancer had not been confirmed (by confirmation of the type), it is possible that the original site of the cancer may have been found to be elsewhere, from which it spread to the lung. Only

a minority of deaths in the UK go to autopsy, however (about 25 per cent in England and Wales, and 15 per cent in Scotland), with the lowest percentage in the 65-and-older age group.

Exercise 2.1.2

Reason for Verdict

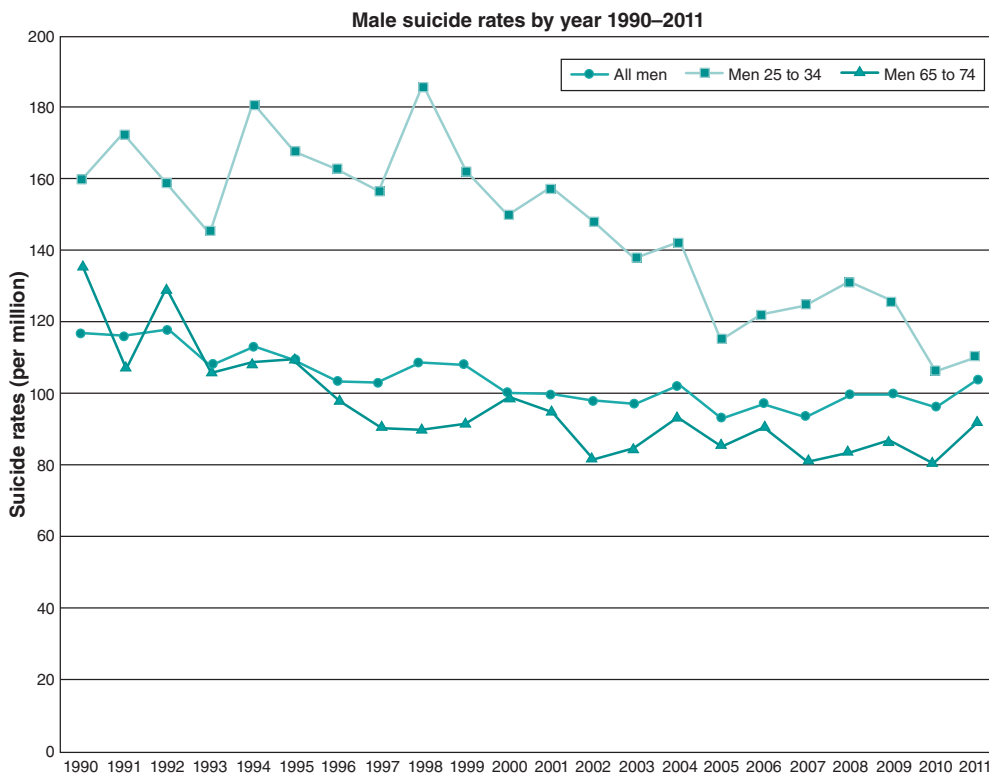
Even though it may seem to us that this was obviously a suicide, the coroner did not return a verdict of suicide because the evidence was not strong enough. If the verdict can be challenged (e.g. by relatives), the coroner tends to return an open verdict or a verdict of accidental death.

Proportion Receiving Verdict of Suicide

In the UK, true suicides receive verdicts of suicide, open (injury undetermined), and (some) accidental death. When studying suicide, we often use the numbers of suicide and open verdicts combined, because this represents the best easily available estimate of the true suicide rate. In England and Wales in 2011, there were 3,644 suicide verdicts and 1,251 open verdicts, making a total of 4,895. On this basis, 74 per cent $[(3644 \div 4895) \times 100]$ of our estimate for all suicides actually received an official verdict of suicide. Thus, only three quarters of suicides may be recorded as such, and the true proportion may be even lower than this. This example illustrates well how social, legal, and other factors can influence the recorded cause of death.

Exercise 2.1.3

Plot of Data for Men:

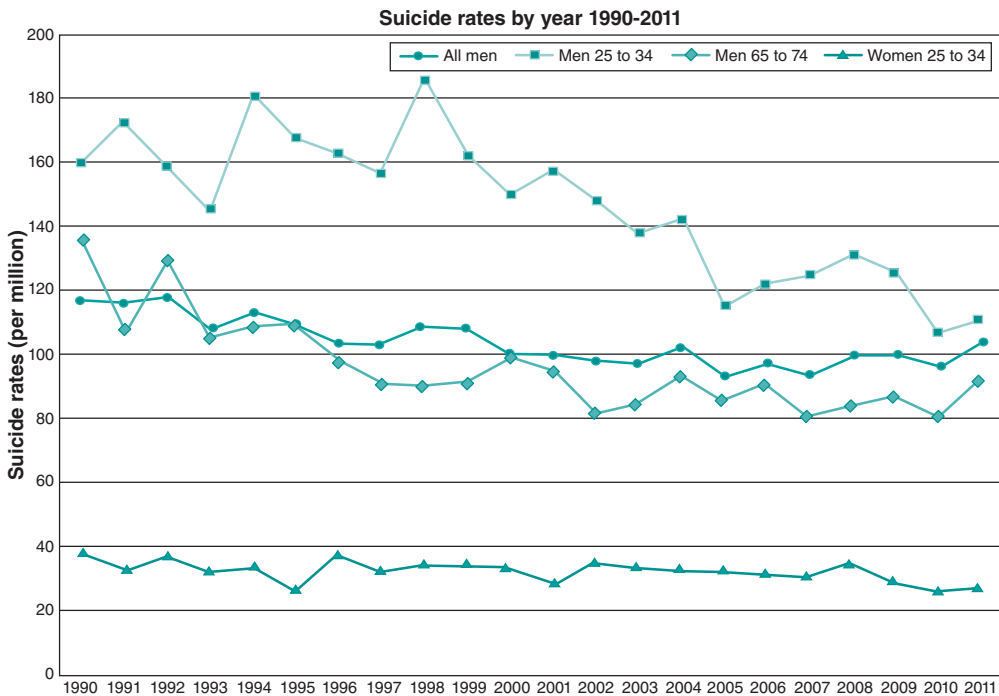


This figure is called a line graph. The graphical presentation of data is covered in more detail in Section 2.4. Suicide rates for all males have gradually fallen since 1990 (it is not shown in our data, but the male suicide rate actually increased in the 1980s) although there were some year-on-year rises. However, the male all-age suicide rate increased in 2011, reaching a rate similar to that seen in 1998–1999. If we look at the age-specific rates, for men in the 25–34 year age group, the rate is considerably higher than that for all men. However, the rate for this age group shows a sharper fall since 1998, although also increasing in 2011. For older men (65–74 years), the rate is much lower than at younger ages, and again we can see a gradual fall during the 1990s and an increase in 2011. Thus rates for all ages (or very wide age bands) may mask very different trends within key age subgroups.

What kind of conclusions can be drawn from this? Biddle *et al.* (2008) examined suicide trends in more detail, looking at specific causes of death (that is, the method of suicide). They suggested that the decline was partly due to a reduction in poisoning with car exhaust gas, because an increased number of cars had catalytic converters; they also noted that there were declines in suicides from all common methods, including hanging, suggesting a more-pervasive effect. In addition, they reported that other risk factors for suicide, such as unemployment and divorce, also decreased during that time period.

Exercise 2.1.4

Combined Male and Female Suicide Rates:



Interpretation

For women, the rate is considerably lower than for any of the male age groups and has declined slightly over the period (although the scale of the graph, which has been set to accommodate the much higher male rates, makes this somewhat difficult to appreciate).

Before moving onto the next section, you might find it useful to spend a little time exploring the National Statistics website and familiarising yourself with the many different types of health data available; this can be found at <https://www.gov.uk/government/statistics>.

Exercise 2.1.5

The ratio for IMRs in groups 7 and 8 compared with that for group 1.1 is $7.9/1.6 = 4.9$; that is, the rate for the most disadvantaged groups is almost five times that for the most affluent.

Infants born to families in the lower socioeconomic groups may be more vulnerable for the following reasons:

- Higher proportion of low birthweight (that is, less than 2500 g), as these babies are more at risk for infections and other problems
- Lower uptake of antenatal care
- Poorer access to high-quality health services due to location, language barriers for ethnic minorities, etc.
- Higher rates of maternal smoking and drug use during pregnancy
- Higher rates of parental smoking in the home environment during the first year of life
- Poorer home circumstances, including housing quality (damp, cold), risk of accidents, etc.

Section 2.2

Exercise 2.2.1

1. The data for both Austria and the UK are for 2012. The numbers of deaths among women aged 75 and older is much greater in the UK, but the UK also has a larger population. However, when rates are calculated, the rates in the UK are still around two and a half times higher than in Austria.
2. When considering the reasons for differences, you should always consider the following three possibilities:
 - **Chance**; that is, it is due to random variation where relatively small numbers are concerned.
 - **Artefact**; that is, the difference is not real but is the result of how the information is collected, coded, or presented.
 - **Real**; that is, the rate for lung cancer really is higher in the UK.

We will look more at how to assess the role of chance and what might lead to artefact in due course (artefact is in the next exercise). In thinking about the possible explanation for a real difference, we need to consider the known causes of lung cancer, which are smoking (by far the most important), and other exposures, including radon gas (a natural radioactive gas that permeates from the ground into houses, mainly in areas where there is granite), air pollution, and asbestos exposure. The difference in these country rates is almost certainly due to smoking, as the individual risk and exposure patterns for the other causes could not account for such a large difference. One aspect of smoking to consider among women of this age is country differences in the time period at which women began smoking. These women would have been in their early twenties during the 1950s, and trends in postwar Europe are very likely to have had a major influence on patterns of smoking among young women in the two countries.

Exercise 2.2.2

The Austrian and UK systems for collecting mortality data are really very similar, with the exception of the percentage of deaths among women in this age range occurring in hospital. This information does not, however, tell us anything about how underlying or predisposing

causes of death are coded in practice; for instance, whether a woman with lung cancer, but dying eventually of pneumonia or heart failure precipitated by the cancer, would be handled in the same way by certifying doctors and coding staff in the two countries.

The higher percentage of deaths in hospital is unlikely to contribute to a more-accurate diagnosis in the UK because, given the nature of lung cancer and the high chance of intensive contact with the health services for treatment and care, the diagnosis normally is made before death.

We do not have information about certification and coding rules, and it would be useful to find out about these in making international comparisons. Nevertheless, the difference in death rates is so great that explanations relating to artefact are unlikely to alter our overall conclusion. In other words, it seems reasonable to conclude that the differences in lung cancer rates between the two countries are real.

Exercise 2.2.3

Food Poisoning Rates per 1,000 Population

1. Calculation of crude rates of cases of food poisoning, and population (thousands), England and Wales, 2001–2009

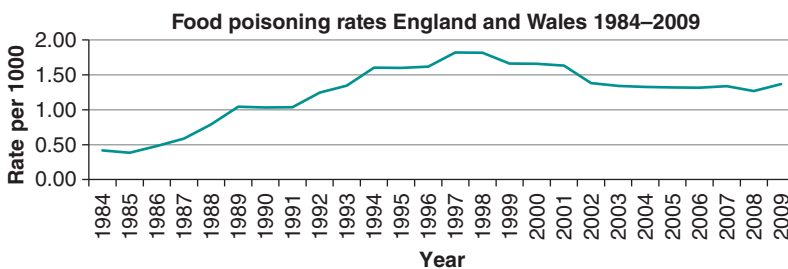
Year	Number of cases ¹	Population (1,000s) ²	Crude rate/1,000/year
2001	85,468	52,360.0	1.63
2002	72,649	52,567.2	1.38
2003	70,895	52,792.2	1.34
2004	70,311	53,053.2	1.32
2005	70,407	53,416.3	1.32
2006	70,603	53,725.8	1.31
2007	72,382	54,082.3	1.34
2008	68,962	54,454.7	1.27
2009	74,974	54,809.1	1.37

Source: ¹ Statutory Notifications of Infections Diseases (NOIDS) available at <http://www.hpa.org.uk>

² Available from <http://www.ons.gov.uk>

Includes notified cases and also cases otherwise ascertained. Cases otherwise ascertained were no longer collected after week 35 of 2010, so it is not possible to compare trends for subsequent years.

2. Plot of food-poisoning rates



3. Comment and possible explanations

There was a steady rise in the food-poisoning rate over the period 1984–1998, during which time there was an almost fourfold increase. However, rates have dropped from 1999 onwards. Again, we should think about *chance, artefact, and real* change.

- **Chance:** Given the numbers involved, and the progressive nature of the change up until 1998, this is unlikely to represent a chance increase. Likewise, the steady progressive decrease through to 2008 does not appear to indicate a chance decline.
- **Artefact:** Could the trends be due to changes in the proportion of cases being recognised and notified (*case ascertainment*)? This is unlikely; although there is typically under-notification of cases of food poisoning, this generally runs at about the same level over time, so observed trends in notifications are unlikely to be an artefact due to differential inaccuracies in reporting. Also, it would be important to verify that there were no major changes to the process of notification, which could explain these trends.
- **Real increase/decrease:** If the increasing trend is real, there is a range of possible explanations, including increases in fast-food outlets, less home cooking, more pre-packed foods, problems with food production (salmonella in eggs and poultry), increasing travel and trade, and so on, that may have contributed to the rising number of cases up until the late 1990s. Conversely, if the decreasing trend is real, greater public awareness of the causes of food poisoning, more hygienic food handling and storage, and more-rapid and more-effective control of outbreaks could all have contributed to these changes in rates.

Exercise 2.2.4

Trends

We cannot be very certain of the trends without a more-thorough examination of how consistently data have been recorded and of the margins of random error. For the current purpose, however, we can note that the numbers of cases are quite substantial and that considerable effort has been made to ensure reasonable comparability of the data recording over time. With this in mind, we observe that

- Acute bronchitis increased between 1955–1956 and 1971–1972, and thereafter remained stable. There is certainly no evidence to suggest a marked reduction.
- Asthma, for which we have no separate data in 1955–1956 (a reflection perhaps of how the condition was viewed at that time), has almost doubled between 1971–1972 and 1981–1982.
- Hay fever, again with no separate data for 1955–1956, has also increased in line with the asthma rates.

Change in Diagnostic Fashion?

This appears to show that as asthma increased, acute bronchitis has remained stable. If the recorded increase in asthma was due to change in diagnostic fashion (and was simply a re-labelling of acute bronchitis cases), the true rate of acute bronchitis in the population would also have increased during the period under consideration. We do not know whether or not this has happened, but the data here show that acute bronchitis has not decreased, and are therefore consistent with the view that the increase in asthma is not simply due to change in diagnostic practice.

What Has Happened to Rates of Asthma?

Although there is some uncertainty about what is happening, it appears that rates of asthma really have increased.

Section 2.3

Exercise 2.3.1

1. Edinburgh centre has the lowest concentration of PM_{10} ($0.4 \mu\text{g}/\text{m}^3$). Bury roadside has the highest concentration ($38.0 \mu\text{g}/\text{m}^3$).
2. The locations that did not record a level of PM_{10} are as follows:

Newcastle centre	Stockport
Bolton	Sheffield centre
Leamington Spa	London Eltham
Haringey roadside	

These are 7 out of the total of 34 urban locations. The percentage is

$$\frac{7}{34} \times 100\% = 20.6\%$$

About 20 per cent, or one in five, of the urban monitoring sites did not record a level of PM_{10} at this time.

Section 2.4

Exercise 2.4.1

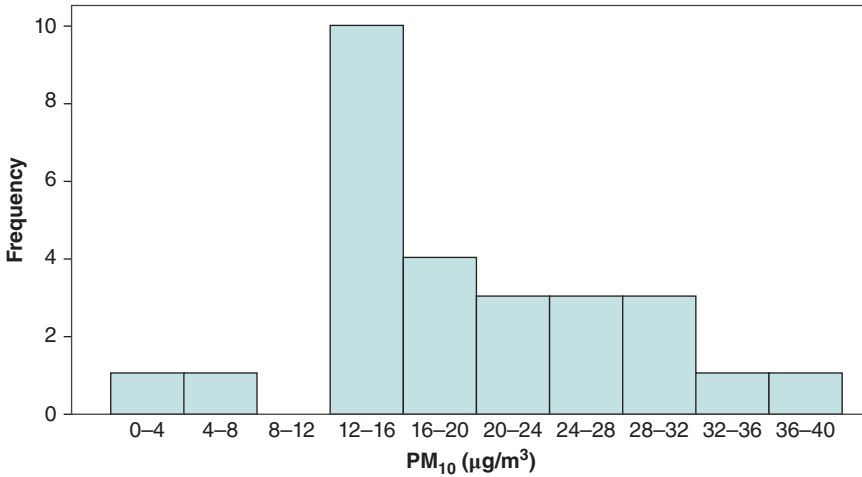
1. Frequency distribution

A convenient way of finding the frequency distribution (in the absence of a computer) is to tally the data, write down the intervals, and then go through the data in the order they are given and make a tally mark for each value next to the appropriate interval. It is then simple to count up the number in each interval to get the frequency distribution. Note that it is also easy to make mistakes with this method, so check your results. Remember not to include the values at the upper limit of each interval; those belong in the next highest interval.

PM_{10}	Tally	Frequency
0–4	/	1
4–8	/	1
8–12		0
12–16	///// /////	10
16–20	////	4
20–24	///	3
24–28	///	3
28–32	///	3
32–36	/	1
36–40	/	1
Total		27

2. Histogram

Your histogram should look like this.



It is important that the axes are labelled and have scales marked on them. The vertical axis can be labelled frequency, or number, or count. There is now a gap in the histogram: There are no values in the range 8–12 µg/m³. There are two low values on their own (<8 µg/m³), followed by a peak for the interval 12–16 µg/m³. Then it is downhill all the way.

This histogram gives the impression of the data being bunched up at lower values and then slowly tailing off towards the higher values. Note that the histogram in Figure 2.4.1 does not show this so clearly. In that picture, the patterns of data each side of the peak (15–20 µg/m³) look more similar to each other.

Exercise 2.4.2

The distribution of PM₁₀ values is unimodal with some right/positive skew. There are no gaps or outliers in the distribution. If you said that the distribution is approximately symmetric, rather than right skewed, that is a good enough approximate description, as the skew is not marked.

Exercise 2.4.3

PM₁₀ concentrations for the London locations

London Bloomsbury	29.4	London Bexley	23.2
London Hillingdon	30.7	London Brent	25.0
Sutton roadside	24.8	London Eltham	–
London Kensington	24.9	Haringey roadside	–
Camden kerbside	32.8		

Mean and Median

The concentration of PM₁₀ was recorded at seven locations in London. The mean concentration is found by adding up the seven values and dividing by 7:

$$\frac{29.4 + 23.2 + 30.7 + 25.0 + 24.8 + 24.9 + 32.8}{7} = \frac{190.8}{7} = 27.25714286$$

The mean concentration of PM₁₀ is 27.26 µg/m³. The data are recorded to one decimal place, so the mean is stated to two decimal places. Any sensible degree of accuracy will do, such as

one, two, or three decimal places, but using eight decimal places shown above would not be sensible. The accuracy of the data does not justify it.

To find the median, we need to order the data:

23.2 24.8 24.9 **25.0** 29.4 30.7 32.8

There are seven values. The middle value is the fourth. The median PM_{10} concentration is therefore $25.0 \mu\text{g}/\text{m}^3$.

Exercise 2.4.4

- Which of the following are correct statements?
 - Yes. This is a picture of all the data, and it shows the shape.
 - No. The mean tells us where the distribution is located but not its shape.
 - No. The mode is the peak of the distribution.
 - No. This is part of the description of the shape, but it is not sufficient. We also want to know whether the distribution is symmetric or skewed and whether there are any outliers.
- Which *two* of the following are correct statements?
 - No. The mean is greater than the median because the high values in the long right tail of the distribution all contribute to the value of the mean.
 - No. Skewed distributions are not necessarily unimodal, although they usually are, because most distributions are unimodal.
 - Yes. The mean is greater than the median, so more than 50 per cent of the observations are less than the mean.
 - No. The definition of left skewness is that the distribution has a long left tail.
 - Yes. Positive skewness is an alternative name for right skewness, which means the right tail is longer than the left.

The two correct statements are (c) and (e).

- At which point (a, b, or c) does the mean lie, and at which point does the median lie?
 - The mean and median are both at point (a).
The distribution is approximately symmetric, so the mean and median have about the same value. This is in the middle.
 - The mean is at (c) and the median is at (b).
The distribution is right skewed, so the mean is greater than the median. Fifty per cent of the data are less than the median, so half the area of the histogram must be to the left of the median.

Exercise 2.4.5

Interquartile Range

There are three observations each side of the median. So the quartiles are the 2nd and 6th values:

$$Q_1 = 24.8; Q_3 = 30.7$$

The interquartile range is therefore $IQR = 30.7 - 24.8 = 5.9 \mu\text{g}/\text{m}^3$.

Standard Deviation

The alternative formula for calculating the variance of the observations is

$$s^2 = \frac{1}{n-1} \left[\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 \right]$$

We need to calculate the sum of the observations and the sum of the squared observations.

x_i	x_i^2
23.2	538.24
24.8	615.04
24.9	620.01
25.0	625.00
29.4	864.36
30.7	942.49
32.8	1075.84
$\Sigma x_i = 190.8$	$\Sigma x_i^2 = 5280.98$

The number of observations is $n = 7$, so

$$s^2 = \frac{1}{6} \left(5280.98 - \frac{(190.8)^2}{7} \right) = 13.38619$$

and

$$s = \sqrt{13.38619} \approx 3.66 \mu\text{g}/\text{m}^3 \text{ (to one more decimal place than the data).}$$

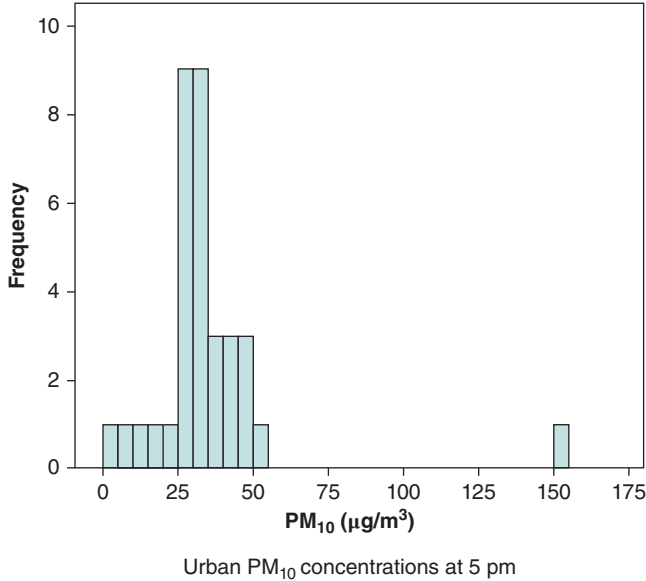
Exercise 2.4.6**Frequency Distribution**

PM ₁₀	Frequency
0–5	1
5–10	0
10–15	0
15–20	0
20–25	1
25–30	9
30–35	9
35–40	3
40–45	3
45–50	3
50–55	1
...	0
160–165	1
Total*	31

*Three locations did not record a value of PM₁₀.

You can tally the data to obtain the frequency distribution. You do not need to include all the empty intervals between the outlier and the rest of the data, but make sure it is clear that there is a gap, as shown in the table above.

Histogram



The distribution of PM₁₀ concentrations is unimodal and right skewed. Most of the data fall within the range 20–60 µg/m³, but there are two outliers with concentrations in the intervals 0–5 µg/m³ and 160–165 µg/m³.

Mean and Standard Deviation

The sum of the 31 values is $\sum x_i = 1160.2$, so the mean is $\frac{1160.2}{31} = 37.43 \mu\text{g}/\text{m}^3$.

The sum of the squared values is so the variance is

$$\begin{aligned}
 s^2 &= \frac{1}{n-1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] \\
 &= \frac{1}{30} \left[62608.5 - \frac{(1160.2)^2}{31} \right] \\
 &= 639.5693118
 \end{aligned}$$

and the standard deviation (s) = 25.29 µg/m³. Since the distribution is skewed, we could have chosen to describe the location and spread with the median and interquartile range. This is perfectly acceptable, but here we will use the mean and standard deviation to make comparisons with the PM₁₀ values recorded at midday. Note that the outlying value of 164.0 greatly affects the mean and standard deviation, and so these measures could be misleading. If we omit 164.0 from the calculations, the mean is 33.21 µg/m³, and the standard deviation dramatically reduces to 9.53 µg/m³.

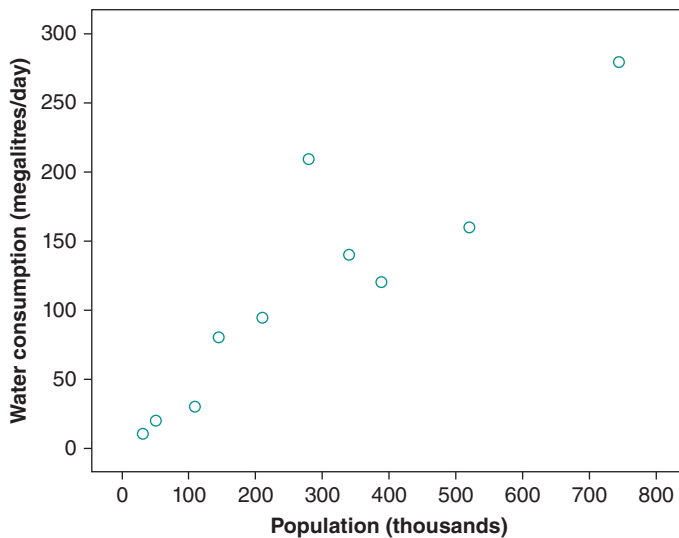
Comparison of PM_{10} Concentrations at Midday and 5 PM

Concentrations tend to be higher at 5 PM than at midday (the mean is $37.43 \mu\text{g}/\text{m}^3$, compared with $19.37 \mu\text{g}/\text{m}^3$ at midday), and they are more spread out (standard deviation of $25.29 \mu\text{g}/\text{m}^3$ compared with $8.27 \mu\text{g}/\text{m}^3$).

The concentrations at 5 PM are more clearly right skewed than those recorded at midday. Most of the data lie between $23.6 \mu\text{g}/\text{m}^3$ and $51.0 \mu\text{g}/\text{m}^3$. There are two outliers of $2.0 \mu\text{g}/\text{m}^3$ and $164.0 \mu\text{g}/\text{m}^3$. The very high value will inflate the mean and standard deviation. However, it is clear from the histogram that even without this high value, concentrations are higher, on average, at 5 PM than at midday.

Exercise 2.4.7

- The relationship is approximately linear (there is no obvious nonlinear pattern).
- It is a negative relationship: Neonatal mortality decreases with increasing percentage of births attended.
- The relationship between neonatal mortality and percentage of births attended is moderately strong.

Exercise 2.4.8

Water consumption and population for Scottish regions, 1995

The scatterplot, shown above, indicates that there is a strong, positive, approximately linear relationship between water consumption and population size.

Calculation of correlation coefficient is provided for reference. Labelling population x and water consumption y , we have

$$\begin{aligned} \sum x &= 2833 & \sum x^2 &= 1279523 & \sum y &= 1174 \\ & & \sum y^2 &= 208634 & \sum xy &= 496696 \end{aligned}$$

The correlation coefficient is therefore

$$r = \frac{496696 - 2833 \times 1174/10}{\sqrt{(1279523 - 2833^2/10)(208634 - 1174^2/10)}} = 0.89$$

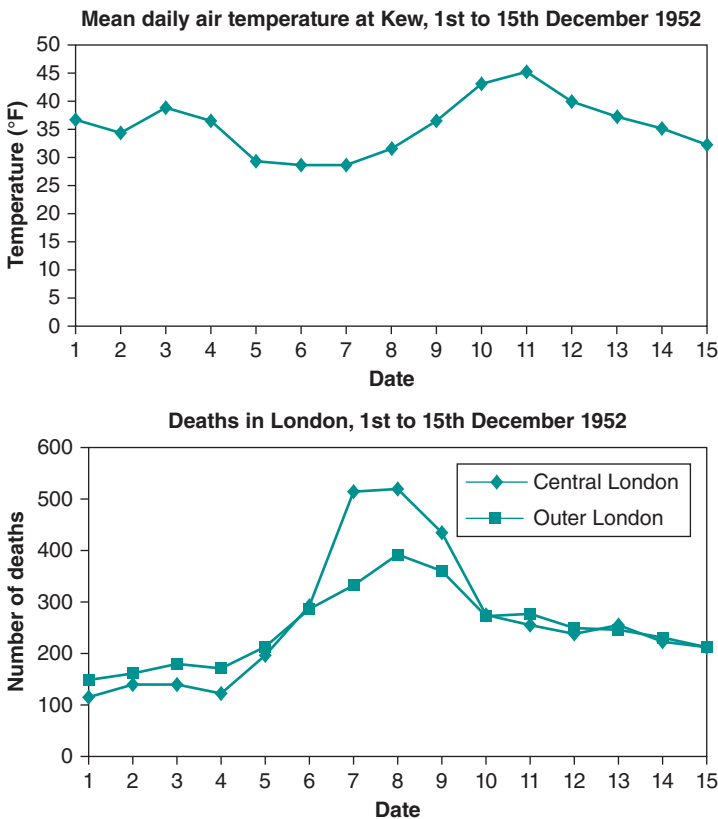
The high value (close to +1) confirms that the relationship is strong and positive. You may have noticed that all the points on the scatterplot except one (Central Region) lie very close to a straight line. A point that is distant from the rest of the data and does not follow the general pattern is called an **outlier**. This is the same definition as for data on one variable (see Section 2.4.3). There is no reason to think that the values for Central are erroneous (they are almost certainly correct). It may be of interest to investigate why water consumption in Central does not follow the same pattern as in the rest of the Scottish regions.

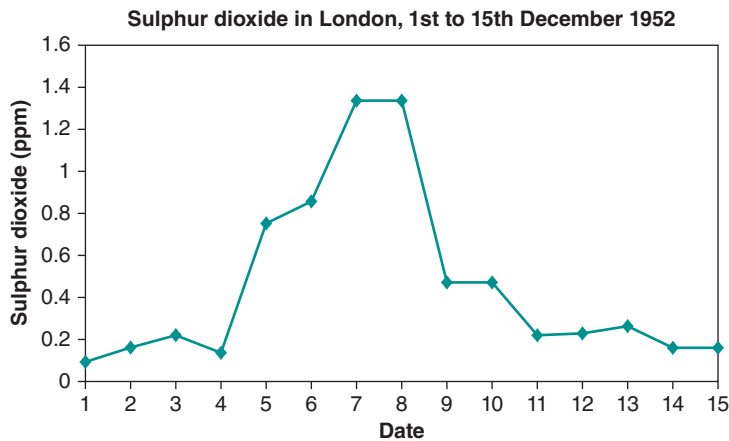
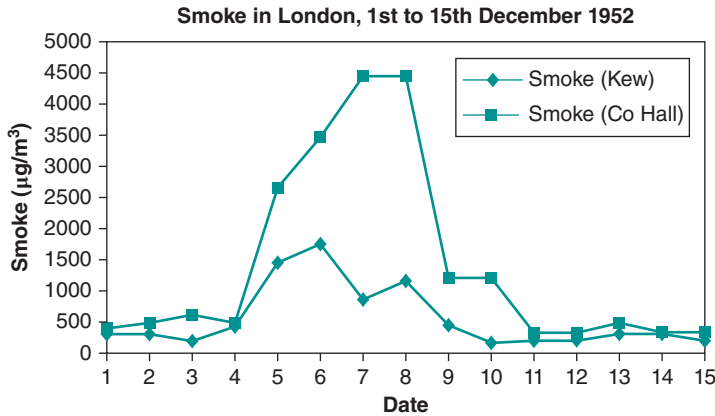
The presence of an outlier reduces the correlation between the variables. If we calculate the correlation coefficient again, excluding Central, we obtain a value of 0.98, which represents an almost perfectly linear relationship. Note that this calculation is for illustration only; if we are interested in the relationship between water consumption and population, we should not exclude the outlier.

Section 2.6

Exercise 2.6.1

1. Plots of data





2. Findings: deaths

- Outer London: Steady at around 150–180 deaths per day until 4 December, and thereafter steady rise to peak of about 400 on 8 December, and then gradually falls to a little over 200 per day by the end of the observation period.
- Central London: Steady at around 110–140 deaths per day until 4 December (below level for outer London), then steep increase to peak of over 500 on 7–8 December (considerably above level for outer London), and thereafter falling to similar level as outer London.

3. Findings: pollution

- Smoke: County Hall levels began rising on 4–5 December, reaching a peak on 7–8 December, falling rapidly on 9 December, and returning to initial level by 12 December. The initial rise on 4–5 December was very rapid, and it appeared to occur a day in advance of the equivalent rise (and fall) in deaths. The Kew data show a very similar time pattern but at rather lower levels. Note that these values of smoke concentration (up to 4,460 µg/m³) are very high indeed.
- Sulphur dioxide (SO₂): County Hall levels mirror the smoke changes very closely, apart from the less-steep increase on 5–6 December.

4. Findings: temperature

At the same time as the pollution increased after 4 December, the temperature fell below freezing (32°F), as expected with the climatic conditions associated with smogs (cold,

trapped, still air). The temperature began rising again on 8 December, before the pollution began to fall, consistent with the fact that the change in climatic conditions allowed the polluted air to disperse. It can be seen how quite warm air then appeared on the scene over the next few days. Although the temperature data were reported only for the one site (Kew), it seems unlikely that there would have been any substantial variation in air temperature across the city.

5. Interpretation

How can we interpret these findings? We know a number of facts now:

- Deaths began increasing on 4 December, though mainly from 5 December, and increased most in central London.
- The pollution levels increased at the same time, and there is some evidence that the steep rise occurred a day or so in advance of the rise in deaths (especially for smoke), and fell a day or so in advance of the decline in deaths.
- The pollution reached much higher levels in central London, although levels were very high by modern standards in both parts.
- The temperature fell at the time that the pollution increased.

This information strongly suggests that the air conditions could have led to the increase in deaths. Was smoke, SO₂, or temperature responsible, or was it a combination of these? The spatial pattern (central versus outer London) and the time sequence (pollution changes precede changes in deaths) support this being a real and probably causal association.

6. What other information?

We have information on place (different parts of London) and time but not on person (e.g. age, sex). We could strengthen our conclusions with additional information, as follows:

- Break down the deaths by age, sex, and cause to see whether those most likely to be affected by air pollution (e.g. respiratory and heart conditions such as heart failure) showed the most dramatic increase.
- Obtain additional information on hospital admissions (morbidity), which could add numbers and detail.
- It would be useful to see whether a fall in temperature, not associated with marked increases in pollution, led to increased deaths. (In fact, it does, but not on the scale seen here: What happened in London in 1952 was a combined effect with a major contribution from the pollution, mainly due to the smoke component.)

Exercise 2.6.2

A problem can arise because we are assuming that just because the fat consumption (average per capita) is high for a country, it is also high for those individuals in that country who died. For all we know, the people who died could be the ones with average, or even lower-than-average, fat consumption in that country. Without data on the actual fat consumption of individuals, we cannot say whether or not this is the case.

3

Standardisation

Introduction and Learning Objectives

Standardisation is an important and commonly used technique that effectively allows you to make more valid comparisons between sets of data than would otherwise be possible. We are going to explore this method, and you will learn how to carry it out, by reference to comparisons of mortality data from two contrasting communities in Merseyside in the northwest of England.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the purpose of standardisation.
- Describe direct and indirect methods of standardisation, and calculate and interpret examples of each.
- Describe the concept of a confidence interval, and calculate and interpret this for a standardised mortality ratio.
- Describe key aspects of the direct and indirect methods of standardisation, and draw on these to select the most appropriate method for a given data set.

3.1 Health Inequalities in Merseyside

3.1.1 Socio-Economic Conditions and Health

It is important to be able to summarise levels of ill health within a local health system administrative unit (such as the local government area), as well as variations among these areas, for multiple reasons including allocating resources and monitoring health status. We now explore how to make such comparisons in a valid way through the example of *all-cause mortality* among men in two contrasting areas within Merseyside: the Local Authorities (local government areas or LAs) of Liverpool (high levels of unemployment and poverty) and Sefton (more affluent).

Liverpool and its surrounding communities, like all other cities, contains areas of poverty and wealth. These social inequalities are, however, very great, and the reasons lie principally in the history of the city's economic growth and decline. The growth of Liverpool was based on maritime trade; the industries that grew up in the Merseyside area processing raw materials brought

in from overseas, such as sugar and tobacco; and manufacturing industries, such as shipbuilding. The decline of this maritime trade, combined with the general loss of the manufacturing industry in the UK, hit Liverpool particularly badly, with the result that, historically, unemployment has been extremely high. More recently, starting with Objective 1 in 1994, Liverpool has attracted considerable investment through European Union funding and also as a result of its successful bid to become European Capital of Culture (2008). However, the legacy of high levels of unemployment, poverty, and lack of opportunity has been high levels of morbidity and mortality from common conditions such as heart disease and lung cancer. The comparison we will make is between the city of Liverpool and Sefton, which is a more-affluent neighbouring local government area to the north of Liverpool.

3.1.2 Comparison of Crude Death Rates

If we wish to compare the mortality in Liverpool and Sefton, either with each other or with the region or country, we could simply do so with the *crude death rates*. Based on data for the year 2011, the all-cause crude death rates for men were 8.83 per 1,000 per year in Liverpool and 11.34 per 1,000 per year in Sefton; for women the rates were 9.34 per 1,000 per year in Liverpool and 10.72 per 1,000 per year in Sefton. This appears to suggest that for both men and women, mortality is higher in Sefton, although this is not what we would expect given the socio-economic circumstances of the two areas. Let's now consider the consequences of one LA having a higher proportion of older people than the other. We would expect that this factor alone would result in a higher overall death rate in the LA with the older population. Table 3.1.1 shows the *age distributions* for men and women respectively for the two areas in 2011.

Table 3.1.1 Mid-year populations of Liverpool and Sefton LAs, 2011.

(a) Men				
Age group (years)	Liverpool		Sefton	
	Population	(%)	Population	(%)
0–4	13,538	5.87	7,552	5.75
5–9	11,345	4.92	7,044	5.37
10–14	12,106	5.25	7,894	6.01
15–19	16,119	6.99	8,557	6.52
20–24	25,666	11.14	8,216	6.26
25–29	21,176	9.19	7,479	5.70
30–34	17,607	7.64	6,644	5.06
35–39	14,139	6.14	7,096	5.40
40–44	15,467	6.71	9,063	6.90
45–49	15,232	6.61	10,015	7.63
50–54	14,361	6.23	9,857	7.51
55–59	12,710	5.52	8,678	6.61
60–64	12,253	5.32	8,986	6.84

(continued)

Table 3.1.1 (Continued)

(a) Men				
Age group (years)	Liverpool		Sefton	
	Population	(%)	Population	(%)
65–69	8,747	3.80	7,118	5.42
70–74	7,517	3.26	6,049	4.61
75–79	6,135	2.66	5,171	3.94
80–84	3,939	1.71	3,521	2.68
85–89	1,771	0.77	1,698	1.29
90 and older	619	0.27	651	0.50
Total	230,447	100.00	131,289	100.00
(b) Women				
Age group (years)	Liverpool		Sefton	
	Population	(%)	Population	(%)
0–4	12,782	5.43	7,064	4.95
5–9	10,984	4.67	6,746	4.73
10–14	11,917	5.07	7,629	5.35
15–19	16,820	7.15	8,200	5.75
20–24	25,393	10.80	7,722	5.41
25–29	18,849	8.01	7,478	5.24
30–34	15,568	6.62	6,834	4.79
35–39	13,826	5.88	8,067	5.65
40–44	15,286	6.50	9,622	6.74
45–49	16,701	7.10	11,052	7.75
50–54	15,073	6.41	10,485	7.35
55–59	12,872	5.47	9,154	6.42
60–64	12,152	5.17	9,469	6.64
65–69	9,198	3.91	8,144	5.71
70–74	8,655	3.68	7,581	5.31
75–79	7,865	3.34	6,731	4.72
80–84	5,851	2.49	5,418	3.80
85–89	3,565	1.52	3,260	2.28
90 and older	1,852	0.79	2,024	1.42
Total	235,209	100.00	142,680	100.00

Source: <http://www.ons.gov.uk/ons/index.html>



Self-Assessment Exercise 3.1.1

1. Examine these data, and comment on the age distributions of the two areas. You may find it helpful to plot the data in a histogram (this is provided for the male population in the answers in Section 3.5).
2. What are the implications of your observations for any comparison of crude death rates between the two areas?

Answers in Section 3.5

3.1.3 Usefulness of a Summary Measure

The analysis of the age distribution for the two areas shows how important it is to examine **age-specific death rates**. We shall do this shortly, but for many purposes it is very useful to have a summary measure of mortality for all age groups (or perhaps a wide band of age groups that we may be interested in). For example, if we wish to look at mortality for all local government areas in the country (or even in one region), imagine how cumbersome it would be to try to do so with many tables of rates for each age group, especially if all of these were to be included in a report.

What we are looking for is a summary measure of mortality that takes account of the differences in age distribution of the two areas. This is provided by a technique called **standardization**, which allows us to adjust an overall rate (e.g. death rate) for the type of age (and other) differences that we saw in Table 3.1.1. There are two methods:

- The **indirect** method, which provides the **standardised mortality ratio (SMR)** and **indirectly standardised rates**
- The **direct** method, which provides **directly standardised rates**

We will look at the **indirect** method first. Using Liverpool as an example, we will first work through the calculation and interpretation of the **SMR**. You will then have the opportunity to try this out with the Sefton data. Generally, standardisation should be done separately by sex, as this (like age) is such an important factor in health. Accordingly, we will use data for women for the calculations (our explanations and also for the exercises), but we will report the completed standardisation results for men as well, for comparison.

Summary: Why Standardise For Age?

- If two populations differ in their age structure, a simple comparison of overall disease or death rates is misleading.
- This is because age is a powerful determinant of disease and death rates.
- Standardisation provides a means of adjusting for these differences in age distribution, hence of making a more valid comparison.

3.2 Indirect Standardisation: Calculation of the Standardised Mortality Ratio (SMR)

3.2.1 Mortality in Liverpool

Table 3.2.1 provides all of the data that we require for carrying out the standardisation and for calculating the SMR:

- The population in Liverpool in each age group in 2011
- The age-specific annual death rates for a standard population, in this case, England and Wales, for 2011.

The table also provides the number of deaths that occurred in each age group in Liverpool for the year 2011. While it is not necessary to have information on the numbers of deaths in each age group in order to calculate the SMR, we have included these in the table so that you can calculate and compare the local and national mortality rates.

Table 3.2.1 Female deaths, population and death rates (Liverpool), and death rates (England and Wales), for 2011.

Age group (years)	Deaths: Liverpool 'observed'	Liverpool population	Liverpool age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths 'expected' in Liverpool in 2011
0–4	10	12,782	0.78	0.90	11.50
5–9	0	10,984	0.00	0.06	0.66
10–14	0	11,917		0.08	
15–19	2	16,820		0.15	
20–24	6	25,393		0.23	
25–29	5	18,849		0.29	
30–34	8	15,568		0.45	
35–39	16	13,826		0.66	
40–44	24	15,286		1.06	
45–49	38	16,701		1.62	
50–54	47	15,073		2.59	
55–59	68	12,872		4.01	
60–64	99	12,152		6.26	
65–69	135	9,198		9.79	
70–74	208	8,655		16.41	
75–79	320	7,865		28.43	
80–84	388	5,851		52.64	
85–89	402	3,565		96.95	
90 and older	421	1,852		198.15	
Total	2197	235,209			



Self-Assessment Exercise 3.2.1

- We look first at the **age-specific rates** for Liverpool. For age 0–4, this has been calculated as 10 (deaths) divided by 12,782 (population) and multiplied by 1,000 (to express as a rate per 1,000 per year). This comes out at 0.78 deaths per 1,000 population per year.
 - Calculate the rate for the age group 5–9, and ensure that you obtain the correct value of 0.00 per 1,000 per year.
 - Now calculate the missing values for the remaining age groups, and check these against the answers (Section 3.5).

Compare the age-specific rates for Liverpool with those for England and Wales. What do you notice? It has been useful to calculate age-specific mortality rates for Liverpool so that we can compare rates with those of England and Wales, but we do not need to do this in order to calculate an SMR. In fact, we do not need to know how many deaths occurred in each age group in Liverpool, either, but only the overall number of deaths.

- We now need to calculate the numbers of deaths that would have occurred in Liverpool, in each age group, if the England and Wales death rates had applied. This is in the last column and is termed the **expected deaths**. For the 0–4 age group, we multiply the national rate (0.90 per 1,000 per year) by the local population in the same age group (12,782), and obtain 11.50 **expected deaths**:

$$\frac{0.90 \times 12,782}{1,000} = 11.50$$

Of course, we cannot really have fractions of deaths! This is not a problem here though, since we are dealing with a hypothetical situation of 'How many deaths would you expect if national rates applied to the local population?'

- Check now that you can calculate the correct number for the 5–9 age group (0.66 deaths).
 - Finally, calculate the numbers for the other age groups, and check your results with the answers.
- To calculate the SMR, we need the total expected deaths. This is calculated by adding up all the age-specific expected deaths (11.50 + 0.66 + ...). Note, it is *not* calculated by multiplying the total Liverpool population by the all-age (crude) mortality rate for England and Wales.
 - We are now in a position to calculate the SMR. This is defined as

$$\frac{\text{Total number of observed deaths}}{\text{Total number of expected deaths}} = \frac{2197}{1729.35}$$

It is usually expressed as if it were a percentage, and therefore multiplied by 100. So for females in Liverpool, the SMR is 2197 (observed) divided by 1729.35 (expected), multiplied by 100, which comes out at 127.04, or 127 rounded to the nearest whole number.

Answers in Section 3.5

Summary: Calculating the SMR

- An SMR is calculated by the indirect method of standardisation.
- To do this, we need to select a standard population (e.g. the country) and know the rates for each category of the variable we wish to standardise for (in this case, age).
- Standardisation is achieved by applying these standard category-specific rates to the category-specific population in the area to be standardised. This yields expected cases.

- The SMR is calculated by adding up all the expected cases and presenting the ratio of observed to expected, usually as a percentage.

3.2.2 Interpretation of the SMR

The SMR was calculated using the England and Wales rates as a standard. In other words, we are comparing the actual (observed) numbers of deaths in Liverpool with the number that would have occurred (expected) if the local population experienced these national mortality rates in each age group. An SMR of 127 therefore means that, independent of the influence of the age distribution in Liverpool, the overall mortality in that LA was 27 per cent higher than that in England and Wales in 2011. The SMR for males in Liverpool was 125, implying that mortality was 25% higher than that in England and Wales in 2011, and a very similar result to that for females. We are not yet in a position to make any comparison with Sefton, but we will come to that shortly.

3.2.3 Dealing With Random Variation: The 95 per cent Confidence Interval

It can be appreciated that the exact number of deaths occurring in Liverpool in a given year is determined by the level of mortality in that population, but it will inevitably vary from year to year. This variation is essentially due to *chance* (random), so that in one year there may be 2,735 deaths, in the next year 2,681, in the following year 2,700, and so on. This random variation is separate from systematic effects such as a *trend* over time, which may show a gradual increase or decrease in the level of mortality. It is also distinct from variation due to changes in the population numbers or age distribution.

The result of this chance, year-to-year variation is that the SMR is subject to *random error*. This error is greater if the numbers of deaths are small, as with a small population and/or a less-common cause of death. So we have a female SMR of 127, but we also know that this *estimate* for a specific year could, by chance, be higher or lower than the true level of mortality (compared to the standard population). The question is, how much random variation might there be? That is, can we quantify this random variation? The way to answer this important question is by calculating a *confidence interval (CI)*. We will be exploring the uses and calculation of CIs a good deal more in later chapters (including why the confidence interval is usually set at a value of 95 per cent), so at this stage we will focus mainly on the purpose and interpretation. We will, however, calculate the 95 per cent CI for the SMR in this exercise, to allow a more complete understanding of the interpretation of an SMR. The 95 per cent CI for the SMR for a given population can be defined as follows:

The 95 per cent CI for an SMR is the range in which we can be 95 per cent confident that the true mortality value for the population lies.

The calculation of this 95 per cent CI is quite straightforward by the following method (which is adequate for most purposes):

$$\begin{aligned} \text{Upper limit of 95\% CI for SMR} &= \text{SMR} + 1.96 \times \frac{\text{SMR}}{\sqrt{\text{Number of observed deaths}}} \\ \text{Lower limit of 95\% CI for SMR} &= \text{SMR} - 1.96 \times \frac{\text{SMR}}{\sqrt{\text{Number of observed deaths}}} \end{aligned}$$

The derivation of the 1.96 multiplication factor will be covered in Chapter 4 (Surveys). Using this formula, the upper limit is therefore

$$127.04 + \left[1.96 \times \frac{127.04}{\sqrt{2197}} \right] = 127.04 + 5.31 = 132.35$$

There is no point being more precise with the confidence limits than for the SMR, so we round this to 132. The lower limit therefore is $127.04 - 5.31 = 121.73$, which we round to 122.

The term in the formula ($SMR/\sqrt{\text{number of observed deaths}}$) is known as the *standard error* (SE) of the SMR. The SE is a measure of the precision of an estimate, and this applies whether we are talking about an SMR, a mean, a proportion, a measure of risk, etc. We will look at SE in more detail when thinking about the precision of means derived through sample surveys in Chapter 4. At this stage, it is sufficient to understand the concept of SE as being a measure of how big the margin of random error is around an estimate obtained from a sample.

We have now determined that the Liverpool SMR for females in 2011 was 127, with a 95 per cent CI of 122–132. If we apply these figures to the definition given above, we can be 95 per cent confident that the true SMR lies between 122 and 132, or 22 to 32 per cent above the national level of mortality. It is clear, then, that mortality for females in Liverpool for 2011, independent of the age distribution, is still some way above the national average, which is represented by the value of 100. We saw earlier that the male SMR for Liverpool in 2011 was 125. The 95% CI for this SMR is 120–131, so we can be very confident that male mortality in Liverpool was also well above the national average, after allowing for differences in the age distributions.

3.2.4 Increasing Precision of the SMR Estimate

We have been able to estimate the precision of the SMR quite accurately, as Liverpool includes quite a large population. However, this would not necessarily be true if we were dealing with a specific cause of mortality (rather than all-cause as we have done so far), or with much smaller population groups, for example at the electoral ward level. In these circumstances, we may find that our estimate is very imprecise with a wide 95% CI. In this case, the best way to reduce this imprecision is to increase the numbers of events. We usually do this by calculating the SMR for a longer period of time, say, 3 or 5 years.

Summary: 95 per cent CI for the SMR

- The SMR is subject to year-on-year random variation.
- It is very helpful to be able to quantify this variation, and this quantification is provided by the 95 per cent CI.
- The statistical derivation of the CI, the reason for choosing a 95 per cent CI, etc., will all be discussed in later chapters.

3.2.5 Mortality in Sefton

Now it is your turn to calculate the SMR for the more-affluent area of Sefton. Work through the data provided in Table 3.2.2 and the questions that follow. Once we have completed this stage, we will look at how to compare the results with those for Liverpool.

Table 3.2.2 Female deaths, population and death rates (Sefton), and death rates (England and Wales), for 2011.

Age group (years)	Deaths: Sefton 'observed'	Sefton population	Sefton age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths 'expected' in Sefton in 2011
0–4	2	7,064		0.90	
5–9	0	6,746		0.06	
10–14	0	7,629		0.08	
15–19	0	8,200		0.15	
20–24	0	7,722		0.23	
25–29	5	7,478		0.29	
30–34	2	6,834		0.45	
35–39	8	8,067		0.66	
40–44	14	9,622		1.06	
45–49	18	11,052		1.62	
50–54	32	10,485		2.59	
55–59	36	9,154		4.01	
60–64	69	9,469		6.26	
65–69	85	8,144		9.79	
70–74	128	7,581		16.41	
75–79	159	6,731		28.43	
80–84	275	5,418		52.64	
85–89	317	3,260		96.95	
90 and older	379	2,024		198.15	
Total	1529	142,680			



Self-Assessment Exercise 3.2.2

1. Calculate the age-specific mortality rates for Sefton, and put these into the table. Comment on how these rates compare with those for Liverpool (Table 3.2.1).
2. Calculate the expected numbers of deaths from the 2011 England and Wales age-specific rates, and put these into the table.
3. Calculate the SMR for Sefton.
4. Calculate the 95 per cent CI for the Sefton SMR.
5. Interpret the results you now have for Sefton (do not compare with Liverpool yet).

Answers in Section 3.5

This exercise shows that there was no evidence that the female all-cause mortality for Sefton in 2011 differed from that in England and Wales. The SMR for males in Sefton was 110, with a 95% CI of 105 to 116, implying there is good evidence that male all-cause mortality in Sefton in 2011 was higher than that for the country. These results raise questions as to whether the male

mortality is higher than that for females in Sefton, and whether mortality in Liverpool is higher than that for Sefton once differences in age distributions are accounted for. We will now look at how to address these comparisons.

3.2.6 Comparison of SMRs

Starting with the two areas, it might be tempting to make a direct comparison between the SMRs and say that female mortality in Liverpool in 2011 was $127 \div 97 = 1.31$ times higher than for females in Sefton. This, however, would be incorrect, as SMRs should not be compared directly in this way. What we can say, based on the single year of data for 2011, is that female mortality in Liverpool is 22 to 32 per cent higher than the national average. For Sefton, female mortality lies between 8 per cent lower and 2 per cent higher; that is, it does not differ significantly from the national level because the 95 per cent CI includes 100. The reason we cannot make a direct comparison between the two SMRs is that the age-specific rates used for the standardisation have been applied to two different populations. Similar considerations apply to comparison of the 2011 male and female SMRs for Sefton, which we noted to be different. Although we cannot compare them directly, we can say that mortality for males was higher than that for the country (we are 95% confident that mortality is higher by between 5 and 16 per cent), but there was no evidence that female mortality differed from the national rates.

In Section 3.4, we will look at *direct standardisation*. One characteristic of this method is that it does allow a simple comparison between standardised rates.

3.2.7 Indirectly Standardised Mortality Rates

If we wish to express the mortality in Liverpool or Sefton as an overall rate (equivalent to the crude rate, but standardised), it is derived as follows:

$$\text{Indirectly standardised rate} = \frac{\text{SMR} \times \text{crude rate for the standard population}}{100}$$

The female crude death rate for the standard population (England and Wales) in 2011 was 8.75 per 1,000/year:

- For Liverpool, the female *indirectly standardised rate* = 1.2704×8.75 per 1,000/year = 11.12 per 1,000/year.
- For Sefton, the female *indirectly standardised rate* = 0.9739×8.75 per 1,000/year = 8.52 per 1,000/year.

Compare these with the unstandardised crude rates of 9.34 per 1,000/year for Liverpool and 10.72 per 1,000/year for Sefton to see how much difference the age-standardisation has made. In particular, while the crude rate for Liverpool was actually lower than that for Sefton, for the standardised rates, Liverpool is the higher of the two.

3.3 Direct Standardisation

3.3.1 Introduction

We mentioned earlier that there are two main ways of standardising: *indirect* and *direct*. In both methods, the underlying principle is that we break the data down into categories of the

factor we wish to standardise for (e.g. age, social class, ethnic group), and apply category-specific rates to category-specific population numbers to find out how many events to expect if those rates had applied. The difference between the two methods is as follows:

- With the *indirect* method (e.g., for the SMRs), we used age-specific rates from another standard population (England and Wales) and applied these to the population numbers (in the same age bands) for the group we intended to standardise (often referred to as the *index* population). This method is termed ‘indirect’, because the rates are obtained from a population other than the index.
- With the *direct* method, we use the age-specific rates from the group we intend to standardise (the *index* population), and we apply these to the numbers of people (in the same age bands) in a standard population. This method is termed ‘direct’, because the rates are obtained from the index population.

3.3.2 An Example: Changes in Deaths From Stroke Over Time

Suppose that we are helping to plan and develop stroke services in a health administrative area with a relatively high percentage of older residents. In the background review, we looked at the *crude death rates* for women aged 65 and older for the last 20 years, and we were surprised to find that these had increased from 10 per 1,000 per year in 1995 to 12.8 per 1,000 per year in 2015. We are aware that the population has been ageing (that is to say, there is now a higher proportion of older people in the population when compared to 1995) and that it is a popular area for retirement, and we therefore wonder whether this might be the explanation, or at least part of it. The data are as shown in Table 3.3.1.

Table 3.3.1 Population and observed deaths from stroke for hypothetical population, 1995 and 2015.

Age group (years)	1995			2015		
	Population	Deaths observed	Age-specific rate/ 1,000 per year	Population	Deaths observed	Age-specific rate/ 1,000 per year
65–74	5,000	25		4,100	15	
75–84	3,500	35		3,000	31	
85+	1,500	40		2,900	78	
Total	10,000	100	10.0	10,000	124	12.4



Self-Assessment Exercise 3.3.1

In this exercise, we are going to standardise the 2015 data (the index) to find out the expected deaths in 2015, using the 1995 population as the standard.

1. First of all, study the data in Table 3.3.1 and describe what has happened to the population numbers and the observed numbers of deaths from stroke.
2. The first step in direct standardisation is to calculate and study the category-specific death rates (in this case, age-specific rates for 2015). Put these rates into the table in the spaces provided.
3. What do you notice about the age-specific rates?
4. You will now standardise the 2015 rate to the 1995 population. To do this,

- a. Transfer your 2015 age-specific rates to the table below, and apply these to the 1995 age-specific population to calculate the expected numbers in each age group. Enter these results into the table.

Age group (years)	Age-specific rate for 2015	Population 1995	Expected deaths for 2015
65–74		5,000	
75–84		3,500	
85+		1,500	
Total		10,000	

- b. Add up the expected deaths to obtain the total.
 c. Divide the total expected cases by the total population in the standard (which is 1995 in this case) to obtain the age-standardised death rate.
 5. Before checking the answers, have a go at interpreting the result you have obtained.

Answers in Section 3.5

3.3.3 Using the European Standard Population

In this example, we have standardised one group (2015) against the population of the other group (1995). Thus, 1995 has been taken as the standard population, and this allows comparison between the standardised rates. As an alternative, we could have standardised both groups against another standard population, which also allows comparison between the two groups. Although this gives essentially similar results, we will examine one example of how this is done, because it is a method that you are quite likely to encounter.

So, what other population can we use as a standard? We could have used the national population, or one of a number of standard populations available for this purpose. These are not real populations, but they are created to represent the population structure of the area we are dealing with. Thus, we will use the *European standard population* (Table 3.3.2, but there

Table 3.3.2 The European standard population.

Age group (years)	Population	Age group (years)	Population
0 (infants)	1,000	45–49	7,000
1–4	4,000	50–54	7,000
5–9	5,500	55–59	6,500
10–14	5,500	60–64	6,000
15–19	5,500	65–69	5,500
20–24	6,000	70–74	5,000
25–29	6,000	75–79	4,000
30–34	6,500	80–84	2,500
35–39	7,000	85 and older	2,500
40–44	7,000	Total	100,000

are others, such as that for Africa, which has a much younger age distribution than that for Europe.



Self-Assessment Exercise 3.3.2

1. Try standardising both the 1995 and 2015 data from the previous exercise (Table 3.3.1) to the European standard population, which is reproduced above. To do this, you will need to use the population numbers from the relevant age groups in Table 3.3.2 and apply (to these population data) the age-specific rates for 1995, and then those for 2015, to obtain the expected deaths for each year.
2. Does the use of this alternative standard population make any difference to your conclusions?

Answers in Section 3.5

3.3.4 Direct or Indirect: Which Method is Best?

Having looked now at both indirect and direct standardisation, how do we decide which is the better method to use in any given situation? Table 3.3.3 summarises and compares the most important characteristics of the two methods.

Table 3.3.3 Comparison of direct method and indirect method.

Characteristic	Indirect method	Direct method
Age-specific rates in populations to be standardised (index populations)	Not actually required, as we need only the age-specific population data. Thus, this is the only method that can be used if the numbers of deaths (cases) in each age group are not available. If they are available, it is wise to check index populations anyway.	Age-specific rates in index population are required; hence, both the number of cases and the population in each age group must be available (unless actual rates can be obtained).
Precision	Generally more precise, since category-specific rates are those from a standard population, which can be selected to be as precise as possible. This is especially true when numbers of deaths (or cases) are relatively small in the categories of the index population.	Since we have to rely on category-specific rates in the index population, the result may be imprecise if these rates are based on small numbers of cases.
Comparison of standardised rates	Direct comparison is not possible.	Direct comparison is possible.

In making a choice, the most important criterion is likely to be the precision, especially if we are dealing with small numbers, such as cause-specific mortality data or the analysis is at the level of an electoral ward. In these situations, the indirect method may be more appropriate.

3.4 Standardisation for Factors Other Than Age

In the examples used in this section, we have standardised death rates for age, but the method is equally applicable for any rate, such as incidence of disease, and for standardising for other factors that could affect a comparison that we wish to make (e.g. socio-economic classification, geographical area of residence, ethnic group).

Taking socio-economic classification as an example (see Chapter 2, Section 2.1.5), we would break down the data into categories of (group 1: large employers and higher managerial; group 1.2: higher professional; group 2: lower managerial and professional; group 3: intermediate, through to group 8: never worked and long-term unemployed) in exactly the same way as we have used age categories (0–4, 5–14, etc.). Obtaining the category-specific rates and populations that we require may not always be as straightforward as for age, but the principle is the same as that we have worked through for age standardisation.

Summary

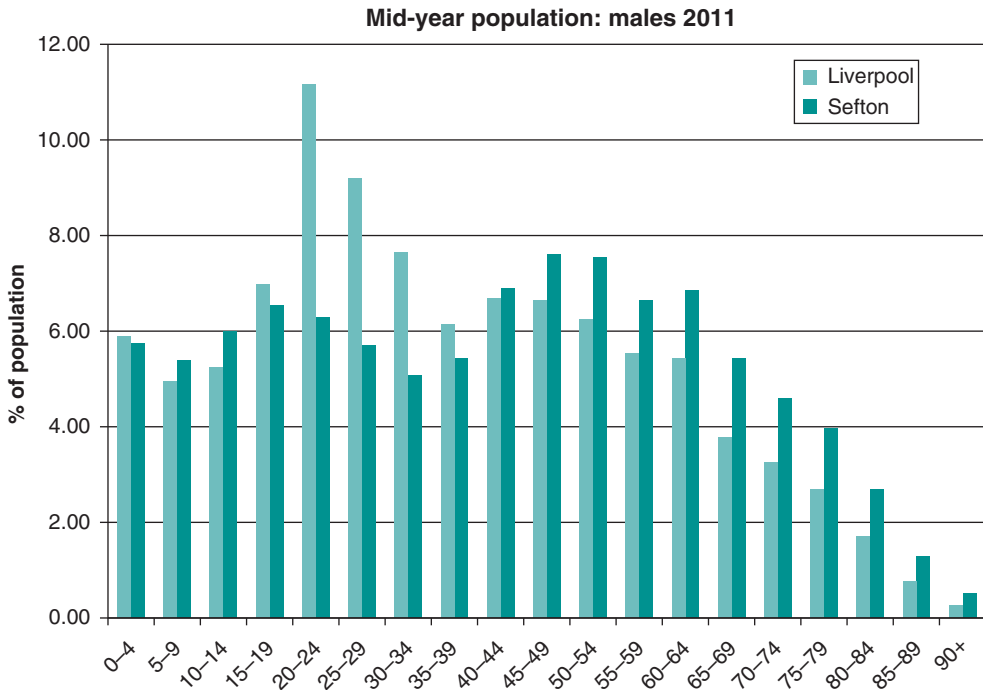
- Standardisation is carried out to adjust rates for the influence of one or more factors that could affect the comparison of those rates. These factors may include age, socioeconomic classification, area of residence, ethnic group, etc.
- There are two main methods of standardisation: indirect and direct.
- Indirect standardisation applies category-specific rates from a standard population to the numbers of people in each category in the index population, and it therefore estimates the rate in the index population if the rates in the standard population had applied.
- Direct standardisation applies category-specific rates from the index population to the numbers of people in each category of a standard population, and it therefore estimates the rate in the index population if the standard population structure had applied.
- Indirect standardisation is generally more precise, does not require category-specific rates in the index population for its calculation (but see the next point), and does not allow direct comparison with other indirectly standardised rates.
- It is always a good idea to check the category-specific rates in the populations being standardised, to see whether there are any important inconsistencies across categories that should be highlighted and that may be masked by an overall standardised rate or standardised ratio.

3.5 Answers to Self-Assessment Exercises

Exercise 3.1.1

1. Population structures

These data clearly show that Liverpool has a younger population than Sefton, for both males and females. A histogram for the male populations is shown below:



2. Implications

All other factors being equal, since Liverpool has a younger population, we would expect a lower death rate. Our aim now is to find out how the two areas compare, independently of the effects of this difference in age distribution.

Exercise 3.2.1**1. Age-specific rates (2011)**

Age group (years)	Deaths: Liverpool observed	Liverpool population	Liverpool age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths expected in Liverpool in 2011
0–4	10	12,782	0.78	0.90	11.50
5–9	0	10,984	0.00	0.06	0.66
10–14	0	11,917	0.00	0.08	
15–19	2	16,820	0.12	0.15	
20–24	6	25,393	0.24	0.23	
25–29	5	18,849	0.27	0.29	
30–34	8	15,568	0.51	0.45	
35–39	16	13,826	1.16	0.66	
40–44	24	15,286	1.57	1.06	
45–49	38	16,701	2.28	1.62	
50–54	47	15,073	3.12	2.59	
55–59	68	12,872	5.28	4.01	
60–64	99	12,152	8.15	6.26	
65–69	135	9,198	14.68	9.79	
70–74	208	8,655	24.03	16.41	
75–79	320	7,865	40.69	28.43	
80–84	388	5,851	66.31	52.64	
85–89	402	3,565	112.76	96.95	
90 and older	421	1,852	227.32	198.15	
Total	2197	235,209			

2. Comparison of age-specific rates with England and Wales

For all age groups over 30 years, the age-specific rates for Liverpool are above those for the country. Note that for younger age groups where the number of deaths are small, rates are highly influenced by just one additional death, illustrating well how unstable these rates are when we are dealing with small areas and numbers.

3. Expected numbers of deaths

Age group	Deaths: Liverpool observed	Liverpool population	Liverpool age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths expected in Liverpool in 2011
0–4	10	12,782	0.78	0.90	11.50
5–9	0	10,984	0.00	0.06	0.66
10–14	0	11,917	0.00	0.08	0.95
15–19	2	16,820	0.12	0.15	2.52
20–24	6	25,393	0.24	0.23	5.84
25–29	5	18,849	0.27	0.29	5.47
30–34	8	15,568	0.51	0.45	7.01
35–39	16	13,826	1.16	0.66	9.13
40–44	24	15,286	1.57	1.06	16.20
45–49	38	16,701	2.28	1.62	27.06
50–54	47	15,073	3.12	2.59	39.04
55–59	68	12,872	5.28	4.01	51.62
60–64	99	12,152	8.15	6.26	76.07
65–69	135	9,198	14.68	9.79	90.05
70–74	208	8,655	24.03	16.41	142.03
75–79	320	7,865	40.69	28.43	223.60
80–84	388	5,851	66.31	52.64	308.00
85–89	402	3,565	112.76	96.95	345.63
90 and older	421	1,852	227.32	198.15	366.97
Total	2,197	235,209			1,729.35

Exercise 3.2.2**1. SMR for Sefton**

Age-specific rates are shown in the table below.

Age group	Deaths: Sefton observed	Sefton population	Sefton age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths expected in Sefton in 2011
0–4	2	7,064	0.28	0.90	
5–9	0	6,746	0.00	0.06	
10–14	0	7,629	0.00	0.08	
15–19	0	8,200	0.00	0.15	
20–24	0	7,722	0.00	0.23	
25–29	5	7,478	0.67	0.29	
30–34	2	6,834	0.29	0.45	
35–39	8	8,067	0.99	0.66	
40–44	14	9,622	1.45	1.06	
45–49	18	11,052	1.63	1.62	
50–54	32	10,485	3.05	2.59	
55–59	36	9,154	3.93	4.01	
60–64	69	9,469	7.29	6.26	
65–69	85	8,144	10.44	9.79	
70–74	128	7,581	16.88	16.41	
75–79	159	6,731	23.62	28.43	
80–84	275	5,418	50.76	52.64	
85–89	317	3,260	97.24	96.95	
90 and older	379	2,024	187.25	198.15	
Total	1,529	142,680			

Apart from a few age groups, each of which has a very small number of deaths, the Sefton age-specific mortality rates are consistently below those for Liverpool. This is important and reassuring, because it is unwise to standardise if there is a less-consistent pattern: Suppose, for example, that the Sefton rates for young/middle-aged people were higher than Liverpool's, whereas those for older people were lower. When calculating an SMR, these two characteristics of the data would be mixed, and the overall SMR would be misleading.

2. Expected deaths

Age group	Deaths: Sefton observed	Sefton population	Sefton age-specific mortality rate per 1,000/year	England and Wales mortality rate per 1,000/year	Deaths expected in Sefton in 2011
0–4	2	7,064	0.28	0.90	6.36
5–9	0	6,746	0.00	0.06	0.40
10–14	0	7,629	0.00	0.08	0.61
15–19	0	8,200	0.00	0.15	1.23
20–24	0	7,722	0.00	0.23	1.78
25–29	5	7,478	0.67	0.29	2.17
30–34	2	6,834	0.29	0.45	3.08
35–39	8	8,067	0.99	0.66	5.32
40–44	14	9,622	1.45	1.06	10.20
45–49	18	11,052	1.63	1.62	17.90
50–54	32	10,485	3.05	2.59	27.16
55–59	36	9,154	3.93	4.01	36.71
60–64	69	9,469	7.29	6.26	59.28
65–69	85	8,144	10.44	9.79	79.73
70–74	128	7,581	16.88	16.41	124.40
75–79	159	6,731	23.62	28.43	191.36
80–84	275	5,418	50.76	52.64	285.20
85–89	317	3,260	97.24	96.95	316.06
90 and older	379	2,024	187.25	198.15	401.06
Total	1,529	142,680			1,570.01

3. Calculation of SMR

The SMR is observed/expected $\times 100$; hence, $SMR = (1529 \div 1570.01) \times 100 = 97.39$. This can be rounded to 97.

4. Calculation of 95 per cent CI

Refer to Section 2.6 to check the formula for the CI. The $SE = 2.49$, so $(1.96 \times SE) = 4.88$, and the 95 per cent CI = 92.51 to 102.27 (93 to 102 when rounded to nearest whole numbers).

5. Interpretation of SMR

We can be 95 per cent confident that the female all-cause mortality for Sefton (2011) lies between 7% below the national level and 2% above it. This range includes 100, which is the value for the standard population (England and Wales), and hence we have no evidence that the SMR for Sefton differs from that for England and Wales.

Exercise 3.3.1**1. Change in population and deaths**

The population has aged (there is now a higher percentage of people represented in the older age group). The numbers of deaths have fallen in the 65–74 age group, are unchanged in the 75–84 age group, and have risen (substantially) in the 85+ year age group.

2. Calculation of age-specific death rates

Age group (years)	1995			2015		
	Population	Deaths observed	Age-specific rate/1,000 per year	Population	Deaths observed	Age-specific rate/1,000 per year
65–74	5,000	25	5.0	4,100	15	3.7
75–84	3,500	35	10.0	3,000	31	10.3
85+	1,500	40	26.7	2,900	78	26.9
Total	10,000	100	10.0	10,000	124	12.4

3. Changes in age-specific rates

Between 1995 and 2015, the death rates have fallen in the youngest age group but have stayed almost constant (risen very slightly) in the older two age groups.

4. Calculation of the directly standardised rate

Age group (years)	Age-specific rate for 2015	Population 1995	Expected deaths for 2015
65–74	3.7	5,000	18.5
75–84	10.3	3,500	36.1
85+	26.9	1,500	40.4
Total	9.5	10,000	95.0

The age-standardised rate for 2015 is $(95.0/10,000) \times 1,000 = 9.5$ per 1,000 per year. This is just a little lower than the crude rate for 1995 (10 per 1,000/year). You can directly compare these by dividing 9.5 per 1,000/year by 10 per 1,000/year, to get 0.95, or, expressed as a percentage, 95 per cent.

5. Interpretation

You will note that after standardisation of the 2015 rate to the 1995 population, far from being higher than the 1995 rate, the 2015 rate is actually a little lower. Thus, it appears that the changing population structure was the main reason for the rise in the crude rate. Please note, however, that the age-specific rates have changed in an inconsistent way, and if we had not examined these, we could have wrongly assumed that the situation had improved for all three age groups given that the 2015 standardised rate is 95 per cent of the 1995 crude rate.

Exercise 3.3.2**1. Standardisation using the European standard population**

The directly standardised rates using the European standard population are as follows:

Age group (years)	European population	1995		2015	
		Rate	Expected	Rate	Expected
65–74	10,500	5.0	52.5	3.7	38.9
75–84	6,500	10.0	65	10.3	67
85+	2,500	26.7	66.8	26.9	67.3
Total	19,500		184.3		173.2

1995: $(184.3/19500) \times 1,000 = 9.45$ per 1,000/year

2015: $(173.2/19500) \times 1,000 = 8.88$ per 1,000/year

Ratio of rates (2015:1995) = 94.0 per cent.

2. Does it make any difference?

You will notice that the actual values of the rates are different when calculated with the European standard population, and the ratio of rates is a little lower (94.0% compared to 95%), but the overall conclusion is very similar. The different results are due to the standardising population being different and, more specifically, the population structure being different; in the original example, the population was much more heavily weighted towards elderly people than the European standard, as this was a retirement area. Thus, it is important to check that the standard population used is appropriate to the study population, even though the process is fairly robust.

4

Surveys

Introduction and Learning Objectives

In Chapter 2 we saw that although we can learn a great deal from the presentation and analysis of routinely collected data, we are inevitably limited to using the information that has been collected for purposes that, to a greater or lesser extent, often differ from our own. Not only may the information be different from what we are after, but it also may have been collected in a way that does not meet the rigorous standards we seek. Surveys provide the opportunity to fill in these gaps; both in the nature of the information collected and in the means by which it is collected.

Learning Objectives
<p>By the end of this chapter, you should be able to do the following:</p> <ul style="list-style-type: none"> • Define concepts of population, sample, and inference, and explain why sampling is used. • Describe the main probability and non-probability sampling methods, including simple random, stratified, cluster, multistage, convenience, systematic, and quota, giving examples of their uses. • Define bias, and give examples relevant to sampling (selection bias). • Define sampling error, and describe its relationship with sample size. • Define, calculate, and interpret standard error for a mean and a proportion. • Discuss the issue of sample size in relation to precision of estimates from a sample. • Define, calculate, and interpret the confidence interval (CI) for a population mean and a population proportion. • Calculate sample size for precision of a sample estimate, for continuous data (mean), and categorical data (proportion). • Describe the principles of valid and repeatable measurement, with examples from interview and questionnaire material. • Describe sources of bias in measurement and some of the ways good design can minimise bias. • Prepare a simple questionnaire, employing principles of good design. • Define categorical, discrete, and continuous data types, providing examples of each type, and present these by suitable display methods.

We will be studying survey methods, mainly using the example of the National Survey of Sexual Attitudes and Lifestyles (Natsal), which provides high-quality information on sexual attitudes and lifestyles that is not available from any other source.

The original survey was carried out in 1990 (Natsal-1) and was repeated in 2000 (Natsal-2) and then 2010 to 2012 (Natsal-3). For the purposes of this chapter, we will be looking mainly at the second follow-up study known as Natsal-3. However, we will also refer to Natsal-1 and Natsal-2, where they include more detail on some important aspects of study design.

Resource Papers

Paper A

Mercer, C.H., Tanton, C., Prah, P., Erens, B., Sonnenberg, P., *et al.* (2013). Changes in sexual attitudes and lifestyles through the lifecourse and trends over time: Findings from the British National Surveys of Sexual Attitudes and Lifestyles (Natsal-3). *Lancet* **382**, 1781–1794.

Paper B

Wellings, K., Wadsworth, J., Field J., Johnson A.M., Bradshaw, S.A., Anderson, R.M. (1990). Sexual lifestyles under scrutiny. *Nature* **348**, 276–278.

4.1 Purpose and Context

4.1.1 Defining the Research Question

We saw in Chapter 1 how important it is to formulate a clear research question. Therefore, one of the first jobs for the research team is to define the purpose of the study as clearly and concisely as possible. This is very important for those carrying out the research, but it is also important that what the research is trying to achieve is absolutely clear to anyone reading the published report. Any assessment you make of the methods, findings, interpretation, and conclusions will be of limited value if the underlying purpose is not clear in your mind. One useful way to formulate a research question is to use the PICO model. Using this model, a research question should clearly identify the patient (or population), the intervention (or exposure), the comparison (if appropriate), and the outcome you are seeking to achieve. Such an approach helps the research team define exactly what they are setting out to do and is used widely in clinical and epidemiological research. It is discussed further in Chapter 9.

Please now read the following Summary (Abstract) and introduction taken from paper A.

Summary

Background

Sexual behaviour and relationships are key components of well-being and are affected by social norms, attitudes, and health. We present data on sexual behaviours and attitudes in Britain (England, Scotland, and Wales) from the three National Surveys of Sexual Attitudes and Lifestyles (Natsal).

Methods

We used a multistage, clustered, and stratified probability sample design. Within each of the 1727 sampled postcode sectors for Natsal-3, 30 or 36 addresses were randomly selected and then assigned to interviewers. To oversample individuals aged 16–34 years, we randomly allocated

addresses to either the core sample (in which individuals aged 16–74 years were eligible) or the boost sample (in which only individuals aged 16–34 years were eligible). Interviewers visited all sampled addresses between Sept 6, 2010, and Aug 31, 2012, and randomly selected one eligible individual from each household to be invited to participate. Participants completed the survey in their own homes through computer-assisted face-to-face interviews and self-interview. We analysed data from this survey, weighted to account for unequal selection probabilities and non-response to correct for differences in sex, age group, and region according to 2011 Census figures. We then compared data from participants aged 16–44 years from Natsal-1 (1990–91), Natsal-2 (1999–2001), and Natsal-3.

Findings

Interviews were completed with 15 162 participants (6293 men, 8869 women) from 26 274 eligible addresses (57.7%). 82.1% (95% CI 81.0–83.1%) of men and 77.7% (76.7–78.7%) of women reported at least one sexual partner of the opposite sex in the past year. The proportion generally decreased with age, as did the range of sexual practices with partners of the opposite sex, especially in women. The increased sexual activity and diversity reported in Natsal-2 in individuals aged 16–44 years when compared with Natsal-1 has generally been sustained in Natsal-3, but in men has generally not risen further. However, in women, the number of male sexual partners over the lifetime (age-adjusted odds ratio 1.18, 95% CI 1.08–1.28), proportion reporting ever having had a sexual experience with genital contact with another woman (1.69, 1.43–2.00), and proportion reporting at least one female sexual partner in the past 5 years (2.00, 1.59–2.51) increased in Natsal-3 compared with Natsal-2. While reported number of occasions of heterosexual intercourse in the past 4 weeks had reduced since Natsal-2, we recorded an expansion of heterosexual repertoires—particularly in oral and anal sex—over time. Acceptance of same-sex partnerships and intolerance of non-exclusivity in marriage increased in men and women in Natsal-3.

Interpretation

Sexual lifestyles in Britain have changed substantially in the past 60 years, with changes in behaviour seeming greater in women than men. The continuation of sexual activity into later life—albeit reduced in range and frequency—emphasises that attention to sexual health and wellbeing is needed throughout the life course.

You should not be concerned if you do not understand all the technical terms used in describing the study. We deal with all of these as we work through this chapter, and we will only assume understanding of concepts and methods that have already been covered or that are being introduced in the section to which the paper refers.

Introduction

Improving sexual and reproductive health remains a public health priority in Britain (England, Scotland, and Wales), as it does globally. A range of factors contribute to a population's sexual health, such as social context and the interplay between behaviour, relationships, and health status. People younger than 25 years are at highest risk for some adverse sexual health outcomes, such as sexually transmitted infections and unplanned pregnancies. However, research into the sexual health and wellbeing of men and women in later life—who now have increasing expectations of sexual fulfilment and make up a growing segment of the population – is a neglected

area. The first National Survey of Sexual Attitudes and Lifestyles (Natsal-1) was done in a probability sample of 18,876 adults aged 16–59 years in Britain in 1990–91. It provided urgently needed population-based data to inform the prevention and prediction of HIV transmission. A second survey (Natsal-2) of 11,161 adults aged 16–44 years in 1999–2001 extended the investigative focus to broader aspects of sexual and reproductive health. Data from these surveys have been widely used to inform sexual and reproductive health policy in Britain. Here, we report data on sexual behaviours and attitudes in Britain from the latest survey, Natsal-3, and the two previous surveys. The combination of data from all three Natsal surveys enables both period and birth cohort analyses; together, the surveys sampled people born between the 1930s and the 1990s. We examine changes in sexual lifestyles throughout the life course and trends over time.



Self-Assessment Exercise 4.1.1

List the reasons given for carrying out the Natsal 2013 study.

Answers in Section 4.8

Aim and Objectives

The purpose of a study is usually described in terms of an *aim* and *objectives*. These can be defined as follows:

- **Aim(s)** is a summary statement of what is proposed and the purpose of the study.
- **Objectives** is a more-specific description of what the various stages and/or components of the research are designed to achieve. The objectives should not be over-complex or too numerous, and they should focus on the questions to answer rather than the methods proposed to answer them.

The aim of the Natsal-3 study could be described as being ‘to provide updated estimates and to assess changes in reported sexual behaviour and attitudes throughout the life course and over time in Britain.’

4.1.2 Political Context of Research

Most, if not all, health research has some political implications. This is important in thinking about how the findings of the research can be translated into policy and how various groups in society will respond. Not surprisingly, sexual behaviour is a subject about which many people hold strong views, including the political establishment. The Natsal study did not escape this, particularly in respect of funding when it was first planned, as described in the following excerpt.

Historically, researchers on both sides of the Atlantic have encountered difficulties in obtaining funds to mount large-scale cross-sectional surveys of sexual behaviour. Fieldwork for the main stage of the British survey of sexual attitudes and lifestyles, set to begin in April 1989, was delayed because of a much-publicised government veto on financial support. (Paper B). Research on this topic is now better accepted by society (at least in the UK and many parts of the western world), but political factors remain very important in terms of governmental support and availability of funding.

Exercise for Reflection

Look through some newspapers (print copy or websites) of your choice for articles based on some research of relevance to health. See whether you can identify some political issues relating to funding, the methods used, or the implications of the results for policy. Because this is an exercise for reflection, no answers or commentary are provided in Section 4.8.

Summary

- In planning your own research, defining a clear aim and a set of objectives is very important.
- Equally, when studying research carried out by others, it is necessary to be clear in your own mind about their aim(s) and objectives.
- The original (1990) Natsal study was carried out to provide accurate estimates of sexual behaviour. Little or no useful information could be obtained from routine data sources, and a survey was therefore necessary. The original study was designed to meet this need and to overcome the limitations of methods used in earlier surveys of sexual attitudes and lifestyles.
- The most recent Natsal study (Natsal-3) was carried out to update this information and to study changes over time.
- All research has a political context that can influence design, funding, reporting, and implementation.

4.2 Sampling Methods

4.2.1 Introduction

Two of the greatest challenges faced by researchers studying sexual attitudes and lifestyles were obtaining *accurate* information and ensuring that the information was *representative* of the general population. These two points essentially sum up what good study design is about: First, is the information accurate? Second, is it possible to infer from sample results to the population of interest, in this case the British population? We examine these two issues by looking first at how the team obtained the *sample* (Sections 4.2–4.5), and second at how they developed methods for obtaining accurate information about the relevant attitudes and behaviours (Sections 4.6 and 4.7).

4.2.2 Sampling

- What is a *sample*, and why do we need one? A sample is a group of individuals taken from a larger population. The *population* is the group of people in whom we are interested and to whom we wish the results of the study to apply. A sample is required because, in most instances, it is not practical or necessary to study everyone in the population. Clearly, then, in taking a sample, it is absolutely vital that it be representative of the population. It is worth remembering that if the sampling is not representative, little can be done about it once data collection is complete. Over the next few sections, we examine how to select and contact a sample and how to check how well it represents the population. There are essentially two ways a sample may be inadequate:

- a. The sample may be too small, and its characteristics are therefore likely to differ substantially from those of the population simply due to chance. This is called *sampling error*, and it can be reduced by increasing the sample size. These issues will be examined in more detail in Section 4.4.
- b. The sample may have been selected in such a way as to under- or overrepresent certain groups in the population. For example, a survey carried out in the UK using telephone land lines would inevitably exclude many young people who rely solely on mobile phones for communication.
- Another example was noted by the Natsal 1990 research team in reviewing previous studies of sexual attitudes and lifestyles: Many of these studies had been carried out on volunteers, who would almost certainly not be typical of the general population. This type of systematic misrepresentation of the population is called *selection bias*, and it cannot be removed by increasing the size of the sample. Selection bias can be avoided only by representative sampling (although adjustment by weighting can be used to reduce bias to some extent if the structure of the sample relative to the population is known, as was done in the Natsal studies).

Bias is a systematic error in sampling or measurement that leads to an incorrect conclusion.

Turning now to the Natsal-3 study, the researchers developed an interview questionnaire that was implemented by trained staff using a combination of computer-assisted personal-interview (CAPI) and computer-assisted self-interview (CASI) techniques. The *population* was all men and women aged 16–74 years in Britain (England, Wales and Scotland), and the task was to obtain a representative *sample* of these people.

Please now read the following two paragraphs taken from paper A: methods, participants and procedure. This account of the sampling is quite complex and includes descriptions of ‘oversampling’ and ‘weighting’, which we have not covered. For now, try to gain a general understanding of the approach to sampling. We discuss the various methods used during the rest of this section.

Methods, Participants and Procedure

Briefly, we used a multistage, clustered, and stratified probability sample design. 1727 postcode sectors (geographical units used for sorting mail) throughout Britain were used as the primary sampling units and were randomly allocated to one of eight periods of fieldwork that took place between Sept 6, 2010, and Aug 31, 2012, with each period lasting about 3 months. Within each primary sampling unit, 30 or 36 addresses were randomly selected and then assigned to interviewers from NatCen Social Research. To allow detailed exploration of behaviours in the age group at highest risk of some sexual health outcomes (e.g. unplanned pregnancy and sexually transmitted infections), we oversampled individuals aged 16–34 years. We randomly allocated addresses to either the core sample (in which all individuals aged 16–74 years were eligible) or the boost sample (in which only individuals aged 16–34 years were eligible). Letters and leaflets giving background information about Natsal-3 were sent to sampled addresses before visits began. Interviewers visited all sampled addresses, identified residents in the eligible age range, and randomly selected one individual to be invited to participate in the survey using a Kish grid technique. Participants then completed the survey in their own homes through a combination of face-to-face interviews with computer-assisted personal interview and a self-completion format

with computer-assisted self-interview. Interviewers were present in the room while participants completed the computer-assisted self-interview and could provide assistance as necessary, but did not view responses. On completion of computer-assisted self-interviews, answers could not be accessed by interviewers. No names or other potentially identifying information was attached to the interviews.

As in Natsal-1 and Natsal-2, we weighted Natsal-3 data to adjust for the unequal probabilities of selection in terms of age and the number of adults in the eligible age range at an address. After application of these selection weights, the Natsal-3 sample was broadly representative of the British population compared with 2011 Census figures although men and London residents were slightly under-represented. Therefore, as in previous surveys, we also applied a non-response post-stratification weight to correct for differences in sex, age, and Government Office Region between the achieved sample and the 2011 Census. We compared data for participants aged 16–44 years in each survey. This age group was common to all three surveys. Information about variables that were compared was derived from identically worded questions. All three surveys had been weighted for differential selection probabilities. Natsal-1 was post-stratified to 1991 Census figures and Natsal-2 to 2001 Census figures, with procedures described for Natsal-3, which allowed us to make comparisons between the three surveys. However, there are minor differences from the weighting schemes used in previous reports.

The key to achieving good representation of the population is *random* sampling. This means that it is purely a matter of chance whether or not each member of the population is chosen to be in the sample – their selection does not depend on their particular characteristics. An important consequence of this is that random sampling ensures that any differences between the sample and the population are due to chance alone. Random sampling and variants of this are the most important and commonly used sampling methods in health research, but they are not the only ones. Table 4.2.1 lists the sampling methods that we review in this section.

Table 4.2.1 An overview of sampling methods.

Random	Non-random
Simple random sampling	Systematic sampling
Stratified random sampling	Convenience sampling
Cluster random sampling	Sampling of people who are hard to contact
Multistage random sampling	Quota sampling

Before looking further at these methods, however, we need to discuss *probability*, which is how we quantify chance.

4.2.3 Probability

Probability is a numerical measure of chance. It quantifies how likely (or unlikely) it is that a given event will occur. The probability of an event has a value between 0 and 1 inclusive. Here are a few examples of the expression of probability:

Probability

- If an event is impossible, its probability is 0.
- If an event is certain, its probability is 1.
- Any event that is uncertain but not impossible has a probability lying between 0 and 1.
- The probability that each of us will die is 1 (a certain event).
- The probability of obtaining a 'head' when a coin is tossed is 0.5, as the outcome can only be a 'head' or a 'tail'.
- The probability that a person weighs -10 kg is 0 (an impossible event).

Suppose we are interested in a population that consists of 10 people. If we choose one person at random (for example, by putting all their names into a hat and selecting one), then each of the 10 people has the same probability of being selected. The probability is 1 in 10, or 0.1. Suppose now that four of the population are men and six are women. What is the probability that the person chosen at random is a woman? The probability of selecting a person with a particular characteristic is

$$\frac{\text{Number of people in the population with that characteristic}}{\text{Total number in the population}}$$

So the probability of selecting a woman is $6/10 = 0.6$. Choosing a sample by random sampling means that we can apply **probability theory** to the data we obtain from the sample. As we work through this chapter, you will see how this enables us to estimate the likely difference between the true characteristics of the population and those of the random sample, that is to say, to estimate how precise our results are.



Self-Assessment Exercise 4.2.1

In a Local Authority (LA) area with a population of 100,000, there are 25,000 smokers, 875 of whom have experienced (non-fatal) ischaemic heart disease (IHD) in the past 5 years. Of the non-smokers, 750 have also suffered IHD in the same period. If a person is chosen **at random** from this LA,

1. What is the probability that this person is a smoker?
2. What is the probability that this person has had (non-fatal) IHD in the last 5 years?
3. What is the probability that this person is a smoker *and* has suffered (non-fatal) IHD?

Answers in Section 4.8

4.2.4 Simple Random Sampling

Simple random sampling is the simplest method of random sampling and is illustrated diagrammatically in Figure 4.2.1. Let's say that we want to sample 10 people from a total population of 60. Each of the 60 people in the population (Figure 4.2.1) can be given an identifier (e.g. a number from 01 to 60), and then 10 numbers can be selected at random, usually by a computer. The most important principle of simple random sampling is that everyone in the population has an equal chance of being selected. This is because each person is chosen independently of the others – if person number 42 is chosen, this does not alter any other person's chance of being selected.

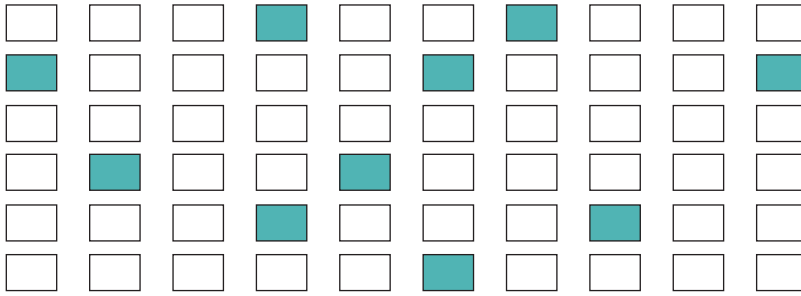


Figure 4.2.1 An example of simple random sampling of 10 subjects from a population of 60 (see text).

Simple random sampling works well because, of all possible samples, most are reasonably representative of the population. However, simple random sampling does not always produce a representative sample. By chance, we may be unlucky and choose a sample that does not represent the population very well. We explore how to deal with this later in the chapter.

4.2.5 Stratified Sampling

Moving on from the simple random sampling example, imagine that we wish to study alcohol-consumption habits in a population of 10,000 adults from a small town, 10 per cent of whom (1,000 people) are from an ethnic minority group (for simplicity we will assume only one ethnic minority group). If a 2 per cent simple random sample is taken, it would contain a total of 200 people (2 per cent of 10,000), about 20 of whom would be from the ethnic minority group and 180 others.

We are interested to know whether the *prevalence* of heavy drinking differs between the ethnic minority group and rest of the population. While 180 people may be enough to estimate prevalence among the majority group with reasonable precision, 20 people from the ethnic minority group is far too few. We could simply take a 20 per cent sample of the whole population, thus obtaining 200 ethnic minority subjects, but we will also then have to survey 1,800 others – far more than necessary and therefore inefficient. A better alternative is to use a *stratified sampling* method, whereby the population is divided into groups (strata), and sampling is carried out within these strata. In this case, although *simple random sampling* is carried out within both strata, a much larger *sampling fraction* is taken from the ethnic minority group than from the others (20 per cent versus 2.2 per cent), as shown in Table 4.2.2.

Table 4.2.2 Example of stratified sampling.

Strata	Total in population	Sampling fraction (%)	Number selected
Ethnic minority	1,000	20%	200
Others	9,000	2.2%	200

This procedure, known as *oversampling*, allows us to measure the prevalence of heavy drinking with similar *precision* in both groups.

Stratified sampling may also be used with a constant sampling fraction to increase the precision of our sample estimate. By carrying out random sampling within strata of, say, age, we can be more confident that the sample is representative of the population with respect to age than

if we just randomly select from the whole population. The key principle of stratified sampling is that the strata are defined by factors thought to be related to the subject under investigation. In the alcohol-consumption example, the strata were defined by ethnicity ('minority' and 'other') because we want to know whether the prevalence of heavy drinking differs between these two groups. In the example of stratifying by age group, we may be investigating an age-related disease. A stratified sample should lead to a more-representative sample than a simple random sample of the same size, and it should lead to more-reliable results. That is, *sampling error* is reduced compared with simple random sampling. Stratification was carried out in the Natsal study, and this is explored further in the section on *multistage sampling*.

4.2.6 Cluster Random Sampling

Suppose we wish to study knowledge and views about nutrition and health among schoolchildren aged 7 and 8 years old (Year 3) in a given town. By far the easiest way to contact and survey these children is through their schools. An example is illustrated in Figure 4.2.2.

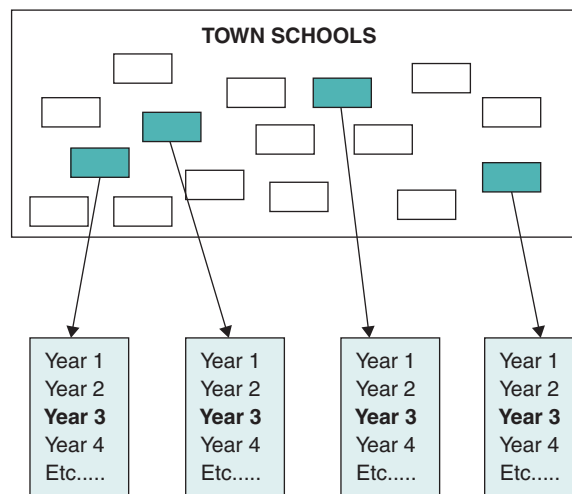


Figure 4.2.2 An example of cluster sampling of Year 3 children in a town. From the 16 primary schools (clusters) in the town, four have been selected. Within each of the selected schools, the children from Year 3 can be surveyed (see text).

Let's say that there are 480 children in the relevant age group in the town, attending 16 primary schools. If we require a sample of 120 children, a simple random sample would present us with the prospect of surveying seven or eight children from each of 16 schools. Would it not be more practical to study whole classes from a representative selection of fewer schools? In fact, a total of four schools would provide the necessary number (with 30 children per class, we require just four schools to achieve a sample of 120), and this would clearly be far more cost-effective. This method is known as *cluster random sampling*, in which the schools (clusters of children) are selected as part of the sampling process.

Ideally, the schools should be selected randomly, although it may be better to *stratify* prior to random selection, for example by religious denomination.

Cluster random sampling reduces the costs and time required for the study, but the penalty is increased *sampling error* compared with a simple random sample of the same size. This is

because individuals in the same cluster tend to be more similar than individuals in the population as a whole; we discuss this concept further in Section 7.5 of Chapter 7 on intervention studies in relation to the design and analysis of cluster-randomised trials. As well as the savings in time and money, cluster sampling has the advantage of not requiring a list of everyone in the population from which to sample. There is not, for example, an easily available list of all UK schoolchildren, so we could not easily select a simple random sample. There is, however, a list available of all schools. So we can select schools (clusters) from the list and then obtain from each selected school a list of pupils from which to sample.

4.2.7 Multistage Random Sampling

This describes a sampling method that is carried out in stages. To return to the example of the school survey, an alternative way of obtaining the 120 children is to randomly select more schools – say, eight – and then randomly select 50 per cent of the children from the specified class (Year 3) in each of those eight schools. This is multistage random sampling: the first stage is to sample eight schools out of a total of 16; in the second stage, 50 per cent of children in Year 3 are selected. Multistage random sampling was used in the Natsal 2013 study, and we will look at this in more detail in Section 4.2.12.

4.2.8 Systematic Sampling

Simple random sampling, as in the first example, requires giving each subject an identifier (such as a serial number), allowing the random selection to be made. There may be situations where this is very difficult to do. Imagine we are working in a low-income country and need to extract information on a representative selection of 250 subjects from a total of 10,000 health records, time is short, and there is no overall list. Simple random sampling would be ideal, but there is no list available to assign serial numbers. We could make such a list, but it would be much quicker to select every 40th record. This is called a *systematic sample*, and although it should be fairly random, there is potential for bias. If we are unlucky, the selection process might link into some systematic element of the way the records are sorted, with the result that we might select a disproportionate number with some characteristic related to the topic under study. The trouble with this kind of bias is that we will probably have no idea that it has occurred.

4.2.9 Convenience Sampling

Where time and resources are very short, or where there is no structured way of contacting people for a given study, it may be satisfactory to use what is known as a *convenience sample*. For example, a study of patients' attitudes to a particular health service might be gathered from people in a waiting room simply by approaching whoever was there and had the time to answer questions. Whereas this type of procedure can be used to gather information quickly and easily, it is unlikely to be representative.

4.2.10 Sampling People Who are Difficult to Contact

Suppose we wish to study the health of homeless men. Although some stay in hostels and are known to various agencies and care groups, others live on the street or in no regular place. Any list of such people is likely to be very incomplete. Moreover, many could be hard to contact or could be unwilling, for a variety of reasons, to take part in a study. Simple random sampling of names either for a postal questionnaire or interview visits, starting from the approach of a

conventional sampling frame, would surely be doomed to failure. In this type of situation, it may be preferable to use an alternative approach that, although not random, is nevertheless likely to identify more of the people we want to contact. **Snowball sampling** is an example of this, and it involves finding people from the population (in this case homeless men and people who care for and support them) who can use their networks and contacts to find other people who fulfil the selection criteria.

4.2.11 Quota Sampling

Quota sampling is another sampling method of convenience. It is used extensively by market research organisations for obtaining views on products, political opinion polls such as we looked at in Chapter 1 (see Figure 4.2.3), and so on.

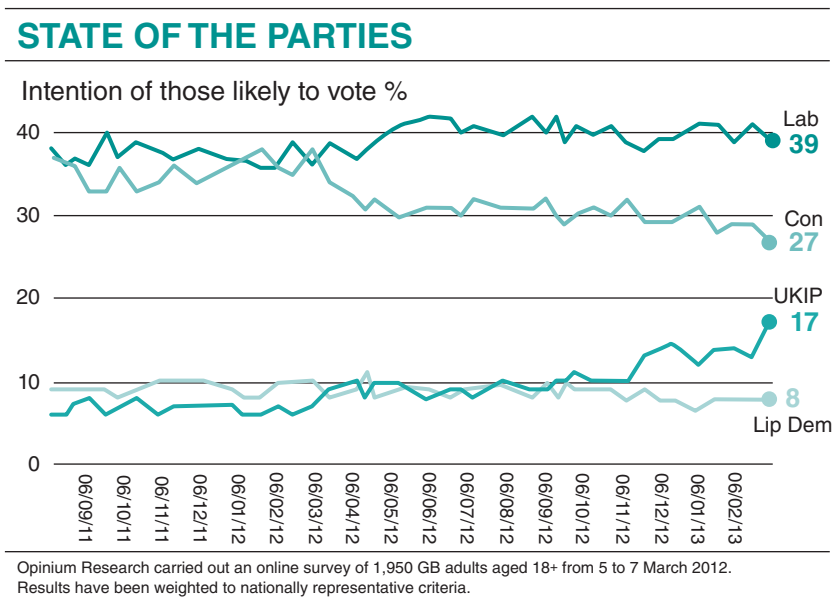


Figure 4.2.3 An example of quota sampling.

The following example illustrates how it works.

A new male contraceptive pill is about to be launched, and we decide to study the views of men and women aged 15–34 years about this product. It is important that the sample is representative of the different age groups among the general population in the area, but because the information is required quickly, there is insufficient time to conduct a random-sampling procedure. An alternative, which requires no numbered list of names and addresses, is known as **quota sampling** and is carried out as follows. First of all, we must work out the number of people required in the sample for each subgroup of the population, in this example defined by age and sex. This would typically be done using information from the census or some other source on local demography. These numbers are called **quotas**, and, having determined these, we can then survey people until the required total in each group is obtained. For example, if we require a sample of 500 people aged 15–34 years with equal numbers of men and women, and the local population is as shown in Table 4.2.3, then a quota of 75 men aged 15–19 years (30 per cent of the 250 men) is needed.

Table 4.2.3 Quota sample table for a local population.

Age group	Men		Women	
	Percentage in population	Number required	Percentage in population	Number required
15–19	30%	75 (30% of 250)	32%	80
20–24	28%	70	28%	
25–29	22%	55	22%	
30–34	20%	50	18%	
Totals	100%	250	100%	250

Although quota sampling is relatively simple and quick, how well the sample represents the population is entirely dependent on the skill with which the survey team identifies the people (respondents) for each *quota*. Here are some questions to help you think more about this sampling method:



Self-Assessment Exercise 4.2.2

In the male contraceptive pill example,

1. What size quota is required for women aged 20–24 years?
2. What would happen if the interviewers fulfilled their quota of women aged 20–24 years solely by calling on homes during the working day (9 A.M. to 5 P.M.)?
3. What could be done to ensure that the quotas are more representative of the population?

Answers in Section 4.8

4.2.12 Sampling in Natsal-3

Having reviewed the different approaches that are available for sampling, we look again at what was done in the Natsal-3 study. Bear in mind that the overall aim was to obtain a representative sample of the population aged 16–74 years in Britain, with the additional objective of ensuring sufficient precision of information about important risk behaviours among the younger age groups (16–34 years). Although the sampling method seems complex, it is in essence a combination of several of the methods we have discussed. We will work through the sampling method, referring again to the excerpt from the methods section of paper A (see Section 4.2.2). Some of the details of the sampling described here are drawn from other Natsal study reports.

1. For the first stage, postcode sectors (the first part of the postcode) were stratified by region and a number of sociodemographic factors. Postcode sectors were therefore the clusters in this sampling method.
2. From all of these postcode sectors, 1727 were selected randomly but with a probability proportional to the number of postal addresses, so that a sector with 3,000 addresses would be twice as likely to be selected as one with 1,500.
3. From each of these postcode sectors, a proportion of addresses were randomly selected, making a total of 59,412 addresses. Of these addresses 26,274 were found to be eligible when visited (i.e. had one or more persons aged 16–74 in the household).

4. Finally, one eligible adult aged 16–74 years was randomly selected from each household using the Kish technique. This technique places the names of all eligible adults onto a grid format and then uses a form of random selection to identify one individual to take part in the study.
5. Individuals aged 16–34 years were oversampled; that is, two further samples were taken of individuals aged 16–34 years (boost 1) and individuals aged 16–29 years (boost 2). This was done to provide enough precision (statistical power) for exploring behaviours among those at the highest risk of a range of sexual health outcomes. We look in more detail at statistical power in Section 5.6.1.

Weighting

The oversampling procedure requires that account be taken of this in estimates derived for the whole sample, such as the proportion of males in Britain reporting a specific behaviour. This is carried out by weighting. The next exercise should help clarify how weighting, which is an important technique, operates.



Self-Assessment Exercise 4.2.3

The Natsal studies have used oversampling for a number of purposes, including for greater precision among younger adults as described above in Natsal-3 and for London residents in Natsal-2. To see how oversampling affects prevalence estimates of behaviour and why weighting is needed, this exercise is based on data from the London oversampling in Natsal-2. London oversampling was done because the prevalence of HIV risk behaviours had been shown to be higher in London than elsewhere, yet they were still comparatively rare. These data are shown in Table 4.2.4.

Table 4.2.4 Oversampling of the London population to measure prevalence of a risk behaviour.

Area	Number in sample*	Sampling ratio	Respondents reporting the behaviour of interest	
			Number	Percentage
Inner London	900	3.0	225	25%
Outer London	1,000	2.0	200	20%
Rest of Britain	9,000	1.0	630	7%
Total	10,900		1,055	

*The numbers in this exercise are modified from Natsal-2 for simplicity.

1. If we use the data in the 'Total' line of Table 4.2.4 and ignore the sampling ratios, what is the overall percentage of people in the sample reporting the behaviour of interest?
2. Why is this result incorrect?
3. How many people would be in the inner London sample if the sampling ratio had been the same as for the rest of Britain?
4. How many people from the inner London sample would have reported the behaviour if the sampling ratio had been the same as for the rest of Britain?
5. Try working out the percentage reporting the behaviour in the country after allowing for the oversampling.

Answers in Section 4.8

This final exercise in this section should help consolidate the ideas and techniques we have covered so far on sampling methods.



Self-Assessment Exercise 4.2.4

Which sampling method would be most appropriate for

1. A general population study of cardiovascular disease risk factors in men and women aged 20–59 years resident in a Local Authority?
2. A study of patterns of drug treatment among people aged 75 years and older living in residential care homes in a city?
3. A study of the health-care needs of female intravenous drug users who are also single mothers living in an inner-city area?

Answers in Section 4.8

Summary

- The single most important principle of sampling is that the sample be representative of the population.
- Random selection is generally the best way of obtaining a representative sample.
- In some situations, random selection is not possible or practical within the constraints of time and resources, and other methods may give a useful sample. In experienced hands, methods such as quota sampling are very effective.

4.3 The Sampling Frame

4.3.1 Why Do We Need a Sampling Frame?

After we decide on the type of sample required, it is necessary to find a practical method of selecting and then contacting the people concerned. Ideally, we would like an accurate list of all the people in the population of interest, which includes various characteristics such as age and sex; indeed, this information is vital if we wish to stratify the sample (see Section 4.2.4). We also need contact details such as address and telephone number. This type of list is called a *sampling frame*.



Self-Assessment Exercise 4.3.1

1. What potential sampling frames can you think of?
2. What are the advantages and disadvantages of the sampling frames you have identified?

Answers in Section 4.8

4.3.2 Losses in Sampling

Despite our best efforts, the sampling frame might not comprise the entire population of interest. Some individuals may not be included in the sampling frame because, for example,

they are not registered to vote or registered with a general practitioner (GP). There may also be individuals in the sampling frame who are not members of the population: for example, people who have recently moved to another part of the country but have not yet been removed from the old GP list. The following example summarises the various stages of selecting and contacting a sample and the ways people can be lost. This is an example for the purposes of illustration, but it is based on findings from previously published examples. In this study of the prevalence of hip pain among older adults (older than 60 years), the sampling frame comprised all those aged 60 years or more who were registered with four GP practices in the town of Harrogate, England (Table 4.3.1).

Table 4.3.1 Sampling frame for a GP population of adults over 60 at four GP practices in Harrogate, West Yorkshire.

Population to sample	Number in sample	Comments
Population of interest	43,173	Adults older than 60 years in Harrogate
Sampling frame	4,126	All adults older than 60 years at four GP practices in Harrogate
Eligible sample	4,080	46 exclusions of patients considered by GPs unsuitable to take part in study
Available sample	3,956	29 deaths or departures from practice 68 questionnaires returned by the postal service as 'not at this address' 27 people with memory problems
Sample on whom data were obtained	2,989	175 people who declined to participate 69 people who stated ill health prevented participation 723 people from whom no response was received

Starting from a total of 4,126 people on the sampling frame, the investigators ended up with 2,989 respondents, having lost the remainder for a variety of reasons, including, of course, that some people did not wish to take part in the study. In terms of the sampling process, it is notable that in this example, 29 were known to have died or left the practice, but a further 68 invitations were returned by the post office. Presumably, these people had moved (or died), but this information was not correct on the GP list (the sampling frame). Problems such as these are inevitable with any sampling frame.

If we take the available sample as the denominator, then $2,989 \div 3,956$ responded, a response rate of 75.6 per cent. One marked advantage of using the GP list for this study is that the characteristics of responders and nonresponders can be compared using a range of information available to the practices. This will facilitate the assessment of *response bias*, an important issue that we will return to shortly.

Summary

- The sampling frame must provide enough information about the population to allow sampling and to contact the people selected.
- No sampling frame is without its problems, including inaccurate information on who is in the population, as well as the information required to contact those selected (e.g. addresses).
- It is very important to obtain and record details of losses in sampling so that an assessment can be made about how representative the sample is.

4.4 Sampling Error, Confidence Intervals, and Sample Size

4.4.1 Sampling Distributions and the Standard Error

Any quantity calculated from the sample data, such as a proportion or a mean, is a statistic. For the purposes of our explanation, imagine for a moment that we have randomly selected several samples of the same size from a population and calculated the value of the same statistic from each one. We would expect, by chance, to obtain different values, and we would expect these values to tend to differ from the population value. This variation is the **sampling error** introduced in Section 4.2.1: the difference between the **sample** and **population** that is due to chance.

Each possible random sample of a given size that could be selected from the population results in a value of the statistic. All these possible values together form the **sampling distribution** of the statistic, which tells us how the value of the sample statistic varies from sample to sample. The following example should help in understanding this.

Suppose our friends A, B, C, D, and E obtain the following percentage marks in their end-of-module assessment:

35 52 65 77 80

In this example, the five friends are the population, so the mean mark of the population is:

$$\frac{35 + 52 + 65 + 77 + 80}{5} = 61.8\%$$

Now suppose that we do not know the marks for all five students in the population, but we want to estimate the population mean mark from a random sample of two students. If the sample chosen has marks of 35 and 65 per cent, the sample mean is 50 per cent – some way from the population mean of 61.8 per cent. If, however, we happen to choose a sample of students with marks 52 and 77 per cent, the sample mean is 64.5 per cent – quite close to the population value we want to estimate. This shows how our estimate of the population mean from a single sample is determined by the play of chance: rather a poor estimate for the first sample, but quite good for the second one.

It is as well to remember at this point that in carrying out research we generally only take one sample, and we have no way of knowing whether we were lucky or unlucky with that single sample. Returning to our example, there are 10 possible samples of size 2, and the marks and sample mean for each of these are as follows (Table 4.4.1):

Table 4.4.1 Means for size-2 samples.

Sample marks	Sample mean mark
35, 52	43.5
35, 65	50.0
35, 77	56.0
35, 80	57.5
52, 65	58.5
52, 77	64.5
52, 80	66.0
65, 77	71.0
65, 80	72.5
77, 80	78.5

The values in the right column (all possible values of the mean that can be calculated from a sample size of two) form the *sampling distribution* of the sample mean mark. We can display this distribution in a frequency table (Table 4.4.2) or a histogram (Figure 4.4.1):

Table 4.4.2 Frequency table of the sampling distribution for Table 4.4.1.

Sample mean mark	Frequency
40.0–49.9	1
50.0–59.9	4
60.0–69.9	2
70.0–79.9	3
Total	10

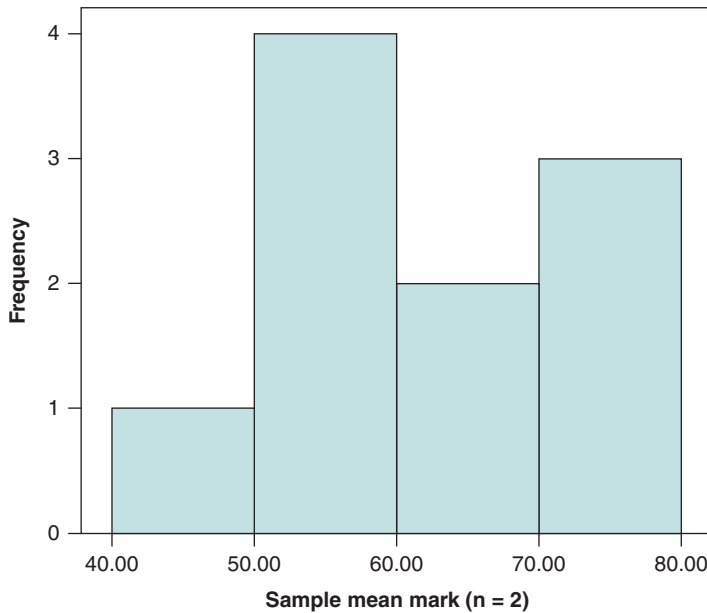


Figure 4.4.1 Histogram of the sampling distribution for Table 4.4.1 (using the intervals shown in Table 4.4.2).

The mean of this distribution is $\frac{43.5 + \dots + 78.5}{10} = 61.8\%$.

That is, the mean of the sample means exactly equals the population mean, although note that this is only the case because we have taken all possible samples.

4.4.2 The Standard Error

The spread of all possible values of the sample mean shown above is measured by the standard deviation of the sampling distribution. This is called the *standard error* of the sample mean, and it tells us how spread out the sample means are. It is important for you to be clear that the *standard error* is the measure of spread of repeated sample means, and it is in fact the

standard deviation of those values of the sample means. We have already encountered the standard deviation of the population (the measure of spread of values of the variable of interest in the population) and the standard deviation of the sample (the measure of spread of values of the variable of interest in the sample).

In this case, the standard error is 11.4 per cent. We have seen how the possible values of the sample mean are spread out around the population mean, and the smaller the standard error is, the more closely bunched the sample means are around the population mean. This means that it is more likely that the mean of any particular sample we choose will be close to the population value; that is, it will be a good estimate of the population mean. A smaller standard error is therefore desirable. We now look at how this can be achieved by increasing the sample size.

Reducing the Standard Error

Suppose we increase the sample size to three in order to estimate the population mean mark. There are 10 possible samples that could be chosen (not listed), and the sampling distribution of the sample means is as follows (Figure 4.4.2 and Table 4.4.3):

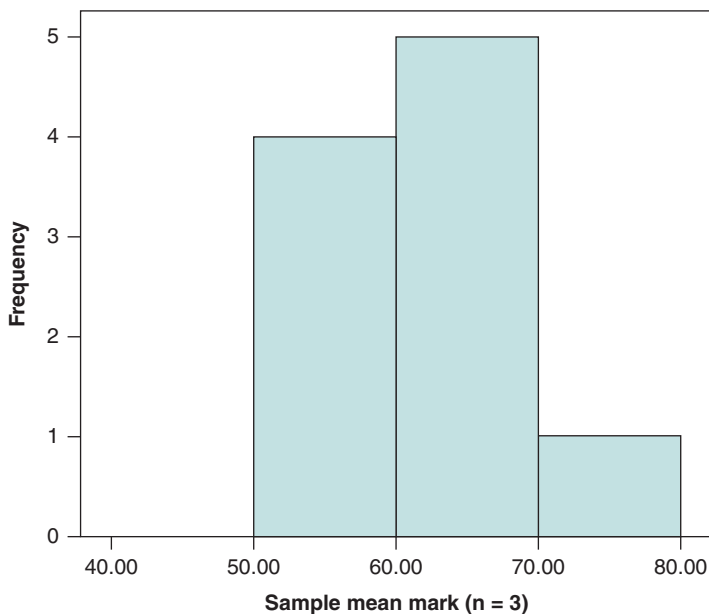


Figure 4.4.2 Histogram of the sampling distribution with a larger sample size (using the intervals shown in Table 4.4.3).

Table 4.4.3 Frequency table of a larger sampling distribution.

Sample mean mark	Frequency
40.0–49.9	0
50.0–59.9	4
60.0–69.9	5
70.0–79.9	1
Total	10

The mean of the sampling distribution is again 61.8 per cent, and the standard error is now 7.6 per cent, considerably less than the value of 11.4 per cent with a sample size of only two. So this distribution is less spread out – the sample means are bunched more closely around the population mean.

This example has illustrated that the larger the sample, the closer the sample mean is likely to be to the population mean: The standard error decreases with increasing sample size. The standard error of the mean also depends on the standard deviation of the population. This makes sense intuitively, since we would expect that the more spread out the population values are, the more spread out the sample means will be.

The formula for calculating the standard error for the mean is as follows:

$$\text{Standard error of the sample mean} = \frac{\text{standard deviation of the population}}{\sqrt{n}}$$

for a sample of size n

Note that if you try calculating this for the exam marks example, you will find that the answers differ somewhat from the standard errors quoted earlier. This is because the formula above strictly applies to random sampling from an infinitely large population, and the examples given are corrected for a large **sampling fraction** by using what is termed the finite population correction factor (see the reference section below). The sampling fraction is the fraction that the sample is of the whole population. In the case of our sample of three marks, this is $3/5 (\times 100) = 60$ per cent, which is a very large sampling fraction. In most situations, we can consider our population of hundreds or thousands or millions to be, in effect, infinite without incurring any significant error, provided the sampling fraction is relatively small. If the sampling fraction is no more than about 10 per cent, or we are sampling with replacement (that is, after selection, a subject goes back into the population and could in theory be selected again), there is no need to make any adjustment to the standard error.



RS – Reference Section on Statistical Methods

The finite population correction factor allows for the fact that, in reality, populations are not infinite, and in some cases they may not be very large compared to the size of the sample. We achieve the correction by multiplying the standard error of a sample mean by the correction factor. This factor is given by the formula

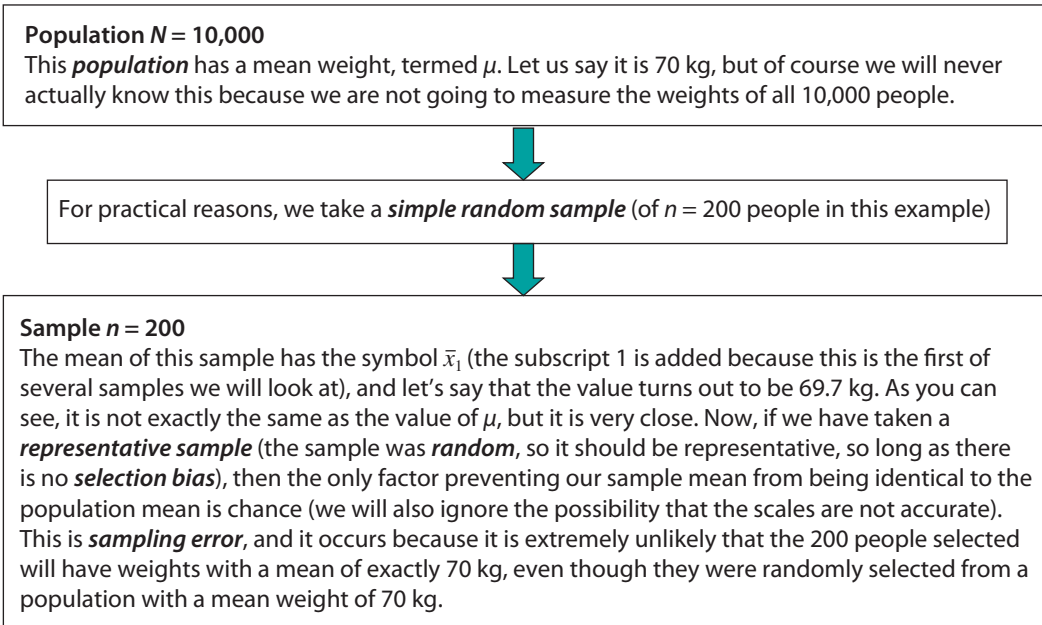
$$\text{Correction factor} = \sqrt{\frac{N-n}{N-1}}$$

where N = population size, and n = sample size. Hence, with our example of a sample of size $n = 3$, drawn from a population of size $N = 5$, the correction factor = $\sqrt{0.5} = 0.707$. We would then apply this correction by multiplying the standard error (calculated as $s \div \sqrt{n}$) by 0.707.

In our example of sampling from a very small population, we have seen the effect of sample size on the precision of an estimate. Keep in mind, though, that in practice, we take only one sample. We do not know whether we have selected a sample with a mean close to the population mean or far away. We do know, however, that the larger the sample, the more likely it is that we obtain a sample mean close to the population mean.

Now let's look at a more realistic example, with a large population. Let's say we are interested in finding out about the weight of people living in a small town with a population of 10,000. It is not practical, or necessary, to measure everyone, so we are going to take a random sample of 200

(sampling fraction of 2 per cent), and find the sample mean weight, termed \bar{x} (pronounced ‘x bar’). The mean that we are estimating is, of course, the population mean, for which we use the Greek symbol μ (pronounced ‘mew’), and it is important to remember that our survey of 200 people will not tell us what this population mean weight actually is; it will only estimate the population mean with the sample mean (\bar{x}_1). The following diagram summarises what we are doing:



Imagine now that we started all over again, and took another sample of $n = 200$ and calculated the mean of that group. Let’s call this \bar{x}_2 , because it is the second sample. By chance, will be a bit different from and from μ . Let’s say that the mean in this second sample comes out at 71.9 kg. If we repeated this exercise many times, we would end up with many estimates of the population mean μ , which we can term $\bar{x}_1, \bar{x}_2, \bar{x}_3$, and so on.

There are many, many possible samples of size 200 that could be chosen from a population of 10,000 – millions of them, in fact. If we prepare a histogram of the sampling distribution of all the possible sample means ($\bar{x}_1, \bar{x}_2, \bar{x}_3$, and so on), the intervals are very narrow and the histogram has a smooth shape. The *sampling distribution of the sample mean* is shown in Figure 4.4.3.

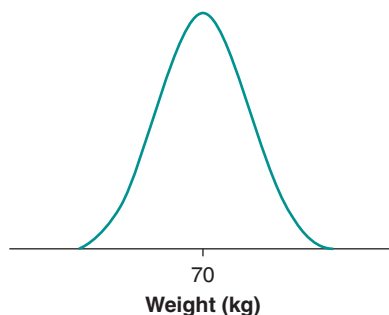


Figure 4.4.3 The sampling distribution of the sample mean from a population with mean weight 70 kg.

The distribution in Figure 4.4.3 shows the **normal distribution** that we introduced in Section 2.4.3 of Chapter 2. The mean of the sampling distribution is 70 kg, the population mean weight, and the standard deviation of the sampling distribution (which we now know as the **standard error** of the mean) depends on the population standard deviation, for which we use the Greek symbol σ (pronounced ‘sigma’) and the sample size $n = 200$.

If we repeated this exercise with repeated samples with different sample sizes ($n = 20$, $n = 50$, or $n = 100$, etc.), we would obtain a set of sampling distributions, all with the same mean (70 kg) but with different standard errors (Figure 4.4.4).

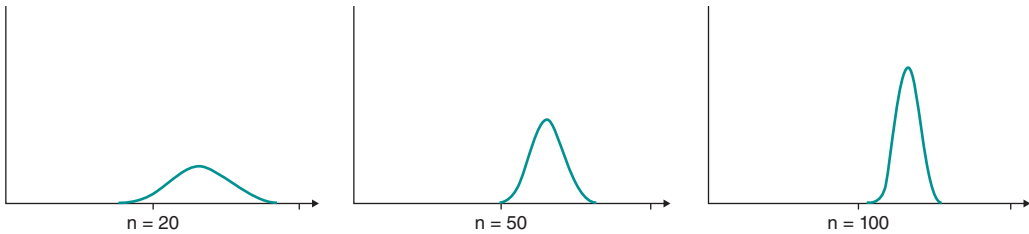


Figure 4.4.4 Sampling distributions of the sampling mean for different sample sizes.

All of these sampling distributions have the same basic shape (a normal distribution), but they are more and more closely centred on the population mean as sample size increases. What is surprising (and very useful), though, is that if the sample size is large enough (at least 30), the sampling distribution of the mean always has a normal distribution (or close enough to a normal distribution for this property to be safely assumed), no matter what the shape of the population distribution; this is known as the Central Limit Theorem. This rather convenient phenomenon is illustrated in Figure 4.4.5, which shows the distribution of weekly earnings for women working in the UK.

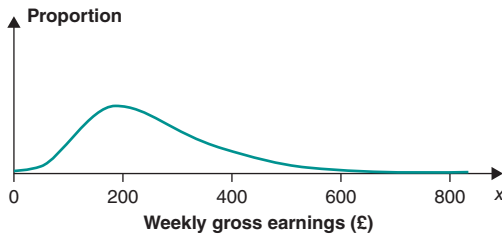


Figure 4.4.5 Weekly gross earnings of all women in the UK 2013. Source: IDS Employment Law Brief. Private sector distribution of gross weekly earnings for full-time employees. Accessed 2015.

The population distribution for these data is clearly right skewed. However, if we were to collect data on earnings from a random sample of the population, we would find that as the sample size increases, the sampling distribution would become more and more bell shaped and symmetric (normal distribution). This is illustrated in Figure 4.4.6. It also becomes more peaked and narrower as the standard error decreases with increasing sample size.

As a rough guide, if the sample size is larger than 30, the sampling distribution of the sample mean can be assumed to have an approximately normal distribution whatever the shape of the population distribution. This is important, because we often do not know the shape of the population distribution.

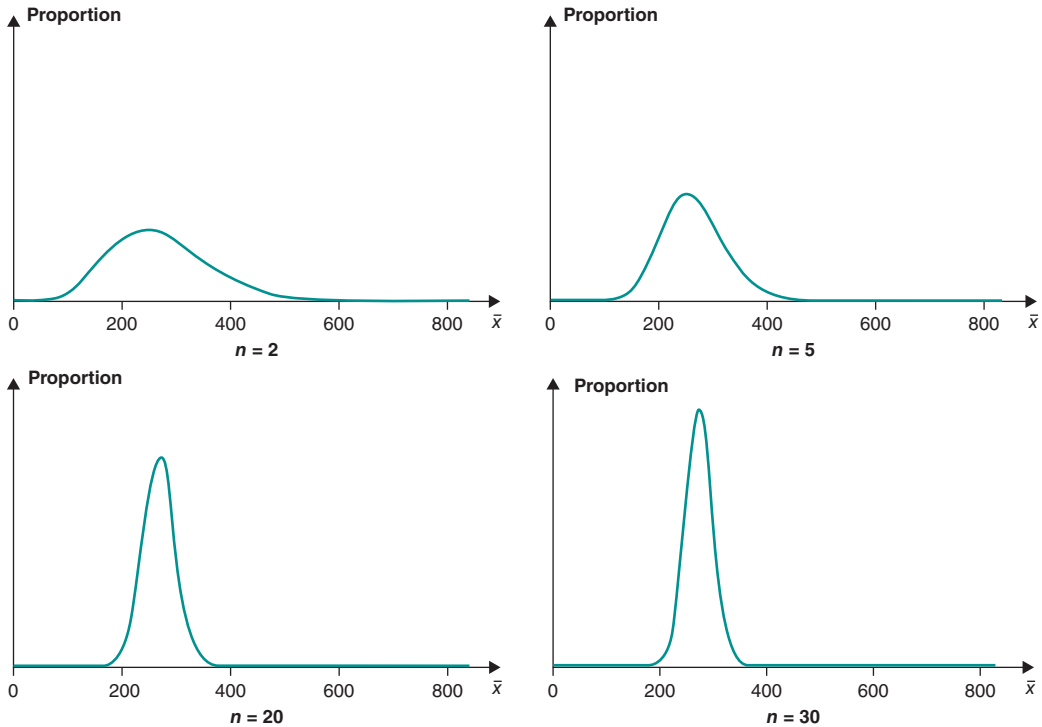


Figure 4.4.6 Sampling distribution of the sample mean from a skewed population.

Summary: Sampling Distributions and Standard Error of the Sample Mean

1. We label the population mean μ and the population standard deviation σ . We do not usually know these values, and we estimate them from a sample of size n .
2. There are many different samples of size n that could be chosen. Each sample has a particular mean and standard deviation s . In practice, only one sample will be chosen.
3. The distribution of all the possible sample means is called the sampling distribution of the sample mean.
4. The standard deviation of the sampling distribution is called the standard error of the sample mean. It is a measure of how precisely a sample mean estimates the population mean.
5. The standard error depends on the population standard deviation (σ) and on the sample size n . The standard error $= \sigma / \sqrt{n}$.
6. Where the sampling fraction is greater than (about) 10 per cent, adjustment to the standard error should be made by using the finite population correction factor.
7. The sampling distribution of the sample mean is typically approximately normal for sample sizes larger than about 30, whatever the shape of the population distribution.
8. If the population distribution is normal, the sampling distribution of the sample mean is normal even for small samples ($n \leq 30$).

4.4.3 Key Properties of the Normal Distribution

Although we refer to the normal distribution, there are many different normal distributions – in fact an infinite number. They all have the same symmetric bell shape and are centred on

the mean, but they have different means and standard deviations. The normal distribution is defined by a formula, so it is possible to calculate the proportion of the population between any two values. This is a very important property, which, for example, allows us to say that for any population with a normal distribution,

- about 68 per cent of the values lie within one standard deviation of the mean ($\mu \pm 1\sigma$);
- about 95 per cent lie within two standard deviations of the mean ($\mu \pm 2\sigma$); and
- almost the whole of the distribution (99.7 per cent) lies within three standard deviations ($\mu \pm 3\sigma$) (Figure 4.4.7).

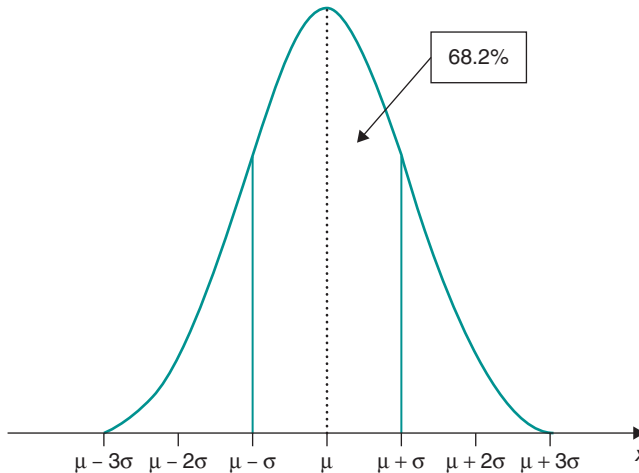


Figure 4.4.7 The normal distribution; 68.2% of all values lie between $\mu - \sigma$ and $\mu + \sigma$ (see text).

In fact, exactly $\mu \pm 1.96$ standard deviations includes 95 per cent of the values, and this explains where the figure of 1.96 that you have started to become familiar with in earlier sections comes from.

Summary: The Normal Distribution

- A normal distribution is bell shaped and is symmetric about its mean value.
- A normal distribution is defined by its mean and standard deviation.
- About 68 per cent of the values lie within ($\mu \pm 1\sigma$), about 95 per cent lie within ($\mu \pm 2\sigma$), and almost the whole of the distribution (99.7 per cent) lies within ($\mu \pm 3\sigma$).
- Exactly 95 per cent of the values lie within ($\mu \pm 1.96\sigma$).

4.4.4 Confidence Interval (CI) for the Sample Mean

Our discussion of sampling error has been built around the theoretical notion of taking repeated samples of the same size from a given population. In reality, of course, we do not take repeated samples: In the weight example, we have one estimate derived from 200 people, and we have to make do with that.

We do not know whether we had a lucky day and this single estimate has a value very close to the population mean (say, 69.7 kg, only 0.3 kg out), or a bad day and we got the value 71.9 kg

(1.9 kg out), or worse. What we can do, however, is calculate a range of values, centred on our sample mean, which we are fairly confident includes the true population mean. This is called a **confidence interval (CI)**, and it is much more informative than just quoting the sample mean.

It is common to calculate a **95 per cent CI for the population mean**. With such an interval, 'we are 95 per cent confident that the specified range includes the population mean'. We have seen that approximately 95 per cent of a normal distribution falls within two standard deviations of the mean, and exactly 95 per cent falls within 1.96 standard deviations of the mean. Since, as shown in Section 4.4.2, the distribution of sample means is normal with mean μ and standard deviation (the standard error), there is a 95 per cent chance that we choose a sample whose mean lies in the interval $\mu - 1.96 \sigma/\sqrt{n}$ to $\mu + 1.96 \sigma/\sqrt{n}$. This is illustrated in Figure 4.4.8.

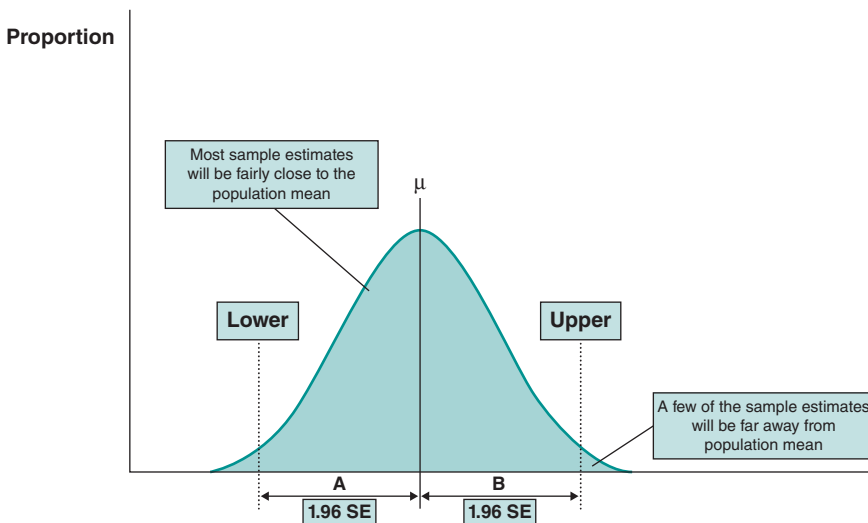


Figure 4.4.8 The sampling distribution of the sample mean.

The range covered by the arrows A and B between the upper and lower dotted lines is ± 1.96 standard errors from the true population mean and includes 95% of possible sample means.



Self-Assessment Exercise 4.4.1

1. On Figure 4.4.8, mark a value between the upper and lower dotted lines. Is it true that the range $\bar{x}_1 \pm 1.96$ standard errors includes the value μ (that is, the population mean that we want to know)?
2. What percentage of sample means lies in the area between the dotted lines?
3. Mark a value \bar{x}_2 either to the left of the lower line or to the right of the upper line. What percentage of values lie outside these lines?
4. Is the value μ included in the range $\bar{x}_2 \pm 1.96$ standard errors? What is the implication of your answer to this?

Answers in Section 4.8

To summarise, in order to calculate this CI, we first need the sample mean of a random sample, the sample size n , and the population standard deviation σ . We know the first two after collecting information about the members of the sample, but we do not know the standard deviation of the population. However, provided the sample is reasonably large (again, more than about 30), we can use the sample standard deviation, which is given the symbol s , to estimate the population standard deviation σ . A 95 per cent CI for the population mean μ is therefore shown as follows:

95 per cent CI for a Population Mean


$$\bar{x} - 1.96 \times s / \sqrt{n}, \bar{x} + 1.96 \times s / \sqrt{n} \text{ or } \bar{x} \pm 1.96 \text{ SE}$$

where SE stands for 'estimated standard error of the sample mean': that is, the standard error with σ replaced by s . The CI is normally stated to the same accuracy (number of decimal places) as the sample mean.

For our weight example, suppose our sample of 200 people has a mean weight of 69.7 kg and standard deviation of 15 kg. Then a 95 per cent CI for the mean weight of the population (10,000 people) is

$$69.7 \pm 1.96 \times 15 / \sqrt{200} = (67.6, 71.8)$$

Hence, we are 95 per cent confident that the mean weight of the population lies between 67.6 and 71.8 kg. In this example, the CI does include the population mean (70 kg), but of course we would not actually know this. Although 95 per cent of all the possible random samples of size 200 will result in a CI that includes the population mean, 5 per cent will not. Put another way, we do not know for certain whether or not the CI for our sample has captured the population mean, but we are 95 per cent sure it has.


Self-Assessment Exercise 4.4.2

The following summary statistics for the lifetime number of partners for women aged 45–54 years were obtained from the Natsal-3 study.

Number of lifetime partners for women aged 45–54 years	
Mean	6.8
Median	4
Standard deviation	11.8
Sample size (weighted)	1443

Using these summary statistics, find a 95 per cent CI for the mean number of lifetime partners for all women aged 45–54 years.

Answer in Section 4.8

Summary: CI for the Population Mean

- A CI (for a mean) is a range of values that is likely to include the population mean.
- A 95 per cent CI for the population mean is $\bar{x} \pm 1.96 \text{ SE}$, where $\text{SE} = s/\sqrt{n}$. This formula requires the sample to be large ($n > 30$) and random.
- 95 per cent of such CIs will include the population mean.
- For any particular sample, we do not know for sure whether or not the 95 per cent CI includes the population mean. There is a 5 per cent chance that it does not.

4.4.5 Estimating Sample Size

The 95 per cent CI allows us to quantify the precision of a population estimate measured in a sample. Being able to quantify the precision of a sample estimate in this way also means that it is possible to calculate the *sample size* required to estimate a population value to a desired level of precision. Now that we have calculated the 95 per cent CI for the sample mean, it is a relatively simple step to work out the sample size needed to estimate the sample mean with a precision that we can specify.

4.4.6 Sample Size for Estimating a Population Mean

We have seen that for a large ($n > 30$) random sample there is a 95 per cent chance that the sample mean lies in the interval $\mu - 1.96\sigma\sqrt{n}$ to $\mu + 1.96\sigma\sqrt{n}$ (for small samples with $n < 30$ we need to use the t -distribution, which is introduced in Chapter 5). As sample size n increases, the standard error σ/\sqrt{n} decreases, and the sample mean is a more-precise estimate of the population mean. To determine the sample size needed, we first specify the error range we will allow, termed ϵ (epsilon). If we want to be 95 per cent certain of achieving the specified error, drawing on the formula for a 95 per cent confidence interval, we set the error (ϵ) to $1.96\sigma/\sqrt{n}$. The sample size required is therefore as follows:

Sample size (n) for estimating a population mean to within $\pm \epsilon$

$$n = \frac{1.96^2 \sigma^2}{\epsilon^2}$$

The derivation of this formula from that for the 95 per cent CI is quite simple and is shown in the reference section below.

**RS – Reference Section on Statistical Methods**

We start with the expression for the 95 per cent CI, which is $\pm 1.96\sigma/\sqrt{n} = \epsilon$ (note that the full range of the 95 per cent CI is $\pm \epsilon$). To extract the sample size, which is n , we square the whole expression:

$$\frac{1.96^2 \sigma^2}{n} = \epsilon^2 \dots \text{which is equivalent to} \dots n = \frac{1.96^2 \sigma^2}{\epsilon^2}$$

To calculate the sample size, we need to estimate σ , the population standard deviation, which we do not know. We have to substitute an estimate of σ from data already available, expert

opinion, or previous experience, or we have to carry out a pilot study to obtain a rough estimate. Fortunately, this estimate need not be very precise in order to calculate the approximate size of sample required. An example and exercise will show how all this works in practice.

Example

In a study of the effects of smoking on birthweight, it was required to estimate the mean birthweight of baby girls to within ± 100 g. Previous studies have shown that the birthweights of baby girls have a standard deviation of about 520 g. The required sample size is therefore

$$n = \frac{1.96^2 \times 520^2}{100^2} = 103.88.$$

So, a sample of 104 baby girls is required.

A calculated sample size should only be used as a guide, and we would not generally use the exact number calculated. Thus, for this example, we would be cautious and decide on, say, 120. We would then need to allow for parents of some of the selected babies refusing to participate in the study. If we expected a 20 per cent refusal rate, we would need to select a sample of $120 \times 100/80 = 150$ babies.



Self-Assessment Exercise 4.4.3

1. Using data from the example on the birthweight of baby girls, calculate the sample size required to estimate the mean to ± 50 g.
2. What sample size is required to estimate the mean birthweight to ± 25 g? What do you notice happens to the sample size when the error is halved?

Answers in Section 4.8

4.4.7 Standard Error and 95 per cent CI for a Population Proportion

We now look at how these concepts and methods for standard error, 95 per cent CI, and sample size for precision of an estimate apply to proportions. In fact, although the formulae differ, the principles are the same.

Suppose that we want to estimate the proportion P of a population with some characteristic, such as a disease; that is, we want to estimate the disease prevalence in the population. Here we introduce some more symbols that are typically used when explaining proportions. The estimate of P is the proportion of the sample with the characteristic. If there are n individuals in the sample, and r of them have the characteristic, the estimate is conventionally written as $p = r/n$. When we are referring to the sample proportion, we use lower case for proportion (p) and upper case for the population proportion (P). To calculate the 95 per cent CI for a sample proportion, we first need the standard error, which is derived as follows:

The *standard error of a sample proportion* is

$$\sqrt{\frac{P(1-P)}{n}}$$

If we have a large sample (more than 30), the sampling distribution of the sample proportion is approximately normal, and a 95 per cent CI for the population proportion P is derived as follows:

95 per cent CI for a population proportion

$$p \pm 1.96 \text{ SE}$$

where SE is the estimated standard error of p , which is

$$\sqrt{p(1-p)/n}$$

With this formula, you will see that we have – as with the standard error of the mean – used the value from the sample because we do not know the standard error for the population.



Self-Assessment Exercise 4.4.4

The following summary statistics have been taken from the Natsal-3 study and represent the percentage of men aged 16–24 years who have had anal sex in the past year:

Percentage of men aged 16–24 who have had anal sex in the previous year

Percentage	18.5%
Sample size (weighted)	1,238

Calculate the 95 per cent CI for the percentage of men aged 16–24 years who have had anal sex in the previous year.

Answer in Section 4.8

4.4.8 Sample Size to Estimate a Population Proportion

To find the sample size required to estimate the population proportion to within a specified precision, we follow the same reasoning as we used for the mean, but we use the standard error of a proportion. To estimate a population proportion P to within a specified level of error $\pm\epsilon$, we require:

$$1.96 \sqrt{\frac{P(1-P)}{n}} = \epsilon$$

and the required sample size is as shown in the box. This formula is derived in the same way as for the mean, by squaring the expression and extracting the sample size term, n . See the Reference section on statistical methods in Section 4.4.6.

Sample size for estimating a population proportion to within $\pm\epsilon$

$$n = \frac{1.96^2 P(1-P)}{\epsilon^2}$$

An important point to note when calculating the standard error is that the quantity of interest, P , may be expressed as a proportion (as we have done so far), as a percentage (rate per 100),

or a rate per 1,000, and so on. The standard errors of the sample estimates are then as follows (Table 4.4.4):

Table 4.4.4 Standard errors of sample estimates.

P	Standard error of p
As a proportion	$\sqrt{P(1-P)/n}$
Per 100 (as a percentage)	$\sqrt{P(100-P)/n}$
Per 1000	$\sqrt{P(1000-P)/n}$

We must be careful to ensure that the error term is in the same units. As with calculating the sample size for a mean, we generally do not have the population value P , but we can use and estimate from a sample; that is, p

Example

Suppose we want to estimate the percentage of smokers in a population to within ± 4 per cent of the true value. If the population percentage is 30 per cent, the required sample size is:

$$n = \frac{1.96^2 \times 30 \times 70}{4^2} = 504.21$$

A sample of 505 people is required. Note we have used the formula for percentages (rather than proportion) and that the units for the percentage smoking and the error are the same (per cent).



Self-Assessment Exercise 4.4.5

1. If the true prevalence of a specific but uncommon sexual behaviour is estimated to be 1 per cent, what sample size is required to estimate it to within ± 0.2 per cent?
2. If we have a sample of size 500, to what precision can the prevalence of this behaviour be estimated?

Answers in Section 4.8

Assumptions

The sample size calculations presented here are based on certain assumptions:

1. The sample is selected by simple random sampling.
2. The sample size is large ($n > 30$).
3. The sampling fraction is less than 10 per cent (see Section 4.4.2 for explanation of the sampling fraction).

It is important to remember that the calculations shown here are based on simple random sampling. Stratification increases precision compared with a simple random sample of the same size, so a smaller sample can be selected. Cluster sampling reduces precision compared with a simple random sample of the same size, so a larger sample is needed. We do not look further here at methods for these adjustments to sample size, but the approach to adjustment for

clustering is discussed further in Section 7.6 of Chapter 7. In general, it is recommended that advice on calculating sample size should be sought from a statistician. If the calculation results in a very small sample, assumption (2) above does not hold, and alternative methods must be used to calculate sample size. If the calculation results in a large sample size relative to the population size (large sampling fraction), assumption (3) does not hold, and a modification of the calculation is required.

Summary: Sample Size Calculations for Precision of Sample Estimates

1. To estimate a population mean to within $\pm\varepsilon$ of the true value with a 95 per cent CI, we need a sample of size of

$$n = \frac{1.96^2 \sigma^2}{\varepsilon^2}.$$

2. To estimate a population proportion to within $\pm\varepsilon$ of the true value with a 95 per cent CI, we need a sample of size of

$$n = \frac{1.96^2 P(1 - P)}{\varepsilon^2}.$$

4.5 Response

4.5.1 Determining the Response Rate

Having selected a representative sample, we now have to think about who actually ends up being studied. As we have seen, there are in fact two groups of people who, although selected for the sample, do not contribute to the study:

- People who cannot be contacted, because, for instance, they have moved on from the selected address.
- People who are spoken to or receive a postal questionnaire but who decline to take part in the study.

The percentage of the selected sample that does take part in the study is called the **response rate**. Note that response can be calculated in two different ways. The **overall response rate** usually excludes people who could not be contacted. Although it is desirable to achieve a high response rate, the most important question is whether those who do not take part, the **non-responders**, are different from those who do, the **responders**. If the non-responders differ substantially from the responders, especially with respect to some of the factors under study, then this results in **non-response bias**. We now return to the Natsal-3 study; please read the following excerpt from the results section of paper A (this refers to a separate publication on Natsal-3 methods by Erens *et al.* (2013), which we do not study in detail in this chapter).

Results

According to the methodology paper for Natsal-3 (Erens *et al.* 2013), from the 59,412 addresses visited, 26,274 households were identified with an eligible resident aged 16–74 years. At 4,143 addresses, eligibility was not known (e.g. no contact was achieved). Interviews were completed with 15,162 respondents of whom 7508 were men and 7654 were women. The response rate was

57.7%. Response rates were higher in the over-sampled younger age groups (64.8% and 67.3% in the 16–34 and 16–29 age groups, respectively). Response rates were calculated after estimating the likely proportion of ineligible in the 4,143 households, where there was no information about residents (estimated ineligibility was 1,229 residents).



Self-Assessment Exercise 4.5.1

Do you think the response rate in the Natsal-3 study was adequate in relation to representing the population of interest?

Answer in Section 4.8

4.5.2 Assessing Whether the Sample is Representative

We have said that it is important to find out whether the sample is representative of the population, or, more realistically (since no sample is perfect!), how representative it is. To do this, we need to find some way of comparing the sample with the population.

The key to this is finding some characteristic(s) about which information is available from the sample and the population. This can usually be done, although it is often not as simple as it may appear at first sight. Among other things, we have to be sure that a given characteristic has been measured in the same way in both the sample and in the data source we are using for the population.

For the Natsal-3 study, the team compared estimates for key sociodemographic characteristics with data from the 2011 census because this was (quite reasonably) considered to be the most reliable external source. Some of these characteristics are displayed in Table 4.5.1.



Self-Assessment Exercise 4.5.2

Compare the characteristics of the population from the Natsal-3 sample with those from the census population. What do you notice?

Answer in Section 4.8

Differences between characteristics of the responders to a study and the population sample can lead to non-response bias. Non-response bias occurs when there are systematic differences between non-responders and responders. Another way to assess the potential for non-response bias, which can complement the population comparison method discussed above, is to compare characteristics in the sample with those from non-responders. The same cautions apply in terms of whether relevant and comparable data are available for both groups.

4.5.3 Maximising the Response Rate

Although the sample in Natsal-3, even with a response rate of 57%, appears to match the characteristics of the national population quite closely in most respects, in general we can say that achieving a good response rate is important for keeping *response bias* to a minimum. This

Table 4.5.1 Natsal-3 distributions compared with the 2011 population census.

Characteristics	Natsal 2013 sample*		Census data 2011	
	Men	Women	Men	Women
Age	%	%	%	%
16–19	7.1	6.7	7.1	6.7
20–24	9.4	9.1	9.4	9.1
25–29	9.3	9.2	9.3	9.2
30–34	9.0	8.8	9.0	8.8
35–39	9.0	9.0	9.1	9.0
40–44	9.9	10.0	9.9	10.0
45–54	18.8	18.9	18.8	18.9
55–64	15.9	16.1	15.9	16.1
65–74	11.5	12.2	11.5	12.2
Government region				
North East	4.2	4.3	4.2	4.3
North West	11.5	11.5	11.5	11.5
Yorkshire and the Humber	8.6	8.6	8.6	8.6
East Midlands	7.4	7.4	7.4	7.4
West Midlands	9.1	9.0	9.1	9.0
South West	8.6	8.5	8.6	8.5
East	9.4	9.4	9.4	9.4
Inner London	4.9	5.5	5.6	5.5
Outer London	8.6	8.1	8.0	8.1
South East	13.9	13.9	13.9	13.9
Wales	5.0	5.0	5.0	5.0
Scotland	8.7	8.9	8.8	8.9
Marital Status				
Single, never married	38.1	32.3	41.0	34.3
Married, living with spouse	50.0	49.4	46.7	47.1
Separated, divorced, widowed	11.6	17.7	12.0	18.3
Civil partnership living with partner	0.4	0.6	0.3	0.2
Self-reported general health				
Very good/good	81.1	80.7	82.1	81.3
Fair	14.9	14.3	12.5	13.3
Bad/very bad	3.9	4.9	5.4	5.3

*After final weighting

is because there may be unmeasured characteristics that vary between responders and non-responders, and research teams may be unaware of these. Research teams have some control over the response to any given study, and the following list summarises some of the ways of improving this.

- For a postal survey, include a covering letter signed by a respected person.
- Explain the reason(s) for the study, conveying the relevance to the target audience.
- If possible, arrange for the age, sex, and ethnicity (including language spoken) of interviewers to be appropriate to the audience and subject matter.
- If possible, contact subjects in advance to inform them that a study is planned and they may/will be contacted shortly to ask whether they would be prepared to take part. This may be more practical and relevant in, for example, studies of health service users.
- Reminders are routinely used and are an acceptable technique. These can be sent up to two (rarely three) times after the initial contact, at roughly 1- to 2-week intervals. It may also be acceptable to telephone courteously and to make a personal visit in certain circumstances.
- Always respect the individual's right not to take part, and ensure that procedures agreed with an ethical committee are not exceeded.



Self-Assessment Exercise 4.5.3

Draft a (brief) cover letter that could be used to accompany a postal questionnaire for a research project on asthma among men and women aged 20–59 years that you are carrying out in collaboration with a general practice.

See Section 4.8 for a specimen letter

We have seen how a sample is selected and contacted, and how to maximise and then assess the response rate. We are now ready to think about asking these people questions, taking measurements, and so on, to collect the data that we are interested in.

Summary

- It is important to determine the response rate in surveys.
- Although achieving a high response is obviously desirable, the most important issue is to minimise non-response bias (systematic differences between characteristics of people who respond to surveys and the population sample).
- Non-response bias can be assessed by comparing the characteristics (e.g. age, social class, marital status) of responders with the same characteristics among non-responders, so long as such information is available for non-responders.
- Non-response bias can also be assessed by comparing the sample of responders with independent sources of information about the population the sample has been designed to represent.
- There are a number of practical steps that can be followed to increase the response in any given situation.
- The right of individuals not to take part in research must always be respected.

4.6 Measurement

4.6.1 Introduction: The Importance of Good Measurement

In planning for the Natsal-1 study, the research team recognised that

Sexual behaviour is regarded as intensely personal by most people: reported accounts of behaviour have to be relied on, there are few possibilities for objective verification, and disclosure may include acts which are socially disapproved of, if not illegal. The twin issues of veracity (truthfulness) and recall particularly exercise those who have doubts about the validity of sexual behaviour surveys. (Wellings, K. *et al.*, paper B)

These comments illustrate well the measurement problems faced by the Natsal team. Although especially pertinent to a study of sexual behaviour, the issues they mention of truthfulness, recall, the opportunity for objective verification, and so on, are principles that apply to the measurement of much more mundane behaviours such as smoking and drinking. Furthermore, while we are concentrating here on good measurement in relation to aspects of knowledge and human behaviour, these principles apply to measurement in general, including other characteristics of human populations (such as height, blood pressure, and presence or absence of a given disease), as well as characteristics of the environment (such as air pollution and quality of housing). Measurement issues relating to some of these other characteristics are explored later in this chapter and in other chapters.

There are a number of factors to consider in the design of their survey that can improve the accuracy of measurement. Paper B from the Natsal-1 study includes a good discussion of these issues in respect of the assessment of behaviour, and it suggests that the following are important in ensuring accurate measurement:

- A non-judgmental approach.
- Thorough development work, leading to the choice of meaningful language in the wording of questionnaires.
- Careful consideration of when to use specific data-collection methods, including interviews, self-completion booklets, and computer-assisted methods, the more private methods being used for the most sensitive information.
- Use of precise definitions.
- Recognition of the problems that people face in accurate recall and the inclusion of devices to help people remember as accurately as possible.

4.6.2 Interview or Self-Completed Questionnaire?

In Natsal, the team used both interviews and self-completed questionnaires. Self-completed questionnaires can be completed either in the form of a booklet, as was done in Natsal-1, or by computer-assisted self-interview (CASI), as in Natsal-2 and Natsal-3. Interviews can be conventional pencil and paper interviews (PAPI), as in Natsal-1, or by computer-assisted personal interview, as in Natsal-2 and Natsal-3. So, for any given situation, how do we decide which method to use and which will give the best results? As in much of research, the decision is a trade-off between the best scientific method for obtaining the information required and practical considerations of cost and time. The principal advantages and disadvantages of interviews and self-completed questionnaires are as shown in Table 4.6.1.

Table 4.6.1 Interview or self-completed questionnaire?

Method	Advantages	Disadvantages
Interview	<ul style="list-style-type: none"> • The interviewer can ensure questions are understood • Can explore issues in more depth • Can use more <i>open</i> questions (discussed in Section 4.6.3) • If a computer-assisted personal interview is employed, it can combine the benefits of having an interviewer present with the advantages of self-completion, as in Natsal-2 and Natsal-3. 	<ul style="list-style-type: none"> • Greater cost and time • Need to train interviewer • The interviewer can influence the person's answers; this is called <i>interviewer bias</i>
Self-completed questionnaire	<ul style="list-style-type: none"> • Quicker and cheaper in booklet format, because large numbers can be mailed or distributed at the same time • Avoids <i>interviewer bias</i> • Self-completion, including CASI, allows respondents to record their responses to sensitive issues privately and record directly on a computer 	<ul style="list-style-type: none"> • Research team may be unaware that respondent has misunderstood a question • If, when questionnaires are checked, it is found that some questions have not been answered, it might not be possible to go back to the respondent • Need to rely more on <i>closed</i> questions (see Section 4.6.3) • The questionnaire must not be long or complex

4.6.3 Principles of Good Questionnaire Design

Good questionnaire design is achieved through a combination of borrowing other people's ideas (especially their good, proven, or validated ideas), your own innovation, and paying attention to the rules of good practice. We will look briefly at some of these rules, but if you are involved in questionnaire design, you may also wish to look at specialist references on the topic (e.g. Bradburn *et al.*, 2004).

Question Type: Open and Closed Questions

One of the principal distinctions in question type is the extent to which these are *open* or *closed*. A closed question is one in which we specify in advance all of the allowable answers to which the respondents' comments have to conform. This is done to ensure consistency in the range of answers that are given, as well as for ease of analysis, as each response can be given a predetermined code number. An example of a closed question to respondents who are known to be unemployed is shown in below:

Why are you unemployed?	<i>Tick one box.</i>
(1) I was made redundant.	<input type="checkbox"/>
(2) I am unable to work for health reasons.	<input type="checkbox"/>
(3) I cannot find a job.	<input type="checkbox"/>
(4) I do not wish to work.	<input type="checkbox"/>

This rigid channelling of answers into predetermined categories may be convenient for coding and data handling, but it does not encourage respondents to tell us much of interest, and by the same token it provides very little to go on if we want to understand more about the real reasons for their response. If we wish to capture more of this information content of the response, we can ask respondents to answer in their own words. This is called an open question, and here is the same example expressed in an open way:

Please explain in the space below, why you are unemployed.

As noted in Table 4.6.1, self-completed questionnaires generally require questions that are more closed, but that does not mean that open questions may not be used. Note also that questions can lie anywhere on a spectrum from fully closed to fully open, and the two approaches can be combined: For instance, we could add to the closed question example above an additional line such as, ‘Please write in the space below any further information you wish to provide about the reason(s) you are not employed.’

Length of the Questionnaire

Acceptable questionnaire length depends on the audience and on how specific and relevant the subject matter is to their experience. In general though, it is important to make the questionnaire as short as possible, so avoid including anything that is not really needed. As a rough guide, an interview should not exceed 20 to 30 minutes (but longer interviews are certainly possible and not uncommon), and a self-completion questionnaire should not exceed about 10 to 12 pages. The acceptability of the questionnaire length should be assessed as part of *pretesting* and *piloting*, which are described in more detail in Section 4.6.4.

General Layout

This is most critical for self-completion questionnaires, although it is also important that an interview schedule be well laid out for the interviewer to follow. In self-completion questionnaires (particularly), it is important to

- avoid crowding the questions and text;
- have a clear and logical flow;
- avoid breaking parts of questions over pages.

Ordering of Questions

Question order is determined, to a great extent, by the range of subject matter under study. However, there are some useful guidelines on question order:

- It is wise to avoid placing the most sensitive questions in the early part of a questionnaire.
- Use ordering creatively. For example, in the Natsal study design, respondents were asked about first sexual experience before later experiences, to help them to trigger recall of experiences that are less easily remembered.
- A variation on the above example is known as *funnelling*, whereby questions are constructed to go from the general to the specific, and thereby focus (funnel) the respondent’s attention on the detailed information that is really important.


Phrasing and Structure of Questions

The way questions are phrased and structured is critical, especially when dealing with sensitive material. In the Natsal study, for example, great care was taken to find the best language to use. The way the question is structured can also help. Suppose we wish to ask people how much alcohol they drink, and are especially interested in people who drink heavily. Since this is medically and (to some extent) socially disapproved of, it is helpful to phrase the question in such a way that the respondent senses that the study is not a part of this disapproval.

In the following example (for spirits; similar questions would be given for other types of alcoholic drink), the categories of consumption have been designed to cover the full range of consumption from none to extremely heavy, in such a way that a relatively heavy drinker does not appear to be giving the most extreme response (the very heaviest drinkers would, but that is unavoidable). Thus, since the recommended limit for all alcohol consumption is (currently) two units a day (a unit is 8 g of alcohol), moderate and heavy drinkers (say, those drinking 5 to 6 or 7 to 9 units per day) are still in the upper-middle categories of response, not at the extreme:

On average, how many units of spirits do you drink each day (one unit is equivalent to a single measure of whisky, gin, etc.)?			
Please tick one box only.			
None	<input type="checkbox"/>	5–6 units	<input type="checkbox"/>
1 unit	<input type="checkbox"/>	7–9 units	<input type="checkbox"/>
2 units	<input type="checkbox"/>	10–14 units	<input type="checkbox"/>
3–4 units	<input type="checkbox"/>	15 or more units	<input type="checkbox"/>

There is more to good phrasing of questions, however, than not offending people over the wording of embarrassing subjects. Whatever the subject matter, sensitive or mundane, it is vital to avoid ambiguities in the wording, which can confuse, annoy, convey different things to different people, and so on. Two questions with problems of this type are given in Self-Assessment Exercise 4.6.1.

	Self-Assessment Exercise 4.6.1
See whether you can spot some problems with these questions, and try rewording them to avoid the problem(s). (Note: these are just examples; you may well find other problems!)	
1. How important is it to you to take exercise and eat healthy food?	
	<i>Tick one.</i>
Very important	<input type="checkbox"/>
Fairly important	<input type="checkbox"/>
A bit important	<input type="checkbox"/>
Not at all important	<input type="checkbox"/>
2. These days, more and more people are turning to alternative medicine such as homoeopathy. How effective do you think homoeopathic medicines are?	
Answers in Section 4.8	

Skip Questions

The use of skip questions is part of the ordering and flow of a questionnaire, but we consider this separately because it requires some further emphasis, especially for self-completion questionnaires. A skip question is used when we want to direct respondents past questions that are not relevant to them. In the example of alcohol consumption, it is pointless to expect lifelong non-drinkers to plough through detailed questions on their spirit-, beer-, and wine-drinking histories. So, having established that they are lifelong non-drinkers, it is necessary to design the questionnaire in such a way as to direct them to the next section without anything being missed. This requires clear layout, numbered questions, and simple, clear instructions about which question to go to. Arrows and boxing off the parts to be skipped can also be useful, although complicated, fussy layout should be avoided.

4.6.4 Development of a Questionnaire

The process for developing, *pretesting*, and *piloting* an interview or self-completion questionnaire is summarised in Figure 4.6.1. Although it is important not to cut corners in the development of a measurement instrument, there is no point in reinventing perfectly good ones.

Summary on Questionnaire Design

- Interviews and self-administered questionnaires each have their advantages and disadvantages, and one method often is better suited to a given situation and research purpose than the other.
- Principles of good questionnaire design should be employed. Preparing a questionnaire is a process requiring research, development, and testing. If other languages are needed, ensure that colloquial terms are known, and have the questionnaire back-translated (by a third party) to ensure meaning has not been changed.
- There is no need to reinvent questionnaires when there are already well-used, tested instruments that meet your requirements, such as the SF-36 for measuring health status or the GHQ for measuring psychological distress.

4.6.5 Checking How Well the Interviews and Questionnaires Have Worked

We have seen how, in the Natsal studies, the research team designed what they believed would be a representative sampling method, and then, having obtained the sample, they checked their sample against a range of information sources. We now examine how the team assessed the success of their measurement techniques; that is, the interviews and self-completion questionnaires.

At first sight, it may seem that the team would have a problem with this. How else can one 'measure' sexual behaviours and knowledge other than by asking people questions? They have, if you like, already used the best available method. All is not lost, however, as there are a number of ways the success of the measurement methods can be assessed. We shall also be returning to this concept of comparison with an objective gold standard, in exploring how we can quantify the quality of measurement in Section 4.6.6 as well as in Chapter 10 when we discuss screening

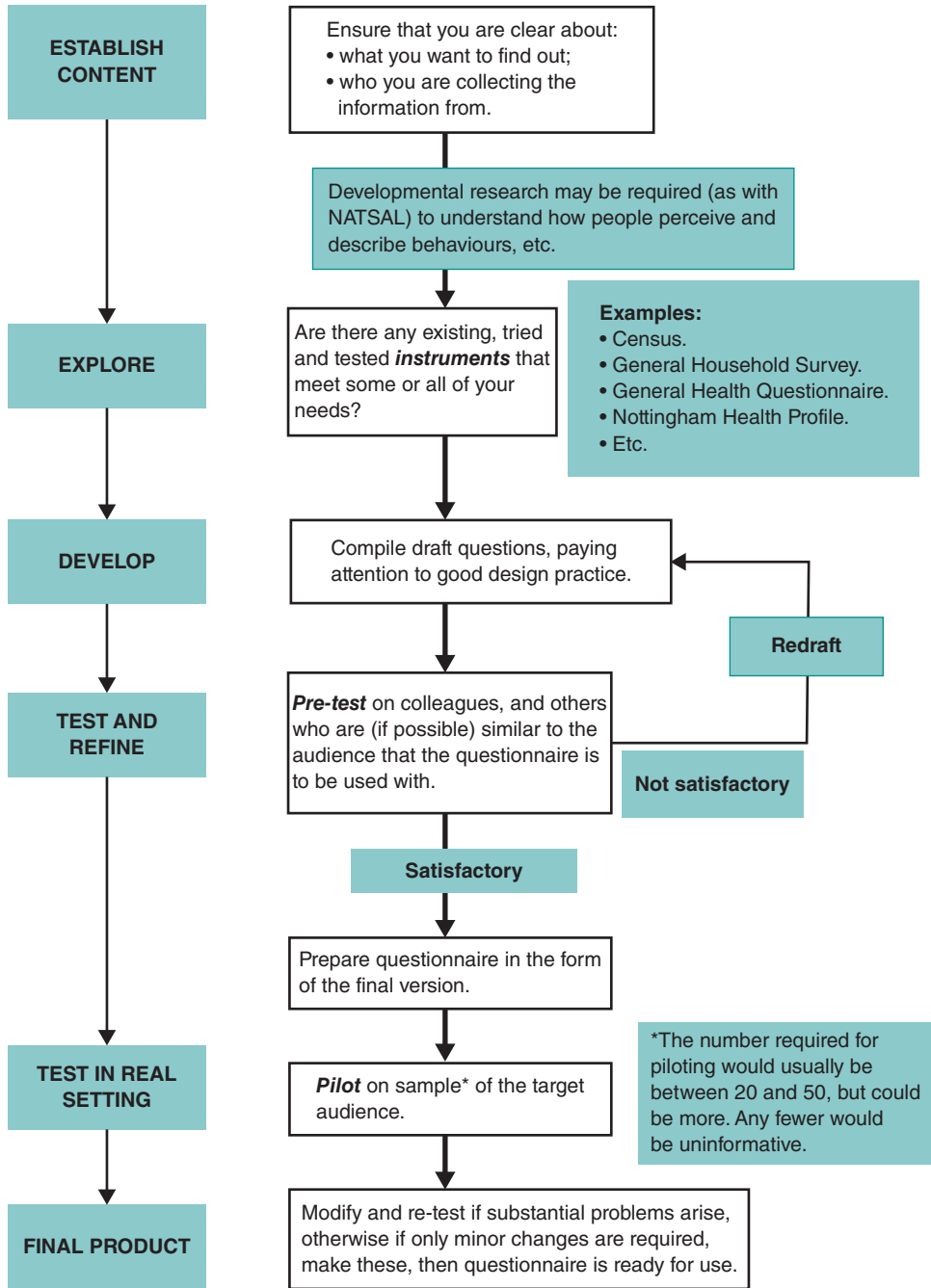


Figure 4.6.1 Summary of process for questionnaire development.

tests. These checks were described in considerable detail for the Natsal-1 study, so we look first at that. These checks can be divided into two categories:

- Comparisons of some characteristics measured in the study with independent sources of information about the same characteristics derived from routine data or other surveys. This is known as checking *external consistency*.
- Comparisons made within the study data of information on the same issue collected in different ways. Thus, information on a certain topic might be obtained initially in the interview and for a second time by self-completion booklet (or CASI in Natsal-2 or -3). This is known as checking *internal consistency*.

Comparison With Independent, External Data Sources

Since one of the main reasons for doing the original Natsal study was the lack of existing information about sexual attitudes and behaviours in a representative sample of the British population, the team was obviously not going to find an adequate alternative source against which to compare the most important data on sexual behaviours. The study did, however, include some less-specific – but nevertheless quite relevant – items that could be compared, including reported therapeutic abortions. Comparison of rates of reported abortions with routinely available national abortion statistics provided one means of checking external consistency. Table 4.6.2 shows data used for this comparison in the Natsal-1 study.

Table 4.6.2 Rates of abortions in the last year per 1,000 women reported from the NATSAL 1990 study and from national figures derived from records of abortions for the same year.

Age (years)	Great Britain: rates per 1000 women, 1990	NATSAL 1990 study (women aged 16–44 years)	
		Rate per 1,000 women	95% CI
16–19	25.2	23.5	14.5–35.0
20–24	27.1	21.4	14.1–31.5
25–29	18.1	14.5	8.9–22.0
30–34	12.0	8.9	4.8–16.2
35–39	7.5	6.1	2.3–12.0
40–44	2.7	4.0	1.3–9.5
Total	15.3	12.7	10.2–15.5

Source: Wadsworth 1993. Reproduced with permission of Journal of the Royal Statistical Society.



Self-Assessment Exercise 4.6.2

1. Describe the rates reported in Natsal-1 in comparison with the national data.
2. What do you conclude about the accuracy with which this sensitive topic has been recorded in Natsal-1?
3. What are the implications for measurement of other sensitive information in the Natsal-1 study?

Answers in Section 4.8

This comparison may seem inadequate in a number of ways. It is not the most important information that is being checked, and comparisons are made with data collected and/or measured in different ways. No comparison was available for men. These points are all fair criticism, but it is often the best that can be done, and it is better than making no checks at all. Furthermore, these external comparisons are only one of a number of ways the adequacy of data collection has been checked in the study. This leads to the second approach used to check consistency.

Checking Internal Consistency

Full advantage was taken of CASI (e.g. compared with a paper self-completion questionnaire) to build in complex filtering (so that respondents are filtered past questions which do not apply to them, given their response to earlier questions) as well as a number of range and consistency checks to ensure high data quality. For example, when asking respondents for the number of sexual partners they had in the last year, the programme checked that the number keyed in was not greater than the number of partners they said they had in the last five years. When this did happen, the CASI programme would prompt the respondent to check that they had correctly keyed in their answers to the questions which were inconsistent. Respondents could then change their response to one of the questions or they could elect to leave their responses so the apparent inconsistency would be retained.

This measure of validity of data is an important means of checking internal consistency between responses to different questions by giving respondents more than one opportunity for disclosure in relation to specific questions and providing a measure of validity of data as shown below. Natsal-3 included 137 internal consistency checks. Table 4.6.3 summarises the percentages of men and women with no inconsistencies, one or two, and three or more:

Table 4.6.3 Inconsistencies in Natsal-3 responses.

Number of inconsistencies	Men	Women
None	78.6%	82.4%
One or two	19.2%	16.2%
Three or more	2.2%	1.4%

The researchers worked on the assumption that people were more likely to report more-sensitive information in the self-completion component (CASI), which suggests that the responses in the CASI were closer to the truth than those given in the face-to-face interview. It does not prove it, of course, because the 'truth' is not available. Note that internal checks such as these can be built into an interview or a self-completion questionnaire, and it is not necessary to have one method (interview) to check the other (self-completion). If, on the other hand, both data-collection methods are required to meet the study aims, as was the case with Natsal on account of the sensitive information required, then it makes sense to make comparisons between them.

Summary

- A variety of methods are available for assessing the accuracy of measurement techniques.
- It is important to anticipate how these checks might be made at the design stage, particularly with internal consistency, since these comparisons cannot be made unless the necessary questions are built in to the interviews and/or self-completion questionnaires.

4.6.6 Assessing Measurement Quality

We have seen from this review of Natsal that the *instruments* used to collect information can result in measurement error in a number of ways. Applying the principles of good design, backed up by developmental research and testing, should result in instruments that keep error to a minimum. Nevertheless, it is important to be able to assess (measure) how well these instruments perform, over time, in different settings, or with different groups of people, and so on. The two key components of this performance are *validity* and *repeatability*.

Validity

Validity is a measure of accuracy of a test or instrument and is defined as follows:

Validity is the capacity of a test, instrument, or question to give a true result.

An instrument is said to be valid after it has been demonstrated to be accurate after being tested, repeatedly, in the populations for which it was designed. Different forms of validity are commonly described, and examples of these are presented in Table 4.6.4 in relation to questionnaire design in the assessment of back pain.

We describe the statistical measurement of validity (including sensitivity, specificity, and predictive value) in the context of screening tests in Chapter 10.

Repeatability

Repeatability describes the degree to which a measurement made on one occasion agrees with the same measurement made on a subsequent occasion. Synonymous terms are *reproducibility* and *reliability*. We have all probably experienced the frustration of making a measurement – for instance, on a piece of wood we plan to saw – and then gone back to check it, only to get a slightly different result and then wonder which one is right. In this situation, we can be fairly sure that it is our carelessness that has caused the error, but in health research the situation is rather more complex. The following example helps to illustrate why this is so. Table 4.6.5 shows the mean systolic and diastolic blood pressures of a group of 50 middle-aged men who were screened in a well-man clinic in a general practice and then followed up on two further occasions. The blood pressures were taken by a variety of different staff in the practice (doctors and practice nurses).

So, why are the mean values different on the three occasions? There are a number of possible explanations for this:

- *Chance*: We must consider chance variations, but this is unlikely to be due to the pattern of a progressive decrease in both systolic and diastolic. We could calculate the 95 per cent CIs around these mean values to see whether the changes are within the range of random variation. The 95 per cent CIs in fact show that chance is very unlikely to be the explanation.
- *Real Change*: The men's blood pressures are actually lower because they are getting used to the procedure and the staff: This is a real change in the men's measured blood pressures, but it is one related to the measurement setting.
- *Artefact or Bias*:
 - The screening and follow-up blood pressures were taken by different staff, and their technique of blood-pressure measurement could have varied. This type of inconsistency between people making measurements is known as *observer variation*. This may occur *within observer* (for instance, the same doctor uses a different technique on two

Table 4.6.4 Definitions of validity and examples in respect of a questionnaire designed to measure disability (restriction of activities of daily living) associated with back pain.

Validity type	Description	Example
Face	<p>The subjective assessment of the presentation and relevance of the questionnaire: Do the questions appear to be relevant, reasonable, unambiguous, and clear?</p>	<p>During questionnaire development, a focus group is held, including five patients with back pain (including both acute and chronic symptoms). The group is used to ensure face validity, ensuring it is easily comprehensible and unambiguous to potential respondents.</p>
Content	<p>Also theoretical, but more systematic than face validity. It refers to judgements (usually made by a panel) about how well the content of the instrument represents the full scope of the characteristic or domain it is intended to measure.</p>	<p>For content validity, another focus group is held, including a physiotherapist, a rheumatologist, and an occupational therapist with an interest in back pain, together with additional patients. They ensure that the questionnaire covers the full spectrum of disabilities that might be related to back pain, and, specifically, that it includes disability items from the relevant domains, such as physical, psychological/emotional, and self-care.</p>
Criterion	<p>This relates to agreement between the measure and another measure that is accepted as valid (referred to as the gold standard). This is often not possible as there are no gold standards for topics such as quality of life, so proxy measures are used instead. Criterion validity is usually divided into two types: concurrent and predictive validity:</p> <ul style="list-style-type: none"> ● Concurrent validity is the independent corroboration that the instrument is measuring what it is designed to. ● Predictive determines if the instrument can predict future changes in key variables in expected directions. 	<p>Questions relating to a restricted range of movement are assessed for criterion validity as follows: 50 patients with acute and chronic back pain are given the questionnaires to complete. Each then receives a physical examination to measure the range of back movement, using a standardised technique. The level of disability assessed by questionnaire is then compared with the measurements from the physical examination to assess <i>concurrent</i> criterion validity.</p>
Construct	<p>This is the extent to which the instrument tests the hypothesis or theory it is measuring. There are two parts to construct validity</p> <ul style="list-style-type: none"> ● Convergent construct validity requires that the measure (construct) should be related to similar variables. ● Discriminant construct validity requires that the measure (construct) should not be related to dissimilar variables. 	<p>Six weeks later, the patients repeat the questionnaire and physical examination. To assess the <i>predictive</i> criterion validity, the change in patients' disability between the original questionnaire score and that six weeks later is compared to the change in range of back movement. A disability score is calculated by combining 25 self-reported disability items for the 50 patients with back pain. These scores are then compared to self-rated severity of low back pain, an item that has also been included in the questionnaire and is measured on a 10-point scale: 1 = minimal pain, 10 = very severe pain. The correlation between these measures of disability and pain severity, which we would expect to be related, allows assessment of <i>convergent</i> construct validity. <i>Discriminant</i> construct validity is assessed by comparing the disability score with a variable we would not expect to be related, such as knowledge of the Internet.</p>

Table 4.6.5 Blood pressure of middle-aged men.

	Mean systolic (mmHg)	Mean diastolic (mmHg)
Screening	162	85
First follow-up	157	82
Second follow-up	152	78

occasions) or *between observers* (for instance, the doctor uses a different technique from the nurse's). Observer variation is a potentially important source of *bias*.

- The sphygmomanometers (blood pressure-measuring equipment) used were in need of maintenance and *calibration*, and some were leaking at the time of the follow-up measurements and consequently gave lower readings. This is another potential source of bias.

In relation to the Natsal studies, it would not make sense to discuss repeatability in terms of findings between studies, because the studies' findings relate to two different populations. The concept of repeatability refers to measurement that has been carried out more than once on the same population. We will therefore look at an example of repeatability from a study on physical activity, the European Prospective Investigation into Cancer Study, Norfolk cohort (EPIC-Norfolk). This study involved the development of a comprehensive questionnaire for measuring physical activity and energy expenditure. To assess repeatability, a group of 399 randomly selected participants were asked to complete a physical activity questionnaire twice within a three-month interval, and the data collected on the two different occasions were then compared (Table 4.6.6).

Table 4.6.6 Descriptive characteristics and questionnaire-derived physical activity variables in the EPIC cohort and repeat sample (data shown are for men only).

Item	Total cohort ($n = 2126$)	Repeat sample	R value
Age (years) ^a	64.6 (8.4)	65 (8.2)	n/a
Body mass Index (kg/m^2) ^a	29.6 (3.6)	26.9 (3.3)	n/a
Activity (home) MET.h/wk ^b	18.1 (9.1,30.4)	20.6 (12.9,32.0)	0.77
Activity (work) (MET.h/wk) ^b	0.0 (0.0,70.4)	0.0 (0.0,15.7)	0.57
Recreational activity (MET.h/wk) ^b	27.8 (14.7,48)	28.4 (14.9,47)	0.69
Vigorous activity ^b	0.0 (0.0,0.8)	0.0 (0.0,0.5)	0.75
Self-reported physical activity index ^b (MET.h/wk)	54.9 (25.0,108.1)	43.0 (20.1,86.2)	0.74

Figures in brackets are ^a mean and standard deviation and ^b median and interquartile range. P value for all R values were ≤ 0.05 .

Source: Wareham 2002. Reproduced with permission of Oxford University Press.



Self-Assessment Exercise 4.6.3

Looking at the data in Table 4.6.6, what conclusions do you draw about the reliability of the physical activity questionnaire that was developed for this study?

Answer in Section 4.8

To conclude this section, we return to the Natsal studies and look at the proportion of men who had vaginal sex in the last month between the Natsal-2 and Natsal-3 studies. We will consider whether the differences are likely to be the result of chance or artefact or are indeed real.

Self-Assessment Exercise 4.6.4

1. Use the following table to compare findings in Natsal-2 and Natsal-3 (with 95 per cent CIs), for having had vaginal sex in the last 4 weeks among men aged 16–44 years.

Behaviour	Natsal-2		Natsal-3	
	%	95% CI	%	95% CI
Vaginal sex in last month (men 16–44 years)	72.2	71.2–73.3	68.0	66.6–69.4

2. Drawing on your learning so far on sources of error and the interpretation of differences, make brief notes on the possible explanations for Natsal-2 and Natsal-3 findings.

Answers in Section 4.8

When you have checked your answer, please read the following section from the discussion section of paper A, in which the authors discuss the possible reasons behind the issue of changes in reporting of sexual behaviours in greater detail.

Now read the following excerpts taken from the discussion section of Paper A.

Discussion

We have presented findings from Natsal-3 on sexual attitudes and lifestyles in Britain in 2010–12. By also including data from the two previous Natsal surveys and thus responses from more than 45 000 people, we could track the sexual lifestyles of successive British birth cohorts back to the 1930s. We have shown that substantial changes have occurred in age at first heterosexual intercourse, numbers of sexual partners, sexual practices, and attitudes towards sex.

We have shown wide variability in sexual lifestyles by sex, age, and birth cohort, and, for the first time, have recorded behaviour patterns and attitudes in individuals aged up to 74 years. Most adults at all ages are sexually active, but sexual frequency and the range of practices reported reduces with age, especially in women. Although many aspects of health behaviour have strong social determinants, we recorded complex and inconsistent patterns, and noted that education is more strongly associated with sexual behaviours and attitudes than is individual socioeconomic status. Area-level deprivation was seldom associated with sexual behaviours.

Although response in the previous Natsal surveys was higher than for Natsal-3 (66.8% for Natsal-1 and 65.4% for Natsal-2), response rates for social surveys in Britain have decreased in the past decade, and different sampling strategies and changing industry standards for calculation of response make direct comparisons with other surveys, including Natsal-1 and Natsal-2, difficult. However, the response in Natsal-3 is in line with other major social surveys completed in Britain around the same time. Nonetheless, we acknowledge that non-response could be a source of bias for our data. We aimed to minimise this bias by weighting the sample so that it was broadly representative of the underlying population with respect to the distribution of the sexes, age,

and regions as used in the census. Furthermore, the sampling strategy used for the Natsal studies means that the target population is specifically the population resident in private households in Britain, and as such excludes individuals living in institutions, whose behaviour could differ from others, such that this strategy is also a potential source of bias.

Caution is needed when interpreting changes in behaviour captured by cross-sectional surveys like Natsal. Behaviour change and differences between men and women should be considered in the context of changing social attitudes and norms, which can affect willingness to report and social desirability bias. The hypothesis that changing attitudes and norms affects willingness to report and social desirability bias has been examined elsewhere. By contrast with comparisons between Natsal-1 and Natsal-2, which suggested that willingness to report might have increased in Natsal-2 because of improvements in methods, we noted little evidence of such a difference in a similar comparison between Natsal-2 and Natsal-3. We partly attribute this finding to fewer methodological differences between the latest two surveys than between Natsal-1 and Natsal-2, because we used computer-assisted personal interview and self-interview for both Natsal-2 and Natsal-3 (which have contributed to low levels of item-non-response; typically 1–3%), together with consistent question wording across all three surveys.

The recorded trends need to be considered against the backdrop of changing social norms, demographic trends, and changing legislation and policy. In Britain, as in many other countries, the position of women in society—particularly their increased social, economic, and reproductive freedom—has continued to change. The proportion of women who were married or cohabiting decreased substantially between the three surveys, and the intervals between first heterosexual intercourse, first cohabitation, and birth of first child have grown. The portrayal of women's increasing independence and choice of diverse sexual lifestyles in the media could have increased both inclination to engage in, and willingness to report, experiences. The demographic and social changes provide new opportunities for women and their sexual lifestyles, as shown by the increased numbers of partners and greater likelihood of same-sex experience reported by Natsal-3 participants.

The decrease in sexual frequency and recent vaginal sex since Natsal-2 also needs to be set in demographic context. The proportion of people not living with a partner has increased since 1996, because of an increase in the proportion who marry late in life or not at all, or who experience breakdown of relationships. However, because sexual frequency in individuals living with a partner also dropped during this time, people in Britain seem to report sex less frequently nowadays, even taking account of changes in the nature of sexual partnerships.

Despite an increase in the proportion of the population not living with a partner and thus potentially seeking to form new partnerships, as well as the new opportunities for people to meet and interact (e.g. via social media and the Internet), we recorded little change in partner acquisition since the increase between Natsal-1 and Natsal-2.

4.6.7 Overview of Sources of Error

The following table summarises the main sources of error we have looked at so far in epidemiological research (Table 4.6.7). We will return to this classification in later sections as we explore some of these sources of error in more detail and look at ways to overcome them. For now, it is important to note that there are two principal types of error in this table:

- **Systematic error** or **bias**, which falsely alters the level of a measurement.
- Chance, or **random error**, which causes the sample estimate to be less precise.

Table 4.6.7 Summary of sources of error in surveys.

Source	Type of error	Comment
Sampling	Sampling error	Random imprecision arising from the process of sampling from a population and quantified by the standard error (if the sampling is random).
	Selection bias	Non-random systematic error arising from a non-representative sampling method or nonresponse bias.
Measurement – instrument	Inaccuracy (poor validity)	Systematic error (bias) in measurement due to inadequate design of the measurement instrument, or poorly calibrated or maintained equipment.
	Poor repeatability (unreliability)	This may be due to variable performance in different situations or with different subjects (which may be mostly random), or it may be a systematic drift over time, which is bias (and in effect a change in accuracy).
Measurement – observer	Between observers	A systematic difference (bias) between measurements obtained by different observers, arising from the way they carry out the measurements, their training, etc.
	Within observers	Measurements by the same observer vary between subjects and over time, due to inconsistencies in technique, etc. This is mainly random variation, but it may also drift over time in the same way as described for instruments.

This is a very important distinction, because we can deal with the two problems in quite different ways. Thus, one of the main ways to reduce random error is by increasing the number of subjects or measurements. This is why sample size estimation is such an important aspect of study design. On the other hand, bias must be avoided (or at least kept to a minimum) by representative sampling and the use of valid (accurate) measurement instruments and techniques, including careful operator training.



Self-Assessment Exercise 4.6.4

A study of nutrition was carried out among 1,000 women attending an antenatal clinic. Women were allocated randomly to see one of two dieticians (A and B), who asked them about their diet during the pregnancy using a standard interview questionnaire. The estimated iron intake based on the questionnaire for the women seeing dietician A was 25 per cent lower than that for dietician B. Discuss how each of the following could explain the observed difference:

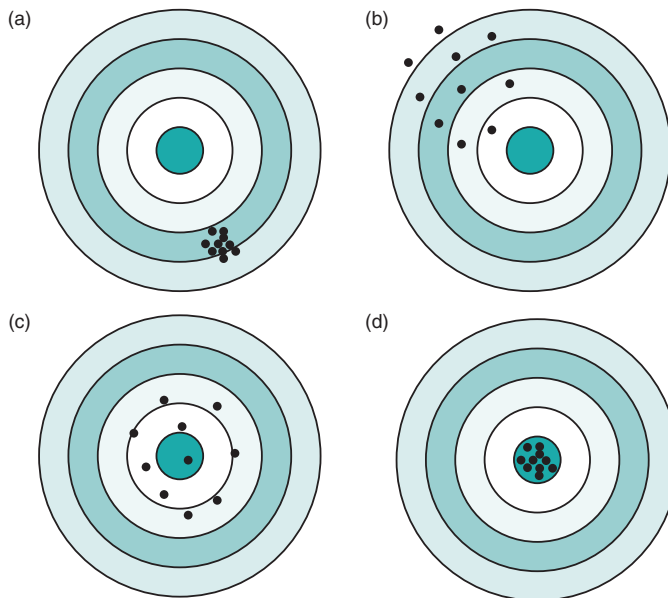
- a. A real difference in iron intake
- b. Between-observer bias
- c. Instrument unreliability

Answers in Section 4.8



Self-Assessment Exercise 4.6.5

The diagram below illustrates four targets in a firing range, each of which has been shot with 10 bullets. For each target (a–d), describe the validity (accuracy) and reliability (repeatability).



Answers in Section 4.8

Summary on Measurement Quality

- The validity of a measurement instrument describes how accurately it measures the true situation.
- The validity of measurement should be checked where possible; both internal consistency checks and external comparisons offer opportunities to do this. Some forward planning is often required to make these checks possible.
- Repeatability describes how consistently the instrument functions on different occasions.
- Error in measurement falls into two broad categories: random and systematic (bias), and it may arise from sampling, from measurement instruments, and from the observers who carry out the measurements.

4.7 Data Types and Presentation

4.7.1 Introduction

A survey such as Natsal generally results in a vast amount of data. In their raw state, the data tell us little. We need to *summarise* the data to make sense of the information they contain,

and we need to present this information in an appropriate way. We need to *analyse* the data to get answers to our questions and make *inferences* about the population. The ways data are summarised, presented, and analysed depend on the type of data. In Chapter 2 we looked in some detail at presenting and summarising *continuous* data. In the Natsal study, we meet other types of data, which require different methods, which are discussed here.

4.7.2 Types of Data

Quantities that vary from person to person, such as age, sex, and weight, are called *variables*. Measured values of these variables are *data*. For example, weight is a variable. The actual weights of five people (say, 62.5, 68.4, 57.3, 64.0, and 76.8 kg) are data. Values may be measured on one of the four measurement scales shown in Table 4.7.1.

Table 4.7.1 Types of data.

Type	Description
Categorical (nominal)	A nominal scale has separate, named (hence 'nominal') classes or categories into which individuals may be classified. The classes have no numerical relationship with one another; for example, sex (male, female) or classification of disease.
Ordered or ranked categorical (ordinal)	An ordinal scale has ordered categories, such as mild, moderate, or severe pain. Numbers may be used to label the categories, such as 1 = mild, 2 = moderate, 3 = severe, and so on, but this is only a ranking: the difference between 1 and 2 does not necessarily mean the same as the difference between 2 and 3.
Interval	An interval scale is so called because the distance, or interval, between two measurements on the scale has meaning. For example, 20°C is 30 degrees more than -10°C, and the difference is the same as that between 60°C and 30°C.
Ratio	On a ratio scale, both the distance and ratio between two measurements are defined: 1 kg (1,000 g) is 500 g more than 500 g and also twice the weight. An additional property is that the value zero means just that, whereas with an interval scale such as degrees Celsius, 0°C does not mean 'no thermal energy'.

Interval measurements can be further classified into *continuous* or *discrete*. Continuous measurements can take any value within a range, the only restriction being the accuracy of the measuring instrument. Discrete measurements can take only whole-number values: A household may contain one, two, three, or more people but not two and a half people. Note that interval measurements are also ordered. Discrete interval data may be treated as ordered categories or continuous, depending on the number of discrete values and how they are distributed (we will return to this in Section 4.7.3).

Because there are a number of interchangeable terms for data types, from now on we shall refer to *categorical*, *ordered categorical*, *discrete* (interval), and *continuous* (interval, ratio) data. We shall now look at some of the variables measured in the Natsal 2 study, which include:

- Age
- Marital status
- Sex
- Number of heterosexual partners in the past 5 years



Self-Assessment Exercise 4.7.1

For each of the variables used in Natsal (age, marital status, sex, number of partners), think about how each one is likely to have been measured and so classify them into categorical (including whether or not ordered) or continuous variables (type of scale and whether or not discrete).

Answers in Section 4.8

4.7.3 Displaying and Summarising the Data

We generally start by describing the number of study participants and their basic characteristics, such as age, sex, ethnicity, and so on. We now look at how this is done for categorical variables.

Categorical Data

Often, many of the data collected in surveys are categorical, such as sex (male, female), behaviours such as smoking (never, ex-smoker, current smoker), or a disease classification (present or absent). Data on one categorical variable can be summarised by the number of individuals in each category. This is the *frequency distribution*, and it is simpler to construct for a categorical variable than for a continuous variable (Chapter 2, Section 2.4.2) because we do not have to decide how to divide the range of the data into intervals. Table 4.7.2 shows the frequency distribution of marital status of participants in the Natsal-3 study taken from Table 1 of paper A.

Table 4.7.2 Frequency distribution of marital status in the Natsal-3 sample.

Marital status	Number*
Single	2,115
Cohabiting	936
Married	3,797
Previously married or in civil partnership	660
Total	7,508

*Numbers are estimates based on published percentages.

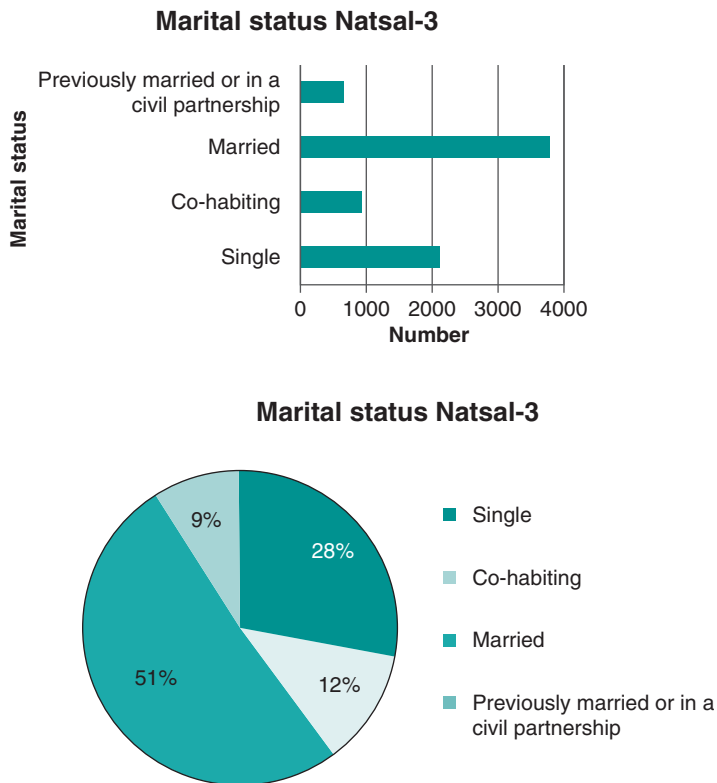
It is easier to compare the relative numbers in each category by percentages or proportions of the total number. These are the *relative frequencies* introduced in Section 2.4.4 of Chapter 2. They are particularly useful for comparing two distributions, as when we compared characteristics of the Natsal sample with the 2011 population census in Section 4.5.2 (Table 4.7.3).

The frequency and relative frequency distributions can be displayed in a *bar chart* or a *pie chart*, as in Figure 4.7.1.

Note that there are spaces between the bars of the bar chart. A histogram does not have spaces between intervals, because they are parts of a continuous scale. In the pie chart, the angle of each sector is proportional to the frequency (or relative frequency). The *modal category* of the distribution is sometimes stated. This is simply the category with the highest frequency or relative frequency. The modal category of marital status in Natsal-3 is 'married'.

Table 4.7.3 Relative frequency distribution of marital status in the Natsal-3 sample.

Marital status	Number	Percentage
Single	2,115	28.2
Cohabiting	936	12.5
Married	3,797	50.6
Previously married or in civil partnership	660	8.8
Total	7,508	100

**Figure 4.7.1** Data on marital status from Natsal-3 presented as (top) bar chart of frequency distribution and (bottom) pie chart of relative frequency distribution.

Discrete Data

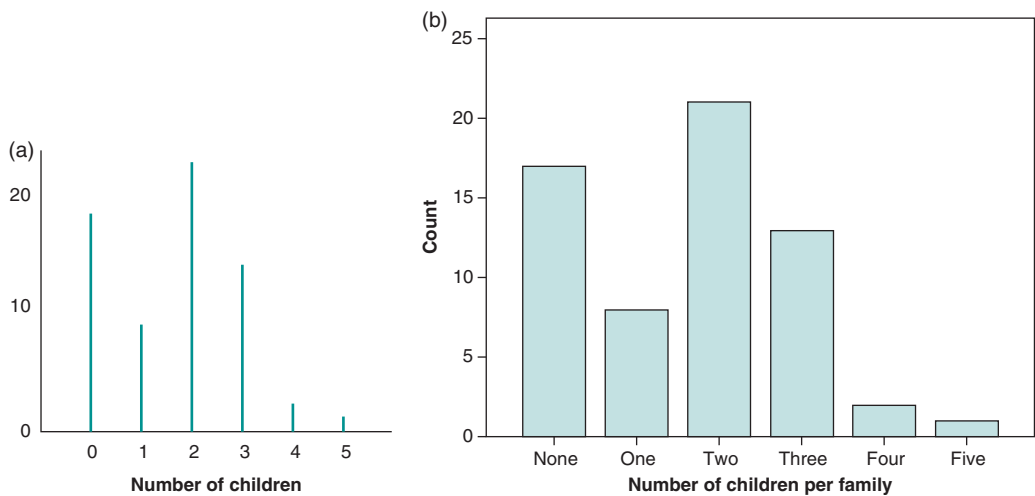
Discrete data are tabulated in exactly the same way as categorical data (Table 4.7.4). Here we illustrate the distribution of numbers of children in each of 62 families. These data are discrete because families can only have whole numbers of children, that is, 0, 1, 2, etc., and not 1.4, 2.1, etc. The frequency distribution shows the frequency of each separate value.

The frequency distribution can be displayed as a *line chart* or a bar chart (Figure 4.7.2).

If there are many different values in the data, they may be grouped. This is not generally into equal-size groups, but they are grouped in such a way that the most frequently occurring values

Table 4.7.4 Hypothetical frequency distribution of number of children in 62 families.

Number of children	Frequency
0	17
1	8
2	21
3	13
4	2
5	1
Total	62

**Figure 4.7.2** (a) Number of children. (b) Number of children per family.

are shown individually, and the less-common values are grouped. For example, the number of heterosexual partners in the past 5 years is grouped in Table 2 of paper A as shown in Table 4.7.5.

Table 4.7.5 Percentage of heterosexual partners in the past 5 years among women in the Natsal-3 sample.

Number of partners	Percentage
0	4.2
1	22.4
2	11.1
3–4	19.2
5–9	23.2
≥10	19.9
Total	100

The location and spread of discrete data may be summarised by the mean and standard deviation or the median and interquartile range, depending on whether the distribution is approximately symmetric or is skewed. The same criteria apply as for continuous data (Chapter 2, Section 2.4.3).

Summary of Data Presentation

Table 4.7.6 provides an overview of methods recommended for presenting and summarising the main types of data.

Table 4.7.6 Methods for presenting and summarising data.

Type of data	Presenting the data	Summarising the data
Categorical (including ordered)	Frequency/relative frequency table Bar chart Pie chart	Percentages Modal category
Discrete	Frequency/relative frequency table Bar chart Line chart	Mean and standard deviation (symmetric distribution) Median and interquartile range (skewed distribution)
Continuous	Frequency/relative frequency table* Histogram Box and whisker plot**	Mean and standard deviation (symmetric distribution) Median and interquartile range (skewed distribution)

*For continuous data, a frequency/relative frequency table would be obtained by dividing the data into subgroups.

**Box and whisker plots are illustrated and explained in Chapter 11, Section 11.3.8.

4.8 Answers to Self-Assessment Exercises

Section 4.1

Exercise 4.1.1

Reasons for carrying out the Natsal-3 study

a. *General background reasons*

- There is no available source of information that reports on sexual health and lifestyle across the British population.
- Improving sexual and reproductive health remains a public health priority in Britain.
- A number of factors contribute to a population's sexual health, including social context and the interplay among behaviour, relationships, and health status.
- Research into the sexual health and wellbeing of men and women in later life—who now have increasing expectations of sexual fulfilment and make up a growing segment of the population—is a neglected area.

b. *Specific needs*

- To combine data from the first two surveys with data from the third Natsal survey to enable both period and birth cohort analyses (we will explore the concept of period and cohort approaches in Section 10.4 of Chapter 10).
- To examine changes in sexual lifestyles throughout the life course and trends over time.

Section 4.2

Exercise 4.2.1 Probability

1. The probability that a smoker is chosen is

$$\frac{\text{Number of smokers in population}}{\text{Total population}} = \frac{25000}{100\,000} = 0.25$$

2. The probability the person has had IHD is

$$\frac{875 + 750}{100\,000} = 0.01625$$

3. The probability that the person chosen is a smoker who has suffered IHD is

$$\frac{875}{100\,000} = 0.00875$$

Exercise 4.2.2 Quota sampling

1. A quota of 70 women aged 20–24 years is required (28 per cent of 250).
2. If they were to do this, it would of course badly underrepresent people in employment or with other reasons to be away from their homes, and the sample would be very biased.
3. It is not very easy to find a quota of, say, 75 males aged 15–19 years who are representative of everyone in that age group. Contrast this with a random sampling method in which a name and address is specified, and the research team has to keep trying (within limits!) until the specified person is contacted and can be offered the opportunity to take part in the study. Nevertheless, in expert hands, quota sampling can be very effective, and it can be made more representative by breaking the quotas down into more subgroups (for instance, men, aged 20–24 years, in employment; men, aged 20–24 years, unemployed; etc.), so that a more-representative structure is imposed on the sampling and quota-finding process.

Exercise 4.2.3 Weighting

1. The overall percentage of people in the sample reporting the behaviour of interest is $105/10,900 = 9.68$ per cent.
2. This is incorrect because it does not allow for the fact that, for example, people from inner London have a much higher prevalence of the behaviour and are three times more likely to be in the sample.
3. $900/3 = 300$ people.
4. $225/3 = 75$ people.
5. The weighted answer would be $(630 + 100 + 75) \div (9000 + 500 + 300) = 8.2$ per cent. This is considerably different from the unweighted answer to question 1 of 9.7 per cent.

Exercise 4.2.4 Sampling methods

1. *Cardiovascular disease*

Simple random sampling would be appropriate. Cluster sampling would probably not be necessary or appropriate in an area such as a district. It did form part of the Natsal sampling method, with postcode sectors as clusters, but this was in part due to the very large area that needed to be covered and visited by interviewers (locations across the whole country). Quota sampling would be a poor method unless relatively simple information was required quickly. It would be a very inappropriate way to approach people if a physical examination and blood tests were required. A convenience sample would be inappropriate and would be very unlikely to yield a representative sample.

2. *People older than 75 years in residential homes*

Simple random sampling would be difficult given that the residents are spread very unevenly in the population, but a cluster sample would be appropriate, because the homes are in effect clusters. We could select the homes randomly and then sample people randomly within homes or survey all residents in the selected homes, depending on the sample size required. A snowball sample is inappropriate and unnecessary, as it is not difficult to identify and contact the subjects, so that a more satisfactory random-sampling technique is quite possible. Quota sampling would also be inappropriate and unnecessary, for the same reasons.

3. *Drug users*

A cluster sample would be inappropriate because the women will be living in many different situations and not in any identifiable clusters. A snowball sample may be a good approach, depending on whether the available networks and contacts could yield a sample with this hard-to-reach group. Simple random sampling is unlikely to work because it would be hard to identify people, and contacting the women via a standard sampling frame (e.g. addresses) would probably be ineffective. Quota sampling would be inappropriate because we have no basis for working out quotas.

Section 4.3

Exercise 4.3.1 Sampling frames

Here are some examples of sampling frames (not including the postcode address files used in Natsal):

Sampling frame	Advantages	Disadvantages
Electoral register	<ul style="list-style-type: none"> Updated each year 	<ul style="list-style-type: none"> Restricted to people of voting age Addresses, but no personal details Losses due to non-registration and mobility
GP lists	<ul style="list-style-type: none"> Considerable amount of information on individuals available (subject to permission). 	<ul style="list-style-type: none"> Loss due to inaccurate addresses and non-registration may be serious; e.g. younger people in urban areas List may also be inflated by people who have actually moved away but have not yet been removed from the list
School registers	<ul style="list-style-type: none"> All children required to attend school, so almost complete and updated Age and sex information available subject to permission 	<ul style="list-style-type: none"> May be difficult to include independent (private) schools
Employee registers	<ul style="list-style-type: none"> Should be complete Personal information available subject to permission Good for studying employment-related issues 	<ul style="list-style-type: none"> Not representative for population survey purposes, as they exclude the unemployed, long-term sick, etc.

In the Natsal study, the team also needed to choose from various options, and, on balance, the postcode sectors and address files were judged best to meet their needs.

Section 4.4

Exercise 4.4.1 Sampling distribution

1. Yes, in this case the range ± 1.96 SE does include μ , as the sampling distribution is centred on the population mean.
2. 95 per cent of sample means.
3. 5 per cent of sample means.
4. No, it is not included. This shows that there is a 5 per cent (1 in 20) chance that the 95 per cent CI around a given sample estimate will not include the true population mean, an important conclusion to be aware of.

Exercise 4.4.2 CI for a mean

For the mean number of lifetime partners for women aged 45–54 years in the Natsal 3 study, we have

$$\begin{aligned}\bar{x} &= 6.8 \\ s &= 11.8 \\ n &= 1443\end{aligned}$$

So the 95 per cent CI = $\bar{x} \pm 1.96$ SE = $6.8 \pm 1.96 \times 11.8 / \sqrt{1443} = (6.19, 7.41)$.

Note that the CI for the mean number of lifetime partners for women is narrower than the one for men: The estimate is more precise. This is because, although the sample size is similar, the standard deviation is smaller, resulting in a smaller SE.

Exercise 4.4.3 Sample size for estimating a mean

1. $n = 1.96^2 \times 520^2 / 50^2 = 415.5$, so the required sample size should be around 420 (not allowing for refusals).
2. If the precision is doubled (ϵ halved), the sample size is multiplied by 4, so the sample size required is $4 \times 415.5 = 1662$. A common mistake in calculating sample size is to think ϵ is the required width of the entire 95 per cent CI, when it is in fact *half* the width as the error range is specified as $\pm \epsilon$.

Exercise 4.4.4 CI for a proportion/percentage

The sample percentage is 18.5 per cent, and the sample size is 1,238:

$$SE = \sqrt{\frac{0.185(1 - 0.185)}{1,238}} = 0.01104$$

95% CI = $0.185 \pm (1.96 \times 0.01104) = (0.1634, 0.2066)$, or in percentages (16.34–20.66 per cent). As with the mean, because the sample size is fairly large, the 95 per cent CI is narrow. We interpret this as we are 95 per cent certain that the percentage of men aged 16–24 years in Britain who had anal sex in the previous year is between 16.43 per cent and 20.66 per cent.

Exercise 4.4.5 Sample size for estimating a proportion

- a. The sample size formula for estimating a proportion using percentages is $n = 1.96^2 P(1 - P) / \epsilon^2$. Substituting $P = 1\%$ and $\epsilon = 0.2\%$, the required sample size is $1.96^2 \times 1 \times 99 / 0.2^2 = 9508$. If the prevalence had been expressed as a proportion, we would use the formula $n = 1.96^2 P(1 - P) / \epsilon^2$ with $P = 0.01$ and $\epsilon = 0.002$.
- b. If we have a sample of only 500 people, $\epsilon = 1.96 \sqrt{[P(100 - P) / n]} = 1.96 \sqrt{[1 \times 99 / 500]} = 0.87$.

That is, using a 95 per cent CI, we have a 95 per cent chance of obtaining an estimate of the prevalence to within ± 0.87 per cent of the true value. This is large relative to 1 per cent: we cannot obtain a very precise estimate from a sample of 500. This emphasises why the sample size for the Natsal studies had to be large. Some subgroups were considerably smaller, however, so these estimates (drug use, for example) have not been measured with the same precision.

Section 4.5

Exercise 4.5.1 Natsal response rate

So, is a response of 57.7 per cent adequate? It is not particularly good (less than 2/3 of original sample), but perhaps not bad given the subject matter of the survey. However, the more relevant question is whether the non-responders differed substantively from the responders, and we will look at that question next.

Exercise 4.5.2 Representativeness of sample

The sample reflects the national distribution of sociodemographic characteristics in most respects, although there are some differences. In terms of age, there is a minor discrepancy for men aged 35–39 years (slightly underrepresented), which the team report is due to trimming of the final weight for this category. For geographical region, men in inner London seem to have been underrepresented and men from outer London, overrepresented. However, the authors explain in their technical report that this is due to merging the cells for men in **inner** and **outer** London. Thus, overall, the proportion of men sampled from London corresponds fairly closely to the proportion of men living in London, based on 2011 Census data.

The team weighted the data to adjust for the unequal probabilities of selection in terms of age and the number of adults in the eligible age range at an address. After weighting, the sample was broadly representative of the British population according to 2011 census data. Where there were minor discrepancies in the proportion of respondents by sex, age, and government office region, this was dealt with by adding a non-response post-stratification weighting to correct for any differences.

Individuals in the Natsal-3 sample appear to be more likely to classify themselves in fair health than bad or very bad health, and they overrepresent people who are married or live with a spouse or partner, whilst underrepresenting men and women who are single. However, Natsal-3 did not include individuals who live in institutions in their study, and these individuals are more likely to be single, so this may explain at least part of this difference.

Exercise 4.5.3 Specimen Invitation Letter

*The Health Centre
Yellow Brick Rd
Liverpool LXX 9XX
Tel: 0111-111-1111*

25 January 2015

Dear *Mr Patterson*

I am writing to ask for your assistance in a study we are carrying out jointly with the University of Liverpool, on the treatment of chest problems such as asthma and bronchitis. The purpose of the study is to help the practice develop and improve the service we can offer our patients.

I would be grateful if you would complete the enclosed questionnaire, and return it in the postage-paid reply envelope provided.

All information will be treated confidentially. The enclosed sheet contains further information about the study and your role, should you agree to take part. Once you have read the information, and if you are happy to participate, please sign to give your consent, and return this form with the completed questionnaire.

If you have any questions about the study, please contact Dr Jenny Smith of the Department of Public Health at the University on 0222-222-2222, who will be pleased to assist you.

Yours sincerely

Anthony Brown

Dr A. Brown

There are of course many ways to phrase such a letter, but here are the key features of good practice incorporated in this one:

- It is from a respected person known to the subject. The fact that the person is known should help, but be aware that it can be inhibiting if (for example) the study is about quality of services and the respondent is reluctant to criticise the GP's service in case the subject thinks that this could affect the way the s/he is treated in the future.
- The letter has a personal touch, with handwritten name and signature.
- The reason for the study and its relevance are explained.
- A postage-paid reply envelope is provided.
- Confidentiality is explained, along with the recipient's role in the study and a form for providing informed consent.
- There is a contact name for assistance, and note that this person is not at the practice. This may be of help provided the respondent has been made aware of why people other than the practice staff are involved.
- The letter is short and clear.

Section 4.6

Exercise 4.6.1 Question phrasing

1. The problem here is that the initial question contains two important but quite different ideas, namely exercise and healthy eating. The respondent may well feel differently about the two and hence cannot answer the question meaningfully. This is called a **double-barrelled question**, and the most obvious solution is to ask two separate questions.
2. In this question, the opening statement implies growing interest, acceptability, and approval of homoeopathy. This could well convey to the respondent that an approving answer is the correct one, or at least expected, and is therefore a **leading question** (it is not neutral, and leads respondents towards one type of answer). A more neutral way of setting the context might be to start by asking respondents whether or not they had used homoeopathic medicines.

Exercise 4.6.2 Comparison of abortion rates in Natsal and Great Britain

1. All the age-specific rates are lower in the Natsal sample, with the exception of the 40–44 year age group. Although the 95 per cent CIs are wide (and all include the national rates), the fact that almost all age groups are lower for Natsal suggests that these differences are systematic rather than due to chance (sampling error).
2. On the face of it, there seems to have been underreporting of abortions in the Natsal study. The authors comment, however, that '[National] abortion statistics may include a slight

excess of temporary residents using UK addresses and women having more than one abortion in a year'. Hence, the difference may in part be due to artefact.

3. It is difficult to say with any certainty what might be going on with other sensitive information, particularly for men for whom no external comparisons were made. Overall, we may conclude that although there is some evidence of underreporting where comparative information was available (therapeutic abortions), this was not substantial and may in any case be largely due to artefact.

Exercise 4.6.3 Repeatability of a questionnaire on physical activity

The men in the total sample and also in the second subsample are of a similar age range. The correlation coefficients for the various indices between the repeat questionnaires were 0.68 or higher, with the exception of work activity, which was 0.57. These are all moderately strong correlations, and they were all significant at the 0.05 level, so we can conclude that the questionnaire is a reliable measure of physical activity. The authors suggest that a three-month time period between the two measurements is relatively long, but they wanted to try to ensure that the participants would not remember their scores, so it is possible that the true repeatability measure was higher. In summary, the authors conclude that the level of repeatability is high, at least in relation to the population on whom it was tested, although this could vary depending on the population chosen for a subsequent study and also whether future studies would be using the questionnaire for precisely the measure for which it was intended.

Exercise 4.6.4

Question 1

The data show that there was a decrease of 4.5 per cent in the proportion of men aged 16–44 years having vaginal sex in the previous four weeks between Natsal-2 and Natsal-3. The 95 per cent confidence intervals do not overlap, suggesting that the difference is unlikely to be due to chance.

Question 2

- | | |
|------------------------------------|---|
| Vaginal sex in the last four weeks | <ul style="list-style-type: none"> • Chance: 95 per cent CIs exclude this. • Artefact: Could changes in reporting methods between surveys have decreased reporting? We know this was not the case, as question wording and delivery were largely identical between Natsal-2 and Natsal-3. Is it possible that there has been an increase in the reluctance to report? If so, we would have expected to see a downward trend across all sensitive issues between the two surveys. This possibility would also appear to be at odds with changing social attitudes and a consensus of increasing willingness to report over the last fifteen years. • Real: If the decrease is not due to artefact and not due to chance, we can conclude this is a real decrease. |
|------------------------------------|---|

*95 per cent CI excludes chance, by convention, although, as we have seen, there is still a 5 per cent chance that the population value lies outside the CI of the sample estimate.

Exercise 4.6.4 Nutrition study

- a. A 25 per cent mean difference is very unlikely to be real, given the random allocation and the relatively large numbers (about 500 for each nurse).

- b. Bias arising from **between-observer variation** is the most likely explanation. When obtaining the dietary information from the women, dietician A tends, on average, to obtain results on iron intake 25 per cent lower than dietician B. Of course, this does not tell us which one is nearer the truth (most valid). The solution to this problem is to ensure standardised training for both dieticians and/or to compare with a more thorough method (a gold standard).
- c. **Instrument unreliability** is unlikely, as both dieticians are using the same standard questionnaire, and even if there is a lot of variation in how it works from subject to subject, this type of variation is mainly random in nature and will not produce such marked systematic differences (bias) between observers.

Exercise 4.6.5 Targets

- a. Validity is very poor (shots are a long way from the bull's-eye, which in this analogy is the 'truth', or measurement we are seeking to make), but the repeatability is very good. This pattern might be caused by poorly adjusted (calibrated) sights on a rifle that was otherwise in very good condition and that was being fired with very consistent technique.
- b. Poor validity and poor repeatability.
- c. Good validity, but poor repeatability: We can see that, on average, these 10 shots are on target, but there is scatter. This scatter is more or less random and could, for example, be caused by poor shooting technique or a worn barrel. Thus, in measurement-analogy terms, we can see this result as an accurate estimate but one with an excessive amount of random imprecision.
- d. The desired result: High validity and high repeatability.

Section 4.7

Exercise 4.7.1

Age

This question is a bit more difficult than it looks. Age can take any value of between 16 and 44 years (in this study), so it is a **continuous** variable, even though it may be measured only to an accuracy of whole years. However, for the purposes of analysis, age is often presented in groups as an **ordered categorical** variable (16–24, 25–34, etc.). It is most likely that age either was recorded as age last birthday or was calculated from the respondent's date of birth and the interview date and then grouped into categories for the purpose of presentation.

Sex

This is a **categorical** variable, with possible values being male and female.

Marital status

There are a number of categories, but these are not ordered in any way, so this is simply a **categorical** variable.

Number of heterosexual partners in the last 5 years

Again, as with age, we do not know how this variable was measured. It is likely that the actual number of partners was recorded, in which case it is (or can be treated as) continuous. The number of partners must, however, be a whole number (including 0), so this is a **discrete** variable. Unlike age (even in years), the number of partners for any given study respondent cannot take a value between whole numbers, such as 3.6, although it is meaningful to present such values as group averages.

5

Cohort Studies

Introduction and Learning Objectives

Survey methods have provided us with a way of studying population groups of our own choice and better-quality information on issues that may not be available from routine data sources. These methods have also allowed us to examine associations in greater detail, but they are not able to provide particularly strong evidence about factors that cause disease.

In chapters 5 and 6, we will study two research designs that offer marked advantages over the descriptive methods examined so far. These are called *cohort* and *case-control* studies. We look at cohort studies first because the design, analysis, and interpretation of case-control studies is more complex. However, many of the design issues that occur in this chapter are also relevant to case-control study design. You may find it useful to refer to Table 2.7.1, which was introduced at the end of Chapter 2, to refresh your memory of the concepts we introduced at that stage. In particular, you should focus on how the cohort study ideas fit into the overall framework of studies so far.

A cohort study may also be referred to as a longitudinal study, a prospective study, an incidence study, or a follow-up study. The defining feature of the cohort study is the follow-up of subjects over time. Obtaining sufficient information from a cohort to enable reliable estimation (e.g. of disease incidence or mortality) often requires a large sample, a fairly long follow-up period (i.e. several years at least), or both.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the purpose and structure of the cohort study design within an overall framework of epidemiological study designs.
- Give examples of the uses of cohort studies, including examples from your own field.
- Describe the strengths and weaknesses of cohort studies.
- Describe the various types of bias that can arise with cohort studies, including measurement and observer error, and how these can be minimised.
- Define relative risk and how this can be calculated from a cohort study.
- Explain the use of, carry out, and interpret the appropriate hypothesis test for categorical data (chi-squared test).
- Explain the use of, carry out, and interpret the appropriate hypothesis test for continuous data (Student's *t*-test).
- Describe the information required to carry out a sample-size calculation for a cohort study, including the concept of power.

- Define what is meant by confounding, and be able to illustrate this with examples.
- Describe the criteria used to assess the strength of evidence for judging whether an association may be causal (using the Hill viewpoints), and give examples.
- Explain the uses of, and interpret, simple linear regression.
- Describe the uses of, and principles underlying, multivariable linear regression.
- Describe how adjustment for confounding by regression methods helps in the assessment of causation, and give examples.

The primary example of a cohort study that we will look at investigates the association between physical activity and risk of cancer in middle-aged men (Paper A). This study was part of a much broader programme, the British Regional Heart Study (BRHS), designed primarily to investigate cardiovascular disease (CVD) in Great Britain. We will also refer to the original BRHS publications where necessary to provide more details of the methods used (papers B and C). Although originally set up around 1980, this study still provides a very useful example of cohort design. We will also refer more briefly to a second study, based on a birth cohort in Brazil, because this illustrates the practical issues in following up a cohort in a setting different from the UK.

Resource Papers

Paper A

Wannamethee, S.G., Shaper, A.G., Walker, M. (2001). Physical activity and risk of cancer in middle-aged men. *Br J Cancer* **85**, 1311–1316.

Paper B

Shaper, A.G., Pocock, S.J., Walker, M., Phillips, A.N., Whitehead, T.P., Macfarlane, P.W. (1985). Risk factors for ischaemic heart disease: the prospective phase of the British Regional Heart Study. *J Epidemiol Community Health* **39**, 197–209.

Paper C

Walker, M., Shaper, A.G. (1984). Follow up of subjects in prospective studies based in general practice. *J R Coll Gen Pract* **34**, 365–370.

Paper D

Hallal PC, Wells JCK, Reichart FF, *et al.* (2006). Early determinants of physical activity in adolescence: prospective birth cohort study. *Br Med J* **332**, 1002, doi: 10.1136/bmj.38776.434560.7C

5.1 Why Do a Cohort Study?

5.1.1 Objectives of the Study

We will begin our examination of cohort studies by finding out exactly why the team chose this design to investigate their research question. Please now read the summary and introduction section of Paper A, which are given below:

Summary

A prospective study was carried out to examine the relationship between physical activity and incidence of cancers in 7588 men aged 40–59 years, with full data on physical activity and without cancer at screening. Physical activity at screening was classified as none/occasional, light, moderate, moderately vigorous or vigorous. Cancer incidence data were obtained from death certificates, the national Cancer Registration Scheme and self-reporting on follow-up questionnaires of doctor-diagnosed cancer. Cancer (excluding skin cancers) developed in 969 men during mean follow-up of 18.8 years. After adjustment for age, smoking, body mass index, alcohol intake and social class, the risk of total cancers was significantly reduced only in men reporting moderately vigorous or vigorous activity; no benefit was seen at lesser levels. Sporting activity was essential to achieve significant benefit and was associated with a significant dose–response reduction in risk of prostate cancer and upper digestive and stomach cancer. Sporting (vigorous) activity was associated with a significant increase in bladder cancer. No association was seen with colorectal cancer. Non-sporting recreational activity showed no association with cancer. Physical activity in middle-aged men is associated with reduced risk of total cancers, prostate cancer, upper digestive and stomach cancer. Moderately vigorous or vigorous levels involving sporting activities are required to achieve such benefit.

Introduction

There is increasing evidence that physical activity is associated with altered risk of total cancers and certain specific types of cancer, especially colon and prostate (Lee, 1995; Oliveria and Christos, 1997; Gerhardsson, 1997; McTiernan *et al.*, 1998; Moore *et al.*, 1998; Shephard and Shek, 1998). In particular, the evidence strongly supports the role of physical activity in reducing risk of colon cancer (Lee, 1995; Oliveria and Christos, 1997; Gerhardsson, 1997; McTiernan *et al.*, 1998; Moore *et al.*, 1998; Shephard and Shek, 1998; Giovannucci *et al.*, 1995; Slattery *et al.*, 1997); the findings relating to prostate cancer have been inconsistent (Albanes *et al.*, 1989; Thune and Lund, 1994; Oliveria and Lee, 1997; Hartman *et al.*, 1998; Giovannucci *et al.*, 1998; Liu *et al.*, 2000). A few prospective studies have suggested an inverse relationship with lung cancer (Lee and Paffenbarger, 1994; Thune and Lund, 1997; Lee *et al.*, 1999). Data on physical activity and other types of cancer are limited and the amount and type of physical activity required to confer protection remains unclear for many of the cancers. There is some indication that a high level of activity (vigorous) is required to achieve benefit for prostate cancer (Giovannucci *et al.*, 1998). The inconsistent findings between studies for the various cancer types may relate to the different levels of activity in the populations studied. This raises the question of how much activity is required to achieve benefit and in particular, whether light to moderate physical activities have any effect on diminishing risk. This paper examines the relationship between physical activity and the incidence of total cancers and some site-specific cancers and assesses the type and amount of activity required to achieve benefit in a prospective study of middle-aged men.



Self-Assessment Exercise 5.1.1

1. Make a list of all of the research issues and the aim of the study noted in this passage.
2. Which of these do you think could be addressed adequately by descriptive and survey methods, and which might require a more sophisticated research design (such as is offered by a cohort study)?

Answers in Section 5.10

5.1.2 Study Structure

Before looking at the study methods in detail, we will consider the overall structure of a cohort study. This will help to tie in what we have identified about the research issues and aim of the study with the research design chosen to address them. Figure 5.1.1 summarises the structure of a cohort study with a 5-year follow-up period.

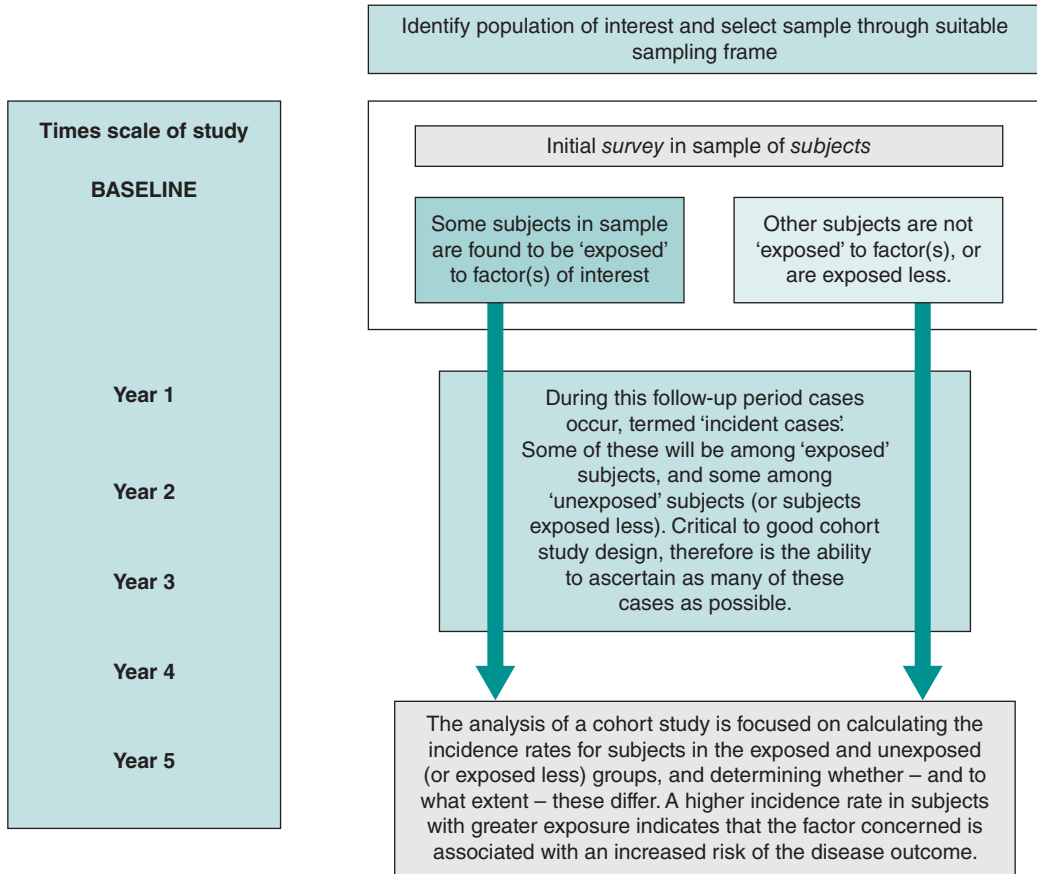


Figure 5.1.1 Overview of the structure of a cohort study with a 5-year follow-up.

5.2 Obtaining the Sample

5.2.1 Introduction

We have seen that the first stage in this study design is to obtain a sample of people to study. You will recall from Chapter 4, Section 4.2.2, that a sample is taken to provide a representative group of subjects, and this group must be large enough to allow conclusive results. We will not examine the statistical methods for determining sample size in a cohort study until Section 5.6, but we look now at how the sample was chosen, and we identify which population the sample was designed to represent.

Please now read the sections below taken from Paper A entitled 'Subject and Methods' and from Paper C, which provides more detail about how the towns and subjects were selected.

Subjects and Methods

The British Regional Heart Study (BRHS) is a prospective study of cardiovascular disease comprising 7735 men aged 40–59 years selected from the age–sex registers of one group general practice in each of 24 towns in England, Wales and Scotland initially examined in 1978–80. The criteria for selecting the town, the general practice and the subjects as well as the methods of data collection, have been reported (Shaper *et al.*, 1981). Research nurses administered a standard questionnaire that included questions on smoking habits, alcohol intake, physical activity and medical history. Height and weight were measured and body mass index (BMI) defined as height/weight. The classification of smoking habits, alcohol intake, social class and physical activity have been reported (Shaper *et al.*, 1988, 1991). The men were classified according to their current smoking status: never smoked, ex–cigarette smokers and current smokers at four levels (1–19, 20, 21–39 and ≥ 40 cigarettes/day).

In brief, 24 towns were primarily selected from those with populations of 50,000–100,000 (1971 census). They were chosen to represent the full range of cardiovascular mortality and water hardness, and towns in all the major standard regions were included. One general practice in each town was selected after consultation with the District Medical Officer who listed practices apparently fulfilling the required criteria. These included a patient list greater than 7500, an interested group of doctors and a social class distribution in the practice that reflected the social class distribution of the men of that town. All practices on the shortlist were visited and the most suitable group selected. If there was no up-to-date age–sex register, one was established for the practice. From each age–sex register, about 420 men aged 40–59 years were selected at random to produce five-year age groups of equal size. The list of names was reviewed by the doctors in the practice, who were asked to exclude those whom they considered could not participate because of severe mental or physical disability. Close scrutiny of the returned annotated lists reduced the exclusions to approximately 6–10 names per practice.

The remaining subjects were invited to take part in the study by a letter signed by their general practitioner. A response rate of 78 per cent was achieved, and 7735 men – approximately 320 men per town – were examined over a period of two and a half years at the rate of one town per month. The 22 per cent of non-responders comprised men who

1. Did not reply to the invitation and one reminder, but as far as was known lived at the address supplied by the practice.
2. Were not available to attend the examination in the two-week period offered because of work commitments.
3. Refused without reason.

The sampling procedure was designed to provide a representative sample, but women were not included.



Self-Assessment Exercise 5.2.1

1. What population was this study designed to investigate?
2. Why do you think the research team restricted their investigation to men?
3. What sampling frame was used?

4. From the information given in the two excerpts you have just read [Paper A (Subjects and Methods; Paper C (Sampling))], including the reported response rate and exclusions, how representative do you think the sample was of the defined population?

Answers in Section 5.10

5.2.2 Sample Size

A total of 7,735 men were included in the study. This may seem like a large number, but in fact it is not particularly large for a cohort study. We have to bear in mind that in this type of study, it is necessary to wait for cases of disease to occur in the original sample (Figure 5.1.1), and unless the sample is large, it would be necessary to wait a very long time for enough cases to occur, even with a relatively common condition such as cancer. In the BRHS, the size of the sample would have been calculated to provide adequate numbers of cases within a reasonably manageable follow-up period (although as noted, the study was designed originally to study CVD rather than cancer outcomes). In Section 5.6, we look at the importance of sample size calculation and how to calculate this for a cohort study.

Summary: Sampling

- The aim of sampling is to obtain a representative group of subjects suitable for addressing the study research question(s).
- The sample also needs to be of sufficient size to meet the objectives (calculation of an appropriate sample size is addressed later), but it must not be so large as to be impractical or too costly for the budget available. Sometimes a pragmatic balance among these various demands needs to be struck.
- For the BRHS, sampling through general practice had the important advantage of preparing the system that was to be used for the follow-up. This is studied in Section 5.4.

5.3 Measurement

5.3.1 Importance of Good Measurement

One of the other very important issues we identified in Chapter 4, Section 4.6 was the importance of good measurement and of avoiding the bias that can arise with poor questionnaire design and inconsistent administration of interviews, for example. A wide range of measurements were made in the BRHS, and considerable care was taken to avoid bias during the survey and follow-up periods.

5.3.2 Identifying and Avoiding Measurement Error

From reading the subjects and methods section of Paper A, you noted that a number of characteristics of interest were measured. Some data were collected by means of a nurse-administered questionnaire and some by physical examination. When designing the study, the authors needed to give considerable thought to ensuring that each of the different variables was measured accurately (precisely) to reduce random error and bias. There is more to ensuring accurate measurement than you might think. For example, let's consider the measurement of blood pressure, a routine measurement taken in many epidemiological studies.

5.3.3 The Measurement of Blood Pressure

Most, if not all, of us have had our blood pressure measured, and no doubt it seems a straightforward procedure. If you break down the various components of blood pressure measurement, however, you can appreciate that it is actually rather complex and subject to all kinds of error, which has implications for both clinical and research settings.



Self-Assessment Exercise 5.3.1

1. Think about how blood pressure is measured, and make a list of the equipment used and activities that go on during this procedure.
2. Against each item in your list, write down all the sources of error that you think could arise during the process of measurement.
3. Against each source of error you have identified, suggest ways the error could be minimised.

Answers in Section 5.10

This exercise emphasises how complex and open to error an apparently simple measurement procedure can be. This example has identified issues that apply, to a greater or lesser extent, to any type of measurement in research or in clinical practice. These sources of measurement error, their nature, and the means of addressing them can be summarized as shown in Table 5.3.1.

Table 5.3.1 Types of measurement error and bias.

Type of error	Nature	Comment
Observer error	Systematic differences between observers	Systematic observer error (bias) can best be avoided by careful standardisation of techniques through training personnel and checking performance. In addition, for some measures such as blood pressure, special measurement equipment may help. At the time of the BRHS, machines that hid the readings from the operator until the measurement was completed were common in research work. Since then, automated devices have become available that help to reduce observer/operator effects.
	Variation within observers, e.g. from day to day; this variation is mainly random	An individual observer's performance varies from day to day. Because there is usually no consistent pattern with this type of error, it is mostly random and just adds noise to the data, effectively increasing the standard error. We will return to this issue.
Instrument* error	Systematic differences between instruments	These systematic differences may result from design factors, faults, poor maintenance, etc. Wherever possible, such differences should be detected and removed or reduced.
	Variation in how the instrument performs from day to day; may be random or systematic	This variation applies principally to reliability of equipment, since for questionnaires this variation arises from the observer (interviewer). Equipment unreliability may be random, but watch out also for systematic effects developing over time (drift). Random variation increases the standard error (as with the within-observer error), but systematic drift results in bias that increases over time.

*The term *instrument* refers to the tool used to make a measurement, and it can include a machine, self-administered questionnaire, interview schedule, etc.

In fact, for the BRHS, even these precautions could not overcome the problem completely. The research team felt that the residual **observer bias** they were able to detect was large enough to warrant adjustment of all the blood pressure data before they were used in analysis. We will look at what is meant by adjustment in Section 5.8.

In the physical activity and cancer study, the researchers developed an instrument (a set of questions) to measure levels of physical activity among participants (see page 1311 of Paper A). A total physical activity score was calculated for each participant based on frequency and intensity of physical activity reported, and a physical activity index with six levels (an ordered categorical variable) was created that ranged from ‘inactive’ to ‘vigorous exercise’. Such measurement of physical activity is open to error (including bias) in a number of ways. The researchers therefore compared the use of the index, using heart rate and forced expiratory volume (a measure of lung function) measurements as a means of validation.

5.3.4 Case Definition

The set of criteria used to determine what is, and what is not, a case of cancer is known as **case definition**. Identifying cases is known as **case finding**. Overall, this process is termed **case ascertainment**. Please now read the following section from Paper A, entitled ‘Ascertainment of cancer cases.’

Ascertainment of Cancer Cases

Cancer cases up to December 1997 were ascertained by means of (1) death certificates with malignant neoplasms identified as the underlying cause of the death (ICD140–209); (2) Cancer registry: subjects with cancer identified by record linkage between the BRHS cohort and the National Health Service Central Register (NHSCR); and (3) postal questionnaires to surviving members in 1992, 1996 and in 1998. In each survey the men were asked whether a doctor had ever diagnosed cancer and if so, the site and year of diagnosis. Smoking related cancers were regarded as cancers of the lip, tongue, oral cavity and larynx (ICD codes 140, 141, 143–149), oesophagus (ICD 150), pancreas (ICD157), respiratory tract (ICD 160–163), bladder (ICD 188) and kidney (ICD 189).



Self-Assessment Exercise 5.3.2

How effective do you think the case ascertainment was in this study? If you are unsure of the processes of death certification and the generation of mortality statistics, refer back to Chapter 2, Section 2.1.

Answers in Section 5.10

This exercise shows the care that needs to be taken in defining, and then finding, cases. In working through these questions, you have seen that for the cancer study, information was required from a number of sources: the official death registration system, the cancer registration system, and direct contact with study participants (by questionnaire).

In the next section, we look further at the methods commonly used in cohort studies for following up participants and finding cases, while also seeking to meet the criteria demanded by the case definition. Having precise and comprehensive case definitions is clearly important, but if the follow-up system can only find 50 per cent of the people who have developed the

outcomes (e.g. cancer), the results may be very misleading. In other words, good case finding is just as important as good case definition in minimising the possibility of bias.

Summary: Measurement

- Good measurement is vital, because inadequate measurement can lead to bias and/or an increase in random error.
- Measurement error can arise from a variety of sources, including the instrument (equipment, questionnaire, etc.) and the observer (the person making the measurements).
- Measurement error can be reduced in a number of ways, including
 - observer training and calibration against standards,
 - clear, unambiguous case definition,
 - careful design and testing of questionnaires,
 - good maintenance and regular calibration of equipment,
 - special measurement equipment.
- These precautions should reduce measurement error, but they rarely remove it altogether.

5.4 Follow-Up

5.4.1 Nature of the Task

Think about the problem for a moment. The research team has identified and studied a group of almost 8,000 men who live in 24 towns throughout the country. These men responded to an invitation to come into a survey centre in their town, and then they went home and got on with their lives. Somehow the team had to follow up the men over the next 15 to 20 years to identify those who developed some form of cancer. The men were all registered with a general practitioner (GP) at the time of the baseline survey, but of course some might move to other parts of the country and change GP, and a few might emigrate to other countries. A cohort study depends very much on the ability to carry out such a demanding logistical exercise, and this is why these studies are complex, time-consuming, and expensive. The BRHS provides a good illustration of the complexities of the follow-up procedure in cohort studies and of how these can be successfully addressed.

The aim of good *case ascertainment* is to ensure that the process of finding cases, whether deaths, illness episodes, or people with a characteristic (e.g. smoker vs. non-smoker), is as complete as possible. In a number of countries (including the UK), there are systems that make tracing deaths relatively straightforward. We begin by looking at the follow-up of deaths (mortality), and then move on to the more difficult task of finding the non-fatal cases (morbidity).

5.4.2 Deaths (Mortality)

The system used for tracing deaths among the study sample relied on records held by the National Health Service (NHS) Central Register (currently part of the Health and Social Care Information Centre). Following certification, every death is notified to the NHS register, so this provides a convenient and very complete means of finding out about the deaths and obtaining information on the cause of death. When carrying out a study, it is possible to flag individuals who have been recruited to the study so that if a death occurs among one of the study participants, the research team is notified.

5.4.3 Non-Fatal Cases (Morbidity)

Finding the non-fatal cases is not so simple. Fortunately, in the cancer study, the researchers were able to utilise a national database of cancer morbidity, the Cancer Registry. Even so, the team also used questionnaires to contact participants and find out whether any cases had slipped through the net of cancer registration.

For many diseases, there is no central register in the same way as for deaths or cancer, so the only way to find these events is to pick them up through the general practices and hospitals concerned and by asking the subjects themselves. For example, in the original arm of the BRHS, the researchers needed to identify all cases of non-fatal ischaemic heart disease. In that study, co-ordinators were identified from each of the group practices with responsibility for mailing and updating morbidity reports, notifying the study centre of all deaths and address changes, and reviewing the medical records at set intervals to ensure that no cases had been missed. Even a common occurrence such as change of address requires a well-organised system to prevent study subjects from being lost to follow-up. The methods used by the BRHS team relied on the systems used for administration of general practice and on the transfer of records when a person leaves one practice and later registers at another one. These procedures are described in Paper C, and they are worth reviewing because this method clearly demonstrates the attention to detail required to achieve effective follow-up in cohort studies.

5.4.4 Challenges Faced with Follow-Up of a Cohort in a Different Setting

Obtaining information on morbidity was demanding in the BRHS. It can be even harder and more time-consuming in cohort studies carried out in countries without systems in place that at least help a bit with this task. The excerpt below from Paper D describes the follow-up methods used in a cohort study on the determinants of physical activity in adolescents in Pelotas, a city in southern Brazil.

Methods

Pelotas (population 320 000) is located in southern Brazil in a relatively developed part of the country. In 1993, mothers of all hospital-born children were invited to join a birth cohort study. Home births account for less than 1% of all deliveries. Mothers were interviewed soon after delivery for personal, socioeconomic, and behavioural variables.

Follow-Up Visits

The cohort has been followed on several occasions. In the present analysis we use data from four follow-up visits.

One Year and Four Years

At follow-up visits at one and four years, all low birthweight (<2500 g) children ($n = 510$) and a systematic sample of 20% of the remainder were sought; 1363 children were seen at one year and 1273 at four years. Analyses were weighted to compensate for the over-sampling of low birthweight children. Weight gains (kg) from birth to 1 year, 1–4 years, and 4–11 years were categorized into quartiles. Overweight at 1 and 4 years was defined as weight for height Z scores greater than 2 according to the reference standard of the National Center for Health Statistics.

Behavioural Sub-Study at Four Years

A randomly selected subsample of 634 children followed up at four years was visited. At this visit the mother completed the child behaviour checklist questionnaire. We used two variables on the basis of the mother's self-report in the present paper: the child's level of physical activity compared with children of the same age (below average, on average, above average) and how well the child performed at sports activities (below average, on average, above average).

10–12 Years

In 2004–5 we sought all cohort members through a school census, as well as a population census in which about 100 000 households in the urban area were visited in search of adolescents born in 1993. Detailed data were collected on physical activity, including mode of transportation to and from school, physical education classes, and leisure time activities



Self-Assessment Exercise 5.4.1

1. Who was recruited to join the study?
2. How many follow-up visits were used for the analysis reported in this paper?
3. How were children and their parents contacted at 10 to 12 years?
4. Table 1 of Paper D (Table 5.4.1) reports that 87.5% of 5249 children were followed up at 10 to 12 years, which includes 141 deaths; the table also provides a breakdown of the percentage of children followed up, according to categories of a number of sociodemographic and anthropometric variables. Review this information and describe, with reasons, how adequate you feel the follow-up was.

Table 5.4.1 Comparison between those followed up at 10–12 years and original cohort in terms of sociodemographic and anthropometric variables. (Table 1 taken from Paper D (Hallal *et al.* 2006).)

Variable	No in original cohort (n = 5247)*	% located†	P value (χ^2 test)
Sex:			
Boys	2580	86.9	0.18
Girls	2667	88.1	
Family income (No of minimum wages per month):			
≤1	967	88.3	<0.001
1.1–3.0	2260	88.7	
3.1–6.0	1204	88.9	
6.1–10.0	433	79.9	
>10.0	385	82.6	
Maternal education at birth (years):			
0	134	82.1	<0.001
1–4	1338	88.7	
5–8	2424	89.9	
≥9	1350	82.5	

(continued)

Table 5.4.1 (Continued)

Variable	No in original cohort (n = 5247)*	% located†	P value (χ^2 test)
Birth weight (g):			
<2500	510	89.8	0.16
2500–3499	3361	86.9	
≥3500	1361	87.9	
Prepregnancy body mass index:			
<20.0	1147	87.6	0.004
20.0–24.9	2811	86.6	
25.0–29.9	894	90.3	
≥30	245	92.2	
Overall	5249	87.5	

*Numbers vary owing to missing values.

†Includes 141 deaths.

Answers in Section 5.10

5.4.5 Assessment of Changes During Follow-Up Period

Returning to the BHRS study, we saw that the cohort study began by selecting a sample and then carrying out a survey to measure all of the social, lifestyle, health, and other baseline characteristics of interest. This survey provides information on levels of **exposure** to risk and possible protective factors that will be used in the analysis, such as whether a person exercises and how much. A man reporting that he exercises moderately at the time of the baseline survey, however, might not continue to exercise at the same rate during the follow-up period. Indeed, he might give up the day after the survey, or, conversely, he might exercise more over the ensuing years. If the research team has no further contact with the study subjects after the initial survey, or if there are no records providing valid and representative information on subsequent changes in risk factors, then it has to be assumed that the information obtained at baseline applies for the entire period of follow-up.

Random changes in risk factors over the follow-up period would simply increase random error in the analysis, resulting in less **precision** but not necessarily **bias**. On the other hand, bias arises if there are undetected differential changes in levels of exposure within the sample. For example, suppose the research team wishes to understand how socioeconomic circumstances influence the impact of physical activity on risk of cancer. If men in higher socioeconomic groups are more likely to increase the amount that they exercise during the follow-up period than men in lower socioeconomic groups, and no further assessment of activity is made after the baseline, this will bias the results. This is because the observed benefits in the higher socioeconomic group will be higher than expected for the level of physical activity they reported at baseline. This is illustrated in the example in Figure 5.4.1.

In this scenario (Figure 5.4.1), the association between exercise and cancer is being studied. Exercise level and frequency are measured at the baseline survey, and a lower prevalence of vigorous exercise is found among subjects in lower socioeconomic groups. However, during the follow-up, a greater proportion of ‘none or occasional exercisers’ in the higher socioeconomic group begin to exercise vigorously. Whereas at baseline, members of lower

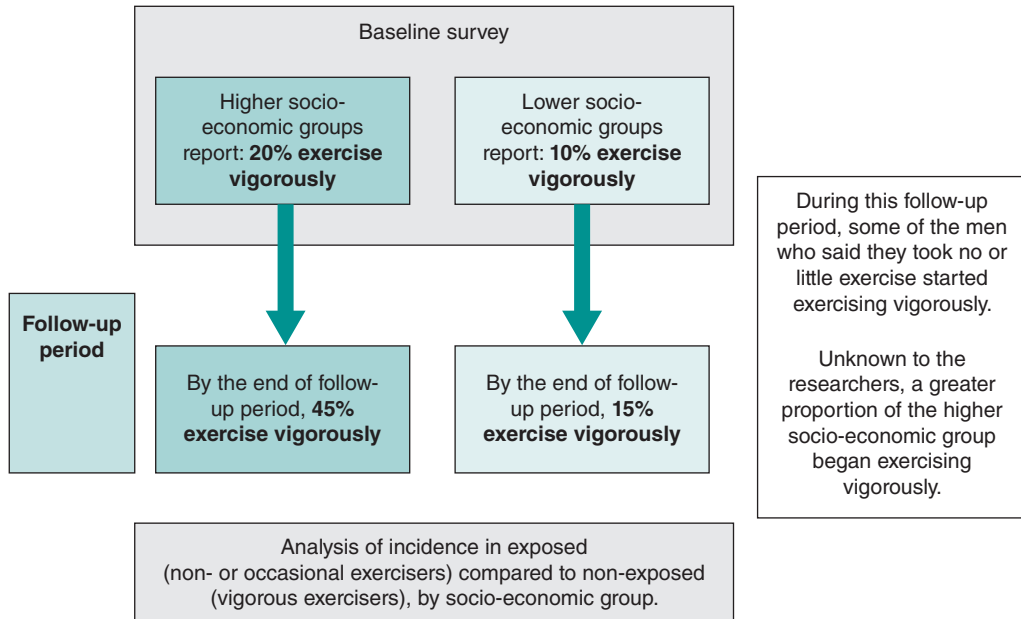


Figure 5.4.1 Hypothetical scenario in a cohort study with undetected changes in exercise behaviour during the follow-up period, which vary by socioeconomic group.

socioeconomic groups were half as likely to exercise vigorously (20 per cent versus 10 per cent), by the end of the study they were three times less likely to be vigorous exercisers (45 per cent versus 15 per cent). If the researchers do not know about this, bias will occur in analysis of the impact of exercise on cancer by socioeconomic group. In this example, the benefit (assuming exercise does reduce the risk of cancer) will be overestimated in the higher socioeconomic group, because the observed reduction in cancer risk will be associated with a lower prevalence of vigorous exercise than is actually the case. Consequently, it is a good idea to carry out repeat measurements of exposure during the course of a cohort study, especially if the follow-up period is quite long.



Self-Assessment Exercise 5.4.2

1. What further exposure information was collected during the follow-up period of the physical activity and risk of cancer study (Paper A)?
2. What might be the impact on the results if those who exercised more vigorously had been more likely to give up smoking during the follow-up period?

Answers in Section 5.10

Summary: Follow-Up

- The follow-up (including case ascertainment) is perhaps the most critical and demanding part of a cohort study.
- Generally, follow-up for mortality is more straightforward than for morbidity, especially where there are systems available such as the flagging of records at the NHS Central Register. Where

such systems are not available, intensive follow-up (such as the household census in the Pelotas study in Paper D) may be needed.

- The percentage of the sample lost to follow-up should be kept to an absolute minimum, and differential loss to follow-up should be noted (for example, if more subjects with a risk factor for the disease are lost from the exposed compared to the unexposed group).
- Changes in the level of exposure to the main factor of interest, or to other important risk factors, after the initial survey and during the follow-up period, may increase random error and/or bias. It is valuable to carry out one or more interim assessments, such as the 5-year questionnaire in the BRHS.

5.5 Basic Presentation and Analysis of Results

5.5.1 Initial Presentation of Findings

We have completed the follow-up component of the study and can now look at the findings of the baseline survey to see how these are related to the subsequent deaths and non-fatal cases that occurred. Referring back to Figure 5.1.1, which summarises the cohort study design, we see that analysis involves calculating how the *incidence rate* varies by level of *exposure* to a *risk factor*. This exposure level is expressed in different ways, according to the type of data:

- As the presence or absence of the risk factor (e.g. whether or not there is a family history of a specific cancer), which are dichotomous categorical variables.
- As categories (which may be ordered), such as never, ex-, or current smoker.
- As an actual numerical value, such as body mass index and age, which are continuous variables.

Incidence of Study Outcome (Cancer)

Table 5.5.1 (Table 2 from Paper A) shows, in column 4, the incidence rates of cancer for each of the different categories of physical activity. You will recall that in order to calculate the incidence, we need a *numerator* (number of cases), a *denominator* (number of people at risk), and

Table 5.5.1 Physical activity and age-adjusted cancer rates/1,000 person-years and adjusted relative risks for all cancers (excluding skin cancer) in 7,588 middle-aged British men.

Physical activity	No.	Cases	Rates/1,000 person-years	Relative risks (95% CI)	
				A	B
None/Occasional	3017	424	8.4	1.00	1.00
Light	1749	234	7.8	0.93 (0.79, 1.09)	0.95 (0.80, 1.11)
Moderate	1196	163	8.1	0.95 (0.80, 1.14)	1.04 (0.86, 1.24)
Moderate-vigorous	1113	101	5.8	0.68 (0.54, 0.84)	0.78 (0.62, 0.97)
Vigorous	513	47	5.7	0.65 (0.44, 0.88)	0.76 (0.56, 1.04)
<i>Test for linear trend</i>				<i>P</i> < 0.0001	<i>P</i> = 0.02

A = age-adjusted.

B = adjusted for age, cigarette smoking, BMI, alcohol intake, and social class.

Source: Wannamethee 2001. Reproduced with permission of Nature Publishing Group.

a *time period*. In this study, the authors have calculated *incidence density*, expressed as a rate per 1,000 person-years (p-y).

The final columns show adjusted relative risks (and 95 per cent CIs). We look at relative risk in Section 5.5.2, and we consider confounding and the methods used for adjustment in Sections 5.7–5.9.



Self-Assessment Exercise 5.5.1

Use the incidence density rates from Table 5.5.1 (Table 2 from Paper A), column 4, to answer the following questions:

1. What is the incidence in the 'none/occasional' group?
2. What is the incidence in the 'vigorous' group?
3. How much lower is the incidence in the 'vigorous' group than in the 'none/occasional' group? What does this tell you?

Answers in Section 5.10

5.5.2 Relative Risk

The Concept of Relative Risk

We now look at how we define and calculate the risk of disease associated with a given factor such as smoking, cholesterol, or physical activity. First, here are some everyday examples of risk:

'I have smoked 20 cigarettes a day for the last 20 years. What are the chances of my getting lung cancer compared to my friend who never smoked?'

'If I wear a helmet when I go cycling, how much does this reduce my chance of a serious head injury if I am knocked off my bike in traffic?'

Relative risk describes how the risk of a disease varies according to the level of exposure to a risk factor, such as smoking relative to not smoking or wearing a cycle helmet relative to not wearing one at the time of an accident.

This idea is the same as that introduced in Exercise 5.5.1 when we looked at the way the incidence of cancer varied among those who took no exercise or only occasional exercise compared to those who exercised vigorously. Indeed, it is the comparison of *incidence rates* that provides us with the measure known as *relative risk (RR)*:

$$\text{Relative risk} = \frac{\text{Incidence of disease in } \textit{exposed} \text{ group}}{\text{Incidence of disease in } \textit{unexposed} \text{ group}}$$

The term *relative risk* is synonymous with *risk ratio*, which you may also see used for this purpose.

Often, relative risk is calculated as a comparison of rates in more exposed and less exposed, rather than the absolutes of exposed and unexposed. Returning to our data in Exercise 5.5.1, the relative risk for those who exercised vigorously compared with those who took no or only occasional exercise was the ratio of the two incidence rates, which we calculate as $5.7 \div 8.4 = 0.68$, or 68 per cent. A relative risk of 1.0 (or 100 per cent as a percentage) would mean that the risk of cancer among the two groups was the same. A relative risk of less than 1.0 implies the exposure has a protective effect in relation to the outcome. In other words, the relative risk of

cancer among vigorous exercisers is 32 per cent less or 0.68 times the risk of those who take no or only occasional exercise.

You will see that the relative risk of 0.65 for vigorous versus none/occasional exercise quoted in Table 5.5.1 is slightly different from the value of 0.68 that we have just calculated. This is because these have been adjusted for the confounding factors, listed at the bottom of the table, using regression methods. These methods are introduced in Sections 5.8 and 5.9.

In the next exercise, we look at some more relative risks using data on cardiovascular disease from the BRHS. For this investigation, the research team was interested in the relative risk of heart disease among smokers compared to non-smokers, data for which are shown in Table 5.5.2 (from Paper B).

Table 5.5.2 Incidence of ischaemic heart disease (IHD) among non-smokers, ex-smokers, and current smokers, after an average of 4.2 years of follow-up.

Cigarette smoking	Cases	Total men	Percentage	Rate/1000/year
Never	18	1819	0.99	2.36
Ex-smoker	76	2715	2.80	6.66
Current	108	3185	3.39	8.07

Source: Shaper *et al.*, 1985 (Paper B).



Self-Assessment Exercise 5.5.2

1. Calculate the relative risks of ischaemic heart disease (IHD) for current smokers and ex-smokers compared to never-smokers.
2. Given that we have taken the never-smokers as the **reference group**, have a go at interpreting the values of relative risk you have just calculated for the ex-smokers and the value for the current smokers.
3. One explanation for smokers having a higher incidence of IHD than never-smokers is that smoking causes IHD. Can you think of any other possible explanations for this finding?

Answers in Section 5.10

Summary: Relative Risk

- Another term for relative risk is 'risk ratio'.
- With relative risk, we can express the risk of an outcome (e.g. IHD or cancer) for men with a level of a risk factor (e.g. current smoker or vigorous exercise), relative to men with another level (e.g. never smoked or no/occasional exercise).
- We can place a value on this risk; that is, we can quantify the risk.
- If an exposure category has a relative risk greater than 1.0, it means that the people in that category have a higher risk of the outcome than the people in the reference category.
- If the relative risk is less than 1.0, the people in that exposure category have a lower risk of the outcome than the reference category.
- If the relative risk is 1.0, the people in that exposure category have the same risk as the people in the reference category.



Self-Assessment Exercise 5.5.3

Look again at Table 5.5.1 (Table 2 reproduced from Paper A) (Section 5.1.1).

1. What group was taken as the *reference group* for calculating relative risk?
2. Why was this group appropriate to use as a reference?

Answers in Section 5.10

5.5.3 Hypothesis Test for Categorical Data: The Chi-Squared Test

Real Association or a Chance Finding?

It appears from the cancer study that exercise is associated with a decreased risk of cancer in this sample of men. From Exercise 5.5.2, we found that the relative risk for ‘none or occasional exercisers’ compared with those who exercise vigorously was 0.68 (before adjustment for confounding factors). This is lower than 1.0, the relative risk if the risk of cancer is the same for both groups. In fact, the risk of cancer among vigorous exercisers is 32 per cent less than the risk of cancer among people who take no or only occasional exercise.

But does the observed association in this sample of men mean that cancer and exercise are associated in the population of all middle-aged British men, which is the question the research set out to answer? If we studied another sample from the population, we would almost certainly obtain a different value for the relative risk (though it may not differ by much). Any of the many samples that we could have taken will provide an estimate of the relative risk in the population, and the estimate will vary from sample to sample: this is the *sampling distribution* we introduced in Chapter 4 and now apply to sample estimates of relative risk.

To address the research question about exercise and cancer in British men, we need an objective way of deciding whether the association between cancer and exercise that we have observed in this particular sample is evidence of a real association in the population of middle-aged British men. This is the process of *inference* that was referred to in Chapter 1. This objective assessment of the evidence provided by the sample is called an *hypothesis test*. We now look at *hypothesis testing* and carry out a test to investigate whether cancer incidence and level of physical activity are statistically significantly associated; that is, the association we observe in this study sample is very unlikely to be the result of chance arising from sampling error, and we can therefore conclude that the sample findings do represent an association between exercise and cancer in the population.

Hypothesising About the Population

We have said that we want to use these data from the sample to answer the question, ‘Is there an association between exercise and cancer in middle-aged British men?’ We do this by analysing the sample data to see whether an assumption of no association between exercise and cancer is reasonable. If this assumption turns out to be unreasonable on the basis of the information we have from the sample, we will conclude that there is evidence of an association between exercise and cancer for middle-aged British men. Assumptions about populations are called *hypotheses*, and so this analysis is an *hypothesis test*.

The assumption of no association is called the *null hypothesis*, abbreviated to H_0 , and the alternative, that there is an association, is called the *alternative hypothesis*, H_1 . These must be stated precisely, so that we know exactly what we can conclude after we have tested the null

hypothesis. The hypotheses we are interested in may be written as follows (RR is relative risk and the symbol \neq means 'not equal to'):

H_0 : There is no association between exercise and cancer for middle-aged British men, and any observed association has arisen by chance ($RR = 1$).

H_1 : There is an association between exercise and cancer for middle-aged British men ($RR \neq 1$).

The hypotheses must clearly state the variables between which there may be an association, and the population about which we are hypothesising.



Self-Assessment Exercise 5.5.4

For H_0 and H_1 as given above,

1. State the (outcome and exposure) variables we are investigating.
2. State the population about which we are hypothesising.

Answers in Section 5.10

If the null hypothesis is unreasonable – that is, the data are not in agreement with such an assumption – we reject the null hypothesis in favour of the alternative hypothesis. We then say there is evidence that exercise and cancer are associated for middle-aged British men.

We can assess the strength of evidence against the null hypothesis of no association (H_0) provided by the data because we can work out how likely the observed frequencies are to have arisen if there is no association. The measurement of 'how likely' is known as probability. The usual convention is that if the probability of the observed frequencies arising, assuming no association, is less than 0.05 (5 per cent, or 1 in 20), then it is sufficiently unlikely that H_0 is true. Therefore, we can conclude that there is an association.

We are now ready to apply this notion of hypothesis testing to the exercise and cancer data. The appropriate hypothesis test for these data is called the *chi-squared test*.

First Step for the Chi-Squared Test: Contingency Tables

To start our explanation of the chi-squared test, we will take another look at the risk of cancer for those who never or occasionally exercise and those who exercise vigorously, building on what you have already done in previous exercises. The relevant data from Table 5.5.1 (Table 2 from Paper A) are given in Table 5.5.3.

Table 5.5.3 Physical exercise and cancer (data from Table 5.5.1, Paper A).

Exercise category	Cancer				Total
	Cases		Other men		
	No.	%	No.	%	
None/occasional	424		2,593		3,017
Vigorous	47		466		513
Total	471		3,059		3,530

Source: Wannamethee 2001. Reproduced with permission of Nature Publishing Group.



Self-Assessment Exercise 5.5.5

1. Table 5.5.3 is defined by the two variables *cancer* and *exercise*. What values can each of these two variables take? Are they categorical, ordered categorical, discrete, or continuous variables?
2. Insert the percentage of 'none/occasional' exercisers who became cases and the percentage who did not become cases in Table 5.5.3.
3. Do the same for vigorous exercisers and for everyone in the sample (the Total row).
4. Interpret the findings.

Answers in Section 5.10

Table 5.5.3 is defined by two categorical variables, also called **factors**, and each of the four possible combinations of the variables is called a **cell**. The figure in each cell is a frequency, or count. A cross-tabulation of frequencies such as this is called a **contingency table**.

Testing the Null Hypothesis using the Chi-Squared Test

We now work through the hypothesis test using the data in Table 5.5.3, and then we summarise the procedure. Although hypothesis tests are done routinely as part of data analysis by computer, it is worth going through the calculation of this test step by step (and later in this chapter, the other most commonly used test – the *t*-test), because this will help your understanding of the application of the test and the assumptions used.

The first step is to find the frequencies we would expect if the null hypothesis (no association) were true. These are calculated for each of the four cells by multiplying together the totals of the row and column the cell is in and dividing by the total number in the sample.

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{total sample size}}$$

For example, taking the top left-hand cell (cases who take none/occasional exercise), if there really is no association between exercise and cancer, then of the sample of 3,530 men we would expect 402.55 men to be cases of cancer in that cell. The calculation of this expected number is shown below:

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{total sample size}} = \frac{3017 \times 471}{3530} = 402.55$$

Of course, it is not possible to have 402.55 cases among men taking no or occasional exercise. By the 'expected' frequency, we mean that if we were to observe lots of different samples of 3,530 men, then the average number of men in each sample who took none/occasional exercise and were cases would be 402.55 if there is no association between exercise and cancer. This calculation of expected frequency is based on the fact that if there is no association between exercise and cancer, we would expect the same proportions of none/occasional exercisers and vigorous exercisers to develop cancer.



Self-Assessment Exercise 5.5.6

Calculate the expected frequency for the other three cells of Table 5.5.3 and complete Table 5.5.4, including the column and row totals, and the overall total.

Table 5.5.4 Physical exercise and cancer: expected frequencies.

Exercise category	Cancer		Total
	Cases	Other men	
None/occasional	402.55		
Vigorous			
Total			

Answers in Section 5.10

Note that the expected frequencies add up to the same totals as the observed frequencies. This is always true (apart from small rounding errors), so this can be used to check whether the expected frequencies are correct.

The next step is to calculate the differences between the observed and expected frequencies in each cell. These are called the *residuals*. For cases of cancer among men who took no or occasional exercise (top left cell), the residual is

$$424 - 402.55 = 21.45$$

The residual is always calculated this way round; that is, the observed frequency minus the expected frequency.



Self-Assessment Exercise 5.5.7

Complete the table of residuals (Table 5.5.5).

Table 5.5.5 Exercise and cancer: residuals.

Exercise category	Cancer		Total
	Cases	Other men	
None/occasional	21.45		
Vigorous			
Total			

Answers in Section 5.10

The row and column totals of the residuals should always be zero, apart from small rounding errors. The sign (+/−) of a residual shows whether the observed frequency is larger or smaller than expected, and the size of the residual shows how large the difference is. We now combine

these residuals to obtain an overall measure of the difference between what we have observed and what we expect under the null hypothesis of no association. The measure we use is the sum of the squared residuals each divided by the corresponding expected frequency:

$$\sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \sum \frac{\text{residual}^2}{\text{expected frequency}}$$

This expression is generally written as:

$$\sum \frac{(O - E)^2}{E}$$

This is called the **chi-squared statistic** (*chi* is pronounced 'ky'). Thus, in our example, for cases of cancer who took none/occasional exercise, the first value contributing to the chi-squared statistic is

$$\frac{(21.45)^2}{402.55} = 1.14$$

It is convenient to put these values in another table, called the chi-squared table, which is illustrated after the following brief reference section on the chi-squared statistic.



RS – Reference Section on Statistical Methods

This statistic is called the chi-squared statistic because it can be proved that all the possible values it can have form a distribution with a particular shape, called the chi-squared distribution. This is often abbreviated to the χ^2 distribution, using the Greek letter *chi*. This is a theoretical probability distribution, like the normal distribution, the shape of which can be described by a known formula. The features and applications of the most common probability distributions, including the chi-squared distribution, are further described in Chapter 11.

In this final exercise on the chi-squared hypothesis test, we calculate the statistic and then interpret the result.



Self-Assessment Exercise 5.5.8

Complete the chi-squared table (Table 5.5.6) and calculate the value of the chi-squared statistic.

Table 5.5.6 Cancer and exercise: chi-squared table.

Exercise category	Cancer		Total
	Cases	Other men	
None/occasional	1.14		
Vigorous			
Total			

Answers in Section 5.10

You should have obtained a value of the chi-squared statistic for these data of 9.08. What does this mean? This value is a measure of how the observed frequencies differ from the frequencies we would expect to occur if the null hypothesis, that there is no association between cancer and exercise, were true.

If the observed and expected frequencies are similar (that is, the data are consistent with the null hypothesis), then the differences ‘observed frequency – expected frequency’ will be small, and the value of the chi-squared statistic will be small. Conversely, if the observed frequencies are very different from the expected frequencies, either smaller or larger (that is, the data are not consistent with the null hypothesis), then the value of the chi-squared statistic will be large. So a large value is evidence against the null hypothesis.

If the value is large enough, it is *very unlikely* that we could have obtained the observed frequencies if H_0 is true, and we reject H_0 and conclude that cancer and exercise are associated. We will now look at how we decide whether this value of 9.08 for the chi-squared statistic is large enough to reject the null hypothesis in this case.

Interpreting the Test Result: Degrees of Freedom and the p -value

To decide whether the chi-squared value is large enough to reject H_0 , we use computer software or published tables to find the probability of obtaining such a value, or a larger one, when H_0 is true. This probability is called the **p -value**. As already noted, we generally consider a probability of less than 0.05 to be small enough for the observed data to be so unlikely that H_0 must be untrue.

If we are looking our p -value up on a table of critical values of chi-squared (see below), we first need to calculate **degrees of freedom**. As noted above, the chi-squared distribution is a probability distribution (see Chapter 11, Section 11.1.3) and can have different shapes, depending on the size (how many rows and columns) of the contingency table. The quantity that defines the size of a contingency table is called the degrees of freedom (df), and each value of degrees of freedom defines a different chi-squared distribution. The degrees of freedom are calculated by multiplying the number of rows minus one by the number of columns minus one.

‘Degrees of Freedom’ of a Contingency Table

If a contingency table has r rows and c columns, the degrees of freedom are

$$df = (r - 1)(c - 1)$$

Such a table is called an $r \times c$ contingency table.

In our case, we have only two rows (‘None/occasional’ and ‘Vigorous’ exercise) and two columns (‘Cancer case’ and ‘Other’), so the degrees of freedom are $(2 - 1) \times (2 - 1) = 1$. We then look up our p -value on a table of critical values of the chi-squared statistic under one degree of freedom. Whilst computer software will automatically calculate this for us, it is useful to refer to the tables because this helps understand how critical values of the chi-squared statistic vary with degrees of freedom and the level of probability. An excerpt of the table, with values of degrees of freedom up to 10, is shown in Table 5.5.7; the relevant value for $df = 1$ and a probability of 0.05 are shaded.

The level of probability below which we consider that the sample provides sufficient evidence to reject the null hypothesis is called the **significance level** of the test. It is usually written as a percentage, so if we choose a probability of 0.05, we say the hypothesis test is carried out at the 5 per cent significance level. If the null hypothesis is rejected, we may say that H_0 is rejected at

Table 5.5.7 Upper-tail critical values of chi-square distribution with ν degrees of freedom.

ν	Area to the right of the critical value (p -value)				
	0.10	0.05	0.025	0.001	0.0001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588

the 5 per cent significance level, or that the value of the chi-squared statistic obtained is significant at the 5 per cent level. An alternative name for a hypothesis test is a **test of significance**.

The required significance level should always be decided before carrying out an hypothesis test. The value of 5 per cent is very commonly used but is arbitrary. In different situations, we may require either more evidence (a lower significance level, e.g. 1 per cent) or less evidence (a higher significance level, e.g. 10 per cent, although this is uncommon) in order to reject the null hypothesis.

Result for the Chi-Squared with One Degree of Freedom

As shown in Table 5.5.7, the critical value of the chi-squared statistic associated with a 0.05 probability with one degree of freedom ($df = 1$) is 3.841. Our chi-squared statistic was 9.08, and there is a probability of only 0.003 (three in a thousand) of obtaining a chi-square value as large as this on one degree of freedom, or a larger value, if H_0 is true. This is a probability much smaller than 0.05. If we refer back to Table 5.5.7, our chi-squared value of 9.08 lies between the critical values for 0.99 and 0.999, so $0.0001 < p < 0.001$; this is consistent with the actual p -value calculated on the computer of 0.003.

Consequently, we can conclude that it is very, very unlikely that we would observe the frequencies in Table 5.5.3 if the null hypothesis of no association between exercise and cancer were true, but it is not quite impossible: the p -value is greater than 0! We can therefore conclude that the null hypothesis is unreasonable and reject it, and we can state that there is evidence that exercise and cancer are associated in middle-aged British men.

How Should We Interpret a Non-Significant Result?

If the chi-squared statistic had turned out to be smaller than 3.84 on one degree of freedom (and the p -value consequently greater than 0.05), then we could not reject the null hypothesis at the 5 per cent significance level. However, neither would we accept the null hypothesis and conclude that there is no association between exercise and cancer. This is because we can never know whether the null hypothesis is true unless we observe everyone in the population. Our conclusions in this case would be that we do not reject H_0 at the 5 per cent significance level and that there is insufficient evidence of an association. One explanation for our not finding

evidence of an association is because one does not actually exist in the population. It is also possible that an association does exist in the population, but perhaps our sample was not large enough to detect it. This introduces the concept of power, which is discussed further in Section 5.6 on sample size for a cohort study.

Test Assumptions and Summary

The chi-squared test is used to determine whether or not, on the basis of information in a *random sample*, there is evidence of an association between two categorical variables in the *population* from which the *sample* is drawn. Use of the test is subject to a set of assumptions, which are summarised below.

Summary: The Chi-Squared Test

1. The chi-squared test is based on statistical theory that requires certain assumptions to be fulfilled for the test to be valid. These assumptions are as follows:
 - (a) The two variables must be categorical.
 - (b) The groups defined by the variables must be independent. This means a person can be in only one group – for example, a person takes ‘none/occasional exercise’ or takes ‘vigorous exercise’ – but cannot be in both categories.
 - (c) The sample must be large enough. The usual rule for determining whether the sample is large enough is that 80 per cent of the expected frequencies in the contingency table must be greater than 5 and all the expected frequencies must be greater than 1.
2. It does not matter which variable defines the rows of the contingency table and which defines the columns.
3. The hypotheses and conclusions are about the population from which the sample is taken.
4. Although we used an example of a 2×2 table, the chi-squared test can be applied to any size of contingency table, that is, one with more than two rows and columns. The method for calculation is exactly the same. Also, with larger tables, it is more likely that the conditions described in 1(c) for a large enough sample are not fulfilled. If there are too many small expected frequencies, variable categories (that is, rows and/or columns) can be combined to ensure the conditions are satisfied.
5. If calculating by hand, it is usually sufficient to calculate expected frequencies and chi-squared values to two decimal places.

Interpreting the Chi-Squared Test

Although the chi-squared test uses sample information to determine whether or not there is evidence of an association between two categorical variables in the population, it does not tell us the direction of any association. Thus, in our example, the test result (Section 5.5.3) does not tell us whether men who exercise are more or less likely to develop cancer than those who never or only occasionally exercise, but only that there is evidence of an association. Nor does the chi-squared test tell us how strongly the variables are related: The chi-squared statistic is not a measure of the strength of an association.

To determine the direction and strength of the association, we need to return to the observed frequencies and percentages to describe how they are related. In Exercise 5.5.4 we found that the percentage of none/occasional exercisers who developed cancer was larger than the percentage of vigorous exercisers who became cases. Now with the result of the chi-squared test, we can say that this is evidence that vigorous exercise is associated with a lower risk of cancer in the population of middle-aged British men. Note, we are not yet saying that vigorous exercise actually prevents cancer, because this would assume that the relationship is causal. We

discuss the various factors that can help us decide whether or not this association is causal in Section 5.7 by reference to the Hill viewpoints.

We can quantify the difference in risk by looking at the difference between the percentages of men developing cancer in the vigorous and none/occasional exercise categories, and we can give a range of likely values of this difference in the population by calculating a 95 per cent confidence interval (95% CI) for the difference. In general, this is how the results of an analysis of a 2×2 contingency table are summarised. However, in the particular case of a cohort study with a categorical outcome (cancer: yes/no), the primary aim is usually to estimate the relative risk, so it is more appropriate to summarise the results in terms of relative risk (Section 5.5.2), rather than as percentages. The relative risk is a measure of the strength of association.

Summary: Categorical Data and Hypothesis Testing

- A contingency table is a table of frequencies, defined by two categorical variables.
- An hypothesis is a statement about the population of interest.
- We use an hypothesis test to decide whether the information we have from a sample is in agreement with an hypothesis about the population.
- A chi-squared test is used to test the null hypothesis of no association between two factors of interest defining a contingency table.
- The appropriate value of the chi-squared statistic used for this test is determined by the number of degrees of freedom (df), which is calculated as $(r - 1) \times (c - 1)$, where r = number of rows and c = number of columns in the contingency table.
- The probability of obtaining the observed test result (or a more extreme one) is the p -value of the test.
- If the p -value is small (e.g. less than 0.05), we reject the null hypothesis.
- The value below which the p -value must be in order to reject H_0 is the significance level of the test.
- A chi-squared test does not tell us the strength or direction of any association between factors; it is necessary to refer back to the sample data for this information.

Testing for Significant Effects with Continuous Data

Having worked through the application of the most common test for categorical data, we now look at hypothesis testing for continuous data. It will be reassuring to know that the underlying principles are essentially the same as those we have just introduced for categorical data, although we do have to use a different type of test. This test, which you probably have come across in its most familiar form, the *t-test*, is based on the ideas we covered in Chapter 4, Section 4.4, on the standard error of sample means.

5.5.4 Hypothesis Tests for Continuous Data: The z-Test and the t-Test

Although a number of continuous variables were investigated in the physical activity and cancer study, such as age and body mass, the values of these variables were not reported in Paper A. We therefore refer to the original BRHS report on heart disease to provide an example of carrying out an hypothesis test on continuous data (Paper B).

During that study a number of potential risk factors for ischaemic heart disease (IHD) were investigated, including cholesterol levels, blood pressure, smoking, and body mass index. As reported in Paper B, after a mean follow-up of 4.2 years, 202 cases of IHD had been identified. We now look at whether there was a statistically significant difference in systolic blood pressure (SBP) among men who experienced IHD and men who did not.

Table 5.5.8 Means of risk factors in cases and other men (continuous variables) (Paper B).

Risk factor	Cases ($n = 202$)	Other men ($n = 7533$)	t -value
Age (years)	52.8	50.2	7.1
Systolic blood pressure (mmHg)	155.4	144.9	6.1
Diastolic blood pressure (mmHg)	87.9	82.1	5.1
Body mass index kg/m ²	26.44	25.46	4.1
Total cholesterol (mmol/l)	6.78	6.29	6.0
Triglyceride* (mmol/l)	2.10	1.73	4.1
HDL-C (mmol/l)	1.08	1.15	-3.5

*Geometric mean used (the geometric mean is explained in more detail in Section 11.2.2).

Source: Shaper 1985. Reproduced with permission of BMJ Publishing Group Ltd.

Table 5.5.8 is reproduced from Table 3 of Paper B and shows the mean SBP for cases (of IHD) to be 155.4 mmHg and for non-cases to be 144.9 mmHg. The table also gives a t -value – this is the result of the hypothesis test, a t -test. There is a substantial difference between the two means: In this group of men, systolic blood pressure was, on average, 10.5 mmHg higher among men who suffered from IHD. Can we infer that this is evidence of an association between high systolic blood pressure and IHD in all middle-aged British men?

Comparing the Means of Large Samples: The Two-Sample z -Test

We will see shortly that the t -test can be used for both large and small samples, and when an hypothesis test for comparing means is carried out using computer software, the t -test is most commonly calculated. For large samples such as the BRHS, however, we can also use the two-sample z -test, although (as we will demonstrate) the conclusion is almost identical to that produced by the t -test. Since the z -test relates directly to our discussion of sampling distributions and the standard error in Chapter 4, and the t -test requires consideration of some additional factors (degrees of freedom), we start our explanation of hypothesis testing for continuous variables with the z -test.

As mentioned above, the basic principle of hypothesis tests for continuous data is the same as for the chi-squared test. In summary, we formulate an hypothesis, calculate a test statistic, and use the result to determine the likelihood of observing these sample means if the null hypothesis were true. If it is very unlikely, we reject the null hypothesis.

Stating the Hypotheses

The question of interest is, ‘In middle-aged British men, does average systolic blood pressure differ significantly between those who become IHD cases and those who do not?’ The null hypothesis may be stated as follows:

H_0 : there is no difference, on average, between the systolic blood pressure of middle-aged British men who become IHD cases and the systolic blood pressure of middle-aged British men who do not become IHD cases.

This can be written more concisely by symbols. We use μ_1 to denote the mean systolic blood pressure in the population of middle-aged British men suffering IHD, and we use μ_2 to denote the mean systolic blood pressure in the population of middle-aged British men who do not suffer IHD. The null hypotheses and alternative hypothesis are then

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Testing the Null Hypothesis

We have evidence of a real difference (that is, the null hypothesis is unlikely to be true) if the difference between the sample means is large enough. Labelling the sample means $\bar{x}_1 = 155.4$ and $\bar{x}_2 = 144.9$, the observed difference is $\bar{x}_1 - \bar{x}_2 = 10.5$ mmHg. Whether this is large enough depends on how much this difference can vary from sample to sample; that is, the **precision** of this estimate of the difference between means. This precision is measured by the **standard error of the difference between means**, described below.

The Standard Error (SE) of the Difference Between Means

The estimated standard error of $(\bar{x}_1 - \bar{x}_2)$ for use with the z -test is

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s_1 and s_2 are the standard deviations of the sample of systolic blood pressures of cases and the sample of systolic blood pressures of non-cases, respectively.

Moving now to the hypothesis test, the test statistic is called z . It is the difference between the sample means relative to the standard error of the difference between the means:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

To calculate the standard error of the difference between means, we require an estimate of the standard deviations, which we obtain from the sample. Paper B does not report the sample standard deviations of systolic blood pressure, so to illustrate the z -test we shall use some plausible values. If the standard deviation of the sample of men with IHD is $s_1 = 26.1$ mmHg, and that of the sample of men without IHD is $s_2 = 24.1$ mmHg, we then have:

	Cases	Non-cases
\bar{x}	155.4	144.9
s	26.1	24.1
n	202	7533

The estimated standard error of the difference in means is:

$$\sqrt{\frac{26.1^2}{202} + \frac{24.1^2}{7533}} = 1.8573$$

and the value of the z -statistic is

$$\frac{155.4 - 144.9}{1.8573} = 5.653$$

Obtaining the p -value for the Test

You will normally carry out hypothesis testing on a computer, which produces the relevant p -value, but to find the p -value associated with the value of z that we have calculated here, we can also look up the critical value of z in a table of the normal distribution. We looked at a table of critical values for the chi-squared statistic in the section above, and examples for some other tests are also provided and discussed in Chapter 11, but since you will generally obtain hypothesis test results including p -values by computer, we will not use tables anymore in this

chapter. The critical value for the z -statistic at the 0.05 probability level is 1.96. For our z -statistic of 5.653, the relevant p -value is $p < 0.001$. This is the probability that this value of the z -statistic could have arisen by chance.

Interpreting the Result of the Test

This p -value is small, very small in fact, as it indicates a probability of less than 1 in 1,000. If we adopt the convention of rejecting the null hypothesis if the p -value is less than 0.05 (1 in 20, or the 5% significance level), then we reject H_0 and conclude that there is a statistically significant difference between the mean systolic blood pressure of men who suffer IHD and those who do not suffer IHD.

Summary: The Two-Sample z -Test

The two-sample z -test is used to test the null hypothesis that the means of two populations are equal. As with the chi-squared test, there are a number of assumptions that should be met. So, to use the z -test,

- The data must be continuous.
- The samples must be random.
- The samples must be independent (i.e. a person cannot be in both samples).
- The samples must be large (each exceed about 30).

1. State the hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

defining μ_1 and μ_2 , the population means.

2. Decide on the significance level. Call this α . (Typically $\alpha = 0.05$).
3. If calculating the z -statistic by hand, use the formula:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

where $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, \bar{x}_1 and \bar{x}_2 are the sample means, and s_1 and s_2 are the sample standard deviations.

4. Then compare the value of z with the normal distribution and determine the p -value; if using a computer, identify the p -value in the output (though as noted above, computer software generally carries out a t -test).
5. If $p < \alpha$, (as stated above, most commonly 0.05), reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

Comparing the Means of Smaller Samples: The Two-Sample t -Test

So far, in the z -test, we have assumed that the samples from which we wish to make inferences about the population are large. Smaller samples (by which we mean less than 30 values) still give us information about the population, but we need different techniques to make these inferences. If we want to use small samples to investigate whether population means differ, we should use a **two-sample t -test**, provided we can reasonably assume that the two samples are from populations with (roughly) normal distributions and that the two populations have

the same (or at least similar) standard deviations. If these assumptions are not reasonable, we can use a non-parametric test: these tests are described in Chapter 11.

Calculating the t -test

We calculate a test statistic, called t , in very much the same way as we did in the z -test,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

However, the standard error of the difference between means (SE in the formula) differs from that used for the z -test, as explained in the reference section below.



RS – Reference Section on Statistical Methods

The Standard Error for Difference Between Means (t -test)

Since for the t -test we are assuming that the populations have the same (or at least similar) standard deviations, we use information from both samples to estimate this common standard deviation. The estimate is called the **pooled standard deviation**, s_p , and its square, the estimated variance, is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Where n_1 and n_2 are the sizes of the two samples, and s_1 and s_2 are the respective standard deviations of the samples. The estimated standard error of the difference between means is

$$SE = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

To calculate the t -test result for our example, we can start with the difference between the means – which is the same as for the z -test (10.5 mmHg). The standard error, using the formula stated above, is 1.722, and the t -statistic is 6.097. We are now ready to determine the p -value, but, to do so, we have to know the degrees of freedom. As with the chi-squared distribution, the t distribution is a probability distribution, and there are different distributions, each defined by a number of degrees of freedom. For the t -statistic, the degrees of freedom are the total number in both samples minus 2:

$$\text{df (degrees of freedom)} = n_1 + n_2 - 2 = 7733$$

The probability of obtaining a value of the t -statistic of 6.1 on 7733 degrees of freedom (or a value more extreme) is <0.001 , the same result that we obtained using the z -test.

Use of the t -test or z -test?

We saw that the results of using t -test and z -test were in effect identical with the large sample in the BRHS, so how should these two tests be used? A t -test can be used whenever the populations are considered to have normal distributions, whatever the sample size. So it is never wrong to use the t -test in this situation. However, it would be wrong to use the z -test for small samples, that is, smaller than the (arbitrary) cut-off of 30 suggested as a guide in most situations. As noted, computer software generally calculates only a t -statistic, so this is most commonly quoted in journal articles. The z -statistic is easier to calculate without a computer. The t -test,

like the other tests we have discussed, is subject to a number of assumptions, which are listed below as part of a summary of the procedure.

Summary: The Two-Sample *t*-test

This is used to test the null hypothesis that the means of two populations are equal, including when the samples are small ($n < 30$ values in each sample). To use the *t*-test:

- The data must be continuous.
- The samples must be random.
- The samples must be independent (i.e. a person cannot be in both samples).
- The samples must be from populations with approximately normal distributions.
- The populations must have the same (or similar) standard deviations. A useful general rule is that if the larger of the sample standard deviations divided by the smaller of the sample standard deviations is < 2 , then we can assume sufficiently similar population standard deviations.

1. State the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

defining μ_1 and μ_2 , the population means.

2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. If working by hand, calculate the square of the pooled standard deviation

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where s_1 and s_2 are the sample standard deviations.

4. Then calculate the *t*-statistic,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE} \quad \text{where } SE = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}},$$

and \bar{x}_1 and \bar{x}_2 are the sample means.

5. Finally, compare the value of *t* with the *t*-distribution on $n_1 + n_2 - 2$ degrees of freedom and determine the *p*-value; if using a computer, identify the *p*-value in the output.
6. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
7. State the conclusion and interpret the result.

5.6 How Large Should a Cohort Study Be?

5.6.1 Perils of Inadequate Sample Size

In Chapter 4, we discussed the importance of sample size for precision of estimates of means and proportions, and we looked at the information needed to calculate these sample sizes. Sample size is just as important when designing an analytic study or trial, but we need to introduce additional components when comparing values (e.g. means, rates) between groups if we are to avoid making what are termed type I and type II errors. These are explained in Table 5.6.1 in the context of a cohort study.

Avoiding these errors is clearly very important. We now look at the method for calculating the sample size for a cohort study, which incorporates components for setting the significance level of the hypothesis test and the power of the study. Given the type of outcome in the

exercise study, namely cancer (present or absent), we focus here on sample size for a categorical outcome. Calculation of sample size for a continuous outcome variable is covered in Chapter 7, Intervention Studies.

Table 5.6.1 Type I and type II errors.

Error Type	Explanation and examples
Type I error	<p>In a cohort study, the effect of exposure to (for example) smoking is being studied. For our example, let's say smoking does not in reality affect the risk of the disease. By chance, however, the study shows a relative risk of 1.25 (25 per cent increase in risk). We would be making a type I error if we were to conclude that smoking really does increase the risk of the disease.</p> <p>We reach a conclusion by carrying out an hypothesis test (e.g. a chi-squared test). A type I error occurs if we reject the null hypothesis when it is in fact true (that is, there is no increased risk, as in this example). The chance, or probability of a type I error is called the significance level of the test and is usually denoted by α, and as we have seen in Section 5.5, we typically choose a significance level of 5 per cent; thus, the probability of a type I error occurring is 5 per cent or 0.05. Let's say the result of the hypothesis test is $p = 0.086$, and as a result we do not reject the null hypothesis (because the p-value is greater than 5 per cent, it is in fact 8.6 per cent). The hypothesis test, set at the 0.05 level, has helped us to avoid making a type I error.</p>
Type II error	<p>In another situation, let's say the exposure (smoking) really does increase the risk of the disease, and the true relative risk is 1.25 (25 per cent increase in risk). A type II error occurs if this genuine increase in risk is not recognised as being real. The probability of a type II error is denoted by β. This happens if the study is too small, relative to the amount of sampling error. Let's say that when the hypothesis test is applied the result is $p = 0.086$ (non-significant), and consequently we do not reject the null hypothesis.</p> <p>A study must therefore be designed to be large enough to be fairly certain of avoiding this mistake. This is described as ensuring that the study has enough statistical power. The power is the probability that we reject the null hypothesis (conclude there is evidence of an association) when it is false (there really is an association); in other words, we arrive at the right conclusion. Since the probability of the type II error is β, the power of the study (which is the probability of avoiding this error) is $1 - \beta$. Typically, we choose a power of 80 per cent or 90 per cent, so β, the probability of a type II error occurring, is then 20 per cent or 10 per cent, respectively.</p>

5.6.2 Sample Size for a Cohort Study

The information required to calculate sample size for a cohort study, for which we will use OpenEpi (see the 'Cohort/RCT' option in the menu), is as follows:

Sample Size Parameters for a Cohort Study

- The level of significance of the hypothesis test (α) – typically set at 0.05.
- The power ($1 - \beta$) – typically around 80 per cent.
- The expected frequency of the outcome (characteristic, disease, etc.) in the unexposed group (e.g. the percentage of subjects getting the disease over the projected follow-up period) – for example, 1 per cent (this would be equivalent to an incidence of 10 in 1000 per year for a 1-year study).
- Ratio of the number of unexposed to the number exposed – this is in effect the prevalence of exposure. For example, 50 per cent exposed is a 1:1 ratio, and 10% exposed is a 9:1 ratio.
- The relative risk that is to be detected – for example, 2.0.

The choice of levels of the parameters to be used to calculate a sample size (for example, the relative risk you wish to detect) should ideally be based on previous experience using information from the published literature of studies investigating similar topics. In reality, however, it is not always possible to obtain such information, especially if the research topic is new and values will have to be estimated. We will see how changes in parameters entered into a sample size calculation affect the required sample size in Section 5.6.3.

It is strongly recommended that you always seek statistical advice about this crucial aspect of study design. It is, however, valuable for you to be able to understand the type of information that is required because it will be you (the researcher) rather than the statistical consultant who will need to make the decisions about what is important.

5.6.3 Example of Output from Sample Size Calculation

The information in Table 5.6.2 shows the values for the parameters of the sample size calculation we listed above that were entered into OpenEpi.

Table 5.6.2 Information for sample size in a cohort study.

Information needed for calculating sample size	Values for this example
Two-side significance level ($1 - \alpha$)	95
Power ($1 - \beta$, or % chance of detecting)	80%
Ratio of sample size, unexposed/exposed	1
Percentage of unexposed with outcome	5
Risk ratio (RR) or relative risk – closest to 1.00	1.5

The output from the sample size calculation using OpenEpi as it appears on the screen, which also shows how the input data are displayed, is shown in Table 5.6.3.

Using the continuity-corrected (CC) calculation in the final column, we find that a total of around 3,100 subjects are required for a significance level of 0.05, power of 80 per cent, 5 per cent disease incidence in the unexposed, 1:1 ratio of unexposed to exposed, and a relative risk to be detected of 1.5. A continuity correction provides a more conservative estimate due to the use of the normal distribution to approximate the binomial distribution in this calculation – see Chapter 11 for more discussion of these distributions – and we recommend that you use this value for estimating sample size.

Changing the input parameters affects the sample size, and in Table 5.6.4 you can see the implications for the study size of changing all of the different values we chose initially (we have changed these one at a time). Some of the changes we have made result in an increase in the sample size, for example reducing the significance level to 1 per cent, increasing the power to 90 per cent, or having an outcome that is less common (in the unexposed group), because these are all more demanding in terms of the amount of data required to detect the specified relative risk. On the other hand, making the outcome more common, and increasing the effect size (relative risk) that we wish to detect, reduce the required sample size.

As with sample size for precision of estimates, we should make allowances for the expected initial response rate. Additionally, because cohort studies involve follow-up, it is important to make allowance for losses to follow-up that can arise from losing contact with study subjects (e.g. migration), refusals to continue in the study, and deaths from causes that may not be

Table 5.6.3 Example of output from sample size calculation for a cohort study (OpenEpi).

Two-sided significance level(1-alpha):	95		
Power(1-beta, % chance of detecting):	80		
Ratio of sample size, Unexposed/Exposed:	1		
Percent of Unexposed with Outcome	5		
Percent of Exposed with Outcome	7.5		
Odds Ratio:	1.5		
Risk/Prevalence Ratio:	1.5		
Risk/Prevalence Difference:	2.5		
	Kelsey⁽¹⁾	Fleiss⁽²⁾	Fleiss with CC
Sample Size – Exposed	1475	1474	1553
Sample Size – Unexposed	1475	1474	1553
Total Sample Size:	2950	2948	3106

Notes:

(1) Refers to value calculated using method proposed by Kelsey *et al.* (Methods in Observational Epidemiology, 2nd edition 1996; Oxford University Press).

(2) Refers to value calculated using method proposed by Fleiss *et al.* (Statistical Methods for Rates and Proportions, 3rd Edition 2003; Wiley & Sons). The third value (Fleiss with CC), refers to inclusion of the continuity correction, which provides a more conservative estimate.

relevant to the study. As a result, the eventual required sample size will always be larger than that obtained from the statistical calculations such as those shown here.

If our cohort study had a continuous outcome, such as minutes of physical activity, as in the Brazilian adolescent study (Paper D), we would use the ‘Mean Difference’ option in the OpenEpi menu. We work through an example of this calculation in Chapter 7.

Table 5.6.4 Impacts on sample size estimation with variation in the value of the various parameters used for the calculation.

Parameter changed	$1 - \alpha$	$1 - \beta$	Ratio of numbers unexposed to exposed	Unexposed with outcome (%)	Risk ratio	Total sample size*
Initial example	0.05	80%	1:1	5	1.5	3106
Significance level (%)	0.01	80%	1:1	5	1.5	4544
Power (%)	0.05	90%	1:1	5	1.5	4102
Ratio of unexposed to exposed	0.05	80%	1:2	5	1.5	3408
	0.05	80%	1:3	5	1.5	3984
% unexposed with the outcome	0.05	80%	1:1	2.5	1.5	6386
	0.05	80%	1:1	10	1.5	1444
Relative risk	0.05	80%	1:1	5	1.25	10734
	0.05	80%	1:1	5	3.0	950

*Fleiss method with continuity correction.

5.7 Assessing Whether an Association is Causal

5.7.1 The Hill Viewpoints

Determining whether an association is causal is a very important issue in the interpretation of research evidence. Our ability to make judgements about causation depends on the type and extent of the evidence, and there are a number of pointers that can help us reach a conclusion. In 1965, the eminent medical statistician Sir Austin Bradford Hill published a landmark paper discussing this issue, in the context of environment and disease, drawing particularly on experience from occupational exposures (Hill, 1965). This led to what are termed the ‘Hill viewpoints’, and with relatively minor qualifications (see for example: Howick *et al.*, 2009), these have stood the test of time as a valuable way of assessing the strength of causal inference where an association has been demonstrated. Table 5.7.1 provides a summary of the Hill viewpoints, with some examples (some drawn from Hill’s paper) and discussion.

Table 5.7.1 Hill viewpoints with explanations and examples.

Viewpoint	Explanation and example of application
1. Strength of association (including the role of confounding)	<p>The stronger an association, the more likely it is to be causal. A good example is the association between smoking and lung cancer, for which the relative risk (RR) has generally been found to be at least 10, and even higher for very heavy smokers. This effect is so strong that it is extremely unlikely to be anything other than causal. It is important to distinguish between the strength of the association (as estimated by a relative risk), and the p-value that arises from an hypothesis test used to examine that association. Thus, a very large study could well identify a modest relative risk of something like 1.05 with a p-value of <0.001: This is a highly statistically significant finding of a very weak association.</p> <p>Hill discusses confounding (where a factor is related to both exposure and outcome and may therefore explain an observed association between the exposure and outcome, discussed in more detail in Section 5.7.2) in the context of the strength of the association. He compares the examples of smoking and lung cancer (a very strong association) and smoking and heart disease (a weaker association, with a RR of 2–3). He notes that while it is difficult to see how confounding by one or more other risk factors could explain the very strong association seen for the former, it could certainly contribute – perhaps substantially – to the latter. It is therefore very important to be able to demonstrate that the association is not explained by confounding factors. Matching in case–control studies (Chapter 6) and randomisation in trials (Chapter 7) are used to avoid confounding, and multiple-regression techniques are used in analysis to adjust associations for confounding. Even with modern statistical methods, however, some humility is still required in the face of confounding with observational study designs: There may well be confounding factors that have not been considered or that have not been measured adequately enough to fully adjust for, resulting in <i>residual confounding</i>.</p>
2. Consistency of evidence from studies in different settings	<p>We would expect that a genuine causal association would show up in studies of different types and in different populations. Hill refers to the association being repeatedly observed ‘by different persons, in different places, circumstances and times’. Of course, some causal effects may be dependent on, or modified by, genetic or other factors associated with an ethnic group, but a degree of consistency across studies in a variety of settings, occasions, and by multiple research groups can nonetheless be expected. Hill does provide an example of an industrial exposure (nickel) causing cancer from which firm conclusion could reasonably be drawn without multiple studies, but this is rather an uncommon situation.</p>

Table 5.7.1 (Continued)

Viewpoint	Explanation and example of application
3. Specificity of the association	Hill noted that a causal association might be specific; for example, a disease arising among certain groups of workers exposed to a particular process or chemical. Although there are undoubtedly examples of this, the specificity viewpoint is now considered to be less useful, as many risk factors are linked to multiple disease outcomes, smoking, exercise, obesity, air pollution, etc., being good examples. Hill does recognise this, and he cautions that we 'must not over-emphasize the importance of this characteristic'.
4. Temporality: exposure preceding the onset of the disease	If a factor is causing a disease, it stands to reason that the exposure to that factor must have occurred before the disease process began. Also, the exposure must be shown to have occurred long enough before the onset of disease to allow for the <i>latency period</i> , this being the time it takes for the exposure to lead to overt disease in a human. An example is the relationship between radiation exposure and leukaemia: It takes several years after exposure to radiation before leukaemia appears. An important advantage of the prospective study design (cohort studies) is that information is usually available on the time sequence of exposure and disease onset.
5. Biological gradient (dose–response relationship)	If an association is causal, we would expect that the greater the exposure (dose), the greater the effect on the outcome (response). If, for example, lifetime heavy smokers had a risk of lung cancer similar to, or lower than, that of light smokers or people who smoked for only a few years, we would have to question whether smoking could cause lung cancer. Although there may be threshold effects for some exposures (e.g. above a certain level there is no further increase in risk), it is usual to find evidence of a dose–response relationship across most of the range of exposure encountered.
6. Plausibility	It would be unwise to suggest that an association is causal if it does not make any sense given what we know of the biological mechanisms involved. Equally, if it is plausible, that would strengthen our view of causality. For example, if we have observed an association between bladder cancer and a certain carcinogenic (cancer-causing) substance, and we know this substance is excreted in the urine, it would be reasonable to use this biological evidence to strengthen our case for this being a causal association. On the other hand, if the toxin is handled by the liver and converted to metabolites with no known carcinogenic effect, we may be more cautious in assuming causation. Alternatively, we may not know enough about the biological mechanisms involved to exclude the possibility of causation. Other aspects of biological plausibility are evidence from experiments with animals and evidence from laboratory-based studies of disease mechanisms at cellular and other levels. The topic of animal research does create a good deal of controversy, not only from the animal-welfare point of view but also because of the relevance of animal models to humans. If an effect is seen in rats or monkeys, how do we know it applies to humans? And just because an effect is not seen in animals, does this mean that it cannot occur in humans? Despite these uncertainties, experience has shown that animal evidence can be relevant, and where it exists it should be taken into account.
7. Coherence	If a factor is causal, it would be expected that populations with a high level of exposure would have a higher incidence of the disease in question than populations with lower levels of exposure. For instance, if high serum cholesterol really is an important cause of IHD, then populations such as Japan with low levels should have correspondingly low IHD incidence rates, as is indeed the case. If on the other hand, Japan had high IHD incidence, or we found populations with high cholesterol levels and low IHD incidence, we would have to question whether cholesterol really is causal. Another perspective discussed by Hill is temporal coherence, with for example, the rise in incidence of lung cancer historically following (allowing for the latency period) the rise in smoking prevalence.

(continued)

Table 5.7.1 (Continued)

Viewpoint	Explanation and example of application
8. Experiment: removal of the exposure reduces risk	Perhaps the most powerful evidence of causation comes from experiments that measure the effect of reducing exposure to the suspected causal factor. In the ideal form (from a scientific point of view), the randomised, blinded, controlled trial (see Chapter 7) has the capacity to study the effect of changing the level of exposure, in a comparison where only the factor of interest differs between the groups being studied.
9. Analogy	Where there is good evidence of, for example, causal effects of an exposure during pregnancy on an adverse birth outcome such as low birthweight or some malformation, it is reasonable to assume causality of another exposure operating in a similar way on such outcomes. Hill uses the examples of thalidomide (a drug used for morning sickness) and rubella (German measles) to suggest that association between other drugs or viral infections in pregnancy might also cause adverse pregnancy outcomes. To which we might add, knowing the causal effect of smoking by the pregnant mother on birthweight, we might expect that lower levels of exposure from second-hand smoking and outdoor air pollution could have similar (though smaller) impacts on the same outcome.

We review these viewpoints again in later chapters when looking at the results from other study designs. By then, we will have encountered most of the points discussed in one form or another, and it will be a matter of bringing them together to see how they can help in assessing the strength of evidence for causal inference. Useful though these viewpoints are, Hill emphasises that these should not be applied in a rigid way, nor should all be required to be satisfied in any given situation. He wrote, ‘None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non* (something that is absolutely needed)’.

One of the most important of the Hill viewpoints that we began to think about in our earlier discussion of cohort studies is strength of association (Viewpoint 1), a critical aspect of which is that the association should be independent of *confounding*. This is a suitable point to look in more detail at what confounding means and to begin thinking about how to deal with it.

5.7.2 Confounding: What Is It and How Can It Be Addressed?

The conditions necessary for a factor to confound an association are as follows: for a factor (X) to *confound* an association between a *potential cause* (A) and the *outcome* (B), it must

- be associated with the potential cause (A)
- be a predictor of the outcome (B)
- not be in the causal pathway for the effect of (A) on (B).

To illustrate this, we return to the example of smoking and ischaemic heart disease (IHD). We are interested in the question of whether smoking is *causally* associated with IHD, and whether that association is being *confounded* by other factors such as level of exercise or blood cholesterol.

The factors that may confound the smoking–IHD association must themselves be *associated* with smoking. Compared to non-smokers, people who smoke would therefore have to take less exercise, have higher blood cholesterol, and so on. This is often the case, certainly among some population groups.

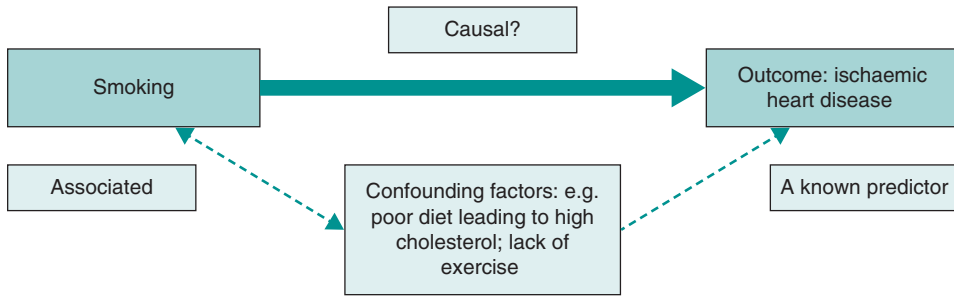


Figure 5.7.1 Conditions required for factors (e.g. lack of exercise, poor diet) to confound an observed association between smoking and IHD.

For factors such as exercise and blood cholesterol to confound the smoking–IHD association, the factors must themselves be predictors of the outcome. Thus, lack of exercise and high cholesterol would need to be risk factors for IHD, and we also know this to be the case. These conditions are illustrated in Figure 5.7.1.

It has been stated that confounders should not be in the causal pathway, and an example may help explain why this is important. Smoking in pregnancy causes low birthweight. One of the key mechanisms for this effect is through smoking causing increasing levels of carbon monoxide (CO) in the blood. The CO combines with haemoglobin (Hb), the substance in red blood cells that transports oxygen, to produce carboxyhaemoglobin (COHb). COHb is not so good at delivering oxygen to the tissues and organs of the body, including to the placenta, so the foetus receives less oxygen, and its growth is restricted. Let us now consider this in our model of confounding through the following two diagrams. Figure 5.7.2(a) shows an hypothesised model of confounding, and Figure 5.7.2(b) shows how CO fits into the causal pathway.

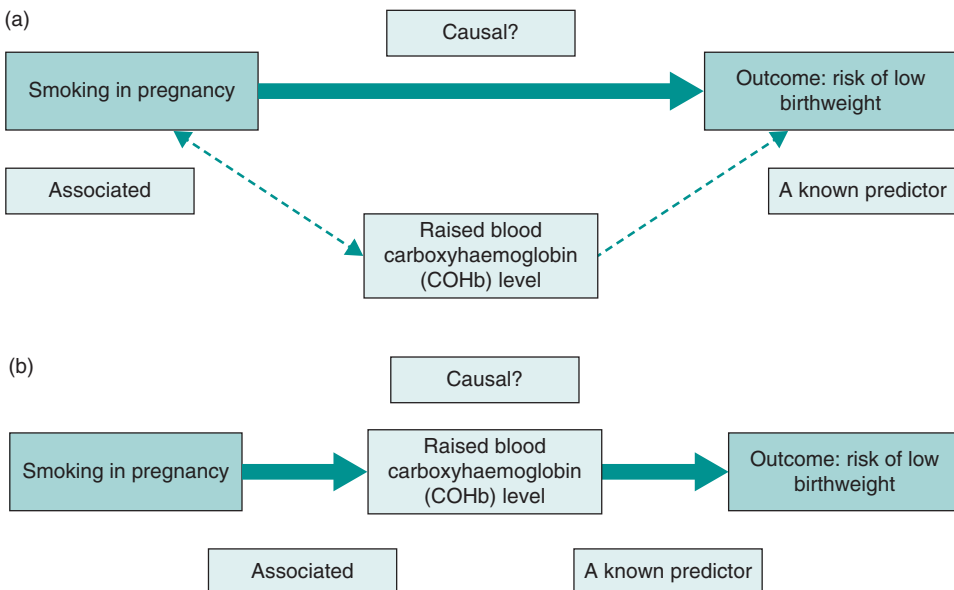


Figure 5.7.2 (a) Hypothesised pathways for raised COHb to confound the relationship between smoking and low birthweight. (b) Alternative pathways, with raised COHb in causal pathway between smoking and low birthweight.

The question then is, does COHb act as a confounding factor in the association between smoking in pregnancy and low birthweight? The answer to this is no, it does not. Raised COHb is associated with smoking, and it does cause low birthweight, but it is *not* confounding the association because it is one of the main mechanisms by which smoking has this effect on birthweight. It is in the *causal pathway*. This is an important conclusion because if we were to ‘allow’ (adjust) for the effect COHb by removing the contribution of smoking to raised COHb, we would as a consequence remove (most of) the effect of smoking as well. We would then erroneously conclude that smoking is not the problem.

5.7.3 Does Smoking Cause Heart Disease?

It is now accepted that smoking does indeed cause a wide range of diseases, including IHD. In our discussion of confounding of the smoking–IHD association, we suggested that factors such as cholesterol and lack of exercise might be confounders. So are they? Well, in some population groups, they certainly could partly confound an observed association between smoking and heart disease, but they do not explain that association completely. That is to say, even after taking into account the effects of these confounding factors, smoking still has a strong independent association with IHD, which we believe to be causal.

We can summarise this conclusion as follows. The observed association between smoking and IHD is mainly due to the fact that smoking causes IHD. In some situations, it is also partly due to the fact that smokers have other characteristics (such as taking less exercise or eating less healthily) that also increase the risk of IHD. In these situations, the observed association between smoking and IHD is partly confounded by other lifestyle factors. When we allow for these other lifestyle factors in the analysis, the independent effect of smoking is still clearly apparent, but it may be reduced.

5.7.4 Confounding in the Physical Activity and Cancer Study

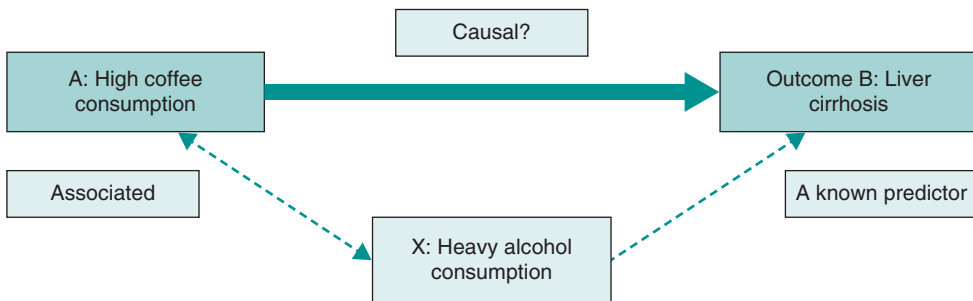
The authors of Paper A were also concerned about the possibility of confounding in their study. For example, some factors associated with the amount of physical activity an individual takes might also be risk factors for cancer, such as age, cigarette smoking, body mass index, alcohol intake, and social class. To adjust for these potential confounders, it was necessary to collect data on them for each individual in the study. So how did the authors then carry out this all-important analysis to demonstrate the independent effect of physical activity on the risk of cancer? This was done by using a method of analysis called *regression*, which we introduce in the next section of this chapter. The following exercise will help to consolidate your understanding of the concept of confounding and the differing extents to which confounding can operate. This will be useful background for when we look at regression methods in Section 5.8.



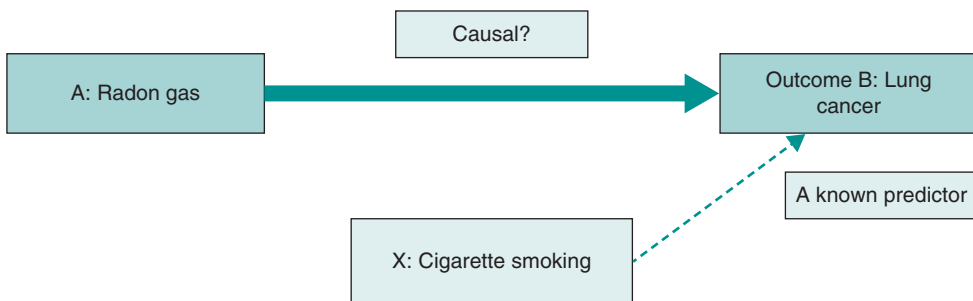
Self-Assessment Exercise 5.7.1

Based on the information given in the following diagrams, which of the following three models of association between factor A and outcome B (Figure 5.7.3), could be confounded by factor X?

Model 1: In populations where high levels of coffee drinking are associated with heavy alcohol consumption, could alcohol confound an association observed between coffee drinking and liver cirrhosis?



Model 2: Could cigarette smoking confound an association between radon gas exposure and lung cancer? (Radon gas is a naturally occurring product of uranium in the ground, especially in areas of the country with granite. It collects in the house and leads to human exposure.)



Model 3: Could poverty confound an association between damp housing and respiratory illness?

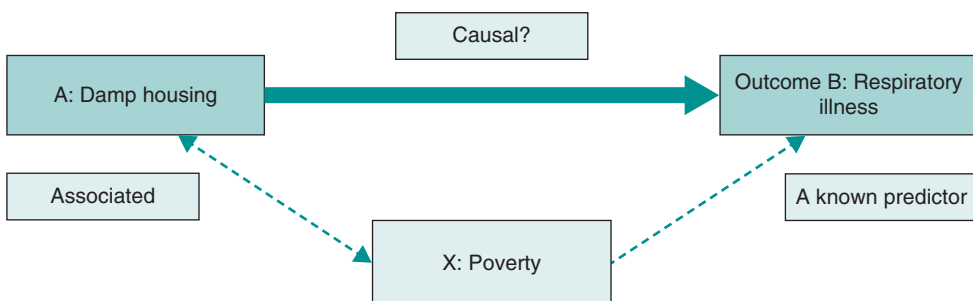


Figure 5.7.3 (a) Model 1 – Relationships among high coffee consumption, heavy alcohol consumption, and liver cirrhosis; (b) Model 2 – Relationships among radon gas, smoking, and lung cancer; (c) Model 3 – Relationships among damp housing, poverty, and respiratory illness.

Answers in Section 5.10

5.7.5 Methods for Dealing with Confounding

There are a number of other aspects of study design and analysis that we need to consider before we can cover this comprehensively, but it is worth considering in general terms how we set about dealing with confounding. There are essentially two approaches:

- We can design the study in a way that minimises the effect of confounding factors. We will return to this when we look at matching and stratification in case–control studies (Chapter 6) and randomisation in intervention trials (Chapter 7).
- We can use statistical methods for adjusting for the effects of confounding. The main methods of doing this are *standardisation* (which we looked at in Chapter 3 when we adjusted for the effects of differing age structures in two populations), post-stratification (described in Chapter 6), and *multivariable analysis* using regression techniques, which is introduced in Section 5.9 of this chapter and further developed through later chapters.

Summary: Confounding – Main Messages

- There are often complex interrelationships among variables being studied.
- These interrelationships can result in confounding if the potential confounder is associated with the possible cause, is itself a predictor of the outcome, and is not in the causal pathway.
- Confounding is often partial; that is to say, it partly explains an association between a risk factor and an outcome.
- The potential for confounding must be recognised in planning research, assessed, and, if necessary, allowed (adjusted) for before meaningful conclusions can be arrived at.
- Confounding can be minimised at the study design stage by matching or randomisation (we will look at these techniques in Chapters 6 and 7), and/or adjusted for in the analysis by standardisation (Chapter 3), stratification (Chapter 6), or multivariable regression analysis (Section 5.9).

5.8 Simple Linear Regression

5.8.1 Approaches to Describing Associations

One key aim of these next two sections is to explain how the authors of Paper A addressed the potential for confounding in the exercise-and-cancer association. To start our exploration of how this was done, we will go back to an example used in Chapter 2, where we looked at the methods for describing the relationship between two continuous variables (birthweight and gestational age), each measured on the same individuals. At that point we introduced correlation, and you will recall that

- Two continuous variables each observed for the same individuals can be displayed on a *scatterplot*.
- We describe any relationship between the variables in terms of its *form*, *strength*, and *direction*.
- The *correlation coefficient* is an objective numerical measure that summarises the strength and direction of a *linear association* between the variables.

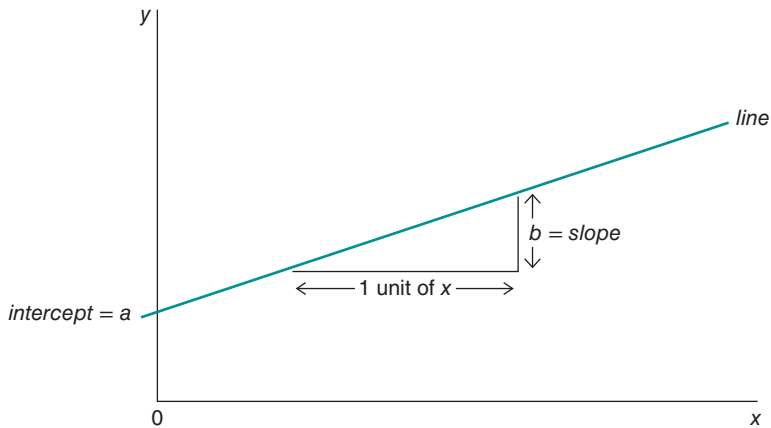


Figure 5.8.1 Components for the equation of a straight line.

We now look at what is generally the next stage in investigating such relationships; that is, to describe how one variable (the **outcome** or **dependent** variable) depends on the other variable (the **explanatory** or **independent** variable). For consistency, we restrict ourselves to the use of the terms *explanatory variable* and *outcome* from now on.

In the birthweight and gestational age example, an association we studied with correlation, having observed that there is an approximately linear relationship between the variables, we may wish to describe how birthweight **depends** on gestational age.

Birthweight is the outcome variable and is given the notation ‘y’, and gestational age is the explanatory variable, with the notation ‘x’. The term **explanatory variable** is appropriate because it (in this case, gestational age) explains, to some extent, how the outcome (birthweight) varies. You may recall that the simplest description of a relationship between two continuous variables is a straight line. Mathematically, a straight line is defined by an **intercept** (where the line crosses the y axis) and a slope (the **gradient**).

The intercept is the value of y when x is zero, and the slope is the amount y increases for each unit increase in x. Thus, the equation for the line is

$$y \text{ (outcome variable)} = a \text{ (intercept)} + b \text{ (slope)} x \text{ (explanatory variable)}$$

Returning to our example of birthweight and gestational age, we saw that there was a fairly strong, positive, linear relationship between the two variables, Figure 5.8.2.

Clearly, a straight line cannot exactly describe this relationship, as few of the values could lie on such a line. A line that is close to the data values, however, can summarise the relationship. It tells us by how much birthweight increases, on average, for an increase of 1 week in gestational age. The statistical method of estimating the relationship between variables is called **regression**.

You will recall variables can be continuous or categorical, and the two variables in our current example are both continuous. When we describe the relationship between one continuous outcome variable and one explanatory variable by a straight line, the relationship is called **simple linear regression**. In the remainder of this chapter, we look at examples with continuous outcomes to explore simple and multiple linear regression; the basic method for categorical outcomes is known as logistic regression, and this is described in Chapter 6 for case–control studies.

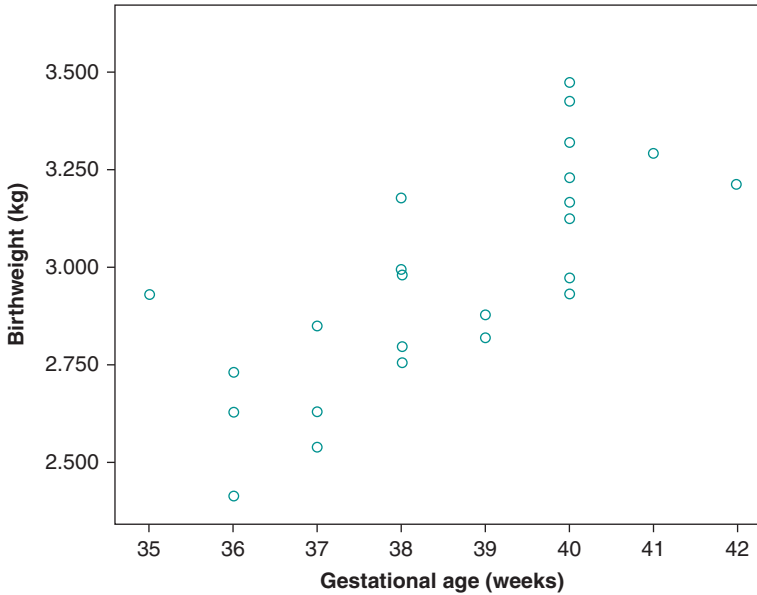


Figure 5.8.2 Birthweight and gestational age of babies.

5.8.2 Finding the Best Fit for a Straight Line

How do we decide which is the straight line that best summarises the relationship? That is, what do we mean by a line that is ‘close’ to the data? There are various ways of defining this closeness. The most common definition involves the differences between the observed y -values and a straight line; that is, the vertical deviations from the line (Figure 5.8.3).

We choose the line that results in the smallest possible value when all the deviations are squared and added together. This method of choosing the ‘best’ line to describe the relationship is called *least squares regression*. We will not go into the statistical basis for this calculation, which is beyond the scope of this book. The least squares regression line summarising the relationship between birthweight (y) and gestational age (x) is calculated as $y = -1.4850 + 0.1155x$.

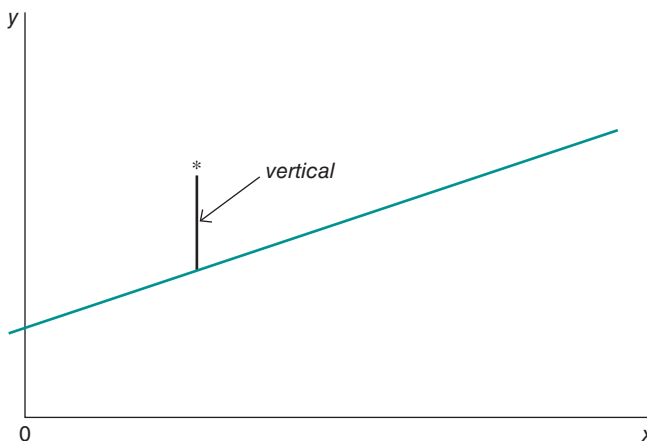


Figure 5.8.3 Difference between observed y -value and a straight line.

Plotting the line $[y = -1.4850 + 0.1155x]$ on the scatterplot shows us how much the data vary about the line (Figure 5.8.4).

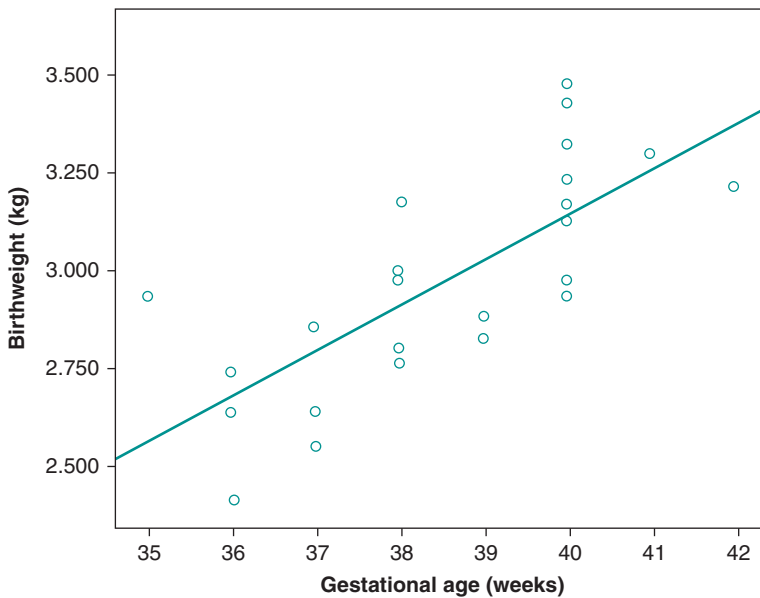


Figure 5.8.4 The regression of birthweight on gestational age.

Exercise 5.8.1 involves using the regression equation to calculate some values of birthweight and see whether they fit the line.



Self-Assessment Exercise 5.8.1

1. Use the regression equation to calculate the value of birthweight (in kilograms) for gestational ages of 37 weeks and for 41.5 weeks.
2. Do your answers lie on the line in Figure 5.8.4?
3. Do you think it would be reasonable to use this line to calculate the birthweight of a baby of 28 weeks' gestation? Make a note of why you think doing this is, or is not, reasonable.

Answers in Section 5.10

5.8.3 Interpreting the Regression Line

The equation we have calculated is a summary of the relationship. It does not tell us the actual birthweight of a baby born at a particular gestational age. We know that birthweight varies for any particular gestational age, because gestational age is only one of several factors associated with birthweight.

The equation shows us the average birthweight of babies of a particular gestational age. It also shows us how birthweight changes with gestational age over the range of 35–42 weeks: For each extra week of gestational age, birthweight increases by 0.1155 kg (about 116 g) on average.

Note that the intercept (that is, where the line crosses the y -axis, which is for a gestational age of 0 weeks) is negative (-1.4850). This implies a negative birthweight for a gestational age of 0

weeks. Of course, this does not make any sense, and clearly somewhere between 35 weeks' gestation and conception, the relationship we have described with this regression equation changes. It is therefore very important to note that the regression equation only describes the relationship within the range of the observed x data. We should not use this equation to describe how birthweight and gestational age are related for ages of less than 35 weeks.

5.8.4 Using the Regression Line

We have used the regression equation to

1. **Describe** the average level of the outcome variable associated with each level of the explanatory variable.

We can also use it to

2. **Predict** values of the outcome variable for new observations of the explanatory variable, within the range of the x data.
3. **Adjust** measurements of the outcome variable for the effects of the explanatory variable, before comparing individuals.

Here are some examples to make this clearer:

- **Predict:** If another baby, not in the original data set, is to be born at 40 weeks, we predict that the weight will be $-1.4850 + 0.1155 \times 40 = 3.135$ kg. In fact, prediction is what you did in Exercise 5.8.1 (Question 1), and is our best guess at the weight of this individual baby.
- **Adjust:** If we know the relationship between weight and height for men, we can take into account a man's height in deciding whether he is overweight.

5.8.5 Hypothesis Test of the Association Between the Explanatory and Outcome Variables

We can calculate a regression equation for any two continuous variables, so it is possible to come up with an equation even if the outcome is not in fact related to the explanatory variable. So how do we know if the explanatory variable really is associated with the outcome?

As a first step, we should always look at the data on a scatterplot, just as we did for correlation in Chapter 2. This will show whether there is any obvious non-linear pattern, in which case linear regression is not appropriate. It will also give an idea of whether there may be no relationship between the variables. This is subjective though – it is easier to see a relationship than the absence of one.

- An objective way of investigating whether the outcome is really associated with the explanatory variable is to carry out an **hypothesis test**. We have said that to carry out linear regression, the outcome variable must be continuous and there should be an approximately linear relationship. However, to test hypotheses, we also need to be able to assume that The outcome y has a population with an approximately normal distribution.
- For each value of the explanatory variable x , the variation of y about the regression line is approximately the same.

These are reasonable assumptions to make in many situations – we shall comment on them further after looking at how to carry out an appropriate hypothesis test.

If the outcome and explanatory variables are really not associated, then the slope b of the regression line should be zero. However, we may happen to estimate a non-zero slope from the particular sample of data we have. So we want to test the null hypothesis H_0 : slope = 0

in the population from which the sample was drawn. We can use the t -distribution for this hypothesis test.

Testing the Null Hypothesis of No Association Between Y and X (H_0 : Slope = 0)

We calculate the test statistic

$$t = \frac{b}{SE}$$

where b is the estimated coefficient of x and SE is the standard error of b . The statistic is compared with the t distribution with $n - 2$ degrees of freedom to find the p -value of the test (n is the number of subjects). The values of b and SE are calculated and displayed by computer software, along with the result of the hypothesis test, as shown below. This explanation should help you to understand how this result is obtained and what it means.

If the assumptions given above are not satisfied, the estimate of the regression equation will not usually be seriously affected, but the standard error and consequently the p -value will not be correct. So if it appears that the assumptions may not be satisfied, we should proceed with caution – particularly if the p -value is close to a critical value such as 0.05 – and seek advice if necessary. There are methods for checking these assumptions and dealing with departures from them, but they are beyond the scope of this book. For the regression of birthweight on gestational age, software (e.g. SPSS) gives

	Coefficient	Standard error	t -value	p -value
Gestational age	0.11553	0.02210	5.23	0.000

You can see how the t -value is derived by dividing the slope (coefficient) by the standard error, as in the formula above. The p -value is stated as 0.000 due to the limit (3) on the number of decimal places in this output, but it is not in fact zero. We can, however, say that $p < 0.0005$. The conclusion is that there is very strong evidence of a relationship between the birthweight and gestational age of babies. The probability of the observed relationship arising if H_0 is true is very small.

5.8.6 How Good is the Regression Model?

So far we have seen how to find the line of best fit, but it is also important to assess how well this line fits the actual data, known as the *goodness-of-fit* of the model. This is a relatively straightforward concept in simple linear regression, but, as we shall see in Section 5.9, it becomes slightly more complicated when we include more than one explanatory variable in the model (multivariable regression).

Calculating Sums of Squares

We estimated the line of best fit for our model (regressing birthweight on gestational age) by calculating the line that results in the smallest sum of squared differences between observed values and the line. To gauge the contribution of our model (gestational age) to predicting the outcome (birthweight), we can calculate several different sums of squares, and we now look at these.

First, we can calculate the sum of squared differences for the simplest model available, the mean value of y (birthweight). Clearly, this is not a good model of a relationship between two variables, but this step is useful, as we will see later. Using the mean as a model, we can calculate the squared differences between the observed values and the values predicted by the mean and

sum them. This sum of squared differences is known as the **total sum of squares** (SS_T) because it is the total amount of differences present when the most basic model is applied to the data.

Second, if we fit the model representing the line of best fit (the regression line) to the data, we can again work out the differences between this new model and the observed data. Although this regression line is ascertained by the method of least squares (as described previously), there will still be some inaccuracy represented by the differences between each observed data point and the value predicted by the regression line. As before, these differences are squared and then summed. The result is known as the **sum of squared residuals** (SS_R). This value represents the degree of inaccuracy when the best model is fitted to the data.

Third, we can now use these two values (SS_T and SS_R) to calculate how much better the regression line (the line of best fit) is than just using the mean as a model. The improvement in prediction resulting from using the regression model rather than the mean is estimated by calculating the difference between SS_T and SS_R . This improvement is known as the **model sum of squares** (SS_M).

Figure 5.8.5 shows each of these sums of squares graphically.

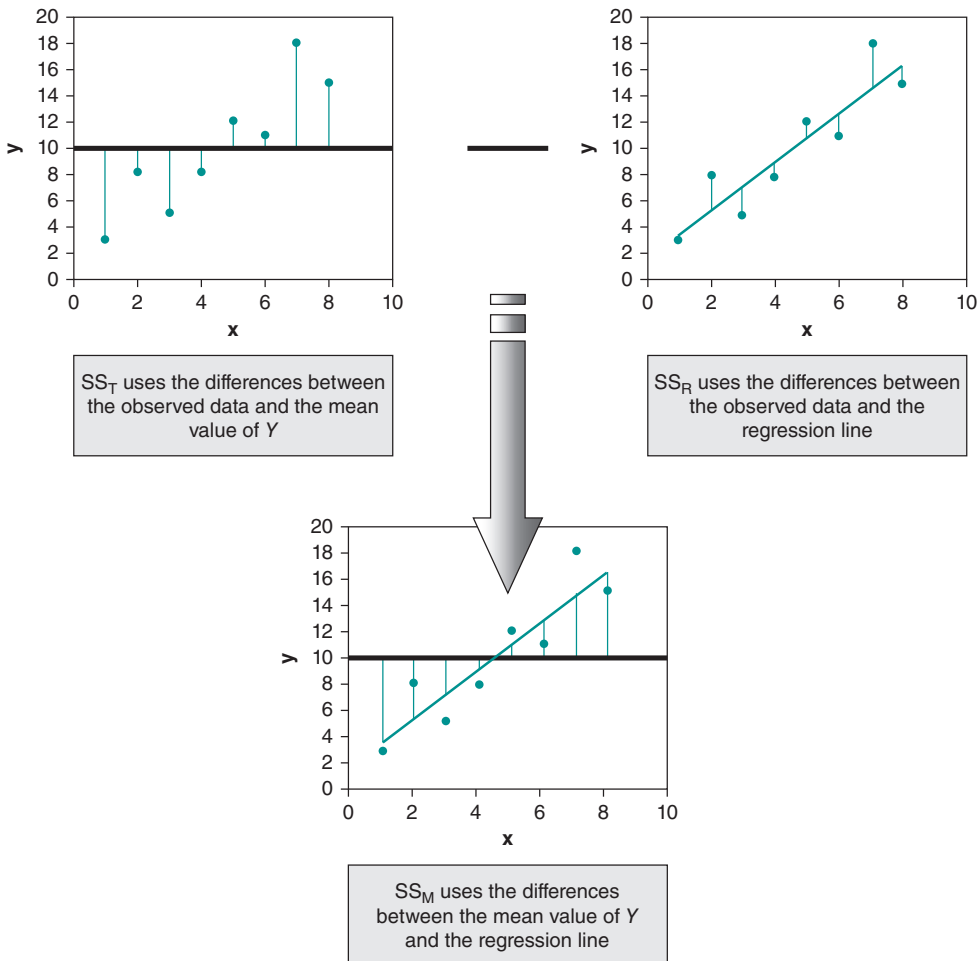


Figure 5.8.5 The three sums of squares used in assessing the goodness of fit of a regression model. *Source:* Field 2001. Reproduced with permission of (last edition).

If the value of SS_M is large, then the regression model is very different from using the mean to predict the outcome variable. This implies that the regression model has made a substantial improvement to how well the outcome variable can be predicted. However, if the SS_M is small, then using the regression model is little better than using the mean. We will look at how to formally assess whether SS_M is large or small shortly.

Measuring Improvement Due to the Model – r^2

One useful outcome from calculating these sums of squares is the ability to assess the amount of improvement due to the model. This is calculated by dividing the sum of squares for the model (SS_M) by the total sum of squares (SS_T). The resulting value is called r^2 ; this can be multiplied by 100 to give a percentage, and represents the amount of variation in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T). The value of r^2 is calculated as follows:

$$r^2 = SS_M/SS_T$$

The concept of r^2 was introduced in Section 2.4 as the coefficient of determination when using Pearson correlation, which is equivalent to the r^2 described here for simple linear regression.

A Measure of Goodness of Fit – F -Ratio

A second use of the sums of squares in assessing the goodness of fit of the regression model is through the F -test. Briefly, the test is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). For this test, rather than using the sums of squares themselves, we take the *mean sums of squares* (referred to as mean squares or MS). To work out the mean sum of squares, we need to divide by the *degrees of freedom*. These are calculated as follows:

- For SS_M , the degrees of freedom are the number of explanatory variables in the model.
- For SS_R , the degrees of freedom are the number of observations minus the number of *parameters* being estimated. The parameters are the number of *beta coefficients* in the model (in the case of simple linear regression, there is just one such coefficient: the single explanatory variable) and the intercept (often referred to as the *constant*).

The results are the mean squares for the model (MS_M) and the residual mean squares (MS_R). The F -ratio is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model:

$$F = MS_M/MS_R$$

If a model is good, we expect the improvement in prediction due to the model to be large (so, MS_M will be large) and the difference between the model and the observed data to be small (so, MS_R will be small). In short, a good model should have a large F -ratio (greater than 1 at least). We can also obtain p -values to estimate the significance of the F -ratio, which are produced by software when a regression analysis is run on a computer, or by using tabulated critical values for the F -distribution (another probability distribution discussed further in Chapter 11) for the relevant degrees of freedom.

5.8.7 Interpreting SPSS Output for Simple Linear Regression Analysis

To demonstrate how we interpret the output from simple linear regression with SPSS, we will use the low back pain data set (described in the Preface) to study the relationship between two

continuous variables: height and weight. The simple linear regression exercise will consider the effects of height on weight for the 775 employees composing the data set. When carrying out linear regression analysis, SPSS generates four tables.

5.8.8 First Table: Variables Entered/Removed

The first table indicates what outcome variable has been selected (weight) and lists the explanatory variables (only one in this example of simple linear regression, namely height) entered into the model to predict values of the outcome variable, Table 5.8.1.

Table 5.8.1 Variables entered/removed (SPSS output) for regression of height on weight (the footnotes are part of the SPSS output).

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	height in cm ^b	.	Enter

^aDependent Variable: weight in kg.

^bAll requested variables entered.

Second Table: Model Summary

The second table provided by SPSS is a summary of the model, Table 5.8.2. This summary table provides the value of r and r^2 for the model that has been calculated. Note that SPSS uses upper case R and R^2 in the tables, although the convention is to use lower case when describing these values; to avoid confusion here, we will use upper case R and R^2 when describing the output shown in the SPSS tables.

Table 5.8.2 Model summary (SPSS output) for regression of height on weight (the footnote is part of the SPSS output).

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.546 ^a	.298	.297	10.3798

^aPredictors: (Constant), height in cm.

Because there is only one predictor (height), R represents the simple correlation between weight and height, that is, the Pearson correlation coefficient introduced in Chapter 2, Section 2.4.

This table includes a value for the 'Adjusted R square'. This is essentially the R^2 , but adjusted for the number of explanatory variables in the model. It is slightly more conservative, and the more explanatory variables there are, the more it will differ from R^2 . Keep this in mind, but in practice the adjusted value is rarely used and we will not discuss it further.

Returning to the table output, we see that the Pearson correlation coefficient between height and weight is 0.546, a moderately strong positive association. The value 'R square' (coefficient of determination) is 0.298, which tells us that height accounts for 29.8 per cent of the variation in weight. In other words, there might be many factors that can explain the variation in weight,

but our model, which includes only height, can explain almost 30 per cent of this variation. Put another way, 70 per cent of the variation in weight cannot be explained by height alone.

Third Table: The ANOVA Table

The next part of the SPSS output reports **analysis of variance** (ANOVA), Table 5.8.3. This summary table shows the various sums of squares (indicated below) and the degrees of freedom (df) associated with each. From these values, the mean squares are calculated by dividing the sum of squares by the respective df. The most important part of the table is the F -ratio (the calculation of which was described in Section 5.8.6) and the associated significance value of the F -ratio.

Table 5.8.3 ANOVA table (SPSS output) for regression of height on weight (the footnotes are part of the SPSS output).

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33436.217	1	33436.217	310.339	.000 ^b
	Residual	78650.918	730	107.741		
	Total	112087.135	731			

^aDependent Variable: weight in kg.

^bPredictors: (Constant), height in cm.

For these data, F is 310.339, which is significant at $p < 0.0005$. This result tells us that there is a very small chance that an F -ratio this large would happen by chance alone. Therefore, we can conclude that our regression model results in significantly better prediction of weight than if we used the mean value for weight.

Fourth (Final) Table: The Coefficients Table

Valuable though it is assessing the overall performance of the model, the ANOVA table does not tell us about the individual contribution of variables in the model. That said, in the case of simple regression, there is only one variable in the model and so we can infer that the variable is a significant predictor of weight. The coefficients table provides details of the model parameters (the beta coefficients, or the slopes of the regressions), the statistical significance of these values, and (if specified for the output) the 95% confidence interval, Table 5.8.4.

Table 5.8.4 Coefficients table (SPSS output) for regression of height on weight (the footnote is part of the SPSS output).

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-34.659	5.818		-5.957	.000	-46.082	-23.237
	height in cm	.614	.035	.546	17.616	.000	.546	.682

^aDependent Variable: weight in kg.

We saw earlier that the intercept for the simple linear regression equation was denoted by the symbol α , also termed the constant. Note that (somewhat confusingly!) in the SPSS output, this is labelled B for the constant, and it has the value of -34.659 . The slope of the regression is also labelled B, for the explanatory variable (height [cm]), and it has the value 0.614 . Thus, the regression equation is

$$\text{Weight (in kg)} = -34.659 + 0.614 (\text{height in cm})$$

We can think of the beta coefficient (slope) as representing the unit change in the outcome variable associated with a unit change of the explanatory variable. Therefore, if our explanatory variable is increased by one unit (if height is increased by 1 cm), then our model predicts that weight will be increased by 0.614 (95% CI: 0.546 to 0.682) kg. You will note that if height = 0, the weight is -34.7 kg, which is impossible, and this again emphasises that the relationship defined by the regression equation only applies across the range of data from which it has been derived.

A poor model will have a regression coefficient (slope) close to zero for the explanatory variable. A regression coefficient of zero means that (a) a unit change in the explanatory variable results in no change in the predicted value of the outcome, and the gradient of the regression line is zero, implying that the regression line is flat. The coefficients table also provides the result of the hypothesis test (t -test) of whether the slope of the regression differs from zero, with the value of t and probability in the last two columns, respectively. This shows that the probability that this coefficient could have arisen by chance is <0.0005 . This is very unlikely, and we can reject the null hypothesis of no association and conclude that height makes a significant contribution to predicting weight.

Finally, the coefficients table also shows standardised coefficients, which may be used when the model includes more than one explanatory variable. These are the β coefficients (slopes) for each explanatory variable that have been standardised (by using the respective values of variance), so that the effect sizes of two or more such variables (which are likely to have different units) can be directly compared. SPSS produces these values in the Coefficients table, but a number of other common software packages no longer do so.

Summary: Simple Linear Regression

- An approximately linear relationship between two variables with a continuous outcome can be summarised by the equation of a straight line.
- The most commonly used way of choosing a line that is close to the observed data is least squares regression.
- We use the regression equation to
 - a. describe the relationship
 - b. predict values of the outcome variable
 - c. adjust values of the outcome variable for the effects of the explanatory variable
- The regression equation is only valid within the range of the observed x (explanatory variable) data.
- To test the null hypothesis that the outcome and explanatory variables are not related, we compare b/SE with the t distribution on $n - 2$ degrees of freedom (n is the number of subjects), and obtain a p -value.
- We can measure the improvement due to the model by calculating the r^2 (R^2 in SPSS tables) value, which tells us the proportion of variation in the outcome variable that is explained by the explanatory variable(s).
- We can also measure the goodness of fit of a linear regression model by calculating the F -ratio, a summary of how much our model has improved the prediction of the dependent variable over the most basic model (the mean of the dependent variable) and obtain a p -value.

5.9 Introduction to Multiple Linear Regression

5.9.1 Principles of Multiple Regression

We have seen that we can use simple linear regression to summarise the relationship between an outcome variable and a single explanatory variable. However, as we know from our discussion of confounding and the BRHS cohort studies, things are rarely so simple. The value of an outcome variable is usually affected by many factors, as we noted in the example of birthweight and gestational age: Gestational age is only one of the factors associated with birthweight.

To summarise the simultaneous relationship between a continuous outcome variable and a number of explanatory variables, we need to use *multivariable linear regression*. The equation that allows us to study the effect on the outcome variable of a number of explanatory variables is really just an extension of that for simple (one variable) linear regression, and is as follows:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

where x_1 is the first explanatory variable, x_2 is the second, and so on. When we have more than one explanatory variable, however, we can no longer think of the regression equation as a straight line: it is a mathematical description (a model) of the average relationship. However, in mathematical terms, this is still a linear relationship – that is, we are not allowing for any non-linearity in the relationships between the explanatory and outcome variables. For example, we may wish to explore how birthweight is related to gestational age *and* amount smoked during pregnancy. The regression equation is:

$$\text{birthweight} = a + (b_1 \times \text{gestational age}) + (b_2 \times \text{amount smoked})$$

with each variable measured in suitable units, such as weeks for gestational age, number of cigarettes smoked per day for amount smoked, and grams for birthweight. The beta coefficients (b_1 and b_2), which define the relationships between the explanatory and outcome variables, do not change across the full range of values of gestational age and amount smoked, so these relationships can be considered linear. There are other forms of regression that do accommodate non-linear relationships, but these are not considered further in this book.

5.9.2 Using Multivariable Linear Regression to Study Independent Associations

Multivariable regression allows us to look at relationships between variables, allowing for the effects of other variables, as illustrated in the example above. In this way, we can investigate whether a risk factor (which we term an explanatory variable in regression analysis) is *independently* associated with an outcome, after allowing for the effects of a confounding factor. In order to illustrate how multivariable regression works in practice, we will look at another example using SPSS. This involves both simple (only one explanatory variable) and multivariable analysis, and it will help to familiarise you with the interpretation of the output.

For this example, we will again use data from the low back pain dataset, which includes 775 male and female employees aged 18–75 years working in a variety of manual occupational settings in northwest England.

5.9.3 Investigation of the Effect of Work Stress on Bodyweight

For this example, we are interested in identifying the relationships between a range of physiological variables and workplace exposures (the explanatory variables) and recorded weight

in kilograms (the outcome variable). In particular we would like to measure the independent effect of stress in the workplace on bodyweight. Let us hypothesise that stress may increase body weight, though of course there may be a variety of mechanisms by which this might happen – for example, stressed employees may eat more food or drink more alcohol. The variables in the low back pain data set that are of interest for this investigation include one outcome variable and four explanatory variables.

Outcome Variable

- *Weight* (kg) – continuous data.

Explanatory Variables (Including Potential Confounders)

- *Stressful work* (4-point scale based on proportion of work shift experienced as stressful) – this is an ordered categorical variable (ordinal) but will be treated here as a continuous variable in the linear regression (the increasing scale is proportional). In Chapter 6 we look at an alternative method for including variables with more than two categories in regression, using dummy variables based on the response categories.
- *Height* (cm) – continuous data.
- *Age* (years) – continuous data.
- *Sex* – a dichotomous (two values, M or F) categorical variable.

Is There a Univariate Association?

Before constructing a multivariable model, we need to examine the association between the explanatory variable of interest (stressful work) and the outcome variable (weight in kilograms) by carrying out univariate (one explanatory variable) analysis by simple linear regression. The main SPSS output for this association is shown in the ANOVA and coefficient tables, shown in Tables 5.9.1 and 5.9.2.

Table 5.9.1 The ANOVA table (SPSS) for univariate regression analysis of the relationship between work stress and body weight (the footnotes are part of the SPSS output).

ANOVA ^a						
	Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	745.135	1	745.135	4.884	.027 ^b
	Residual	111829.214	733	152.564		
	Total	112574.348	734			

^aDependent Variable: weight in kg.

^bPredictors: (Constant), is work stressful.

Table 5.9.2 The Coefficients table (SPSS) for univariate regression analysis of the relationship between work stress and body weight (the footnote is part of the SPSS output).

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients		95% Confidence Interval for B		
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	66.204	.784		84.483	.000	64.666	67.743
	is work stressful	1.168	.528	.081	2.210	.027	.130	2.205

^aDependent Variable: weight in kg.

ANOVA Table

As we discussed earlier, the ANOVA table gives an estimate of the *goodness-of-fit* of the model. The most important part of this table is the *F-ratio*, a measure of how much the model has improved the prediction of the outcome (please refer back to Section 5.8.7 if you need to refresh your understanding of this). We previously saw that the *F-ratio* is calculated by dividing the mean squares for the model (MS_M , in this case 745.135) by the residual mean squares (MS_R , in this case 152.564). In our simple model investigating the effect of stress on weight, therefore, the *F-ratio* is 4.884, which is significant at $p = 0.027$. This result tells us that there is less than a 5 per cent ($p < 0.05$) probability that an *F-ratio* this large would happen by chance alone, and the model has significantly improved prediction.

Coefficients Table

The coefficients table provides details of the model parameters, that is, the beta values (slopes of the regressions) and their statistical significance. For stress, the beta coefficient is 1.168 (95% CI: 0.130–2.205), with a significant p -value of 0.027. Thus, as the explanatory variable (stress) is increased by one unit, our model predicts that the outcome (weight) will be increased by 1.168 kg. We can therefore conclude that in a *univariate analysis*, stress is significantly associated with weight; the longer employees report working in a stressful environment during a shift, the heavier their weight.

What About Confounding?

To sum up, we have shown with the univariate linear regression that stress at work is significantly associated with weight, and for each increase of one unit in the stress score, weight increases by 1.168 kg. This is an interesting finding, but it would be a mistake to immediately conclude that the relationship is causal. It seems likely that this association could – at least in part – be explained by confounding. For example, perhaps it is older people who are more prone to work stress, and weight does tend to increase with age. Or maybe men are more stressed in today's job market, and they are, on average, heavier than women.

To look at the *independent* effect of work stress on weight, adjusting for potential confounding, we can create a multivariable model by including the other explanatory variables in the regression analysis. Before doing this, we should consider the univariate associations of all these other variables with weight. A number of criteria can be used to determine whether or not a variable should be included in the multivariable model:

- It is a factor likely to have an important influence (e.g. sex of the subject) – this might be determined from previous experience and the published literature.
- In univariate analysis, it is significantly associated with the outcome below a specified threshold (e.g. $p < 0.1$).
- When added to the multivariable model, the variable reduces variance of the model by more than a specified amount (e.g. 5%).

The variables that we plan to include in the model are all significantly associated with the outcome (back pain), with $p < 0.05$; additionally, sex can be considered an important factor in its own right.

Briefly, then, our model will estimate the independent effects of each explanatory variable (work stress, height, age, and sex) on the outcome variable (weight in kilograms). The main SPSS output for this multivariable regression model is shown in the ANOVA and coefficients tables below.

ANOVA Table

We can see from the ANOVA table (Table 5.9.3) that the F -ratio for the model including the four explanatory variables is 101.278, which is again highly significant at $p < 0.0005$. This result tells us that there is (much) less than a 5 per cent chance ($p < 0.05$) that an F -ratio this large would happen by chance alone. Therefore, we can conclude that our regression model results in significantly better prediction of weight (kg) than if we used the mean value of weight alone.

Table 5.9.3 ANOVA table (SPSS) for multivariable regression of work and personal factors on body weight (the footnotes are part of the SPSS output).

ANOVA ^a						
	Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	40084.379	4	10021.095	101.278	.000 ^b
	Residual	71735.845	725	98.946		
	Total	111820.224	729			

^aDependent Variable: weight in kg.

^bPredictors: (Constant), height in cm, is work stressful, age, sex.

Coefficients Table

The coefficients table (Table 5.9.4) provides details of the model parameters (the beta values) for the four explanatory variables with their significance (p) values. Interestingly, following this *adjustment* for the effects of the other explanatory variables, the B (beta) value for stressful work is now considerably smaller, the gradient of the regression line being 0.665 (95% CI: –0.176–1.506) kg. This *adjusted regression coefficient* means that if stress is increased by one unit, our model predicts that weight will be increased by 0.665 kg. We can also see from the table that the p -value for stressful work has increased to 0.121 (the result is no longer statistically significant), whereas previously it was $p = 0.027$.

This result illustrates the effects of *confounding* on the relationship between stress and weight. As we suspected, one or more of the other three variables (height, age, and sex) explain

Table 5.9.4 Coefficients table (SPSS) for multivariable regression of work and personal factors on body weight (the footnote is part of the SPSS output).

Coefficients ^a							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	–9.141	9.232		–.990	.322	–27.265	8.984
is work stressful	.665	.428	.046	1.553	.121	–.176	1.506
sex	–6.119	1.062	–.234	–5.764	.000	–8.203	–4.035
age	.176	.031	.174	5.698	.000	.115	.236
height in cm	.477	.046	.425	10.289	.000	.386	.569

^aDependent Variable: weight in kg.

much of the observed univariate association between stress and weight by their mutual association with both factors, though we would have to refer back to the univariate associations to understand which factor, or factors, are most important.

Confidence Intervals (CIs) for the Regression Coefficients

Note that regression beta coefficients, including adjusted regression coefficients, should be quoted with 95 per cent CIs in the same way that we have emphasised previously for means, proportions, relative risks, and so on. You have seen that these are included in the coefficients tables, but they can also easily be derived from the standard errors provided in the same table. Thus, for stress, the 95 per cent CI is

$$\text{Beta (0.665)} \pm 1.96 \times \text{standard error (0.428)} = -0.174; 1.504$$

These figures differ very slightly from those in Table 5.9.4, but this is simply due to rounding errors arising from the value of the standard error (with three decimal places) we have used for the calculation. The results indicate that we are therefore 95 per cent confident that (using the figures from the SPSS output), after allowing for the effects of height, age, and sex in the population that this sample represents, for every increase in one unit of the stress-at-work score, body weight ranges from a decrease by 0.176 kg to an increase of 1.506 kg. The 95 per cent CI includes zero, so the result is not statistically significant. This is consistent with the p -value being greater than 0.05.

5.9.4 Multiple Regression in the Cancer Study

We will now return to the BRHS study of physical activity and cancer risk (Paper A) to see how multiple regression was used to investigate the *independent effect* of physical activity on the risk of cancer after adjusting for a number of other factors. The method used is known as Cox regression, a type of multivariable regression commonly employed in prospective studies where information is available on the time until an event (the cancer) occurs. We will study Cox regression in more detail in Chapter 8 on survival analysis, but for now we can interpret the results of analysis as described in Sections 5.9.1 to 5.9.3 for multivariable linear regression. Table 2 from Paper A is reproduced again below as Table 5.9.5 (initially presented in Table 5.5.1).

Table 5.9.5 Physical activity and age-adjusted cancer rates/1,000 person-years and adjusted relative risks for all cancers (excluding skin cancer) in 7,588 middle-aged British men.

Physical activity	No.	Cases	Rates/1,000 person-years	Relative risks (95% CI)	
				A	B
None/Occasional	3017	424	8.4	1.00	1.00
Light	1749	234	7.8	0.93 (0.79, 1.09)	0.95 (0.80, 1.11)
Moderate	1196	163	8.1	0.95 (0.80, 1.14)	1.04 (0.86, 1.24)
Moderate-vigorous	1113	101	5.8	0.68 (0.54, 0.84)	0.78 (0.62, 0.97)
Vigorous	513	47	5.7	0.65 (0.44, 0.88)	0.76 (0.56, 1.04)
<i>Test for linear trend</i>				$p < 0.0001$	$p = 0.02$

A = age-adjusted.

B = adjusted for age, cigarette smoking, BMI, alcohol intake, and social class.

Source: Wannamethee 2001. Reproduced with permission of British Journal of Cancer.



Self-Assessment Exercise 5.9.1

In Exercise 5.5.1, we compared the incidence rates of cancer among men taking vigorous exercise with those taking none/occasional exercise, and we found this to be 0.68. We showed that this was the relative risk and noted this was not adjusted for any confounding factors. We now look at the results (relative risks) following adjustment with multivariable regression (Cox regression in the case of this study) shown in columns A and B in Table 5.9.5.

1. Why are the relative risks for none/occasional exercise equal to 1.00?
2. What is the relative risk (and 95 per cent CI) for vigorous exercise after adjustment only for age, shown in column A? Interpret this result.
3. How does adjustment for all five potential confounding factors, shown in column B, affect the results, including the 95% CIs? Does this alter your conclusion?
4. Is there any additional information in this table that might lead you to conclude that the association between exercise and reduced cancer risk is causal?

Answers in Section 5.10

5.9.5 Overview of Regression Methods for Different Types of Outcome

Our primary aim in this chapter has been to introduce linear regression methods, firstly with one explanatory variable (simple linear regression) and then for more than one explanatory variable (multivariable linear regression), as shown in Table 5.9.6. We have begun with linear regression because this is intuitively the easiest type of regression to understand, and we have illustrated the method with data from the back pain dataset. You have seen that this method is used where the outcome variable is continuous, for example, weight in kilograms.

Table 5.9.6 Summary of the most commonly used types of regression, the nature of the outcome variable for which these methods are applied, and examples of each.

Type of regression	Outcome variable	Examples
Linear regression (This chapter)	Continuous data (approximating to normal distribution)	The back pain score as used in our examples
Logistic regression: (Chapter 6)	Dichotomous	Used for case–control studies as described in Chapter 6, where the outcome is ‘being a case’ or ‘not being a case’ of a disease or other condition
Poisson regression	Rare/count data	A study of risk factors for incidence of a rare cancer
Binomial regression	Dichotomous (success/failure)	Outcome of a trial of smoking cessation therapy assessed by whether or not subjects quit smoking
Cox regression (Chapter 8)	Dichotomous (time to event)	Outcome of a trial of a new treatment for cancer (not a very rare type), expressed as time to recurrence of the cancer

The published papers we have studied in order to learn about cohort studies, however, have introduced several other regression techniques, including Cox regression (Paper A) and Poisson regression (Paper D), and for now we have suggested that you can interpret the results of these techniques in a similar way.

Although the interpretation of results of different regression methods is similar in general terms, it is very important that you understand the reasons for applying the various methods available to different types of data. These will be explained in subsequent chapters, when we look at logistic regression for categorical outcomes in case-control studies (Chapter 6), and Cox regression for survival analysis where the key characteristic is time to an event (e.g. death) in Chapter 8. These outcomes differ in key respects from continuous outcomes, such as weight (kilograms), for which linear regression was used.

Outcomes also differ in the way that the variable concerned is distributed. This topic is addressed in more detail in Chapter 11 (Section 11.1.3), but it is briefly considered here to help you understand why, for example, Poisson regression was used in Paper D, and circumstances in which other probability distributions such as the binomial may be used.

The regression examples discussed so far have assumed an approximately normal distribution for the outcome variable and are an application of linear regression. For variables that are not distributed normally (or cannot reasonably be approximated to the normal distribution), the linear assumption required for linear regression is not valid, and other types of regression can be used to accommodate these different distributions. Two common types of distribution requiring different regression methods include Poisson and binomial distributions.

The Poisson distribution is used for rates based on events that are relatively rare and occur randomly over time, for example, an unusual form of cancer or serious but uncommon adverse effects of a drug treatment. The binomial distribution is used for situations where we are interested in whether an event has occurred or not, such as the success (or not) of a group of individuals in giving up smoking.

The basic principles of regression analysis as described in this chapter apply to these alternative distributions, although the assumptions, regression equations, and outputs differ. We will not cover the application of these regression methods further in this book.

Summary: Multivariable Regression

- Multivariable regression is used to summarise the relationship between an outcome variable and a number of explanatory variables.
- Multivariable linear regression allows us to determine the relationship between a continuous outcome variable and a number of explanatory variables. These explanatory variables can be continuous or categorical.
- Regression coefficients should be quoted with the 95 per cent confidence interval.
- An important application of multivariable regression is to adjust the estimated effect of a risk factor on an outcome for the effect of other potentially confounding factors, and hence derive an estimate of the independent effect of the exposure of interest.
- Other types of linear regression are commonly used (in simple or multivariable forms), including logistic regression, where the dependent (outcome) is categorical (e.g. presence or absence of disease), and Cox regression, where the outcome is categorical and information is available on time until the event occurs. These methods are studied in Chapters 6 and 8, respectively.

- Where the normal distribution cannot reasonably be used to summarise the distribution of an outcome variable, and linear assumptions are not valid, alternatives including Poisson and binomial may be used as the basis for regression analysis. We do not cover these regression methods further in this book, although more information about these distributions is provided in Chapter 11.

5.10 Answers to Self-Assessment Exercises

Section 5.1

Exercise 5.1.1

Research issues identified	Appropriate research method
There was increasing evidence that physical activity is associated with altered risk of total cancers and certain types of specific cancer.	Descriptive epidemiology and survey methods would probably be inadequate. Thus, as risk factors are thought to have an effect over an extended period of time, a survey would need to rely on recall of exercise patterns extending back many years.
Data on physical activity and a number of types of cancer is limited. Evidence at the time was not clear about the amount and type of physical activity required to alter the risk of cancer.	To address these issues, there is a need for good-quality, well-quantified evidence on risk-factor effects. Prospective data collection on exercise levels (well) before the disease develops provides the best means of achieving this, and a cohort study design is well suited to this requirement.
<i>Specified aims for BRHS cancer study</i> To examine the relationship between physical activity and incidence of total cancers and some site-specific cancers, and to assess the type and amount of activity required to achieve benefit.	A population-based cohort study is the method of choice, at least for common cancers and (as in this case) all types of cancer combined. We will see in Chapter 6 that for rarer conditions, the case-control study design is usually more appropriate.

Section 5.2

Exercise 5.2.1

1. Middle-Aged British Men.
2. There are probably several reasons for this. The original aim of the study was to investigate cardiovascular risk factors, and at the time the study was done, attention was being focused on heart disease in men rather than women since, at any given age, the incidence rates were higher for men. A very practical reason, which follows on from the last point, is that since the disease was less common in women than men (at the same age), a substantially larger sample would have been required to yield enough cases for useful analysis; that is, to achieve the necessary statistical power. However, it is worth noting that at the time the original study was carried out, many epidemiological studies were commonly carried out using male samples, based on the assumption that the findings could equally be applicable to women. We now know that this is not necessarily true, as risk factors may operate differently in a female population.
3. The sampling frames used were the updated age-and-sex registers of the selected general practices in the 24 towns.

4. Some of the points to consider are the following:

- The practices were located in medium-sized towns, chosen purposefully according to pre-specified criteria to cover all regions of the country, but excluding rural and major urban populations.
- The practices were selected from among all those in the town on the basis of certain criteria, and these may have favoured better-organised practices.
- There were exclusions (of individuals), but these were relatively few (6–10 per practice: 1.4 per cent to 2.4 per cent of the men selected from each practice population).
- From the randomly selected sample, 78 per cent responded. This would be considered a high response rate.

Thus, although the sampling frame was not random (towns and practices were selected according to various criteria), it was intended to be representative, apart from excluding rural and major urban areas. From within each practice, the samples of men were selected randomly, exclusions were few, and the response rate was fairly high. A fraction of men didn't participate due to work commitments, which could have led to some bias towards men who did not work, although we don't know how many. Overall, however, this sample can be expected to be fairly representative of the population of middle-aged British men (this was assessed and confirmed in a subsequent publication from the BRHS study).

Section 5.3

Exercise 5.3.1

The following table summarises (question 1) the main components of blood pressure measurement and (question 2) the possible sources of error. Although blood pressure is now commonly taken with an automated device, many of these points still apply:

Component of procedure	Source of error
Person taking the measurements (known as the <i>observer</i>)	<ul style="list-style-type: none"> (a) Training and knowledge of correct procedure, especially interpreting the sounds (Korotkow sounds) heard through the stethoscope. (b) Expectations of blood pressure level, e.g. that this would – on average – be higher in an overweight person. (c) Awareness of the patient's disease state may influence the observations taken by the observer.
Machine used to take measurement	<ul style="list-style-type: none"> (a) Condition of machine, e.g., whether tubing leaks. (b) How well <i>calibrated</i> the machine is; how accurately the machine records the true pressure can lead to error.
Circumstances in which measurement made	Subject's blood pressure is affected by anxiety, posture, recent exercise, etc.

Question 3. We noted in Chapter 4 that measurement error can be systematic (bias) or random. This distinction is very important in terms of the effect the error can have on the data and in respect of what can be done to avoid or reduce the error. The table below identifies the ways the BRHS team attempted to minimise these sources of error and bias that could occur in measurement of blood pressure in the CHD arm of the study.

Type of error	Methods of minimising error in BRHS	Comment
Observer	(a) Diastolic taken at disappearance of sounds (phase V)	(a) This helps standardise readings, because diastolic (the pressure when the heart is between beats) is read by some people at phase IV (when sounds become muffled) and by others at phase V. There is usually about 5 mmHg difference, but it can be a lot more.
	(b) Training with audio tapes	(b) This type of training is a standard procedure.
	(c) Random allocation of observers to study subjects	(c) This allocation means that any observer bias will not, overall, be associated with important variables such as weight, alcohol consumption, etc.
	(d) Use of London School of Hygiene sphygmomanometer	(d) This sphygmomanometer was designed to prevent observers seeing the level of the mercury column (blood pressure level) until the procedure was completed. It also avoids the tendency to round levels to convenient numbers, e.g. 93 to 95, or 90 (this is called <i>digit preference</i>).
	(e) Adjustment for observer error in the analysis	(e) This involves statistical adjustment at the stage of analysis, and is referred to in Paper B.
Machine error	Calibration	The settings of the machine are checked and adjusted at regular intervals.
Circumstances of measurement	(a) Subject seated with arm on cushion	(a) This helps to standardise posture, make the subject comfortable, and keep the arm at roughly the same position relative to the heart – all of which affect the recorded blood pressure.
	(b) Blood pressure measured twice	(b) Successive readings tend to fall as the subject relaxes. The mean of two was used for analysis.

As noted above, the use of automated blood pressure devices avoid some, but not all, of these sources of error. This does, however, remain a useful illustration of sources of error in measurement (not all of which can be fully automated and objective) and the types of methods that can be used to avoid such errors.

Exercise 5.3.2

Death certification should have been traceable for all men who died. There is a possibility, however, that cases could have been missed if cancer had been diagnosed but the death was due to another underlying cause. It is also possible that where the cause of death was due to a different underlying cause, such as a heart attack, for instance, the person had early stages of cancer but it was undiagnosed.

Cancer registration should include all diagnosed cases of cancer. This component of the system should therefore have picked up all known cancer morbidity among the sample. The postal questionnaires were used as a back-up to find any cases that had been missed or any very recently diagnosed cases that were not yet included on the registry database. It is not clear whether the responses given were validated with GP or hospital records, however.

On balance, the system for ascertaining cases appears to have been thorough.

Section 5.4

Exercise 5.4.1

1. Mothers of hospital-born children in Pelotas in 1993; note that less than 1% of births occurred at home.
2. Four: at 1 and 4 years, a behavioural sub-study at 4 years, and at 10–12 years.
3. Through a combination of a school census, and a population census of around 100,000 homes, all of which were visited.
4. The overall figure of 87.5% of children seen at 10–12 years shows that, as a result of the detailed work carried out (100,000 home visits, etc.), follow-up was remarkably complete. Although there were some statistically significant differences in percentage followed up across categories of some variables, the actual differences were quite small and seem unlikely to have resulted in any important bias in the sample available for analysis at 10–12 years, with the possible exception of the trend showing that women with higher pre-pregnancy BMI were better represented in the follow-up sample.

Exercise 5.4.2

1. Paper A does not report that any additional information was collected on either exposure to risk factors (such as BMI and smoking) or on levels of exercise. Under the heading of ascertainment of cases, we are informed that three further questionnaires were sent out in 1992, 1996, and 1998 asking about cancer diagnosis, but there is no discussion of whether assessment was made of levels of risk factors during the follow-up period, so the reader can assume that this information was only collected at the initial screening. A 5-year follow-up questionnaire was, however, used to reassess some risk factors (changes in smoking, drinking, and economic status) for the heart disease component of the BRHS (see Paper A, Chapter 8, Section 8.1.3).
2. If those who took more exercise (when assessed at baseline) had been more likely to give up smoking during the follow-up period, smoking prevalence would have been relatively overestimated in the vigorous exercise group. The observed lower risk of cancer in the vigorous exercise group might then be at least partly due to a much larger true difference in smoking over the follow-up period than was apparent from the baseline data. We might note, however, that we would only expect to see this effect for cancers related to smoking (as you may see from Paper A, this does not appear to be the case), and there would need to be sufficient follow-up time for the growing difference in smoking rates to affect new cases of cancer.

Section 5.5

Exercise 5.5.1

1. Incidence of cancer in the 'none/occasional exercise' group = 8.4 per 1,000 person-years.
2. Incidence in the 'vigorous exercise' group = 5.7 per 1,000 person-years.
3. Incidence in the 'vigorous exercise' group is 0.68 times that in the 'none/occasional exercise' group, that is, 32 % lower.

Exercise 5.5.2

1. The relative risk (RR) for IHD among ex-smokers = incidence in exposed/incidence in unexposed = $6.66/2.36 = 2.82$. Similarly, the RR for smokers = $8.07/2.36 = 3.42$.
2. Interpretation: Smoking is clearly associated with an increased risk of IHD. What is surprising is that ex-smokers had a RR that was not much less than that of current smokers (2.82 vs. 3.42). This gradation of risk from never-smokers through ex-smokers to current smokers

is an example of a *dose–response* relationship in which the greater the exposure to the risk factor (dose), the greater the effect on outcome (response). This finding is one of a number of indications that an association may be causal, and we return to this in Section 5.7.

3. The words we have used so far have been chosen carefully. Thus, on the basis of the BRHS data studied so far, we have said that smoking is *associated* with an increased risk of IHD, but we have not said it definitely causes it. We can see that the RR is quite large (almost 3.5 for smokers), and also that there is some evidence of a *dose–response* relationship, both of which are suggestive of causation. Apart from chance effects, however, there is one other important explanation of an observed association such as this. What if people who smoke more also tend to have higher cholesterol, take less exercise, and have higher blood pressure? Is it possible that it is these factors, and not the smoking, that are contributing to the IHD? This process is called *confounding*, which we look at in more detail in Section 5.7.

Exercise 5.5.3

1. The group with the lowest level of exercise (none/occasional exercise) was taken as the reference group.
2. We do not have to use this group as the reference (any group could be used), but this choice seems entirely appropriate as we can see how increasing exposure to exercise is associated with a reduced risk of cancer. In addition, it is the largest group, so it provides a more-precise reference for the comparisons.

Exercise 5.5.4

1. Cancer is the outcome variable and exercise is the explanatory variable.
2. The population of interest is all middle-aged British men.

Exercise 5.5.5

1. In Table 5.5.3, the variable *exercise* can take only the values ‘none/occasional’ or ‘vigorous’ exercise. The variable *cancer* can take the value ‘case’ and ‘non-case’. They are both *categorical* variables. The physical index variable shown in Table 5.5.1 is an *ordered categorical* variable, with values that range from none/occasional exercise to vigorous exercise.
2. and 3.

Exercise category	Cancer				Total
	Cases		Other men		
	No.	%	No.	%	
None/occasional	424	14	2593	86	3017
Vigorous	47	9	466	91	513
<i>Total</i>	471	13	3059	87	3530

4. We can see that the percentage of ‘vigorous exercisers’ who developed cancer (9 per cent) was less than the percentage of ‘none/occasional’ exercisers who developed cancer (14 per cent). We will conduct an hypothesis test to determine whether this difference is because exercise and cancer are (inversely) associated, or whether it is just the result of chance – due to sampling error.

Exercise 5.5.6**Table 5.5.4** Physical activity and cancer: expected frequencies.

Physical activity	Cancer		
	Cases	Other men	Total
None/occasional	402.55	2,614.45	3,017
Vigorous	68.45	444.55	513
<i>Total</i>	471	3,059	3,530

Exercise 5.5.7**Table 5.5.5** Physical activity and cancer: residuals.

Physical activity	Cancer		
	Cases	Other men	Total
None/occasional	21.45	-21.45	0.00
Vigorous	-21.45	21.45	0.00
<i>Total</i>	0.00	0.00	

Exercise 5.5.8**Table 5.5.6** Physical activity and cancer chi-squared table.

Physical activity	Cancer		
	Cases	Other men	Total
None/occasional	1.14	0.18	1.32
Vigorous	6.72	1.04	7.76
<i>Total</i>	7.86	1.22	9.08

Adding all four values in the cells ($1.14 + 0.18 + 6.72 + 1.04$), we obtain 9.08, this being the value of the chi-squared statistic.

Section 5.7**Exercise 5.7.1**

Model 1: From the information given in this model, heavy alcohol consumption could confound this association. Alcohol is of course a very well recognised and important cause of liver cirrhosis (fibrosis and damage to liver tissue), whereas coffee is not (indeed it may reduce the risk of liver disease, including cancer). Any observed association between coffee and liver cirrhosis is therefore very likely to be the result of confounding by alcohol consumption.

Model 2: From the information given in this model, cigarette smoking cannot confound the radon–lung cancer association. Although there is no question that smoking causes lung cancer, it is not associated with radon exposure; that is to say, we do not have information to suggest that people living in radon-exposed homes smoke, on average, any more than people living in homes that are not exposed to radon.

Model 3: This is a rather more complex situation, but such is life! Damp housing does cause respiratory illness, but that is not the whole story. People living in damp houses are in general poorer and through this are exposed to many other influences that can cause respiratory illness. These would include higher rates of smoking, work exposure, environmental pollution from traffic and industry, poorer nutrition, and so on. This is a situation where the observed association between damp housing and respiratory illness can be explained in two ways:

- In part a direct causal effect through mould, cold, damp air, etc.
- In part due to the association with poverty and the lifestyle, environmental, and employment factors that come with that.

This situation is all too common in health research, and it provides some insight into why establishing causation requires careful study design and interpretation.

Section 5.8

Exercise 5.8.1

1. The predicted values of birthweight are 2.79 kg for 37 weeks and 3.31 kg for 41.5 weeks.
2. These values should (and do) lie on the regression line.
3. It is not appropriate to use this regression relationship to calculate the birthweight of a baby of 28 weeks' gestation because we should never use regression to go beyond the range of the observed x variable data (in this case the x variable is gestational age).

Section 5.9

Exercise 5.9.1

1. The none/occasional exercise group is taken as the reference category, so the relative risk is quoted as 1.00. Note that this does not mean this level of exercise is associated with no risk of cancer: It simply means we can now compare other groups with this group, using the values of relative risk obtained for these from the regression analysis. Selecting this group as the reference is a sensible choice, as it makes the protective effect of higher levels of exercise easy to comprehend. A second, important reason for choosing the none/occasional group as the reference is that it is the largest, so that estimates of risk in this group are the most precisely estimated of all the groups.
2. The RR is 0.65, with a 95 per cent CI of 0.44 to 0.88. Thus, the estimated effect is a 35 per cent reduction in risk, and we can be 95 per cent confident that the population risk is reduced by 12 to 56 per cent. This interval does not include 1.00, so the finding is statistically significant at the 0.05 level. The effect of adjustment for age is small, reducing the relative risk estimate from 0.68 to 0.65.
3. Adjustment for all five potential confounding factors has more effect, attenuating the relative risk estimate to 0.76, equivalent now to a 24 per cent reduction in risk. The 95 per cent CI includes 1.00, and it ranges from a 44 per cent reduction in risk to a 4 per cent increase in risk; hence, it is not significant at the 0.05 level. Thus, adjustment for all these other factors does suggest that part of the effect we originally observed was due to confounding. Although the result for vigorous exercise does not quite reach significance, we can see that the adjusted

estimate for moderately vigorous exercise (which has larger numbers of subjects and hence more power) is similar at 0.78, and the 95 per cent CI does not include 1.00. Overall, these results suggest that the finding of a protective effect of vigorous exercise is independent of confounding, and this is evidence in favour of causality.

4. The statistically significant test for trend provides evidence of a dose–response relationship, further evidence that the association is likely to be causal. The hypothesis test used (chi-squared for linear trend) is discussed further in Chapter 11.

6

Case–Control Studies

Introduction and Learning Objectives

In Chapter 5, we noted that there are two commonly used, analytic, observational study designs, and we have studied the first of these – cohort studies. In this chapter we look at the second design: the case–control study. Although more commonly used than the cohort design, case–control studies are in some respects more complex in terms of methods and interpretation. Your familiarity with cohort studies will help you in understanding the strengths and limitations of case–control studies.

Learning Objectives
<p>By the end of this chapter, you should be able to do the following:</p> <ul style="list-style-type: none"> • Describe the purpose and structure of a case–control study design within an overall framework of epidemiological study designs. • Give examples of the uses of case–control studies. • Describe the strengths and weaknesses of case–control studies. • Describe the most important types of bias that can arise with case–control studies and how these can be minimised. • Describe the purpose of using multiple controls in a case–control study. • Describe the information required to calculate sample size and interpret the output of sample size calculation using Open Epi webtools. • Describe how confounding can be avoided through matching. • Describe the odds ratio (OR) as used in the analysis of a case–control study, and its relationship with relative risk. • Calculate an OR in simple unmatched and matched case–control study analyses. • Describe the approach used to calculate the OR in unmatched and matched case–control studies with multiple controls. • Describe the various approaches to dealing with confounding in the analysis of a case–control study, including stratification and multivariable logistic regression. • Describe the use of, and principles underlying, multivariable logistic regression, including the appropriate circumstances for use of unconditional and conditional methods. • Calculate the adjusted OR from the logistic regression coefficient, and interpret the OR and 95 per cent confidence interval (CI).

We explore case-control studies using material from the following paper. This describes a case-control study investigating the association between maternal sleep practices and risk of late stillbirth, that is, at or after 28 weeks' gestation.

Resource Papers

Paper A

Stacey, T., Thompson, J., Mitchell, E., Ekeroma, A., Zuccollo, J., McCowan, L. (2011). Association between maternal sleep practices and risk of late stillbirth: a case-control study. *BMJ* 342, d3403.

Please now read the abstract for Paper A given below.

Abstract

Objectives To determine whether snoring, sleep position, and other sleep practices in pregnant women are associated with risk of late stillbirth.

Design Prospective population-based case-control study.

Setting Auckland, New Zealand.

Participants Cases: 155 women with a singleton late stillbirth (≥ 28 weeks' gestation) without congenital abnormality born between July 2006 and June 2009 and booked to deliver in Auckland. Controls: 310 women with single ongoing pregnancies and gestation matched to that at which the stillbirth occurred. Multivariable logistic regression adjusted for known confounding factors.

Main outcome¹ measure Maternal snoring, daytime sleepiness (measured with the Epworth sleepiness scale), and sleep position at the time of going to sleep and on waking (left side, right side, back, and other).

Results The prevalence of late stillbirth in this study was 3.09/1000 births. No relation was found between snoring or daytime sleepiness and risk of late stillbirth. However, women who slept on their back or on their right side on the previous night (before stillbirth or interview) were more likely to experience a late stillbirth compared with women who slept on their left side (adjusted odds ratio for back sleeping 2.54 (95% CI 1.04 to 6.18), and for right side sleeping 1.74 (0.98 to 3.01)). The absolute risk of late stillbirth for women who went to sleep on their left was 1.96/1000 and was 3.93/1000 for women who did not go to sleep on their left. Women who got up to go to the toilet once or less on the last night were more likely to experience a late stillbirth compared with women who got up more frequently (adjusted odds ratio 2.28 (1.40 to 3.71)). Women who regularly slept during the day in the previous month were also more likely to experience a late stillbirth than those who did not (2.04 (1.26 to 3.27)).

Conclusions This is the first study to report maternal sleep-related practices as risk factors for stillbirth, and these findings require urgent confirmation in further studies.

¹Note: The authors use the title 'Main outcome measure' for what are in fact 'exposures/risk factors'. The outcome in this case-control study is late stillbirth.

6.1 Why do a Case–Control Study?

6.1.1 Study Objectives

We will begin our examination of case–control studies by establishing why the authors of this New Zealand study chose this design to address their research question. This is described in the introduction section of Paper A, reproduced below.

Introduction

The death of a baby before birth is a tragedy for the family and wider community. In high-income countries more than one in 200 births result in a stillbirth. Stillbirth therefore remains an important public health issue, with little change in its rate over the past two decades. Many studies have examined risk factors for stillbirth, but they have often been population-based retrospective studies that have been unable to explore a broad range of potential risk factors, in particular those relating to maternal lifestyle and personal habits.

Around a third of a person's life is spent asleep, but there has been little research on the potential impact of sleep practices on the developing foetus. Previous studies have reported an association between sleep disordered breathing and pregnancy complications such as pre-eclampsia and preterm birth, but exploration of a potential association with stillbirth has been limited to a single case report. We and others have described a dose-dependent relation between maternal obesity and stillbirth risk, but the mechanisms underlying this association are not understood. Obesity is also associated with sleep disordered breathing. It is therefore possible that sleep-disordered breathing is one of the mechanisms underlying the association between obesity and stillbirth risk.

Supine sleeping position is associated with sleep disordered breathing and in late pregnancy has also been associated with reduced maternal cardiac output, but the impact of position during sleep and risk of stillbirth is not known. There have been no reports of other sleep related practices and risk of stillbirth.

The broad aim of the Auckland Stillbirth Study was to identify potentially modifiable risk factors for late stillbirth (≥ 28 weeks' gestation). We explored a range of factors relating to women's health and behaviour during pregnancy, including general health, socioeconomic factors, diet, exercise, and maternal sleep practices. We hypothesised that sleep disordered breathing and maternal supine sleep position would be associated with increased risk of late stillbirth. We also investigated the relation between risk of late stillbirth and other sleep-related practices, specifically regular daytime sleep, duration of sleep, and getting up during the night.



Self-Assessment Exercise 6.1.1

1. Make a list of all the research issues, including the limitations of previous studies, and the aim and hypothesis of the study noted in the Introduction to Paper A.
2. Why do you think the authors chose to carry out a case–control study instead of a cohort study?

Answers in Section 6.6

6.1.2 Study Structure

It is important to gain an overall understanding of the structure of a case-control study before we get into too much of the methodological detail. For the New Zealand study, this is described in the methods section, summarized below.

Study Subjects

Women who gave birth to a stillborn baby at or after 28 completed weeks of gestation in the Auckland region between July 2006 and June 2009 were invited to participate in the study. Stillbirth was defined as the birth of a baby that died in utero during the antenatal or intrapartum periods. Cases were ascertained weekly from key clinicians in the participating centres (all maternity units in Auckland region) and from hospital birth records checked on a regular basis (by TS). A national system for perinatal data collection started in New Zealand on the same date as recruitment began; cases were compared with this registry to ensure complete ascertainment.

Women were excluded if their baby had died from a congenital abnormality or was from a multiple pregnancy, or if they had not been booked to deliver their baby within the Auckland region (which consists of three district health boards). Two controls were randomly selected from the pregnancy registration list of the district health board in which the stillbirth occurred, with the same exclusion criteria as the cases. Controls were matched to cases by gestation, thus ensuring that the controls were representative of the antenatal population at the same gestation at which the stillbirth occurred.

Collection of Data

Data were obtained through interviewer administered questionnaires in the first few weeks after stillbirth. For the controls, interviews occurred at the equivalent gestation of pregnancy as that of the matched case. Participants were not aware of any of the specific research questions related to risk factors for stillbirth. As there are no validated tools for screening for sleep disordered breathing in pregnancy, we used snoring and daytime sleepiness as proxy indicators for sleep disordered breathing. Participants were asked whether they regularly snored before their pregnancy or during pregnancy. The Epworth sleepiness scale was used to determine the general level of daytime sleepiness.

Specific questions were asked about maternal sleep position both at the time of going to sleep and on waking. Sleep position was classified as left side, right side, back, and other ("other" included front, sitting up, both sides, and unsure or don't remember). The time periods for which data were collected were before the pregnancy and in the last month, week, and night of the pregnancy. The last night was the night before when the woman thought that her baby had died or, for the controls, the night before the interview.

Participants were also asked whether they regularly slept during the daytime in the last month. Further questions were asked about the usual duration of sleep at night during the last month and frequency of getting up to the toilet. The reference duration of sleep was defined as 6–8 hours at night, and sleep duration was therefore categorised as <6, 6–8, or >8 hours. Data were collected on frequency of waking in the night and of getting up to go to the toilet at night. A strong correlation was seen between these two variables, and therefore only getting up to the toilet at night is presented here.

Demographic data and information on other potential confounding factors were collected during the interview, specifically maternal age, ethnicity, parity, smoking status, body mass index at booking (first antenatal visit), and social deprivation level. Ethnicity was self assigned, and a single

ethnicity was applied based on a system of prioritisation as described by the New Zealand Ministry of Health. Smoking status was defined as having smoked at any time during the pregnancy. Maternal body mass index was calculated from the earliest known weight taken in pregnancy and from maternal height measured at interview. Social deprivation level was derived from the address at which the participant lived, based on the 2006 New Zealand census data, with category 1 being the least deprived and category 5 the most deprived. Detailed information about methodology has been reported previously.

There are essentially four key steps in carrying out a case–control study:

1. Identification of **cases**, that is, the people with the disease or outcome. In this study, cases were women in the Auckland region who gave birth to a singleton stillborn baby at or after 28 weeks' gestation.
2. Identification of **controls**, that is, people who do not have the disease or outcome. In this study, they were population controls, or women selected from the pregnancy registration list from a district health board in which the matched case (stillbirth) occurred. We will look at the technique of matching later.
3. Measurement of potential risk factors for stillbirth among the cases and controls. In this study, the risk factors of interest were sleeping position at the time of going to sleep and waking, maternal snoring, and daytime sleepiness. A range of other potential risk factors (including possible confounding factors) were also measured.
4. Analysis of whether or not the **cases** are more likely to have been **exposed** to a risk factor than the **controls**. In this study, we would want to see whether cases (those experiencing a late stillbirth) were more likely to sleep in a position other than on their left side compared to controls. Note: The authors explain in the discussion that sleeping in a supine (lying on back, face up) or right lateral position (compared to left lateral position) inhibits venous return, decreasing uterine blood flow through pressure exerted through the enlarged uterus. Therefore, this could be a mechanism by which sleeping in a position other than on the left side could be associated with late stillbirth.

The basic structure of a case–control study, using the exposure and outcome from the New Zealand study, is illustrated in Figure 6.1.1.

To summarise, study participants were selected in the Auckland area of New Zealand and were at or after 28 weeks' gestation between July 2006 and June 2009. **Cases** were women who gave birth to a stillborn baby (defined as the birth of a baby that died in utero during the antenatal or intrapartum periods) and were identified through weekly checks by clinicians of all maternity units in the Auckland area and through regular hospital birth record checks. (**Case ascertainment** is considered in more detail in Section 6.2.1). **Controls** were pregnant women randomly selected from the three Auckland district health boards' birth registration lists. In this study, **matching** was carried out on duration of gestation. This is done in order to reduce the effect of confounding, and we will look at the why and how of matching in some detail in Section 6.2.2. Finally, we can see that more than one control was selected for each case (two controls were randomly selected for each case). This is done to increase the statistical power of the study, and it is considered further in Section 6.4 on sample size.

6.1.3 Approach to Analysis

We have said that the purpose of the analysis is to determine whether or not cases are more exposed to the risk factor(s) of interest than are the controls. This is illustrated diagrammatically

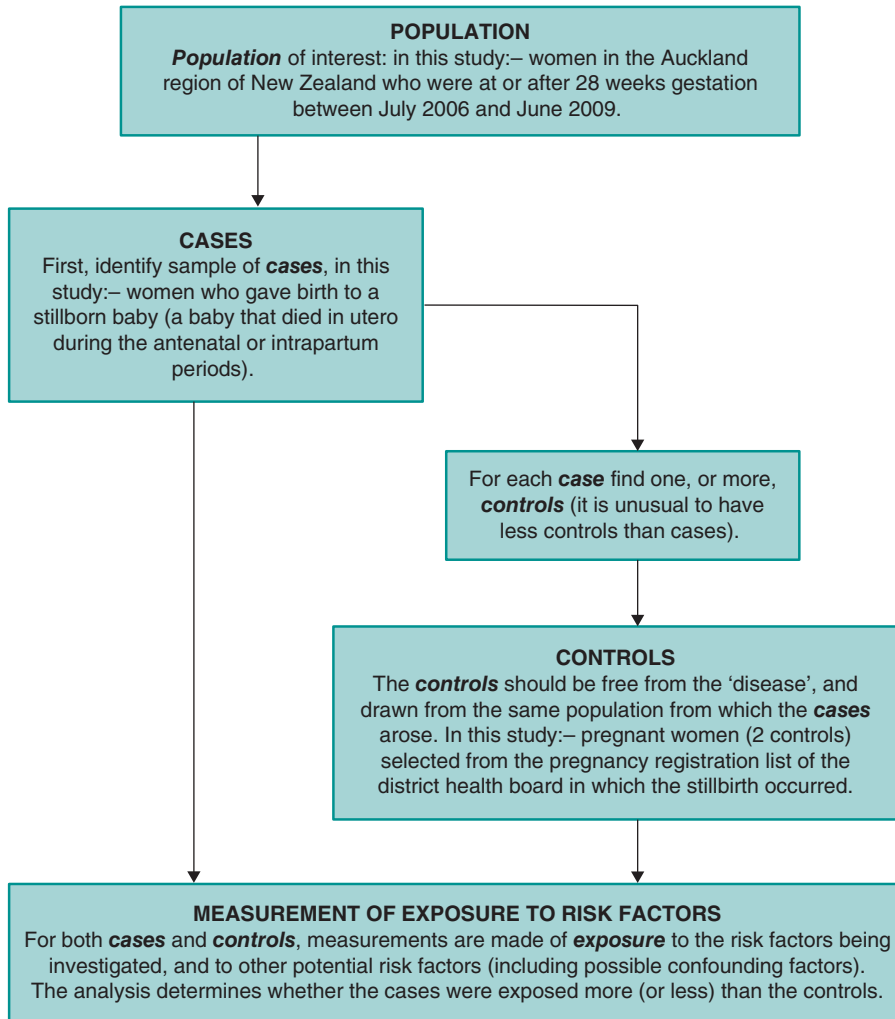


Figure 6.1.1 Overview of the general structure of a case-control study, with an example taken from the New Zealand study (Paper A).

in Figure 6.1.2. We will return to this concept in Section 6.3, when we look in detail at the analysis of case-control studies and how the odds ratio (*OR*) is used to express this difference between cases and controls in the chances (referred to as the 'odds') of being exposed. For the simplified example illustrated in Figure 6.1.2, we keep to the objective of Paper A (risk factors for late stillbirth), but for ease of illustration we restrict this explanation to just 30 cases and 30 controls.

In this example, each square represents a woman, those in the upper box being cases (women experiencing a late stillbirth), and those in the lower box, controls (pregnancy continues). A green shaded square means that the woman was exposed to the risk factor (sleeping on her back during pregnancy). A higher percentage of cases (50 per cent) were exposed to the risk factor than were controls (30 per cent). The analysis of a case-control study concentrates on quantifying this difference in exposure between cases and controls.

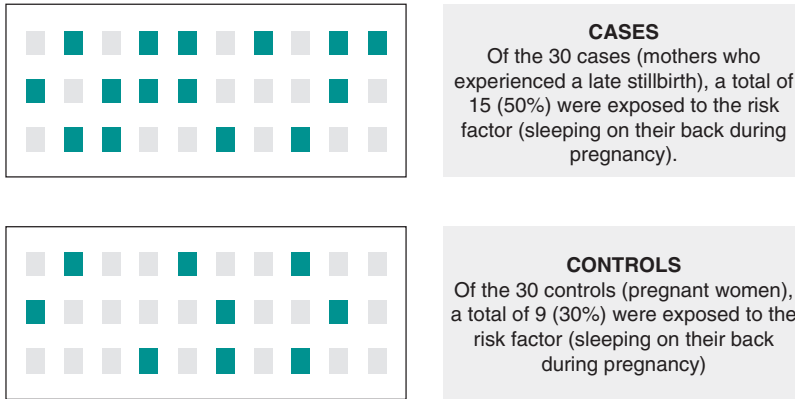


Figure 6.1.2 A hypothetical case–control comparison, investigating the association between exposure to a risk factor (sleeping on back during pregnancy, compared to sleeping on the left side) and late stillbirth.

6.1.4 Retrospective Data Collection

The term *retrospective* is often associated with case–control studies, just as the term *prospective* tends to be associated with cohort studies:

Retrospective means looking backwards in time. If cases are identified now, the exposure must have occurred in the past. We must look backwards in time to find out about that exposure, either by asking the people concerned or (if we are fortunate) finding sufficiently complete and valid records of that exposure.

Prospective, a term we have already mentioned in respect of cohort studies, means looking forward in time. If a survey measuring exposure is done now, we await the occurrence of cases in the future.

These concepts are summarised in Figure 6.1.3.

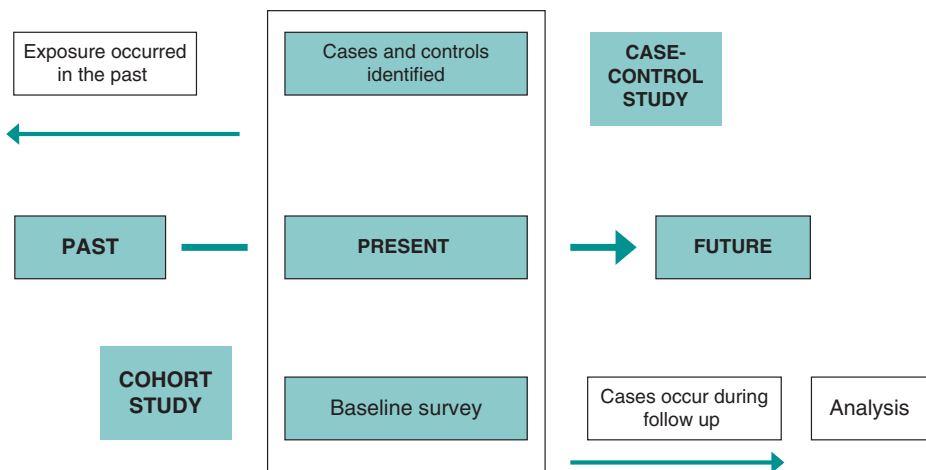


Figure 6.1.3 Time perspectives most commonly seen in case–control and cohort studies.
Note: These perspectives do not always apply as explained in the text.

You may find examples of case-control analyses being carried out on data that have been collected prospectively and cohort studies carried out retrospectively. For example,

- A case-control analysis can be carried out within a prospective cohort study, perhaps for an outcome other than the main one for which the cohort study has been set up. Thus, as cases of this other outcome occur during follow-up, these are selected (and become the cases for a nested case-control study), and controls are selected from other study subjects in the cohort who remain free from that outcome. Information on exposure, other risk factors, and confounders would be available from the baseline survey and potentially also during follow-up (but before the case arose).
- A cohort study can be carried out retrospectively if complete and accurate records exist of exposure, other risk factors, and confounders for a cohort of people, such as employees of a large company, some of whom have now developed the disease or outcome of interest. In effect, information on a baseline can be extracted from records, and since that information was recorded long before cases of disease appeared, there is no risk of bias from knowledge of who did or did not become cases (so long as data extraction is done blind to this knowledge). Information on exposure during the follow-up period can potentially also be extracted from records or by asking the subject, so long as one can be sure it occurred before the disease. Incidence rates for exposed and non-exposed can then be calculated as for a prospective cohort study.

However, most case-control studies are retrospective, and since some of the more important design issues we need to be concerned about arise from retrospective data collection, we will concentrate on this time perspective.

6.1.5 Applications of the Case-Control Design

The case-control design can be used in a wide variety of circumstances. Some of these include the following:

- Understanding the cause of a disease or outcome of interest; for example, risk factors for cancer, death from violent injury, or (as in our example) late stillbirth.
- Measuring the effect of procedures or interventions; for example, the effectiveness of breast screening programmes.
- Investigation of the causes of disease outbreaks; for example, food poisoning among a group at a catered event.

Case-control studies are the design of choice for health outcomes with a long latency period – that is, the time period between exposure and onset of the disease (most cancers) – or that are relatively rare (many specific types of cancer, again). For very rare conditions, case-control studies may be the only option, because cohort studies would be impractical, if not impossible, within any reasonable timescale or budget. The obvious benefits of case-control studies over alternative designs are the savings made in terms of cost and time. However, as we shall see later in this chapter, given the typically retrospective nature of the case-control study design, there are a number of important methodological issues that require careful attention in design and interpretation.

Summary

- The case-control design can provide strong, quantified evidence about the importance of risk factors or the effect of interventions, and it is more practical than a cohort study where the outcome is rare and/or has a long latency period.

- The relative speed and lower cost of carrying out a case–control study also makes it a useful design for initial investigation of a new hypothesis or one for which there is minimal evidence.
- A case–control study essentially involves
 - The identification of cases;
 - The selection of controls, which may be matched to a greater or lesser extent (discussed further in Section 6.3);
 - The measurement of exposure to risk factors of interest, including possible confounders;
 - Analysis of any difference in the odds of exposure between cases and controls.
- Collection of data on the characteristics of cases and controls, including the exposure to risk factors, which is usually done after the disease event has occurred. This is known as retrospective data collection, and it is one of the reasons case–control studies are more vulnerable to bias than cohort studies, where the events occur after the assessment of exposure and other variables (prospective data collection).
- There are circumstances in which assessment of exposure in a case–control study can be done prospectively, such as when it is nested within a cohort study.

6.2 Key Elements of Study Design

6.2.1 Selecting the Cases

Case Definition and Selection

We have noted that the first step in carrying out a case–control study is to identify the cases. As with cohort studies, it is very important to have a clear case definition and to describe carefully how cases were selected. The following section from Paper A describes how case ascertainment was done in the New Zealand study.

Women who gave birth to a stillborn baby at or after 28 completed weeks of gestation in the Auckland region between July 2006 and June 2009 were invited to participate in the study. Still-birth was defined as the birth of a baby that died in utero during the antenatal or intrapartum periods. Cases were ascertained weekly from key clinicians in the participating centres (all maternity units in Auckland region) and from hospital birth records checked on a regular basis (by TS). A national system for perinatal data collection started in New Zealand on the same date as recruitment began; cases were compared with this registry to ensure complete ascertainment.

Women were excluded if their baby had died from a congenital abnormality or was from a multiple pregnancy, or if they had not been booked to deliver their baby within the Auckland region (which consists of three district health boards).



Self-Assessment Exercise 6.2.1

1. Describe how cases were defined, noting also those women who were not included in the study.
2. Describe how and where cases were identified.
3. Comment on the overall quality of the case ascertainment.

Answers in Section 6.6

6.2.2 The Controls

Selecting the Control Group

Having identified a total of 155 women who had experienced a late stillbirth and who consented to participate in the study (over a three-year period), the next step was to select the controls. Selecting the most-appropriate groups of controls can be one of the most demanding aspects of a case-control study. We will look first at how this was done in the New Zealand study and then consider some general points about selecting controls.

The single most important principle in selecting controls is that the controls should be representative of the population from which the cases have arisen, but they should be without the disease or outcome in question. Let's see how this helps to understand the choice made in the New Zealand study.

The first step is to ask ourselves from what population the cases arose. The answer to this can be found in the excerpt from the study methods in Section 6.2.1 describing how the cases were ascertained. The research team was dealing with all (or virtually all) cases of late stillbirth for the Auckland region. The population from which these cases arose was therefore the population of Auckland, but with the qualification that the stillbirths were recorded at maternity clinics and/or hospitals in the Auckland region detected by key clinicians or the lead researcher. Although it can be expected that the maternity clinics and/or hospital would be notified of all late stillbirths, there is a chance that a few cases of stillbirth could have been missed by individuals reviewing the records, although there is no reason to suspect these would differ from those identified through surveillance.

Thus, we would be looking for a method that selected a group of control subjects representative of pregnant women from the Auckland population. Let's now look at the full description of how controls were selected for the New Zealand study.

Two controls were randomly selected from the pregnancy registration list of the district health board in which the stillbirth occurred, with the same exclusion criteria as the cases. Controls were matched to cases by gestation, thus ensuring that the controls were representative of the antenatal population at the same gestation at which the stillbirth occurred.



Self-Assessment Exercise 6.2.2

How well does the selection of controls meet the principle we specified, that is, they should be representative of the population from which the cases have arisen, but without the disease or outcome in question?

Answers in Section 6.6

Bias Arising from Control Selection

Whereas control selection might seem straightforward in the New Zealand study, this is not always the situation in case-control studies.

Let's consider another situation. A case-control study in a developed country is investigating lung cancer in men aged 40–65 years, and cases have been recruited from a hospital. Since this is a very serious disease in relatively young people, we can assume that virtually all cases in the population will end up in hospital, so although the cases are recruited from hospital, they are effectively representative of all cases in the population. Therefore, following our principle for

control selection, we should select controls to represent the population of males younger than 65 years. One approach is to select controls through general practice lists, probably matching on a few variables such as age and area of residence. It might appear counterintuitive to use general practice–based controls for cases selected as inpatients, but it does follow logically from the need to ensure that controls arise from the same population as cases.

If, on the other hand, hospital controls were sought, we would have to ensure that there were no health-care access issues that might prevent the controls being representative of the population, and we would have to ensure that the disease condition for which they were admitted was not related to lung cancer risk factors (e.g. smoking). Sometimes, if it is difficult to decide on which approach will deliver the most-appropriate controls, more than one control group can be used. If the results using (for example) hospital and general practice–based control groups differ markedly, further investigation would need to be made into which approach to control selection is the less biased.

We have mentioned matching on a number of occasions, and we now look at this aspect of case–control design in more detail.

Matching

You will recall the importance of **confounding** when it comes to interpreting associations between **risk factors** and **outcomes**. Figure 6.2.1 shows the model of confounding (introduced in Chapter 5). It has been adapted here to consider the possible causal effect of maternal sleeping position (sleeping on back) on stillbirth after 28 weeks of gestation and the role of confounding due to gestational age (matched for in the New Zealand study).

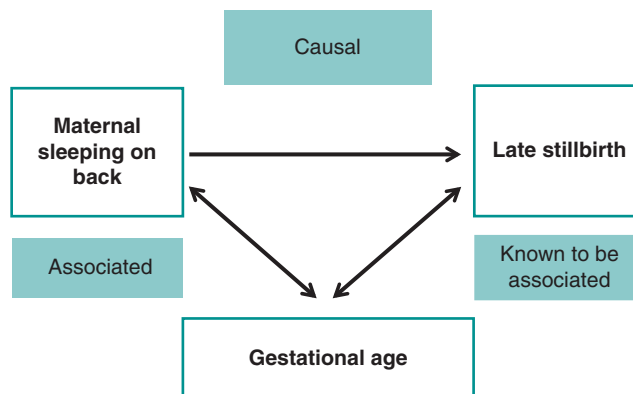


Figure 6.2.1 Diagram summarizing how gestational age could confound the relationship observed between maternal sleeping on the back and late stillbirth.

In this model, increasing gestational age is associated with an increased risk of stillbirth (it is well established that the risk of stillbirth varies with gestational age) and is associated with maternal sleeping on back (as the baby increases in size the pregnant mother might find it more comfortable to sleep on her back). If this is the predominant mechanism at work, it could be that gestational age is mainly responsible for any observed association between maternal sleeping on back and stillbirth after 28 weeks' gestation, and not the sleeping position itself.

Matching overcomes confounding by preventing it from occurring in the first place. This is done by making cases and controls similar with respect to potential confounding factors. In the New Zealand study, matching was not carried out specifically on maternal obesity (a risk factor

for stillbirth), as this would have been very impractical, and we will see how the authors dealt with this in the analysis. Matching was, however, done on gestational age, which can confound the relationship between sleeping position and late stillbirth as explained above.

By matching the cases and controls for gestational age, we know that any effect of this factor on sleeping position and risk of late stillbirth is the same in both cases and controls. Thus, let's say woman A (a case) has a stillbirth at 42 weeks and was sleeping in a supine (on her back) position. If she is matched to a control with the same gestational age, any association between maternal sleeping on back and being a case cannot be due to a relationship between sleeping position and gestational age.

One effect of matching is to remove the possibility of studying the association between the variable matched (e.g. gestational age) and the outcome (e.g. late stillbirth). This does not matter in this example because the association is well established and was not one of the research questions for the New Zealand study. There are a number of related issues that are worth considering at this point, and these are explored through the following self-assessment exercise.



Self-Assessment Exercise 6.2.3

1. If matching helps prevent confounding, why do you think the study team did not match for a lot more possible confounding factors?
2. What would be the effect of a study matching on a variable that was on the causal pathway between the exposure and the outcome of interest?

Answers in Section 6.6

Implications of Matching for Analysis

We have said that matching should be done judiciously, and not overdone. Some studies do not match at all, and the researchers deal with the problem of confounding entirely at the stage of analysis. For example, later in the chapter we shall see how the authors of the New Zealand study did this for other confounding factors. A study that has not been matched will be subjected to *unmatched analysis*. A matched study normally is analysed by *matched analysis* as this is more efficient, but unmatched analysis can also be used. We study both of these analysis methods in Section 6.3.

Multiple Controls

Often in a case-control study, more than one control is selected for each study case. The reason for using multiple controls is to increase the *statistical power* of the study ($1 - \beta$), a concept introduced in Chapter 5. We could of course do this just by increasing the total number of cases and keep to a 1:1 ratio of cases and controls. However, in many situations, cases are harder to identify than controls, so it is simpler and cheaper to increase power by increasing the ratio of controls to cases. We will look in more detail at the multiple control selection when considering *sample size* in Section 6.4.

6.2.3 Exposure Assessment

Assessing the level of exposure to the risk factor(s) under study can also be a challenging component of case-control studies, not least because this is often done *retrospectively*. Refer back

to Section 6.1.2 to the excerpt on how exposure data were collected for the New Zealand study (study structure of Paper A) before completing the self-assessment exercise.



Self-Assessment Exercise 6.2.4

1. Make brief notes on how maternal sleep positioning and sleeping quality were determined.
2. What aspects of these methods do you think could have led to bias?

Answers in Section 6.6

6.2.4 Bias in Exposure Assessment

Having worked through Exercise 6.2.4, we can appreciate the problems of bias that can occur due to *retrospective* data collection. The critical issue, however, is whether the nature and extent of error in obtaining this information differs between cases and controls, as it is this difference that will lead to bias in the main results of the study. Highlighted here are three important possible sources of bias in exposure assessment.

Recall Bias

When asking people in these studies to recall their exposure to a given risk factor, there is always the possibility that the experience of having the disease (being a case) will alter people's actual knowledge of exposure or the importance that they attach to it. This can happen because during the course of treatment the person (case) may have been asked questions about relevant exposure or may be aware of media coverage of emerging theories or 'scares' about what causes the disease. Hence, when asked about the exposure, their description of their own level of exposure may well differ considerably from that of a control, even if in reality there is no difference between them in level of exposure.

Interviewer Bias

In many case-control studies, including the New Zealand study, much of the information is collected by interview. In a retrospective study where cases have already occurred, an interviewer setting out to collect information from the subjects quite often knows who is a case and who is a control. This would likely have been true for the interviewers in the New Zealand study. We have seen in Chapter 5 on cohort studies how, even with training and checking, *observer bias* can still occur. Interviewers, knowing whether they are talking to a case or control, could allow their own knowledge of the disease to interfere with objectivity in eliciting information about exposure. Even if the interviewer cannot be blinded to whether the interviewee is a case or a control, it may well be possible to downplay the influence of knowledge of the hypothesis under test, or to embed this among a number of topics (possible risk factors) being investigated so that it is not obvious which exposure the research team is most interested in.

Bias from Records

Given the problems associated with individual recall of exposure, it would be ideal if valid and unbiased information could be retrieved from records, such as medical case notes or employment records. Again, the key question to ask in this situation is whether the fact of being a case could have influenced the information in the records in a way that differs from records of people used as controls. This becomes less likely the longer the period is between the time the record was made and the appearance of the illness, or some early sign of that illness. In addition, such records may not have recorded the information exactly as required by the study, and some may

be incomplete, but these problems are less likely to lead to serious bias as any such limitations of recorded data are unlikely to be related to the case or control status of the study subjects.

The following exercise presents some case-control study scenarios. See whether you can spot potential sources of bias affecting the assessment of exposure to risk factors.



Self-Assessment Exercise 6.2.5

Study	Hypothesis under study	Methods of assessing exposure
1	Investigation in the USA of whether the oral contraceptive pill (OCP) causes breast cancer.	Telephone interview of women (cases and controls), to determine history of type and duration of OCP use.
2	Investigation of efficacy of BCG vaccine in preventing tuberculosis in a developing country.	District-based clinic records of BCG vaccination.
3	Investigation of whether low birth weight is associated with poor educational performance at age 10 years in the UK.	Health-care records.

Answers in Section 6.6

Consideration of Bias from the New Zealand Study

In a rapid response to the publication of the New Zealand paper (Paper A), one of the stronger criticisms of the results came from authors of a large Danish National Birth Cohort study (Stacey *et al.*, 2011) who *found little indication of an association between sleep-related practices and risk of late stillbirth* in their Danish participants and that the results of the New Zealand study could have been caused by *different sources of bias, including recall bias*. The authors of the Danish study were able to rule out such bias from their own results because they used a prospective cohort study design, so information collected on maternal sleeping practices (exposure to risk factor) preceded information collected on late stillbirth (outcome).

In response to this critique, the authors of the New Zealand study attempted to explain how the potential for *recall bias was reduced as far as possible*, although they could not rule it out. Firstly they used a structured interview with trained interviewers (to minimize interviewer bias), and secondly they ensured *participants were not aware of the study hypothesis* during the study. They also indicated that sleeping position and getting up in the night had not previously been related to stillbirth. This, the authors claimed, would make recall bias less likely and such that any misclassification that did occur (inaccurate recollection of sleeping practices) would be non-differential (similar between cases and controls). This example illustrates how important bias can be in case-control studies and the steps required in order to minimize the potential for such bias.

Summary

- Eligible cases must be clearly defined and ascertained as completely as possible.
- Controls should be representative of the same population from which the cases arise and should be free from the disease or outcome in question.
- Matching helps to overcome the problem of confounding, but not all case-control studies are matched. If matching is not carried out, confounding must be dealt with in the analysis.

- It is possible to overmatch in a case–control study. It is therefore wise, and more practical, to match for a few key confounders whose effects are known already and that are not strongly associated with the potential risk factors under study.
- Obtaining information about exposure to risk factors must usually be done retrospectively in a case–control study. This makes the study especially vulnerable to recall bias and also to interviewer bias and to bias from inconsistent records.
- If possible, the interviewer should not know whether the subject is a case or a control, but in practice this is often difficult or impossible to achieve.
- If possible, the interviewer and participants should not be aware of the hypotheses under investigation, e.g. which risk factors are being studied as possible causes, or at least be unaware of which of a number of risk factors is the subject of the main study hypothesis.

6.3 Basic Unmatched and Matched Analysis

6.3.1 The Odds Ratio (OR)

In the analysis of cohort studies, we introduced the concept of *relative risk* (RR), describing it as a means of expressing the way the risk of disease (or other outcome) varies according to the level of exposure to a risk factor. We defined RR as (incidence of disease in *exposed* group) ÷ (incidence of disease in *unexposed* group). So, to calculate the RR, we need to know incidence rates in exposed and unexposed groups. We saw that a cohort study provides these incidence rates, but in most cases the case–control design does not. This is because the group of exposed people (to deal with that group first) are made up of some cases and some controls, and they are not usually a representative sample of all exposed people in the population. The same can be said of the unexposed group. The exception to this is a population-based case–control study where all or a known fraction of cases in the population are obtained together with a representative sample of controls (we discussed in Section 6.2 the extent to which the New Zealand study is a true population case–control study).

In most case–control studies we do not have incidence rates, and therefore we need another means of expressing and quantifying the concept of relative risk. This is provided by the *odds ratio* (OR), which, as we shall see, in many circumstances can be a good approximation of the relative risk. The OR compares the odds (chances) of a case being exposed to the risk factor with the odds of a control being exposed.

Calculation of the OR – Simple Unmatched Analysis

The New Zealand study incorporated matching for duration of gestation. When matching has been used, it is more efficient – although not required – to use analysis that takes account of that. This is known as *matched analysis*, and we look at the techniques for this in Section 6.3.2. For now, we look at how to calculate the OR in *unmatched analysis* using the results from the New Zealand study. Let's start with a simple example. Illustrated in Figure 6.3.1 is the hypothetical case–control comparison introduced in Section 6.1. This was based on the New Zealand study example, but with just 30 cases and 30 controls.

This diagram can be summarised in a 2 × 2 contingency table as in Table 6.3.1.

The OR is calculated as the odds of being exposed if a case (a:c or a/c) divided by the odds of being exposed if a control (b:d or b/d) and, by transformation, can be expressed as follows:

$$\text{OR} = \frac{a/c}{b/d} = \frac{(a) \times (d)}{(b) \times (c)}$$

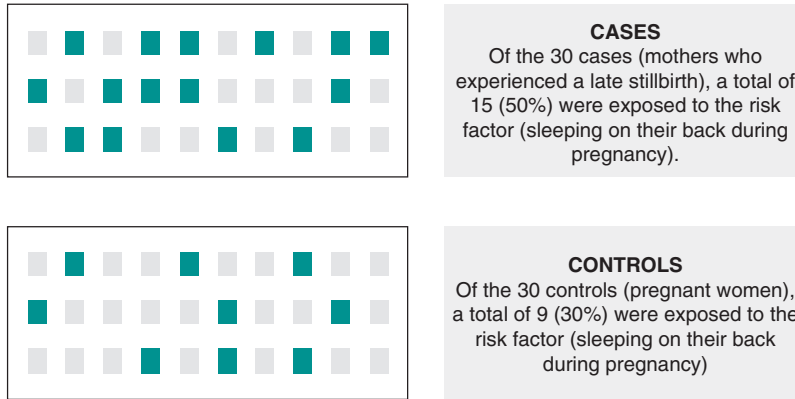


Figure 6.3.1 A hypothetical case–control comparison, investigating the association between exposure to a risk factor (sleeping on back during pregnancy, compared to sleeping on the left side) and late stillbirth.

Table 6.3.1 Outcome of a hypothetical case–control study.

	Cases	Controls
Exposed	15 (<i>a</i>)	9 (<i>b</i>)
Not exposed	15 (<i>c</i>)	21 (<i>d</i>)
Total	30	30

Inserting the figures from Table 6.3.1, we get $(15 \times 21) \div (15 \times 9) = 2.33$. So the OR is 2.33. What does this mean? It means that the odds of exposure to the risk factor (mothers sleeping on their back during pregnancy) among the cases (mothers who experienced a late still birth) is 2.33 times greater than among the controls (pregnant women). Interpreting the OR:

- An OR greater than 1 indicates the odds (or risk) of disease (or outcome) is greater among those who are exposed – that is, a positive association between risk factor and disease.
- An OR less than 1 indicates a reduced odds (or risk) of disease among the exposed – a negative association, or protective effect.
- An OR equal to 1 indicates no association between exposure to the risk factor and disease.

The odds ratio is typically used to approximate the relative risk (RR) and is therefore interpreted in the same way. However, it is important to understand the circumstances under which the OR is a good approximation of the RR and when it is not. Essentially, if the occurrence (incidence or prevalence) of the outcome being investigated is rare (generally accepted to be <10%), the OR will more closely approximate the RR. If the outcome is more common in the population, the OR will exaggerate the RR. This is shown in Figure 6.3.2.

As shown in Figure 6.3.2, as the incidence of the outcome increases (beyond 10%), the OR exaggerates (shows a larger effect than) the RR. This is true for both reductions in risk ($RR < 1.0$) and increases in risk ($RR > 1.0$).

The 95 per cent CI for the OR

A CI to estimate the true OR in the population should be presented, giving more information than simply quoting the sample OR. This is usually obtained by computer (e.g. with SPSS),

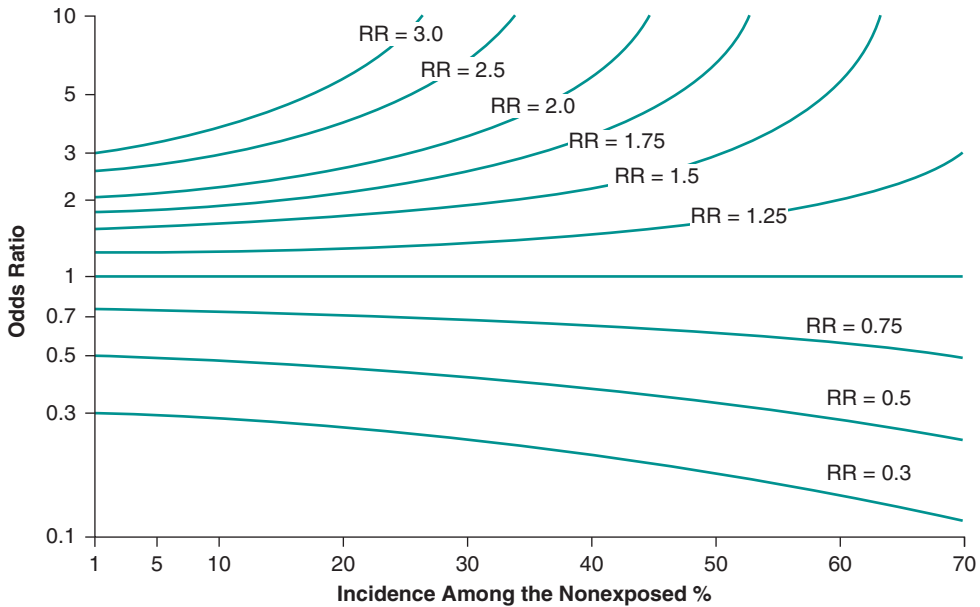


Figure 6.3.2 The relationship between OR and RR by occurrence of the outcome being investigated (incidence among the non-exposed population).

although an approximation can be calculated quite easily, as illustrated in the reference section below.



RS – Reference Section on Statistical Methods

We can estimate the standard error and hence CI using the **log** of the OR (we return to the use of \log_e OR in Section 6.5 when we look at logistic regression methods for case-control studies). The standard error of the log OR:

$$SE[\log_e(\text{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Using the data from Figure 6.3.1

	Cases	Controls
Exposed	15 (a)	9 (b)
Not exposed	15 (c)	21 (d)
Total	30	30

we obtain an OR of 2.3333 (truncated to 2.33 for ease of interpretation). The log OR is therefore $\log_e(2.33) = 0.8459$ with the standard error being

$$SE[\log_e(\text{OR})] = \sqrt{\frac{1}{15} + \frac{1}{9} + \frac{1}{15} + \frac{1}{21}} = \sqrt{0.2921} = 0.5404$$

Provided the sample is reasonably large (views on this vary, but a sample size between 30 and 50 is considered to be sufficient), we can assume that the log OR comes from a normal distribution and thus the approximate 95 per cent CI is

$$\begin{aligned}\log_e(\text{OR}) \pm 1.96 \times \text{SE}(\log_e(\text{OR})) &= 0.8459 - 1.96 \times 0.5404 \text{ to } 0.8459 + 1.96 \times 0.5404 \\ &= -0.2133 \text{ to } 1.9051\end{aligned}$$

To get a CI for the OR itself we must exponentiate (take the antilog):

$$e^{-0.2133} \text{ to } e^{1.9051} = 0.81 \text{ to } 6.72$$

Hence, from the data in Figure 6.3.1, we get an OR of 2.33 with a CI of 0.81 to 6.72. We should note that this CI is not symmetrical around the OR, reflecting the multiplicative nature of the latter (OR for the same risk = 1.0, for half the risk = 0.5; OR for twice the risk = 2, etc.). As with RR, a CI around an OR estimate that includes 1 is not statistically significant. Hence, we cannot be sure (at the 95 per cent confidence level) that our sample OR of 2.33 represents a true increase in risk in the population.

OR and RR – How Similar?

What we have arrived at is a concept similar to that where we used the RR to express and quantify the findings of the cohort study. It is true that this is an approximation of true RR and is calculated differently. As shown in Figure 6.3.2, the OR is a close approximation of the RR if the occurrence of the outcome is rare (<10%), but as this increases, the OR will exaggerate the true RR.

We will now apply the ideas and techniques discussed in the preceding section to the results of the New Zealand study by looking at the associations between sleeping position during the last night of pregnancy (defined as ‘last night before the baby had died’ for cases and ‘night before interview’ for controls) and a late stillbirth. Table 6.3.2, reproduced from Table 1 of Paper A, illustrates these results:

Table 6.3.2 Sleeping position during the last night of pregnancy for women who had experienced a late stillbirth (cases) and the control subjects.

Sleeping position during last night of pregnancy	Case subjects (n = 155)		Control subjects (n = 310)	
	No.	%	No.	%
Position on going to sleep:				
Left side	42	27	132	43
Right side	49	32	84	27
Back	15	10	15	5
Other	49	32	79	25
Position on waking up:				
Left side	31	20	106	34
Right side	45	29	72	23
Back	23	15	37	12
Other	56	36	95	31

Adapted from Stacey 2011.



Self-Assessment Exercise 6.3.1

1. Using the data in Table 6.3.2 relating to *sleeping position on going to sleep* among cases and controls, construct *two* contingency tables comparing (a) sleeping on the right side with sleeping on the left side and (b) sleeping on the back with sleeping on the left side. Refer to Table 6.3.1 if you need to refer back to how the 2×2 table is constructed.
2. Using these tables, calculate the odds ratios (ORs) for sleeping position (sleeping on left side = unexposed; sleeping on right side and back = exposed), as we did in Section 6.3.1. Note that we are ignoring matching for now.
3. Interpret these ORs.
4. Using the formula for the *standard error* of the log of the OR (described in the previous reference section on statistical methods), 95 per cent confidence intervals (CI) for the ORs were calculated as 1.12 to 3.01 and 1.42 to 6.96 for sleeping on right side and on the back, respectively. You may note that these differ from the confidence intervals included in Table 2 of the New Zealand paper where a *matched* analysis was conducted; we'll look at that in Section 6.3.2. Use the formula to calculate this yourself if you wish. How do these results for the 95% CIs help with your interpretation of the OR?
5. Now repeat steps 1 to 4 with *sleeping position on waking up* for (a) sleeping on the right side compared with sleeping on the left side and (b) sleeping on the back compared with sleeping on the left side. The 95 per cent CIs for the ORs are 1.24 to 3.69 for sleeping on the right and 1.10 to 4.10 for sleeping on back relative to sleeping on the left.

Answers in Section 6.6

6.3.2 Calculation of the OR—Simple Matched Analysis

OR with One Control per Case

We noted previously that when matching has been incorporated into the design of a case–control study, this matching should be taken into consideration when the study data are analysed. The authors of the New Zealand study conducted a simple *matched* analysis (presented in Table 2 of Paper A); this was followed by *matched* multivariable analyses (presented in Table 5 of Paper A) taking account of *confounding*. The analytical technique used to calculate ORs for matched multivariable analysis is known as *conditional logistic regression* (as mentioned in the first paragraph of the analysis section in Paper A). We discuss this in Section 6.5, but for now we focus on simple matched analysis conducted by hand for a case–control study.

A simple matched analysis of a case–control study uses data from cases and controls as *matched pairs*. In the example in Table 6.3.3, we have 180 matched pairs, that is, 180 cases, each one of which has been matched to one control.

- In the analysis, there are two categories of matched pair.
- One category consists of pairs of cases and controls that both have the same exposure status: either both exposed [group (e), $n = 12$ in this example] or both unexposed [group (h), $n = 119$]. These are the shaded cells of Table 6.3.3 and are known as *concordant pairs*.
- In the other category of case–control pairs, the exposure status differs: Either the case is exposed and the control is unexposed [group (f), $n = 42$], or vice versa [group (g), $n = 7$]. These are the non-shaded cells of Table 6.3.3, and they are known as *discordant pairs*.

In a case–control study, we are interested in the association between exposure and the disease, so the concordant pairs (same exposure status) do not tell us anything. We therefore do not

Table 6.3.3 Matched pairs in a case-control study.

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	12 (e)	42 (f)	54
Unexposed	7 (g)	119 (h)	126
Total	19	161	180

Note that these are numbers of controls, not numbers of pairs

Note that these are numbers of cases, not numbers of pairs

These are numbers of pairs

include them in the calculation of the OR, because they would only dilute the effect and make the analysis less efficient. In this simple matched analysis, the OR is calculated by dividing the number of matched discordant pairs (different exposure status):

$$OR = \frac{\text{Number of pairs with case exposed and control unexposed}}{\text{Number of pairs with case unexposed and control exposed}} = f/g$$

So in this example, the $OR = 42 \div 7 = 6.00$. If we had ignored the matching in the analysis, we would have the following data (Table 6.3.4), and the $OR = (54 \times 161) \div (126 \times 19) = 3.63$. The estimate of the OR from the matched analysis differs considerably from the estimate based on the unmatched analysis of the same data. Ignoring matching will *bias* the estimate of the OR towards unity (1.00); that is, it dilutes the observed effect.

Table 6.3.4 Unmatched analysis of matched design (see Table 6.3.3 for data).

	Cases	Controls
Exposed	54	19
Not exposed	126	161
Total	180	180

More than One Control per Case

If we have two controls matched to each case, we cannot summarise the data in a 2 x 2 table. We need to take account of eight possible outcomes for each set of three subjects (one case and two matched controls), as shown in Table 6.3.5.

The estimate of the OR is still based on the ratio of discordant sets, and in this case is

$$OR = (n_1 + n_2 + 2n_3) / (2n_4 + n_5 + n_6)$$

The sets of subjects who are all exposed, n_0 , or all not exposed, n_7 , do not tell us about the effect of exposure – these are the concordant sets. This can be extended to more than two matched controls per case, but it becomes a lot more complicated. If there are three controls per case,

Table 6.3.5 Matching two controls.

Case	Control 1	Control 2	Frequency
Exposed	Exposed	Exposed	n_0
Exposed	Exposed	Not exposed	n_1
Exposed	Not exposed	Exposed	n_2
Exposed	Not exposed	Not exposed	n_3
Not exposed	Exposed	Exposed	n_4
Not exposed	Exposed	Not exposed	n_5
Not exposed	Not exposed	Exposed	n_6
Not exposed	Not exposed	Not exposed	n_7

there are $2^4 = 16$ possible outcomes; four controls per case results in $2^5 = 32$ possible outcomes; and so on.

6.3.3 Hypothesis Tests for Case–Control Studies

Simple Unmatched Analysis (Chi-Squared Test)

To investigate whether an OR provides evidence of a statistically significant association, a hypothesis test of the *null hypothesis* that the population OR is equal to 1.0 (no association), is calculated providing a p -value. As with the 95 per cent CI for the OR (if it does not include 1.0), it will show whether there is a statistically significant association at the 5 per cent significance level ($p < 0.05$). Returning to our hypothetical example of an unmatched study from Section 6.3.1, we had (Table 6.3.1) the following (Table 6.3.6):

Table 6.3.6 Unmatched study.

	Cases	Controls	Total
Exposed	15	9	24
Not exposed	15	21	36
Total	30	30	60

We are investigating the association between two categorical variables, women who experienced a late stillbirth compared to currently pregnant women (case or control) and the risk factor (e.g. sleeping on the back during pregnancy compared to sleeping on the left side). As we saw in Chapter 5 on cohort studies, the appropriate hypothesis test for this type of comparison is the chi-squared test. Recall that the null hypothesis can be stated in a number of equivalent forms, including

H_0 : there is no association between late stillbirth and sleeping position.

H_0 : the probability of exposure to sleeping on the back during pregnancy is the same for cases and controls.

A further equivalent form that is meaningful in the context of a case–control study is

H_0 : the OR for sleeping on the back during pregnancy and late stillbirth in the population is 1.0.

The OR from Table 6.3.1 is 2.33, suggesting an increased risk of stillbirth among those exposed. We also calculated the 95 per cent CI (0.81–6.72), and we found that this was quite wide and included the value of 1.0 (no increased risk). To complete the picture, we now need the chi-squared statistic.

Using the procedure to calculate the chi-squared statistic that we introduced in cohort studies (Chapter 5, Section 5.5.3), we obtain a value of 2.50 on one degree of freedom, with a corresponding p -value of $p = 0.11$. Therefore, the value of the test statistic is not statistically significant at the 5 per cent level, and we do not have sufficient evidence to reject the null hypothesis that stillbirth is not associated with sleeping on the back during pregnancy. This is consistent with our conclusion from the 95 per cent CI.

So, although an OR of 2.33 appears to indicate an increased risk, it is in fact, in this small sample, consistent with no association. In other words, this could be a chance finding, or the study is too small (there is insufficient power) to detect what is in fact a real effect (a type II error). A somewhat larger sample with the same OR would have shown a statistically significant result and a 95 per cent CI that did not include 1.0. Clearly, we cannot estimate the OR very precisely from this small sample.

Simple Matched Case-Control Study (McNemar's Test)

To consider the hypothesis test for a matched study, we return to the example from Paper A discussed in Section 6.3.2. We had 180 matched pairs, resulting in the following table (Table 6.3.3 in Section 6.3.2):

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	12 (e)	42 (f)	54
Unexposed	7 (g)	119 (h)	126
Total	19	161	180

The OR was calculated as $f/g = 42/7 = 6.00$. To test the null hypothesis that the population OR is 1, we need a test appropriate for matched pairs. This is called *McNemar's test*. The test statistic is

$$\chi^2 = \frac{(f - g)^2}{f + g}$$

which is compared to the chi-squared distribution with one *degree of freedom* to find the p -value. In this example, the value of the test statistic is

$$\frac{(42 - 7)^2}{42 + 7} = 25.00$$

and the p -value is $p < 0.0005$. The result is highly statistically significant: We can reject the null hypothesis and conclude that there is strong evidence of an association between the outcome and the risk factor. The OR of 6.00 shows that exposure substantially increases the risk of the outcome.

When multiple controls are matched to individual cases, a generalization of McNemar's test (such as the Miettinen statistic) is required to account for this design. (For more information, see Miettinen 1969.)

Summary

- Most case–control studies do not allow calculation of incidence rates in the exposed and unexposed populations that are required for the calculation of the RR; as a result, the OR for exposure in the cases compared with the controls is used. In most situations, the OR is a close approximation of the RR, so long as the occurrence (incidence/prevalence) of the disease or outcome is low (generally <10%).
- The analysis of a case–control study should be appropriate to the study design: matched analysis should be used for a matched design.
- A matched study can be analysed unmatched, but this is generally less efficient and will bias the OR towards 1.0 (no effect).
- For a simple unmatched case–control study, the chi-squared test is used to test the hypothesis of no association between the outcome and a single risk factor.
- For a simple matched case–control study, McNemar's test is used to test the hypothesis of no association between the outcome and a single risk factor, a generalisation of the test being required with multiple controls.

6.4 Sample Size for a Case–Control Study

6.4.1 Introduction

We have already looked at sample size calculation for a cohort study and have seen why this is such an important step in the design. It is equally important for case–control studies. In Paper A, the authors of the New Zealand Study state that ‘The study was powered to detect an odds ratio of 2 with 80% power and significance level of 5%, with a prevalence of the risk factor of $\geq 20\%$ in the control population.’ We now look at how to calculate the sample size for the New Zealand study based on these assumptions and based on some of the findings from the paper.

With cohort studies, the RR was the key measure of outcome that we used for estimating sample size. For a case–control study it is the OR, and we need to decide on a value for the OR that we want to be fairly confident of being able to detect (that is, demonstrate as statistically significant). The authors from the New Zealand study calculated their sample size to detect an OR of 2, that is, a doubling of risk. We look at what needs to be considered in deciding on the OR to be able to detect in the following section.

6.4.2 What Information is Required?

To determine the required sample size for a case–control study, we need

- A decision about the OR that we wish to be able to detect. The authors of Paper A used an odds ratio of 2. We will use some alternative values of the OR from their results to see how much difference this makes. Making a judgment about the value of the OR to be detected is important but often not easy, since this must balance the value of demonstrating the smallest useful effect (or risk) against the costs and logistic demands of dealing with a larger sample size. An OR of 2 – that is, a doubling of risk – is an ambitious target for a study to detect. An OR of 1.50, a 50% increase in risk, would also have been worth detecting but – as we will see shortly – would have dramatically increased the sample size needed. As it happened, the OR of 2 turned out to be justified for the main risk factor in this study.

- Knowledge of the **prevalence** of the risk factor in the population. First, we need to identify a source of information about this. As a general rule, one would try to find previous studies that provide estimates of the population prevalence of exposure. The authors of Paper A chose a ‘prevalence of the risk factor of $\geq 20\%$ in the control population’, although there is no reference to how this value was selected or to the chosen risk factor (e.g. sleeping on the back on the last night of pregnancy).
- A decision on the ratio of controls to cases. As we have seen in the New Zealand study, two controls were selected for each case. We will start by using that assumption, and then we will find out what happens if we adjust this ratio.
- A **significance level** for the appropriate hypothesis test, here taken at 0.05 (and used in most studies).
- A level of statistical **power**, here taken at 80 per cent.

The last two points are exactly the same assumptions as we used for cohort studies.

6.4.3 An Example of Sample Size Calculation Using OpenEpi

As with cohort studies, we will demonstrate the sample size calculation for a case-control study using OpenEpi statistical software. The sample size calculation for case-control studies requires all of the information listed above, but in the format set out in Table 6.4.1. The values based on the authors’ assumptions have been entered into the right-hand column. These were a significance (alpha or α) value of 0.05 (or 5 per cent) (OpenEpi refers to $1 - \alpha$, hence $1 - 0.05 = 0.95$, or 95 per cent), a power ($1 - \beta$) of 80 per cent, a control-to-case ratio of 2:1, an expected frequency of exposure to the risk factor in the control population of 20 per cent, and an OR of 2.

Table 6.4.1 Information for sample size in unmatched case-control study (OpenEpi).

Information needed for sample size calculation	Values for this example
Two-sided confidence interval ($1 - \alpha$), usually 95%	95%
Power (% chance of detecting), usually 80%	80%
Ratio of controls to cases (for equal samples, use 1.0)	2.0
Percentage of exposed controls (between 0.0 and 99.9%)	20%
Odds ratio	2

There are two different approaches to calculating sample size for case-control studies, and these depend on whether we are using a matched or an unmatched design. A matched design requires fewer participants because the variation between cases and controls is likely to be less due to the matching for potential confounders. Although the New Zealand study did carry out some matching (controls were frequency matched to cases with regard to gestation period and area of residence), the researchers did not state whether they considered this in the calculation of their sample size. Even if a study has carried out some matching, it is common to use the more-conservative, unmatched, approach when calculating the required sample size to ascertain the minimum number of study participants. We therefore use the unmatched approach to sample size calculation for this example. Calculation for the unmatched design is often used even if some matching has been carried out, because this provides a more-conservative estimate of the required sample size. In OpenEpi, the only option is sample size calculation for an unmatched case-control (or CC). The output from the OpenEpi program looks like Table 6.4.2.

Table 6.4.2 Example of output from sample size calculation for an unmatched case–control study (OpenEpi).

Two-sided significance level (1–alpha):	95		
Power (1–beta, % chance of detecting):	80		
Ratio of controls to cases:	2		
Hypothetical proportion of Controls with exposure:	20		
Hypothetical proportion of Cases with exposure:	33.33		
Least extreme Odds Ratio to be detected:	2.00		
	Kelsey⁽¹⁾	Fleiss⁽²⁾	Fleiss with CC
Sample Size – Cases	123	126	137
Sample Size – Controls	245	252	274
Total Sample Size:	368	379	411

Notes:

(1) Refers to value calculated using method proposed by Kelsey *et al.* (Methods in Observational Epidemiology, 2nd edition 1996; Oxford University Press).

(2) Refers to value calculated using method proposed by Fleiss *et al.* (Statistical Methods for Rates and Proportions, 3rd Edition 2003; Wiley & Sons). The third value (Fleiss with CC), refers to inclusion of the continuity correction, which provides a more conservative estimate.

So what does Table 6.4.2 show? The first six rows give the information we entered to calculate the sample size for the study with the addition of the hypothetical proportion of cases with exposure that OpenEpi automatically calculates from the data we entered (Table 6.4.1). The rest of the table provides sample sizes required for Cases, Controls, and Overall. Taking the more-conservative estimate provided by Fleiss (with the continuity correction – see Section 5.6 for description of this continuity correction in the calculation of sample size for a cohort study), we need at least 137 cases and 274 controls for a study with 80 per cent power to detect an OR of 2, where exposure in the controls is 20 per cent, with a significance level of 0.05, and a 2:1 ratio of controls to cases. In fact, the New Zealand study (Paper A) had more than enough cases (155) and controls (310) available for analysis to meet these sample size requirements.

Table 6.4.3 shows how varying the input values affects the required sample size.

Table 6.4.3 Examples of sample sizes required for an unmatched case–control study (OpenEpi) with differing input parameters.

Significance (1-alpha):	95	99	95	95	95	95	95
Power (1-beta/%):	80	80	90	80	80	80	80
Controls to Cases:	2	2	2	1	2	2	2
% exposure – Controls:	20	20	20	20	12	20	20
% exposure – Cases:	33.33	33.33	33.33	33.33	21.43	33.33	33.33
Odds Ratio:	2	2	2	2	2	3.28	1.36
Fleiss with CC							
Sample Size – Cases	137	198	181	187	194	45	734
Sample Size – Controls	274	395	362	187	387	90	1468
Total Sample Size	411	593	543	374	581	135	2202

We can see how the sample size increases as the significance level is decreased from 0.05 (95.00 per cent) to 0.01 (99.00 per cent). We can also see how the sample size increases as the power is increased from 80 to 90 per cent. The more precise we require our sample estimate to be, the larger the sample size we will need. The fifth column of the table shows the impact on the sample size of having an equal number of cases to controls: whereas the overall sample size is decreased, we can see that an additional 50 cases are required, and typically it is more difficult to locate cases for a case–control study.

The sixth column shows the effect on the required sample size when the exposure is less common. A reduction in exposure to the risk factor in controls from 20 per cent to 12 per cent (this is the value given in Table 2 of Paper A for sleeping on the back for the position on waking during pregnancy) increases the total sample size by 170 (or a 41 per cent increase) to 581, and this is actually greater than the sample size of 465 for Paper A.

Finally, changing the size of the effect estimate to be measured has a large impact on the required sample size. If we want to be able to detect an odds ratio of 3.28 (Table 2: Position on going to sleep on the last night of pregnancy for sleeping on the back), the sample size is decreased by 67 per cent to 135: A smaller sample is required to detect a larger effect or association. Conversely, for an OR of 1.36 (Table 2: Position on waking up on the last week of pregnancy for sleeping on the back), the sample size is increased by more than five-fold to 2202: A much larger sample size is required to detect a small effect or association. Therefore, when calculating a sample size for a case–control study, a choice is required on what is the most important association to be able to detect (e.g. the primary risk factor of interest with the outcome) to ensure the numbers of participants are sufficient to address this issue.

Summary

- Sample size calculation is as important in case–control studies as it is with any of the other study designs we have been discussing.
- Information on the prevalence of exposure in the population is required, and it is best if there is an estimate available from prior studies, but otherwise, a commonsense estimate may be required.
- Judgement is required to decide what size effect (OR) should be detected, and the OR must balance an assessment of the smallest useful effect against the resources needed to meet the resulting sample size requirement.
- Increasing the number of controls relative to cases is a practical way of increasing power if cases are difficult to identify.
- Increasing the number of cases relative to controls can also be done to increase power, if appropriate, although this is less commonly seen.
- Sample size calculation is only approximate, and allowance must always be made for refusals, dropouts (if the study has a prospective component), and problems of data collection that might mean not all cases and controls can be used.

6.5 Confounding and Logistic Regression

6.5.1 Introduction

In a case–control study, confounding can be dealt with at the design stage, the analysis stage, or both. We saw in Section 6.2.2 how *matching* can be used in the study design to avoid the effects of *known* confounding variables, and we noted that the New Zealand study used this

method to match for gestational age. Another approach employed in the design stage to deal with confounding is **stratification**. This is discussed in the next section. Lastly (and in many ways the most important), we can **adjust** for confounding at the analysis stage by using either post-stratification or multivariable logistic regression. Adjustment in analysis is often used in combination with matching. Logistic regression is discussed in Section 6.5.3.

6.5.2 Stratification

Stratification During Study Design

Stratification is a way of eliminating confounding through the study design by using a stratified sampling procedure. The population is divided into strata (groups) according to one or more confounding factors, and predetermined numbers of cases and controls are selected from each stratum. For example, in an investigation of whether there is a relationship between regular alcohol consumption and myocardial infarction (MI), smoking is suspected of being a confounding factor. Smoking is known to be a cause of MI, and smoking is commonly associated with regular alcohol consumption, as illustrated in Figure 6.5.1.

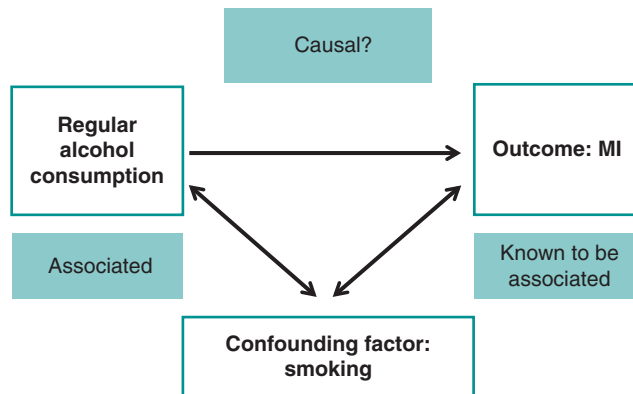


Figure 6.5.1 Confounding factors in myocardial infarction.

We can eliminate the confounding effect of smoking by selecting cases and controls from each of the two groups: smokers and non-smokers. That is, we **stratify** on smoking. The analysis must be appropriate to the design of the study, so we need to use a **stratified analysis**. We will explore this through the following example and exercise.



Self-Assessment Exercise 6.5.1

Suppose that we select cases and controls to include 100 smokers (stratum 1), and 100 non-smokers (stratum 2), making 200 study subjects in total. We obtain the following results with unequal numbers of cases and controls, because it was easier to find cases of MI among smokers:

		Non-smokers		Smokers	
		Cases of MI	Controls	Cases of MI	Controls
<i>Regular alcohol</i>	Yes	11	15	54	36
	No	29	45	6	4
	Total	40	60	60	40

1. Calculate the ORs for non-smokers and smokers, and comment on the results.
2. If we had ignored the stratification, we would have the following combined table.

		Cases of MI	Controls
<i>Alcohol</i>	Yes	65	51
	No	35	49
Total		100	100

Calculate the OR for this combined table, and compare this with the ORs you obtained in question 1. Interpret what you find.

Answers in Section 6.6

Post-stratification

As with matching, stratification during the study design requires that potential risk factors be identified in advance and used during the selection of cases and controls. It is possible, however, to carry out a stratified analysis when the sample was not stratified. This is called **post-stratification**, and it can be used to take account of possible confounders, provided information on these factors has been collected. A disadvantage of post-stratification is that there is no guarantee that all the strata will contain enough cases and controls for calculation of the OR.

Stratification results in a series of ORs, one from each stratum. For example, if we were to study smoking in categories of never-smokers, ex-smokers, and current smokers, we would have three ORs. These ORs are of individual interest, but we generally want an overall summary measure of association with the risk factor as well. This is a **pooled** or **adjusted** OR, adjusted for the effects of the stratifying factor(s).

A pooled, adjusted OR is obtained by the Mantel–Haenszel method, which combines the ORs from each stratum. Although not difficult to calculate, it becomes tedious for more than a few strata. This technique is not covered further in this book. Furthermore, the advent of modern computing means that logistic regression is now the method of choice for dealing with confounding at the analysis stage.

6.5.3 Logistic Regression

Introduction

Confounding may be dealt with at the analysis stage by the method of **logistic regression**. The application of this technique of regression is similar, in principle, to what we described for simple and multivariable linear regression in Chapter 5 on cohort studies. The main difference lies in the way we deal with the fact that in case-control studies the outcome is categorical ('case' or 'not case'), and not a continuous variable such as bodyweight, as we used for the example in Chapter 5.

We enter explanatory variables (exposures of interest and confounding variables) and model these to obtain adjusted ORs that show the independent effect of each risk factor, adjusted for the other variables in the regression. Just as we can carry out simple linear regression to produce an unadjusted (univariate) relative risk, we can do the same with simple logistic regression, although the method differs for unmatched and matched analysis, as described below.

We have seen that it is not usually practical to match on more than a few factors, so the effects of confounders often need to be taken into account in the analysis, irrespective of whether matching has been used in the design. Additionally, factors used to match cases and controls cannot be investigated as risk factors, so if we want to investigate whether a factor is associated with the disease, we should not match for it. Thus, for various reasons, the majority of case–control studies require some adjustment for confounding in the analysis. You will therefore find that adjustment for confounding is carried out during analysis in most published case–control studies.

Two forms of logistic regression are commonly used. These are *unconditional* for unmatched and *conditional* for matched study designs. We will look at these two forms, and the other issues discussed above, in more detail in the remaining parts of this section, but we start with a brief overview of why we need a different type of regression for the categorical outcomes used in case–control studies.

The Logistic Regression Model

Having a categorical outcome presents a problem for regression, which has as one key assumption a *linear association* between the explanatory variable(s) and the outcome. In practice, therefore, to carry out regression with the OR, we need to derive some form of *continuous outcome*. This is done by using – as the outcome – the (natural) logarithm of the *odds* of being a case rather than a control. The log odds provides an appropriate mathematical transformation that gives a continuous linear function and therefore meets the characteristics required for logistic regression analysis. The rationale for the use of the log odds function in logistic regression goes like this:

1. In a given study population, if we call the probability of being a case p , then the probability of being a control (not a case) is $1 - p$. Therefore, the *odds* of being a case is $p/(1 - p)$.
2. If we take the logarithm of this, we have the log odds, which, as argued in (1), is $\log [p/(1 - p)]$.
3. This can have any value: positive, negative, or zero. So we have turned a categorical outcome into a continuous variable, and we can now apply regression methods, both simple and multivariable.

The equation for simple logistic regression, that is, with one explanatory variable, is

$$\log (p / (1 - p)) = \alpha + \beta x$$

The equation for multivariable logistic regression, where we want to look at the independent effects of a number of explanatory variables on the outcome, is

$$\log (p / (1 - p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots$$

We are almost ready now to look at an example of multivariable logistic regression, but there is one more complication to address: how to represent categorical explanatory variables in the regression equation where these have more than two categories. We do this with ‘dummy’ variables.

Dummy Variables

The independent (explanatory) variables may be either continuous (e.g. blood pressure) or categorical (e.g. smoking status). Continuous variables present no problem, as we saw in the example used for multivariable linear regression in Chapter 5. For categorical variables, we need to distinguish between those that have two categories (e.g. sex) and those that have more than two (e.g. never-smoker, ex-smoker, light current smoker, heavy current smoker).

Variables with Two Categories

We need only one variable in the regression equation to represent a categorical variable with two categories. Thus, to include the variable for sex with categories 'male' and 'female', we can use one variable coded, for example, as 0 = male and 1 = female.

Variables with More than Two Categories

We can include these categorical variables in the regression model in one of two ways:

- By the use of *dummy variables*; that is, newly created variables, each of which can have the value 0 or 1 (explained below).
- As a continuous variable if the category values are ordered, and it is reasonable to assume that these values differ in a proportionate way.

Dummy variables are used when a categorical variable with more than two levels cannot be treated as continuous. As we will see from the example in Section 6.5.4, computer programmes such as SPSS create dummy variables automatically when such variables are specified as categorical. It is useful to see how dummy variables work, and the following example based on the New Zealand study illustrates this.

Let us consider the risk factor 'sleeping position' (Table 6.3.2, reproduced from Paper A). There are four possible values for this variable: 'left side', 'right side', 'back', and 'other'. The intervals between these categories are not consistent because this is a 'nominal' variable, so in order to include this risk factor in a multivariable analysis, we will need to define dummy variables. In general, we require one less dummy variable than the number of categories, so we need three: let these be called status 1, status 2, and status 3, and note that each can only take the values 0 and 1.

Status 1 will have the value 1 if sleeping position is on the *right side* and 0 if sleeping position does not fall into this category. Status 2 will have the value 1 if sleeping position is on the *back* and 0 if sleeping position does not fall into this category. Status 3 will have the value 1 if sleeping position is *other* and 0 if sleeping position does not fall into this category. This means that sleeping position on the *left side* will have a value of 0 for status 1 (sleeping position is on the right), 0 for status 2 (sleeping position is on the back), and 0 for status 3 (sleeping position is 'other'). The three dummy variables therefore distinguish the four categories (Table 6.5.1):

Table 6.5.1 Dummy variables.

Sleeping position	Status 1	Status 2	Status 3
Left side	0	0	0
Right side	1	0	0
Back	0	1	0
Other	0	0	1

The value 'Left side', having the value 0 for all three dummy variables, is the **reference category** to which each of the others will be compared.

We now look at an example of multivariable logistic regression to see how all this works in practice.

6.5.4 Example: Multivariable Logistic Regression

To illustrate how multivariable logistic regression works in practice, we now look at an example using our data set investigating occupational risk factors for back pain, analysed using SPSS. This example is used now to describe the process, step by step, and to familiarise you with the interpretation of the output from logistic regression analysis.

Investigating the Effect of Psychosocial Work Environment on Low Back Pain

For this example, we treat the back pain data set as if it were an unmatched case–control study. In total there were 765 employees. Of these, 198 (25.9 per cent) had back pain (cases), and 567 (74.1 per cent) did not have back pain (controls). Therefore, there are approximately three controls for each case, but they are not matched.

We are interested in identifying the relationship between psychosocial work environment, defined as self-reported psychological demands associated with manual work, and low back pain. In particular, we would like to investigate the independent effects of work speed (whether workers find their work hectic or too fast), work monotony (whether workers find their work unstimulating), and work stress (whether the work carried out by employees causes them anxiety or stress).

The study hypothesis is that a poor psychosocial work environment, in terms of these three psychological demands, may increase muscular tension through psychological stress, which in turn increases the risk of low back pain. The *variables* in the database that are of interest for this investigation include the following:

- The outcome (dependent) variable: This is low back pain, a categorical variable with two values (dichotomous) of ‘no low back pain’ and ‘low back pain’.

A number of explanatory (independent) variables, listed in Table 6.5.2.

Table 6.5.2 Features of explanatory variables to be included in the logistic regression.

Domain	Variables	Type of variable
Psychosocial work environment	Hectic work Monotonous work Stressful work	The main variables for exposure of interest. These are categorical variables with four-point response scales (1 = never, 2 = occasionally, 3 = half the time, 4 = always). Each requires three dummy variables in logistic regression.
Psychological distress	General Health Questionnaire	A continuous variable relating to a score on a questionnaire measuring psychological distress. The score ranges from 12 (no distress) to 48 (severely distressed). This variable is included in the logistic regression as a possible confounder of the relationship between psychosocial working environment and low back pain.
Demographic characteristics	Age (years) Sex	Age is a continuous variable, and sex is a dichotomous (categorical) variable. These variables are included in the logistic regression as possible confounders. Neither requires any special treatment for the regression.

Is There a Univariate Association with Psychosocial Work Environment?

Before constructing a multivariate model, we should examine the association between the independent variables of interest (representing psychosocial work environment) and the dependent variable (low back pain) by carrying out univariate logistic regression.

When we carry out logistic regression with SPSS, the first table to take note of is the *categorical variables codings* table that displays how categorical variables are coded into dummy variables.

We noted earlier that with four categories, we would need three dummy variables. These dummy variables are created automatically in SPSS when we specify ‘hectic work’ as a categorical variable and define the reference group (in this case ‘never’). When the regression is run, the SPSS output includes the following table (Table 6.5.3), with the term ‘parameter coding’ with values of 1–3 equivalent to our use of ‘status’ in the explanation of dummy variables above.

Table 6.5.3 Categorical variables codings.

		Frequency	Parameter coding		
			(1)	(2)	(3)
Is work hectic?	never	31	.000	.000	.000
	occasionally	319	1.000	.000	.000
	half the time	223	.000	1.000	.000
	always	191	.000	.000	1.000

We can see that three dummy variables have been created based on the response categories of 2 (occasionally), 3 (half the time), and 4 (always). Thus, the reference category (never) has values of zero (.000 in the table), zero, zero for each dummy variable. The ‘occasionally’ category has values of one (1.000 in the table), zero, zero; and so on. Similar tables would be constructed for ‘monotonous’ and ‘stressful’ work.

Goodness of Fit – How Good is the Model?

The next important output from SPSS to take note of is the Omnibus Tests of Model Coefficients (OTMC) table. In a way similar to the ANOVA table in linear regression (see Section 5.8 of Chapter 5), the OTMC table gives an estimate of the goodness of fit of the model. The *chi-squared* value for the model is interpreted in a way similar to the *F-ratio*: a measure of how much the model has improved the prediction of the outcome compared to the level of random error of the model (think of this as being analogous to standard error). If the model is a good one, then we expect the improvement in prediction due to the model to be large and the difference between the model and the observed data to be small. In short, a good model should have a large chi-square, relative to the degrees of freedom. The number of degrees of freedom is derived from the number of observations and the number of explanatory variables. Thus, the chi-squared statistic should be statistically significant in a good model, and 0.05 is taken as the conventional level of probability for this purpose.

The OTMC table for ‘hectic work’ in univariate analysis looks like this, and shows the chi-square, degrees of freedom, and significance (*p*-value).

Table 6.5.4 Omnibus Tests of Model Coefficients.

		Chi-square	Df	Sig.
Step 1	Step	8.366	3	.039
	Block	8.366	3	.039
	Model	8.366	3	.039

We can see that the chi-squared value for this *univariate model* is statistically significant ($p < 0.05$), so there is less than a 5 per cent chance that a chi-squared value this large would have arisen by chance alone. We can conclude that the univariate logistic regression model with ‘hectic work’ predicts low back pain significantly well. Similar univariate analyses for the two other variables show that these models are also significant.

Univariate Effect Estimates

The final output in this univariate analysis is the variables in the equation table, interpreted in much the same way as the coefficients table when carrying out linear regression in SPSS, but with one important difference. Whereas with linear regression the *beta coefficient* gave us a direct estimate of the effect of any given variable, in logistic regression the beta coefficients are *log odds* and, as such, are difficult to interpret as they stand. To obtain the *OR* for the outcome (low back pain) with exposure to any given variable, it is therefore necessary to take the exponential of the beta coefficient. This is given automatically by SPSS in the variables in the equation table, in the right-hand column with the heading Exp(B). The SPSS table providing the output for the univariate logistic regression analysis of the ‘hectic work’ variable is shown in Table 6.5.5. We can see from the model investigating the association between ‘hectic work’ and low back pain that the OR increases with the amount of time employees spend carrying out work they believe to be too hectic.

Table 6.5.5 Variables in the equation.

		B	S.E.	Wald ⁺	df	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1 ^a	hectic*			7.794	3	.050			
	hectic(1)	.677	.552	1.502	1	.220	1.968	.667	5.808
	hectic(2)	.955	.556	2.950	1	.086	2.599	.874	7.733
	hectic(3)	1.129	.558	4.092	1	.043	3.092	1.036	9.229
	Constant	−1.910	.536	12.703	1	.000	.148		

^aVariable(s) entered on step 1: hectic.

*This row corresponds to the full (4 category) variable for hectic work and therefore has 3 (n−1) degrees of freedom. SPSS automatically calculates the Wald Statistic⁺ and associated *p*-value for the association between the full categorical variable (providing the dummy variables in logistic regression) and the outcome.

⁺The Wald Statistic is the chi-squared value for the null hypothesis of no significant effect, for specified degrees of freedom (calculated as rows-1 x columns-1: see Section 5.5.3 on chi-squared test). The statistic is calculated as β (beta) divided by S.E. (standard error of beta).

For employees who report their work to be occasionally too hectic or fast the *OR* (Exp(B)) is 1.97, with a *p*-value (Sig) of 0.221. This OR of 1.97 means that employees are twice as likely to experience low back pain if they ‘occasionally’ carry out work that is too hectic or fast relative to those who ‘never’ carry out such work – the *reference category*, although because this result is non-significant, we cannot exclude chance as an explanation. We then see that the OR for low back pain increases to 2.6 for employees carrying out hectic work ‘half the time’ and to 3.09 for employees who ‘always’ report their work as being too hectic or fast, relative to those who ‘never’ report such work. However, only the association between low back pain and the fourth level (work is hectic/too fast ‘always’) compared to ‘never’ has achieved statistical significance ($p < 0.05$). Analysis of ‘monotony’ and ‘stress’ shows similar progressive increases in the OR across categories, with generally smaller *p*-values (results are more significant).

Assessment of Potential Confounders

We now need to take into account the effect of confounding factors. The first step is to explore the relationship between possible confounding factors that we know to be associated with the exposure variables (i.e. the three measures of the psychosocial work environment), and the outcome (low back pain). The potential confounders we are interested in are shown in Table 6.5.6.

Table 6.5.6 Confounding factors for consideration in multivariable regression analysis for psychosocial work environment and low back pain.

Variables	Reason for consideration as a potential confounding variable
Psychological distress	In this data set, there is a strong positive relationship between hectic work environment and psychological distress as measured by the General Health Questionnaire score. Such distress has been observed to be related to the experience of pain.
Sex	In this data set there is a weak relationship between female sex and reporting a hectic work environment, with slightly more women than men reporting the most frequent category. Females have been found to report a greater amount of low back pain than males.
Age	Although in these data there is a rather weak, but positive, association with hectic work environment, there is generally such a strong relationship between age and almost all symptoms and health problems that it would be very unwise not to include age in our model.

With the variable for psychological distress (GHQ score), care needs to be taken when considering whether it is a confounder of the relationship between psychosocial working environment and low back pain. This is because it might be considered that psychological distress is on the ‘causal pathway’ from psychosocial working environment to the experience of low back pain (as described in Chapter 5, Section 5.7.2). If this were the case, when we adjust for GHQ we might eliminate the association between psychosocial working environment and low back pain. However, the role of psychosocial working environment in the development of low back pain is thought to be directly related to a theoretical model, the muscular-tension theory, which postulates that increased psychological demands at work, a lack of decision authority or control over one’s work, and a lack of social and managerial support leads to physiological vulnerability, resulting in low back pain. Therefore, whilst psychological distress might be associated with both psychosocial working environment and low back pain, it is not the mechanism by which demand, control, and support at work results in low back pain.

Table 6.5.7 (variables in the equation table in SPSS) describes the association between the first of the potential confounding factors (‘psychological distress’, that is, GHQ score) and low back pain.

Table 6.5.7 Variables in the equation.

								95% CI for Exp(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	psycho	.094	.018	26.249	1	.000	1.098	1.059	1.138
	Constant	-3.225	.438	54.245	1	.000	.040		

^aVariable(s) entered on step 1: psycho.

We can see from the table that the association between psychological distress and low back pain is highly significant ($p < 0.0005$). The OR [Exp(B)] for this association is 1.098. For each unit increase in psychological distress score, the increased risk of having low back pain is 9.8 per cent.

Note here though, that ORs operate in a multiplicative way, so that the increase in risk associated with a 2-point increase in the GHQ score is $1.098 \times 1.098 = 1.206$ (a 20.6 per cent increase), and a 3-point increase is associated with a $1.098^3 = 1.324$ (32.4 per cent) increase. We can therefore conclude that psychological distress is associated with low back pain, at least in univariate analysis. Similar univariate analyses show that ‘sex’ was marginally significantly associated with the outcome (OR = 1.6 for females compared to males; $p = 0.06$), and that ‘age’ is more strongly associated (OR = 1.02 per year; $p = 0.013$). For simplicity, in this example, we include only ‘psychological distress’ and ‘age’ as confounders in the multivariate analysis, excluding ‘sex’ from the model because this was not significant at the $p = 0.05$ level. Note though, as we described in Chapter 5, there are a number of different approaches that can be adopted to choose which variables will be included in a multivariable model. To recap, these might include

- A variable likely to have an important influence (e.g. sex of the subject). This might be determined from previous experience and the published literature.
- In univariate analysis, a variable that is significantly associated with the outcome below a specified threshold (e.g. $p < 0.05$, $p < 0.1$, etc. – this is the approach we have used here).
- When added to the multivariable model, a variable that reduces variance of the model by more than a specified amount (e.g. 5 per cent).

The Multivariable Model

Our multivariable model will determine the independent effects of each explanatory variable on the outcome (low back pain). The SPSS output for this multivariable logistic regression model is now described, starting with the Omnibus test for the overall model (Table 6.5.8).

Table 6.5.8 Omnibus Tests of Model Coefficients.

		Chi-square	Df	Sig.
Step 1	Step	66.763	11	.000
	Block	66.763	11	.000
	Model	66.763	11	.000

The OTMC table tells us how good the model is. We can see from the table that the logistic regression model incorporating the five explanatory variables (the three exposure variables and the two confounders identified in the univariate analysis above) significantly predicts the variation in low back pain in our study sample. The chi-squared value for the model is 66.763 on 11 degrees of freedom, and with a p -value of < 0.0005 we are more than 99.9 per cent certain that obtaining a value this large has not been due to chance. Therefore, we can conclude that our regression model predicts low back pain significantly well.

The variables in the equation table (Table 6.5.9) provide details of the model parameters (the values and the exponential of these values [ORs]) for the five explanatory variables together with their significance (p) values. For this table (Table 6.5.9) we have also specified the inclusion of 95 per cent CIs).

Table 6.5.9 Variables in the Equation.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1 ^a	hectic*			.554	3	.907			
	hectic(1)	.407	.585	.482	1	.487	1.502	.477	4.730
	hectic(2)	.362	.593	.374	1	.541	1.437	.450	4.590
	hectic(3)	.317	.598	.281	1	.596	1.373	.425	4.437
	monot*			10.516	3	.015			
	monot(1)	.333	.307	1.172	1	.279	1.395	.764	2.548
	monot(2)	.783	.313	6.265	1	.012	2.188	1.185	4.038
	monot(3)	.810	.308	6.908	1	.009	2.248	1.229	4.112
	stress*			15.296	3	.002			
	stress(1)	.724	.277	6.858	1	.009	2.063	1.200	3.547
	stress(2)	1.205	.321	14.057	1	.000	3.337	1.777	6.265
	stress(3)	1.189	.384	9.600	1	.002	3.283	1.548	6.962
	psycho	.069	.020	12.317	1	.000	1.071	1.031	1.113
	age	.025	.007	11.353	1	.001	1.025	1.010	1.040
	Constant	-5.251	.811	41.970	1	.000	.005		

^aVariable(s) entered on step 1: hectic, monot, stress, psycho, age.

*These rows correspond to the full (four-category) variable for psychosocial working environment and therefore have 3 (n-1) degrees of freedom.

The Effect of Adjustment

Interestingly, following this *adjustment* for the effects of other variables, the ORs (Exp(B)) for the three levels of 'hectic work' (compared to the 'hectic' reference category) are all attenuated, and none achieve statistical significance. This latter observation is confirmed by the 95 per cent confidence limits that span unity; for example, taking the highest level for 'hectic work' (comparing 'always' with 'never' finding work hectic), the OR estimate is 1.37 but might be as small as 0.43 or as large as 4.48.

This important result illustrates the effects of *confounding* on the relationship between 'hectic work' and low back pain. One or more of the other four variables (stressful work, monotonous work, psychological distress, and age) must explain part of the univariate association that we observed between 'hectic work' and low back pain by their mutual association with both factors.

In contrast, the univariate associations observed for monotonous and stressful work persist in the multivariate analysis, after adjustment for the other explanatory variables. As such, these psychosocial work factors are *independently* associated with low back pain; this is further evidence that GHQ score (psychological distress) was not a major factor in the 'causal pathway' between psychosocial working environment and low back pain. The two levels for monotonous work representing 'half the time' and 'always' are associated with more than a twofold increase in the risk of low back pain relative to 'never' finding work monotonous. The 95 per cent CIs for these ORs do not span unity, and the estimates are therefore significant.

Similarly, all three levels for stressful work are significantly associated with low back pain, with ORs ranging from 2.06 for 'occasionally' finding work stressful to 3.34 for finding work

stressful 'half the time'. Again, the 95 per cent CIs for the ORs do not span unity, and these estimates are therefore significant.

Finally, we can also see from the table that the significant associations observed for the two potential confounders (psychological distress and age) in the univariate analysis have persisted in the multivariate analysis. The ORs are similar to the univariate analysis, and the 95 per cent CIs do not span unity. Thus, psychological distress and age can also be interpreted as being *independent predictors* of low back pain.

In this example, we have seen how multivariable logistic regression is carried out and interpreted. In fact, this was a general form known as unconditional logistic regression, which is the appropriate method for unmatched case–control studies. We now look at the method for matched studies, and finally we review the results of regression analysis in the New Zealand study.

6.5.5 Matched Studies – Conditional Logistic Regression

The form of logistic regression used to analyse matched case–control studies is called *conditional logistic regression*, and as with unmatched analysis, this may be simple or multivariable. The outcome is now defined as a summary of the difference in disease status for each case–control pair. The x variable for each matched pair is the difference between the value of the variable for the case and that for the control. This is the analytic procedure used in the New Zealand study (Paper A). Whereas a different statistical procedure is used to carry out conditional logistic regression (which we do not consider further in this book), interpretation of the output from the analysis is essentially the same.

6.5.6 Interpretation of Adjusted Results from the New Zealand Study

We now consider analysis of the New Zealand study that uses conditional multivariable logistic regression to analyse maternal sleeping practices in relation to late stillbirth, independent of confounding. This brief section from the methods section of Paper A describes the statistical approach used:

Analysis

... We used standard conditional regressions for matched case–control studies A multivariable regression model was developed to include maternal variables reported to be associated with increased risk of stillbirth, based on previous literature (age, body mass index, ethnicity, parity, smoking and socioeconomic status) Statistical significance in multivariable analysis was defined at the 5% level. Global χ^2 statistics were used to assess the significance of variables in the models, and individual level odds ratios were estimated for each category in comparison to a reference category, defined as the category hypothesised to have the lowest risk.

Table 6.5.10 (including data and information extracted from Table 5, Paper A) provides information about which variables were included in the regression models, and the adjusted ORs (and 95 per cent CIs) for maternal sleeping practices. In addition p -values (from chi-squared analysis) are provided, summarizing the overall statistical significance of the association between the sleeping practice variables with the outcome. To judge the statistical significance of individual exposure categories (compared to the reference category) with late stillbirth, we need to focus on the 95% confidence intervals, checking that they do not include 1.0.

Table 6.5.10 Multivariable analysis of relation between maternal sleeping practices and risk of late stillbirth among 155 women who experienced a late stillbirth and 310 controls (based on Table 5 of Paper A).

Practice	Adjusted odds ratio*	95% CI	p value for the difference
Maternal sleeping position in last night of pregnancy:			
Left side	1.0		
Right side	1.74	0.98, 3.01	0.005
Back	2.54	1.04, 6.18	
Other	2.32	1.28, 4.19	
Regular sleep in daytime in last month of pregnancy:			
No	1.0		
Yes	2.04	1.26, 3.30	0.002
Hours of nighttime sleep in last month of pregnancy:			
<6	1.89	0.98, 3.65	0.05
6–8	1.0		
>8	1.71	0.99, 2.95	
Number of times getting up to toilet during last night of pregnancy:			
>1	1.0		
≤1	2.42	1.46, 4.00	0.002

*Adjusted for age, ethnicity, overweight or obesity, parity, social deprivation level, smoking, and the other variables in the table.



Self-Assessment Exercise 6.5.2

1. Describe how the authors of the New Zealand paper decided on which factors were potential confounders for the association between sleeping position and late stillbirth.
2. Looking at the relationship between hours of nighttime sleep in the last month of pregnancy with late stillbirth, why was the middle category (6–8 hours) given an odds ratio of 1.0?
3. Describe the relationship between maternal sleeping position in the last night of pregnancy and late stillbirth.
4. Describe the relationship between regular sleep in daytime in the last month of pregnancy and late stillbirth.
5. Describe the relationship between getting up to go to the toilet during the last night of pregnancy and late stillbirth.
6. Based on these results, is it reasonable to conclude that maternal sleeping practices are *causally* related to late stillbirth?

Answers in Section 6.6

Summary: Confounding and Logistic Regression

- Confounding can be dealt with in the design of the study, the analysis of the study, or (commonly) both.
- Design methods that help to reduce confounding include matched and stratified designs. Stratification can also be carried out during analysis.

- The analysis should be appropriate for the study design: matched analysis for a matched design, stratified analysis for a stratified design.
- Confounding can be dealt with in analysis by using unconditional multivariable logistic regression (for unmatched studies) or conditional multivariable logistic regression (for matched studies).
- The choice of which variables will be included in a multivariable model needs to be clearly explained. A number of criteria can be used to determine whether or not a variable should be included in the multivariable model:
 - it is a factor likely to have an important influence (e.g. sex of the subject); this might be determined from previous experience and the published literature
 - in univariate analysis, it is significantly associated with the outcome below a specified threshold (e.g. $p < 0.1$)
 - when added to the multivariable model, the variable reduces variance of the model by more than a specified amount (e.g. 5 per cent)
- Multivariable logistic regression analysis allows the OR for a putative risk factor to be estimated independently of the effect of other explanatory variables, including confounders.

6.6 Answers to Self-Assessment Exercises

Section 6.1

Exercise 6.1.1

1. Research issues, including limitations of other studies, and aim/hypothesis:
 - Stillbirth rates (1 in 200) have not changed for 20 years.
 - Studies of risk factors are typically retrospective population-based studies, unable to explore a variety of risk factors, including modifiable factors such as maternal lifestyle and habits.
 - Little research on sleep practices and effect on developing foetus: despite evidence of sleep disordered breathing being associated with pre-eclampsia and preterm birth, only one case report has looked at stillbirth.
 - The authors identify that obesity is linked to stillbirth risk but state the mechanisms are unknown and suggest this might be explained by obesity leading to disordered breathing.
 - Supine sleeping is related to sleep disordered breathing and reduced maternal cardiac output (late pregnancy) but there is no evidence for the association of sleeping position and practices with stillbirth.

The main stated aim of the study was to identify potentially modifiable risk factors for late stillbirth (≥ 28 weeks' gestation). These included a number of factors defined as being related to the health and behaviour of women during pregnancy (general health, socioeconomic factors, diet, and exercise, as well as maternal sleeping practices). The main hypothesis was that sleep disordered breathing and maternal supine (on back) sleeping increased the risk of late stillbirth.

2. Choice of study design

Case-control studies are typically chosen over cohort studies for outcomes and diseases that are rare or have a long latency period (waiting for the event to occur after exposure, such as with cancers) where the cohort design would be impractical. For the current study this is not an issue because stillbirths are not very rare and there is not an issue with the latency period. However, the choice of case-control study here was partly due to practicality. To obtain the main benefit of a cohort design, namely, the prospective time frame and assessment of

exposure prior to occurrence of the stillbirth, a population sample of pregnant women would need to be recruited, their sleeping practices assessed during pregnancy, and the sample followed up until sufficient numbers of stillbirths had occurred (approximately 1 in 200 births). We cover sample size calculations in Section 6.4, but based on the figures for the size of risk to be detected ($OR = 2.54$ for sleeping on back relative to sleeping on left, taken from Table 6.5.10) and the occurrence of stillbirth (1 in 200), approximately 5,250 women would need to be recruited, interviewed, and followed up. This would clearly be an issue in terms of the resources required to carry out a cohort study. Another reason for choosing the cheaper, quicker, and more practical case-control study design for this study was the lack of previous research and the testing of a new hypothesis (as described by the authors in the introduction). Before committing the substantial resources required by the more-powerful cohort study, it is worth investigating the hypothesis using a quicker and cheaper case-control design.

Section 6.2

Exercise 6.2.1

1. How cases were defined:
 - a. Women who gave birth to a stillborn baby at or after 28 weeks of gestation in the Auckland region between June 2006 and June 2009.
 - b. Stillbirth defined as 'birth of a baby that died in utero during the antenatal or intrapartum periods.'
 - c. Exclusions: (i) women whose baby had died from a congenital abnormality or was from multiple pregnancy or (ii) women who had not been booked to deliver their baby in the Auckland region (study population).
2. How and where cases were found:
 - a. Weekly checks were done of all (three) maternity clinics in the Auckland region by key physicians.
 - b. Regular checks of hospital birth records by lead researcher.
3. Quality of case ascertainment

Case ascertainment is likely to have been very good. As well as the regular (weekly) checking of all maternity clinics in the study area and hospital birth records, a registry (national system for perinatal data collection) that started in New Zealand at the beginning of the study was used to check identified cases and 'ensure complete ascertainment'.

Exercise 6.2.2

The controls should meet the principle well, for the following reasons. The cases represent the proportion of all women experiencing a late stillbirth and case ascertainment through surveillance of maternity clinic and hospital records, and case ascertainment is likely to be complete in a country such as New Zealand. Hence it should be representative of the population of all cases of late stillbirth. Two controls were randomly selected from the pregnancy registration lists of the district health board (there are three within the Auckland region) where the stillbirth occurred, with the same exclusion criteria as the cases. Controls were then matched to cases by gestation (we consider matching in case-control studies in more detail in Section 6.2.2), but the authors describe the selected controls as being representative of the antenatal population at the same gestation at which the stillbirth occurred. The controls therefore meet the requirements of control selection (i) being representative of the population from which the cases were selected (pregnant women with gestation of 28 or more weeks) and (ii) being free from the condition being investigated (not having had a stillbirth).

Exercise 6.2.3

- There are two important reasons for not matching on too many factors:
 - It would not be practical. If there is a long list of factors that have to be matched, it would be very difficult and time-consuming to find a control that meets all of the criteria for each case.
 - It is quite possible to overmatch in a case–control study. If one or more of the variables used for matching are strongly associated with the risk factor being studied, this could inadvertently lead to matching (at least in part) for the risk factor being studied. In Figure 6.1.2, we saw that the purpose of a case–control study design is to measure the difference in exposure to the risk factor between cases and controls. If by matching we end up making the cases and controls more similar with respect to that risk factor, the very difference we are trying to measure will be reduced. It would be (in Figure 6.1.2) like reducing the percentage of shaded squares from 50 per cent to 45 per cent among cases and increasing it from 30 per cent to 35 per cent among controls, weakening the association. It is therefore wise, and more practical, to match for a small number of key confounders (two or three at most), the effects of which are known already, and that are not associated with the risk factor of interest. Later in this chapter we look at how other confounding factors can be dealt with in the analysis.
- If the matching variable is along the causal pathway between disease and exposure, then matching will contribute bias, reducing or removing the true effect observed when analysing the association between the exposure and disease.

Exercise 6.2.4

All data were obtained through interviewer-administered questionnaires in the first few weeks after stillbirth for cases and at the equivalent gestation of pregnancy for matched controls.

Assessment of sleeping position and sleep quality is provided in the table below, with commentary on possible sources of bias.

Risk factor	How assessed	Comment on possible bias
Sleeping position	Recall of position on (i) going to sleep and (ii) waking. Women asked (a) before pregnancy and (b) last month, week and night of pregnancy.	Recall bias could be an issue, with cases reporting sleeping position differently to controls (either more accurately due to the traumatic event or inaccurately to explain occurrence of stillbirth). The accuracy of reported sleeping position might be poor for longer periods of time (e.g. before pregnancy and/or month or week before interview). Inaccurate recall could simply add noise (random error) in equal measure to both cases and controls: The key issue is whether this error is systematically different between cases and controls and would lead to bias. Although we are not given any information on the potential for recall bias, the researchers state they tried to minimise this by ensuring ‘participants were not aware of any of the specific research questions related to risk factors for stillbirth.’

(continued)

Risk factor	How assessed	Comment on possible bias
		The interviewers could have also introduced bias by eliciting information differently from cases and controls if they were aware of a woman's case/control status (very likely) and of the hypothesis. Training of interviewers is important to minimise this potential source of bias; the researchers do not indicate whether the interviewers were trained.
Quality of sleep	Sleep disordered breathing assessed as snoring and daytime sleepiness (Epworth sleepiness scale)	These were proxy measures for sleep disordered breathing in pregnancy in the absence of previously validated tools. These exposures might not accurately reflect the risk factor (sleep disordered breathing) but, again, with regard to recall bias, the accuracy of responses to these questions would have had to systematically differ between cases and controls. This might have occurred if participants suspected the research hypothesis of the association between quality of sleep and late stillbirth.
	In the last month: (i) regular sleep during the day, (ii) duration of sleep (<6, 6–8, >8 hours), and (iii) frequency of getting up to the toilet.	Recall and interviewer bias as discussed for sleeping position are also relevant for enquiry about the quality of sleep.

Exercise 6.2.5

Study	Potential for bias in the assessment of exposure to risk factor
1	<p>Although there is a risk that telephone interviews could exclude certain groups of the population, in the USA up until the last few years, virtually everyone had a telephone. The more recent issue of using mobile phones in place of landlines may, however, be a complicating factor for recent or future studies utilising such an approach; for example, there are typically no suitable sample frames (lists of mobile phone users) available for sampling within a specific geographical area. There are also two other potential sources of bias to consider:</p> <p>Would the interviewers have known whether they were talking to a case (with breast cancer) or a control (disease free)? If so, could that knowledge have influenced the way they elicited the information about OCP use? If so, that could lead to interviewer bias.</p> <p>Is it possible that the cases, as a result of their experiences and awareness since diagnosis, would be able to give a more-detailed history of OCP use, or perhaps exaggerate their OCP use compared to controls? If so, this would be recall bias.</p>
2	<p>Because this scenario is based on records, we might expect no reason for a difference between cases (children with TB) and controls (disease free) in the recording of this information. However, there are likely to be quite a lot of children with poor records, or no records at all. It is likely that children who are more susceptible to TB live in areas or circumstances where record keeping is poorer. This would lead to bias: For example, we might have a situation in which (assuming BCG was effective in this population) disease-free controls who had in fact been vaccinated did not have this event recorded. This would lead to the study underestimating the efficacy of the vaccine.</p>
3	<p>Birthweight is recorded fairly accurately in the records of virtually every child born in the UK, and this was true 10 years ago. Furthermore, there is no reason to think that cases (children subsequently performing poorly in school at age 10 years) would have any less adequately recorded data than the controls. It is difficult to see how this method of exposure assessment could result in bias in the case-control comparison.</p>

Section 6.3

Exercise 6.3.1

1. It is very useful to go through the process of setting out the data in this way:

Sleeping Position on going to sleep on the last night of pregnancy

	Cases (late stillbirth)	Controls
Exposed (sleeping on right side)	49	84
Not exposed (sleeping on left side)	42	132
Total	91	216

	Cases (late stillbirth)	Controls
Exposed (sleeping on back)	15	15
Not exposed (sleeping on left side)	42	132
Total	57	147

2. $OR(\text{right side}) = (49 \times 132) \div (42 \times 84) = 1.833$, which can be rounded to 1.83.
 $OR(\text{back}) = (15 \times 132) \div (42 \times 15) = 3.143$, which can be rounded to 3.14.
3. The OR is interpreted as follows: Compared to sleeping on the left side, the odds of sleeping on the right side are 1.83 times greater and on the back are 3.14 times greater in cases (women who experienced a late stillbirth) than controls (currently pregnant women). As the prevalence of the outcome is low (in New Zealand rates of stillbirth are approximately 3.5 per 1,000 per year, or 0.35%), the OR provides a close approximation to the relative risk (see Figure 6.3.2). We can therefore interpret the ORs as indicating that women who sleep on their right side and back are 1.83 times and 3.14 times more likely, respectively, to have a late stillbirth than women who sleep on their left *on the last night of pregnancy*.
4. The CI gives us an estimate of the precision of our sample OR in relation to the true population OR. As the CI does not include 1.0 (no increased or decreased risk), we can be 95 per cent confident that our sample ORs of 1.83 and 3.14 represent a true increase in risk in the population. Or put another way, we are 95 per cent confident that the risk of stillbirth among women in the population who sleep on their right side compared to those that sleep on the left side lies in the range 1.12 to 3.01 and the risk of stillbirth among women in the population who sleep on their back compared to those who sleep on the left side lies in the range 1.42 to 6.96.
5. *Sleeping Position on waking up on the last night of pregnancy*

	Cases (late stillbirth)	Controls
Exposed (sleeping on right side)	45	72
Not exposed (sleeping on left side)	31	106
Total	76	178

	Cases (late stillbirth)	Controls
Exposed (sleeping on back)	23	37
Not exposed (sleeping on left side)	31	106
Total	54	143

OR (right side) = $(45 \times 106) \div (31 \times 72) = 2.137$, which can be rounded to 2.14.

OR (back) = $(23 \times 106) \div (31 \times 37) = 2.126$, which can be rounded to 2.13.

The odds of sleeping on the right side and on the back, compared to sleeping on the left, is more than two times (OR = 2.14 and OR = 2.13 respectively) greater in cases (women who experienced a late stillbirth) than controls (currently pregnant women). Alternatively, expressed as the estimate of relative risk, we can say that women sleeping on their right side or back have more than twice the risk of a late stillbirth than women sleeping on their left side *on the last night of pregnancy*. The interpretation of the 95 per cent CIs (1.24 to 3.69 and 1.10 to 4.10) is similar to that in question 4, noting that the interval does not include 1.

Section 6.5

Exercise 6.5.1

1. For non-smokers OR = 1.14, and for smokers OR = 1.00. If chi-squared tests had been carried out, the following values would have been obtained: for non-smokers, $\chi^2 = 0.0780$, $p = 0.78$; for smokers, $\chi^2 = 0.0000$, $p = 1.0$. Both of the OR values are close to (and one is equal to) 1.0, and, not surprisingly, neither reaches statistical significance ($p < 0.05$). Thus, having eliminated the effect of smoking, we see that there is no association between alcohol and MI.
2. If we had ignored the stratification in the analysis (or had not stratified, but obtained these same data), the result would be as follows: OR = 1.78, $\chi^2 = 4.023$, $p = 0.0449$. This OR is (just) significant at the 5 per cent level, so we would conclude that there is a positive association between MI and alcohol. Ignoring the stratification, or failing to take smoking into account in some way, gives a misleading result due to confounding by smoking. While the figures in this example are made up, the results are consistent with research findings.

Exercise 6.5.2

1. Potential confounders included *maternal variables reported to be associated with increased risk of stillbirth, based on previous literature*. Confounders included age, body mass index, ethnicity, parity, smoking and socioeconomic status. This a priori approach to identifying potential confounders is appropriate – to be a *confounder*, the variable needs to be associated with both the exposure (maternal sleeping practices) and the outcome (stillbirth), and the literature has already identified an association with the latter.
2. The authors stated that the *reference category* for each comparison was *defined as the category hypothesized to have the lowest risk*. The reference category (unexposed) is given a value of 1.0 in calculations of the odds ratio to assess the impact of other categories of exposure on the risk of the outcome. In the comparison of hours of nighttime sleep with risk of late stillbirth, the category 6–8 hours was chosen as the reference, reflecting normal sleeping duration (perhaps based on the average duration of sleep). It is therefore possible to estimate the risk of late stillbirth with less than average sleep duration (<6 hours) and more than average sleep duration (>8 hours) compared to normal sleep duration (6–8 hours).
3. There does appear to be an association between maternal sleeping position reported in the last night of pregnancy and late stillbirth. Compared to sleeping on the left side, sleeping on the right is associated with a 74% increase in risk of late stillbirth (although not statistically significant, as the 95% confidence interval ranges from a 2% reduction in risk to a three-fold increase in risk; it spans OR = 1.0). Sleeping on the back or 'other' (described as including front, sitting up, sleeping on both sides, unsure, or don't remember) relative to the left side is associated with a significant increase in the risk of late stillbirth with an OR = 2.54 (95% CI 1.04 to 6.18) and OR = 2.32 (95% CI = 1.28 to 4.19) respectively. It is reasonable to say that maternal sleeping position in the last night of pregnancy is *independently* associated with

risk of late stillbirth due to the multivariable analysis adjusting for other maternal sleeping practices and a range of a priori potential confounders.

4. As with maternal sleeping position, it is apparent that regular daytime sleep in the last month of pregnancy is *independently* associated with a significant increase in the risk of late stillbirth, with approximately a two-fold increase in risk (OR = 2.04; 95% CI = 1.26 to 3.30) after adjustment.
5. Finally, we can see that the number of times the pregnant women got up to go to the toilet during the last night of pregnancy was also *independently* associated with a significant increase in the risk of late stillbirth. Women who reported going to toilet only 1 time or not at all had more than a two-fold increase in risk (OR = 2.42; 95% CI = 1.46 to 4.00) compared to those who went to the toilet more than once, after adjustment.
6. It is not appropriate to say that maternal sleeping practices are *causally* associated with late stillbirth based solely on the results of this case–control study. To assess whether a risk factor is causally associated with an outcome, a number of factors need to be considered, which we described in relation to the Hill viewpoints in Chapter 5 (Section 5.7.1). As we identified in Section 6.2.4, retrospective case–control studies can suffer from bias due to exposure information being collected after the outcome has occurred; this criticism was levied by the authors of the Danish Cohort study, who did not find such an association. In addition, we would need more information about the relationship, including the possible mechanisms, for example, in relation to how sleeping position might lead to late stillbirth.

We would also need to rule out other possible explanations for these associations, and one such explanation could be the presence of reverse causality. There did not appear to be a difference between sleeping positions in cases and controls before pregnancy (Table 2 of Paper A) and therefore no differential pattern of sleeping according through habit. Therefore it is likely that something about the pregnancy affected their maternal sleeping position (and not the sleeping position affecting the pregnancy or pregnancy outcome). One critique of the New Zealand study postulated just this situation of reverse causality (Froen *et al.*, 2011). ‘As the baby grows, the potential for pressure against the vena cava in a supine sleeping position increases, as probably do all other known and unknown mechanisms that make women gradually prefer a (left) lateral tilt in late pregnancy. What would reduce normal progression toward the preference of a more lateral tilt? A smaller belly. An increased proportion of growth restricted babies with oligohydramnios (too little amniotic fluid) among cases would potentially provide exactly the results found in this study, including the appealing finding of greater differences in sleeping position as pregnancy progresses. Indeed, a large proportion of stillborn babies are severely growth restricted and their progress toward death is a gradual one, not a single sudden event. What would explain higher rates of daytime sleep and more nights with less than 6 hours sleep? A sick mom with a complicated pregnancy. Again not a rare finding in stillbirth.’ As the author stated “There is no need to worry about sleeping position in pregnancy – quite yet” or, put another way, more definitive research is needed before it is possible to confirm the association between maternal sleeping practices and late stillbirth.

7

Intervention Studies

Introduction and Learning Objectives

The study designs that we have examined so far have relied on either going to existing data sources or collecting new information on risk factors, health status, and events, and then analysing the associations. In essence, we have been observing what is going on out there, albeit in quite sophisticated ways, especially with the case-control and cohort studies. The key point here is the observation: What we have not done is to deliberately change things; that is, we have not carried out an intervention. The study designs examined so far are therefore described as **observational**.

We now go on to look at **intervention** research. The most important new feature of these studies is that we deliberately intervene to alter the level of one or more factors, the effect of which we are interested in studying. This has important consequences for the strength of the study design, as well as for the ethical implications of the research.

We begin our exploration of intervention studies with a relatively simple example – a drug – in a design where the subjects have an equal chance of receiving, or not receiving, the drug by virtue of *randomisation*. Once you are familiar with the basic design features of a randomised trial, with blinding and placebo control, we will look at studies where the interventions being tested are rather more complex and cannot easily be tested with this ideal design. These alternative designs are important because many aspects of health care and health promotion present considerable challenges for the design of trials. Some retain randomisation, and others do not. The list below presents a simple typology of the intervention designs covered in this chapter, according to whether the allocation to intervention is by randomisation or through some alternative approach.

Typology of Intervention Study Designs Described in This Chapter

Randomised designs:

- Individual randomised, controlled trial
- Cluster randomised, controlled trial
- Cross-over trial

Non-randomised designs:

- Controlled and uncontrolled before-and-after study
- Interrupted time series studies
- Natural experiment

Terminology

All of these study designs are explained in the following sections of this chapter. Throughout, we use the term *intervention* to describe studies designed to assess the effect on health (or a marker of health) of a deliberate change made by the investigation team. The one circumstance in which the change may or may not be deliberate is with the *natural experiment*, where the change (e.g., a nuclear accident, volcanic eruption, ban on smoking in public places) either is a natural occurrence, is unintended, or is deliberate but not on the part of the investigators. The term *intervention* is used to describe the actual change, for example, a drug, surgical procedure, information package, or policy. The term *trial* is often taken as being synonymous with a *randomised, control trial* (RCT), and it is used in this chapter only in respect of RCTs.

It is also possible to investigate the impact of an intervention using observational study designs (cross-sectional, case-control, cohort), but by definition, the investigators have no control over the allocation of the intervention. In that sense, a natural experiment is also an observational design, but it is included in this chapter due to the nature of the change, which may be marked, sudden, or deliberate on the part of policy makers or through natural or other circumstances.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the purpose and structure of various intervention study designs, including their strengths and weaknesses, within an overall framework of epidemiological study designs.
- Describe what is meant by inclusion and exclusion criteria in the context of an intervention study.
- Describe the purpose of randomisation of intervention allocation, and how randomisation can be done in practice, including the practice of concealment.
- Describe the purpose and technique of blinding and the methods for achieving some degree of blinding in situations where the nature of the intervention cannot be hidden.
- Describe what is meant by the term ‘placebo control’ and why this is used.
- Describe what is meant by the terms ‘analysis by intention to treat’ and ‘analysis per protocol’, and explain the application, strengths, and weaknesses of both.
- Carry out an appropriate statistical analysis of a two-group trial with either a categorical or continuous outcome variable(s).
- Identify situations where paired data are generated and describe in general terms the methods for analysis of paired data.
- Describe the important ethical issues in designing and carrying out an intervention study including informed consent.
- Describe what information is required to calculate the sample size of a two-group trial and carry out this calculation using software.
- Describe how to assess whether there are important imbalances, following randomisation, in baseline characteristics that might influence the outcome, and describe when and how to deal with these using multivariable-regression methods.
- Describe the basis for using multi-level modelling in the analysis of a cluster RCT (and in other study designs), and interpret analysis carried out with this method.
- Describe the approach to calculating sample size for a cluster RCT, including the rationale for, and use of, the intraclass correlation coefficient (ICC).
- Describe key aspects of good practice for trial registration and management and the use of the CONSORT statement for reporting.

Resource Papers

We explore intervention studies through the following four papers.

Principal References

Paper A

Bollinger, C. T., Zellweger, J.P., Danielsson, T., van Biljon, X., Robidou, A., *et al.* (2000). Smoking reduction with oral nicotine inhalers: double blind, randomised clinical trial of efficacy and safety. *Br Med J* **321**, 329–333.

Paper B

Day L, Fildes, B., Gordon, I., Fitzharris, M., Flamer, H., *et al.* (2002). Randomised factorial trial of falls prevention among older people living in their own homes. *Br Med J* **325**, 128–134.

Supplementary References

Paper C

Elley, C.R., Kerse, N., Arroll, B., Robinson, E. (2003). Effectiveness of counselling patients on physical activity in general practice: cluster randomised controlled trial. *Br Med J* **326**, 793–799.

Paper D

Biglan, A., Ary, D., Smolkowski, K., Duncan, T., Black, C. (2000). A randomised controlled trial of a community intervention to prevent adolescent tobacco use. *Tob Control* **9**, 24–32.

7.1 Why Do an Intervention Study?

7.1.1 Study Objectives

We begin by looking at the research question identified by the research team for Paper A and seeing why a randomised intervention study design was thought appropriate. The issue that the authors were interested in is described in the short excerpt below:

The best way to prevent the detrimental health consequences of cigarette smoking is to quit, and efforts to date have focused on this strategy. Many smokers, however, find it impossible to quit, even with help, because of their dependence on nicotine, which is a highly addictive psychoactive drug. (Bollinger *et al.* (Paper A))

Smoking is one of the most important contributors to the global burden of disease (Lim *et al.*, 2012), causing an estimated 5.7 million premature deaths globally in 2010. Primary prevention (ways of preventing people from adopting tobacco smoking) is the priority, but we also need to help smokers to quit. As the above quotation from Paper A emphasises, many smokers find this very difficult, so finding effective ways to assist them is important. The first exercise will help you understand why the authors chose an intervention study design.

Please now read the following, which is the abstract from Paper A:

Abstract

Objectives: To determine whether use of an oral nicotine inhaler can result in long term reduction in smoking and whether concomitant use of nicotine replacement and smoking is safe.

Design: Double blind, randomised, placebo controlled trial. Four month trial with a two year follow up.

Setting: Two university hospital pulmonary clinics in Switzerland.

Participants: 400 healthy volunteers, recruited through newspaper advertisements, willing to reduce their smoking but unable or unwilling to stop smoking immediately.

Intervention: Active or placebo inhaler as needed for up to 18 months with participants encouraged to limit their smoking as much as possible.

Main outcome measures: Number of cigarettes smoked per day from week six to end point. Decrease verified by a measurement of exhaled carbon monoxide at each time point compared with measurement at baseline.

Results: At four months sustained reduction of smoking was achieved in 52 (26%) participants in the active group and 18 (9%) in the placebo group ($P < 0.001$; Fisher's test). Corresponding figures after two years were 19 (9.5%) and 6 (3.0%) ($P = 0.012$).

Conclusion: Nicotine inhalers effectively and safely achieved sustained reduction in smoking over 24 months. Reduction with or without nicotine substitution may be a feasible first step towards smoking cessation in people not able or not willing to stop abruptly.



Self-Assessment Exercise 7.1.1

Why do you think the research team opted to carry out an intervention study to test this approach to smoking reduction, rather than an observational design such as a case-control or cohort study?

Answers in Section 7.8

7.1.2 Structure of a Randomised, Controlled Intervention Study

Figure 7.1.1 illustrates the generalised structure of a randomised control trial with two groups. Note that such studies are also commonly described as 'randomised controlled trials', as has been done by the authors of Paper A; we will use the first description throughout this chapter.

One reassuring point to note is that intervention studies have much in common with cohort studies, and many of the methodological issues that we examined in Chapter 5 apply equally to intervention studies here. Make sure that you understand the design features in Figure 7.1.1 before doing the next exercise.

The methods used for the study in Paper A include the following features: randomisation, double blinding, and placebo control. You may already know something about what these terms mean, but we will study each in more detail shortly, together with the reasons for incorporating these features into the study design. Later in this chapter, we look at intervention study designs for which it is not possible to include some or any of these features and consider the implications for interpretation of the results.

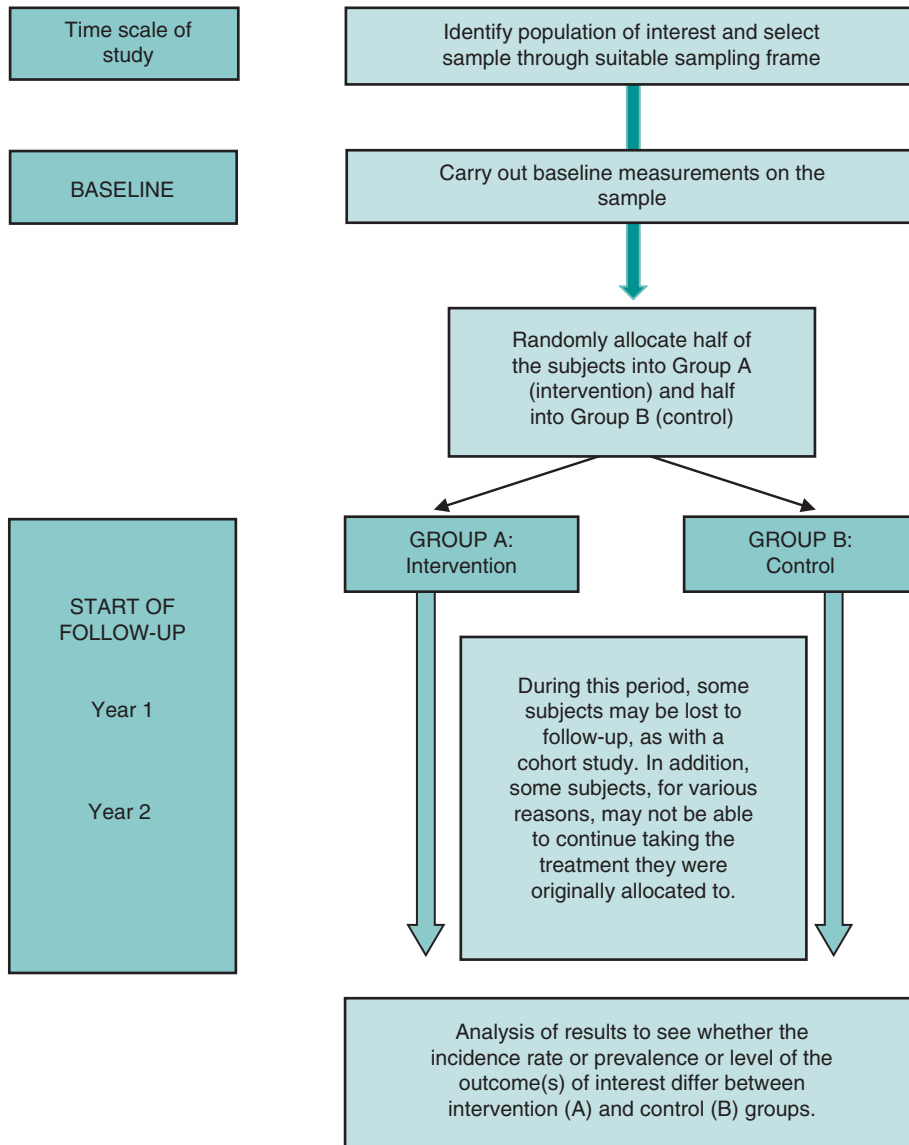


Figure 7.1.1 Generalised structure of a randomised control trial with two groups, testing an intervention in comparison with a control and (in this case) with a 2-year follow-up. Randomisation, blinding, and the use of a placebo are examined in Section 7.2.

At this stage, it is useful for us to look at the rationale and structure of the intervention study design in what might be termed its purest form, the *randomised blind (placebo) control trial*, of which Paper A is an example. This is important because the randomised control trial (RCT) is often regarded as the gold standard when it comes to providing evidence about the *efficacy* of preventative measures or health care. Exercise 7.1.2 will introduce you to more detailed aspects of the design of this type of study. Please now read the following excerpt, which includes parts of the methods section of Paper A relating to study design, participants, and the treatment investigated:

Methods

Study design

Smokers were recruited into this two centre, double blind, placebo controlled, randomised clinical trial through newspaper advertisements that asked for healthy smokers who were unwilling or unable to quit but were interested in reducing their smoking. All participants were given information about possible ways to achieve this goal. Smoking cessation was recommended as the ultimate goal throughout the study.

Participants

Participants in the trial had to be at least 18 years of age, smoke 15 or more cigarettes a day, have a carbon monoxide concentration in exhaled air >10 ppm, have smoked regularly for three or more years, have failed at least one serious attempt to quit within the past 12 months, want to reduce smoking as much as possible with the help of the nicotine inhaler, be prepared to adhere to the protocol and be willing to provide informed consent. Exclusion criteria were current use of nicotine replacement therapy or any other behavioural or pharmacological smoking cessation or reduction programme, use of other nicotine containing products or any condition that might interfere with the study. The ethics committees of the universities of Basle and Lausanne approved the study.

Treatment

Independent pharmacists dispensed either active or placebo inhalers according to a computer generated randomisation list. All smokers received information about the general implications of smoking and its effects on health. Participants were asked to reduce the number of cigarettes smoked daily as much as possible and an initial reduction of 50% was suggested. The active treatment comprised nicotine replacement through an inhalation device (Nicorette Inhaler, Pharmacia and Upjohn). The inhaler consists of a plastic mouthpiece into which a disposable cartridge containing 10 mg nicotine and 1 mg menthol is inserted. At room temperature the total available nicotine content is 4–5 mg per cartridge. The inhaler delivers about 13 g of nicotine per puff (average puff volume of 50 ml), which means that about 80 puffs are required to obtain 1 mg nicotine. The placebo inhalers were identical in appearance and contained only menthol. Both treatment groups were allowed to use the inhalers as needed, with the recommendation to use between six and 12 cartridges over 24 hours. Participants were encouraged to decrease use of the inhaler after four months but were permitted to continue treatment for 18 of the 24 months in the study.



Self-Assessment Exercise 7.1.2

According to information in the abstract (Section 7.1.1) and the methods section (above) of Paper A:

1. What were the objectives of this study?
2. From what population was the study sample selected?
3. How was the study sample selected? Comment on how suitable you think this sample is for addressing the objectives of the study.
4. What were the intervention and control treatments?

Answers in Section 7.8

Summary

- Intervention studies provide a powerful research design.
- The process of selecting a sample and following up the groups has much in common with cohort studies.
- As with any research design, practical considerations (time, money, design feasibility, etc.) influence the choice of population from which the sample is drawn and hence the generalisability of the findings.

7.2 Key Elements of Intervention Study Design

7.2.1 Defining Who Should be Included and Excluded

We have seen that the population for this study was residents in the areas of Basle and Lausanne, Switzerland, qualified by the requirement that they were smokers who were unwilling or unable to quit but were interested in reducing their smoking. Furthermore, since recruitment was via newspaper advertisements, the population was also defined as people who read newspapers or would otherwise see (or be told about) this type of advertisement; this probably includes most of the adult population, but we cannot be sure of this.

The next step is to establish exactly what factors determined whether a subject was included in the sample and who was excluded, known as *inclusion* and *exclusion criteria*. The next exercise looks at these criteria.

For this exercise, we also refer to the first part of the study flow chart (Figure 7.2.1). This diagram maps out all of the stages of the trial and the number of subjects included at each stage. A flow such as this one is an important aspect of the reporting of a trial (see the CONSORT statement in Section 7.7.3).

**Self-Assessment Exercise 7.2.1**

Using the information from the abstract (Section 7.1.1), methods (Section 7.1.2), and the flow chart (reproduced in Figure 7.2.1) of Paper A,

1. List all the criteria used to determine whether or not a subject took part in the trial.
2. Was the selection of study subjects through a random sampling procedure or by some other means of selection?
3. How representative do you think the study subjects are of smokers aged 18 and older in Switzerland?

Answers in Section 7.8

Paper A

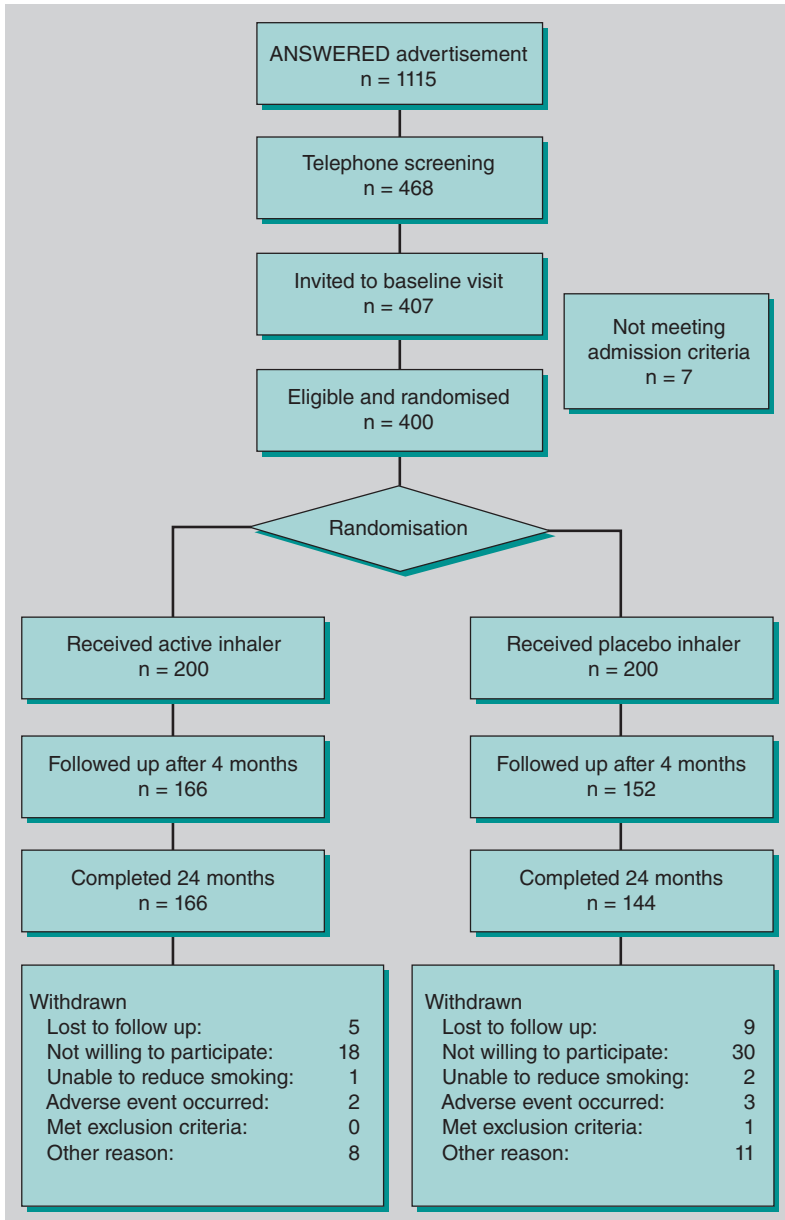


Figure 7.2.1 Study flow chart (paper A): progress of participants in the trial of oral nicotine therapy as an aid for reduction in cigarette smoking. *Source:* Bollinger 2000. Reproduced with permission of BMJ Publishing Group Ltd.

7.2.2 Intervention and Control

Defining the Intervention

In our exploration of intervention studies we consider a range of interventions, from single medications such as the nicotine inhaler (Paper A), through complex care packages for the prevention of falls among the elderly delivered through primary care, to a community-wide health promotion programme on smoking.

For now, though, we will keep to the nicotine inhaler study. The intervention is simple and clearly defined, as follows: the use, as needed, of an inhaler that delivers about 13 grams of nicotine per puff (methods section of Paper A, Section 7.2.1). Both the nicotine and placebo inhalers contained 1 mg of menthol; we turn to the reason for this in the next section.

An intervention study may be designed to test the *efficacy* of a treatment, that is, whether it works (and how well) when delivered in ideal circumstances. Increasingly, however, we are also interested in *effectiveness*, that is, how well the treatment works when used in circumstances that reflect typical care or service delivery; these are known as *pragmatic trials*.

This latter type of intervention study should be designed, insofar as is possible, with a view to providing results that are relevant to wider implementation of the intervention. The intervention being tested and the circumstances in which the trial is done (e.g., sample, exclusions) should be as close as possible to the setting in which it is hoped the intervention would ultimately be applied in the population.

In the case of Paper A, if the trial is successful and the scientific community recommend that the inhaler should be used for smokers ‘unwilling or unable to quit, but interested in reducing their smoking’ the intervention would be applied to the general population in pretty much the same way as it was in the trial. This means that the researchers are testing an intervention that is delivered, packaged, and used almost exactly as it would be if recommended for the general population.

We noted, however, that the sample in this study was selected to ensure good compliance – indeed, this was one of the eligibility criteria. As a result, we might conclude that the findings would be most relevant to well-motivated subjects.

What Should the Intervention be Compared with?

A crucial question arises as to what the new intervention should be compared with. In the case of this study (Paper A), a *placebo* was used. This is a chemically inert substance designed to be indistinguishable from the intervention treatment. So, in this case, the nicotine inhaler is being compared with nothing, although it was a special kind of nothing. We look at what is special about placebo after this next exercise.



Self-Assessment Exercise 7.2.2

Again using information from the methods section of Paper A (Section 7.1.2),

1. Make brief notes on exactly what the nicotine inhaler was being compared with.
2. In what circumstances do you think it is appropriate to compare a new intervention with no active treatment (whether this is nothing at all or a placebo)?

Answers in Section 7.8

Why Placebo?

It is worth restating the key point from this exercise: Ethical practice requires that the control (comparison) group be offered the best existing treatment that is in routine use. If there is no known effective treatment, then the controls can be given no active treatment. If practical, this ‘no active treatment’ may be given as a placebo.

Why should the research team go to the trouble of providing an inactive intervention that appears identical to the real nicotine inhaler? They do so because of the *placebo effect*. The fact of being given something, especially something that looks, feels, smells, and tastes like a medication, can influence the way people perceive their health and may in some circumstances

actually directly affect health outcomes. So part of the effect of an active drug is this influence on perception, and part is due to the pharmacological effect of the chemicals on the body (in the case of nicotine, reducing the craving for a cigarette). To identify the effect that is only due to the pharmacological effect of the drug, a placebo is introduced as the comparison, as shown in Figure 7.2.2:

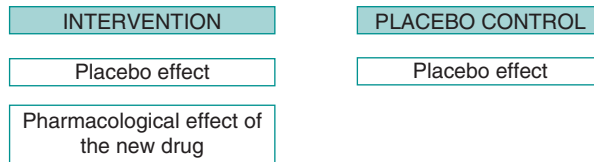


Figure 7.2.2 Comparison of placebo and active treatment in a two-arm trial.

In a RCT, the intervention and control groups should experience an identical placebo effect. Any observed difference in outcome will therefore be due to the pharmacological effect of the active drug. Although it is relatively easy to produce a placebo for a drug, it is much more difficult for other types of intervention. We return to this later in the chapter.

7.2.3 Randomisation

Purpose

One of the most important design features of this type of intervention study is the randomisation of study subjects to receive either the intervention or the control ‘treatment’. The particular strength of randomisation is that, if successful, it avoids the effects of *confounding* factors.

Confounding can operate in an intervention study just as it does in the other study designs we have looked at. Suppose that in the smoking-cessation study, subjects were allocated to intervention and control groups in a non-random way, and as a result those using the active inhaler were lighter smokers or belonged to a socioeconomic group that was more highly motivated to reduce smoking than those on placebo? All other things being equal, we would expect the reduction in smoking to be greater in the nicotine-inhaler group before taking account of any beneficial effect of the drug. This would, of course, be highly misleading.

Randomisation balances the two groups in terms of social, demographic, lifestyle, and other personal characteristics that could have a bearing on the outcome. This is the really big advantage of the randomised intervention design. Until now we have generally had to try to measure confounding factors and adjust for their influence in the analysis. The one exception was the *matching* process in the case-control design, where we selected cases and controls in a way that avoided the influence of potential confounding, although this could only be applied to quite a limited number of confounding variables. The great strength of the randomised trial is that we can arrange the comparison between the intervention and control groups in a way that avoids the influence of confounding factors. This also includes confounders that we do not know about and therefore would not even measure, match, or adjust for in other types of study.

In Section 7.3 we look at how we assess the extent to which randomisation has resulted in a good balance (between intervention and control groups) of factors that could influence the outcome in this way.

Method of Randomisation

You will encounter a variety of methods for actually carrying out randomisation. The most important point is to ensure that the process of random allocation is not influenced by the

characteristics and circumstances of the subject being allocated. The technique for doing this is known as **concealment** and involves preparing the random-allocation listings by someone separate from those individuals who have contact with subjects and the responsibility for implementing the allocation of each subject to either intervention or control.

You can see how this concealment was achieved in the nicotine study, where ‘independent pharmacists dispensed either active or placebo inhalers according to a computer-generated randomisation list’, the creation of which (we can presume) they did not play any part in (please refer back to the text excerpt on the study methods for Paper A, Section 7.1.2).

In many clinical trials, subjects are entered into the study as they present for treatment, with the requisite sample size being built up over a period of months or years. A common procedure in this situation would be to have pre-randomised, numbered envelopes available to the doctor or surgeon carrying out the treatments. When the next eligible person presents, the next envelope is opened, and this informs the doctor about which treatment that person is to receive. Since the allocation decision has been taken in advance, it cannot be influenced by the circumstances of the person when they present for treatment.

7.2.4 Outcome Assessment

The definition and assessment of outcomes is as important in intervention studies as in other study designs, and we will look at how this was carried out in the nicotine trial. Please now read the following excerpt taken from the methods section of Paper A describing assessment and measures of outcome:

Methods

Assessment

After the initial telephone screening and baseline assessment, participants were reassessed at the clinic after one, two, three and six weeks and three, four, six, 12, 18 and 24 months. Counselling on smoking reduction was provided at each visit. Admission criteria and demography, including the Fagerström test for nicotine dependence, reasons for reducing smoking and medical and smoking history were assessed at baseline. At all key visits (baseline and months four, 12, and 24) we measured expired carbon monoxide concentrations, symptoms of withdrawal, smoking status, respiratory function, blood pressure, pulse, weight, plasma cotinine concentration, haematological variables and concentrations of blood lipids and fibrinogen. We also assessed smoking status, intention to quit, compliance, concomitant medications, adverse events and quality of life (with the SF 36 questionnaire). Reported adverse events and plasma cotinine concentrations served as a basis for the safety analysis of concomitant smoking and nicotine replacement therapy.

Measures of Outcome

The primary efficacy measure (success) was defined as self-reported reduction of daily cigarette smoking by at least 50% compared with baseline from week six to month four, the duration for which the study was powered. This reduction was verified by decreased carbon monoxide concentrations at week six and months three and four. Results up to 24 months are presented in this paper. Smoking cessation was defined as not smoking from week six and a carbon monoxide concentration <10 ppm at all subsequent visits. Smoking reduction and cessation are also presented with verification of carbon monoxide concentrations at each time point (point prevalence).



Self-Assessment Exercise 7.2.3

According to the above information from the methods section of Paper A:

1. What is the main outcome (measure of success)?
2. How was this validated?
3. What other important outcome was assessed?

Answers in Section 7.8

7.2.5 Blinding

Blinding means preventing a person involved in the study from knowing which treatment a subject is receiving. In respect of this, there are essentially three groups of people involved in the study that can be blinded:

- the subjects, who therefore do not know whether they are taking the intervention or control treatment;
- the research team who are involved in data collection and analysis;
- the staff carrying out the treatments and patient care, which may include at least some of the research team.

You will encounter the terms single blinding and double blinding (and triple blinding). These terms refer to blinding of one or two (or all three) of the groups described here.

Unfortunately, these terms tend to be used by researchers to refer to different groups or combinations of groups that are blinded, and the best-practice guidance on reporting of trials (known as CONSORT, described further in Section 7.7) recommends that they should be avoided. It is better practice to state clearly which groups have been blinded and how.



Self-Assessment Exercise 7.2.4

Using the information from the abstract (Section 7.1.1) and methods (Sections 7.1.2 and 7.2.4) of Paper A:

1. Make brief notes on the blinding applied to the three groups involved with the study: subjects, researchers, and health-care staff.
2. List some of the reasons you think blinding is important for each group.
3. Drawing (if you wish) on your own field of interest, describe one intervention that could not easily be blinded, and why this is so.

Answers in Section 7.8

7.2.6 Ethical Issues for Intervention Studies

Earlier in this chapter we emphasised the importance of ethical practice in a RCT, particularly in respect of randomisation and the choice of the control ‘treatment’. We also noted that it is necessary to obtain *informed consent* from the subject before randomisation occurs. In Paper A, informed consent was obtained, and ethical approval was gained from the university ethics committees.

Intervention studies, by their very nature, involve something being done to the study subjects. This is very different from the study designs we have considered previously, which have involved **observing** the level of exposure in the past, now, or in the future. Whereas this observation (e.g., the sampling, measurement, and follow-up work) must be done in an ethically appropriate way, in trials the fact of carrying out an intervention demands extra attention to ethical practice.

Some additional aspects of study management and reporting that relate to these ethical considerations, including subject safety, are discussed in Section 7.7.

Summary

- Clarity about criteria for inclusion in, and exclusion from, the study is very important for those involved in carrying out the study and also for applying the results.
- It is important to think about whether the form of the intervention is typical of what would be generally applied. If not, the study may have little practical relevance beyond demonstrating efficacy – that is, the impact of the intervention when delivered in the manner of the study. Such efficacy information may, however, be very important in its own right.
- On ethical grounds, the comparison group should be offered the best existing (routine) treatment and, if there is no evidence of an effective existing treatment, this can be no treatment. Where practical, ‘no treatment’ can be offered as a placebo.
- Randomisation of allocation to the intervention is used to avoid confounding, and it is one of the greatest strengths of a RCT. The random allocation list should be prepared independently from those implementing it with study subjects, a procedure known as concealment.
- Outcome assessment requires the same level of attention to clarity of definition and validity of assessment method as with other study designs.
- Blinding reduces the possibility of bias in assessing outcomes, side effects, and so on, and it can be applied to the study subjects, the research team, and the health-care staff involved in the care of subjects.
- Where the intervention is such that it cannot be blinded, it is important to maintain as much objectivity as possible in the outcome assessment. This can be done by involving staff who are kept unaware of the treatment allocation (e.g., in analysis of laboratory specimens and of outcome data) and by using investigations (e.g., breath carbon monoxide [CO]) that are less likely to be influenced by knowledge of the treatment.
- Good ethical practice is vital for all research on human populations, and this is particularly important in intervention studies. All subjects must give informed consent before they are allocated to intervention and control groups.

7.3 The Analysis of Intervention Studies

The main focus for the analysis of an intervention study is the impact of the intervention on the primary outcome(s). Before getting into that, however, a few preliminary analytic steps need to be taken. First, given the importance of random allocation to the RCT design, we should check how well-balanced the groups ended up in terms of potential **confounding variables** and other factors that could influence the outcome. Second, we are also interested in knowing the extent to which the findings might apply to other populations, that is, the **external validity** of the trial. Both of these issues can be assessed by tabulating values for a range of socioeconomic, demographic, and other variables for the intervention and control groups. This is normally one of the first items in presenting the results of a trial, and we now look an example from Paper A.

Table 7.3.1 Baseline characteristics, including demographic variables and smoking status and history. Values are expressed as means (SD); range (Table 1 from Paper A).

	Placebo (<i>n</i> = 200)	Active (<i>n</i> = 200)
No. of men	104	86
Age (years)	45.8 (10.5); 22–77	46.4 (10.5); 23–79
Weight (kg) – Women	62.9 (10.6); 48–109	64.0 (11.2); 43–120
– Men	81.5 (12.3); 57–121	80.1 (12.6); 58–130
Age when started smoking (years)	17.1 (2.7); 11–35	18.2 (4.4); 12–45
No. of cigarettes smoked/day	30.3 (12.1); 15–70	28.2 (11.4); 15–70
Exhaled CO concentration (ppm)	27.1 (11.1); 10–61	27.1 (11.5); 10–61
FTND score	5.6 (2.0); 1–10	5.5 (2.1); 1–10

CO = carbon monoxide; FTND = Fagerström test for nicotine dependence.

Source: Bollinger 2000. Reproduced with permission of BMJ Publishing Group Ltd.

7.3.1 Review of Variables at Baseline

The usual way to assess how well the randomisation process has balanced potential confounding factors is to tabulate *baseline* information about a range of these variables (age, sex, smoking habits, etc.) for the intervention and control groups. That is, we look at values for these variables before the intervention has been implemented according to the groups to which subjects were then randomised. This has been done in Table 1 from Paper A, reproduced in Table 7.3.1.



Self-Assessment Exercise 7.3.1

1. Which of the variables listed do you think could have an important influence on the main outcomes(s) of the trial?
2. How has variability for each of the variables listed in the table been expressed?
3. Do you think that any of the differences between intervention and control group in the values for the variables listed should cause us concern about the balance of the two groups?
4. Comment on the characteristics of the sample, with a view to assessing the external validity of the trial.

Answers in Section 7.8

So what should be done if it is decided that there is some important imbalance in the two groups, involving one or more variables? First, some effort should be made to establish why this has happened; for example, is it just chance, or can some systematic problem with the randomisation be identified?

Assuming there are no concerns about serious bias in the randomisation process, allowance can to some extent be made for the imbalance by adjusting the intervention effect on the outcome using multivariable regression, with the specific method appropriate to the type of outcome data:

- Continuous data (e.g., body weight): Linear regression or analysis of covariance (see Chapter 5 on cohort studies for description of linear regression; analysis of covariance is not further explored in this book).

- Categorical data (e.g., smoking or not-smoking): logistic regression (see Chapter 6 on case–control studies)
- Time to event data (e.g., months of survival after cancer treatment): Cox regression (see Chapter 8 on survival analysis).

For the current study, the (main) outcome is smoking reduction (yes or no), a categorical variable where time to event is not a feature that would therefore require logistic regression for adjustment. Such adjustment was not carried out by the authors, however, presumably because they did not think there was any important imbalance in the randomisation.

Adjustment for factors that are known to influence the outcome may also be carried out to reduce variance in the analysis, and this is described further in Section 7.3.8.

7.3.2 Loss to Follow-Up

As with any longitudinal study, it is inevitable that some subjects will be lost to follow-up. The study flow chart, reproduced from Paper A as Figure 7.2.1 (in Section 7.2), provides detailed information on the numbers in each group who withdrew for various reasons. Please refer back to this figure to see how these withdrawals and other issues were reported.

Note that there were more withdrawals from the placebo group, particularly for the ‘not willing to participate’ category. If a substantial proportion of the sample is lost to follow-up, this could lead to bias, particularly if the extent and nature of these losses differ in non-random ways between intervention and control groups.

7.3.3 Compliance with the Treatment Allocation

Another issue that arises during the follow-up period of a trial is that some of the intervention group subjects might stop, or change, the treatment to which they were allocated – for example, if they experience an adverse reaction. Similarly, some subjects in the control group may start, or change, the treatment to which they were allocated (which may be the best existing treatment, a placebo, or nothing) if they chose to, or are advised to start a new treatment.

This issue, and the consequences this has for the analysis, is explored in Exercise 7.3.2 and in Section 7.3.4. Please now read the following excerpt that relates to treatment compliance and review Table 7.3.2 from the results section (Table 2 from Paper A).

Results

Treatment Compliance

Inhaler use decreased over time, as expected. Of participants present at week six, 222/368 (60%) used the inhaler every day. Corresponding figures after four, 12 and 18 months were 146/318 (46%), 39/331 (12%) and 30/289 (10%), respectively. Participants in the active treatment group used an average of 4.5 cartridges a day after two weeks and 2.6 a day after 18 months; they reduced their cigarette intake significantly more than participants in the placebo group from week two onward. Table 2 shows inhaler use and reduction in the number of cigarettes smoked and exhaled carbon monoxide concentration in daily users for both the active and the placebo groups.

Table 7.3.2 Number of inhalers used, reduction in number of cigarettes smoked, and exhaled CO concentration in participants using an inhaler every day; active and placebo treatment groups. Values are expressed as means (SDs) and ranges (Table 2 from Paper A).

Time point	Intervention group				Placebo group			
	No of subjects	No of inhalers/day	Cigarettes/day as % of baseline	CO as % of baseline value	No of subjects	No of inhalers/day	Cigarettes/day as % of baseline	CO as % of baseline value
1 week	169	4.3 (2.1); 1–12	53 (19.9); 4.3–105	79.8 (34.3); 12.5–208	162	4.4 (2.0); 1–12	56.4 (18.3); 9.5–100	83.4 (30.6); 29–171
2 weeks	168	4.5 (2.0); 1–10	8.5 (20.0); 0.0–100*	73.6 (31.4); 11.1–182	164	4.9 (2.0); 1–12	54.7 (18.9); 0.0–104*	82 (37.9); 13.6–255
6 weeks	117	4.3 (2.1); 1–12	45.1 (23.8); 0.0–100 [†]	68.4 (31.4); 11.1–155 [‡]	104	4.7 (2.0); 1–14	55.8 (18.2); 1.9–100 [†]	84.1 (50.0); 16.1–450 [‡]
4 months	84	3.9 (2.0); 1–10	42.7 (24.3); 0.0–100 [§]	58.3 (32.1); 5.6–141 [§]	62	4.0 (1.9); 1–10	52.0 (21.7); 0.0–100 [§]	71.1 (26.5); 12.5–170 [§]
12 months	27	3.5 (1.9); 1–7	32.6 (27.3); 0.0–83 [¶]	63.7 (42.4); 14.3–180	12	2.8 (2.1); 0–6	56.3 (33.4); 0.0–133 [¶]	83.9 (61.4); 9.8–250
18 months	22	2.6 (1.7); 0–6	36.2 (29.6); 0.0–100**	71 (58.8); 7.9–222	8	3.9 (2.5); 1–8	67.2 (27.8); 20–100*	81.7 (41.4); 50–177

P values (Wilcoxon's rank sum test) for difference between intervention and placebo: **p* = 0.004; [†]*p* < 0.001; [‡]*p* = 0.003; [§]*p* = 0.01; [¶]*p* = 0.03; ***p* = 0.02.

Source: Bollinger 2000. Reproduced with permission of BMJ Publishing Group Ltd.



Self-Assessment Exercise 7.3.2

Based on the information from the previous text excerpt from Paper A (Results; treatment compliance):

1. Overall, what happened to inhaler use over the course of the study?
2. From Table 7.3.2, describe inhaler use in the intervention and control groups.

Answers in Section 7.8

7.3.4 Analysis by Intention-to-Treat

Table 7.3.2 presents data for only those subjects who used their inhalers every day, and it might be thought best to use these subjects for the main analysis. After all, they are the ones using the intervention that the trial is attempting to assess. However, in the description of the statistical analysis (on p. 330 of Paper A) we learn that

The primary analysis is an intention-to-treat analysis, including all participants who were randomised and received medication.

- The phrase *intention-to-treat* means that the main analysis compares the outcomes for all subjects allocated to the intervention group with all of those allocated to the control group,

regardless of whether they complied or not with their respective treatments. Why have the authors taken this approach? To understand the reasons for this, it is useful to consider what might at first be thought of as a more-efficient approach to the analysis by doing one of the following:

- Exclude anyone from the analysis who did not comply with the treatment that they were originally allocated, or
- Analyse the outcomes for intervention group subjects who stopped using the active inhaler as if they were controls (e.g., add them to the control group) and vice versa for the control group subjects who started to use the active inhaler.

Unfortunately, both of these approaches can lead to bias. The problem is that the intervention group subjects who stop taking the treatment are unlikely to be representative of the whole intervention group. If these people are excluded or are analysed as controls, the two groups (intervention and control) will no longer be balanced for confounding factors, which was the whole point of the randomisation. The same argument applies to studies in which controls who start the intervention treatment (for example, in a study where controls become ill and need to have additional treatment), as they, too, are unlikely to be representative of all controls. Analysis should therefore be by intention to treat. This means carrying out the analysis with subjects in the group to which they were originally allocated, even if they are no longer complying with the treatment for that group. This avoids potentially serious bias and is regarded as the most appropriate method of analysis for randomised control trials.

One consequence of using an intention-to-treat approach is that any real treatment effect will be diluted by having to keep non-complying intervention and control subjects in their original groups. This is unfortunate, but at least the effect detected is likely to only be an underestimate of the true effect and is not biased by the non-compliers. Another important point about analysing the data in this way is that it gives a more realistic idea of the effect that the treatment will have in practice, allowing for the fact that some people will not be willing or able to comply with treatment as intended.

7.3.5 Analysis per Protocol

Section 7.3.4 explains why intention-to-treat analysis is the method of choice. It was noted, however, that if a relatively large proportion of subjects deviate from the allocated intervention (for example, many of those in the intervention group fail to use their nicotine inhaler), the effect estimate will be weakened – assuming that the intervention is effective.

In these circumstances, the trial may be analysed by comparing those who did use the intervention with those who did not use it (ignoring the randomised allocation) but who also fulfil the requirements of the protocol in terms of eligibility and outcome assessment. This is termed *per-protocol analysis*, and it should normally be adjusted using multivariable regression to account for any resulting imbalance in confounding factors.

7.3.6 What is the Effect of the Intervention?

To decide whether the intervention has any meaningful effect on the outcome, we need to measure the effect and then test whether the effect is statistically significant. The *hypothesis test* we use must be appropriate to the question of interest and the type of data: for example, the *chi-squared test* for a categorical outcome or the *t-test* for a continuous outcome.

In the simplest case, we want to determine whether there is a difference between the intervention and control groups in terms of the main outcome measure(s). The hypothesis underlying

our decision is the *null hypothesis*, the opposite (or negation) of the research question; that is, that there is no difference between the two groups. In the nicotine study (Paper A), the main effect is a *categorical* outcome because this has been measured by counting the number of people in each group who have reduced smoking sufficiently to meet the success criterion. In this case, therefore, the appropriate hypothesis test is the *chi-squared* test, so long as the assumptions for this test are met. This approach to the analysis is explored in Exercise 7.3.3. Please now review Table 7.3.3, reproduced from the results section of Paper A (Table 3 of Paper A).

Table 7.3.3 Efficacy results measured as sustained and point prevalence reductions in smoking and point prevalence abstinence rates according to treatment with oral nicotine inhaler (active treatment) or placebo (Table 3 from Paper A).

Definition	Time point (months)	No (%) with active treatment	No (%) with placebo	Odds ratio (95% CI)	P value (Fisher's test)
Sustained					
Reduction*	4	52 (26.0)	18 (9.0)	3.65 (2.04 to 6.19)	<0.001
	12	26 (13.0)	8 (4.0)	3.59 (1.65 to 7.80)	0.002
	24	19 (9.5)	6 (3.0)	3.39 (1.39 to 8.29)	0.012
Point prevalence					
Reduction [†]	4	83 (41.5)	44 (22.0)	2.52 (1.63 to 3.87)	<0.001
	12	59 (29.5)	43 (21.5)	1.53 (0.97 to 2.40)	0.085
	24	55 (27.5)	46 (23.0)	1.27 (0.81 to 2.00)	0.357
Abstinence [‡]	4	13 (6.5)	4 (2.0)	3.41 (1.16 to 10.01)	0.044
	12	16 (8.0)	12 (6.0)	1.36 (0.63 to 2.95)	0.557
	24	21 (10.5)	17 (8.5)	1.26 (0.65 to 2.47)	0.609

*Sustained reduction in number of cigarettes smoked daily by at least 50% from week 6, verified by decreased carbon monoxide concentrations compared with baseline.

[†]Point prevalence reduction of cigarettes smoked daily by at least 50% at months 4, 12, and 24, verified by decreased carbon monoxide concentrations compared with baseline.

[‡]No cigarettes smoked, verified by carbon monoxide concentrations <10 ppm at months 4, 12, and 24.

Source: Bollinger 2000. Reproduced with permission of BMJ Publishing Group Ltd.



Self-Assessment Exercise 7.3.3

1. We will first look at the primary efficacy measure, reported in Table 7.3.3.
 - a. Create a 2×2 contingency table for sustained reduction at 24 months (note: the total numbers that you will need are shown in the flow chart from Paper A [Figure 7.2.1 in Section 7.2.2]).
 - b. The authors use the Fisher's exact test. Can you find any reason why the chi-squared test should not be used for this hypothesis test?
 - c. Interpret the odds ratio (OR) and 95 per cent confidence interval (CI).
2. If you had wanted to compare the mean breath CO (measured in parts per million [ppm]) in the two groups (at any particular follow-up visit), what hypothesis test would you have used?

Answers in Section 7.8

7.3.7 Drawing Conclusions

Based on our assessment of this trial (Paper A), what can we conclude? Do our conclusions differ from those of the authors? We will explore these two questions in the next exercise.



Self-Assessment Exercise 7.3.4

1. Do you agree with the authors' main conclusion, as stated in the Abstract from Paper A (Section 7.1.1), that the nicotine inhaler led to sustained reductions in smoking over 24 months?
2. Assuming you do agree with this conclusion, to what population(s) do you think the findings would apply?

Answers in Section 7.8

7.3.8 Adjustment for Variables Known to Influence the Outcome

We saw previously (Section 7.3.1) that adjustment of the trial results may be carried out if there is concern about imbalance following randomisation. Adjustment may also be done to allow for the influence of factors capable of having a large effect on the outcome, independent of the intervention. Examples of such factors could be sex or level of education, as might plausibly influence an individual's ability to cut down or quit. Here we should be clear that this is not adjustment for confounding, for we are assuming randomisation has dealt with that by balancing the groups for these influential variables. The purpose of this adjustment is to reduce the variance in the model associated with these variables, and hence increase the power available to detect the intervention effect. Standard multivariable regression methods are used, but advice should be sought as to when such analysis is warranted.

7.3.9 Paired Comparisons

The nicotine inhaler trial compared one group using the active inhaler with another group using placebo. These were two separate sets of people, termed *independent groups*. Some types of intervention study begin with a group of subjects, make baseline measurements including the outcome variable, introduce the intervention, and then repeat the measurements in the same group of subjects. These are known as before-and-after studies, and issues relating to the design and interpretation are described further in Section 7.4.8. In this situation, the comparison is made – for each individual – between a measurement after the intervention and one made before it. This provides *paired data*, with each subject acting as its own control.

We now look at a scenario for a study involving the generation of paired data, and we then go on to describe a particular type of RCT that uses paired comparisons, the *crossover trial*.

In a study of the effect of education on patients' management of asthma, a group of asthmatic patients attended a series of talks about asthma and its treatment. The intention was that this would result in greater understanding about their disease and how best to use associated medication with the expectation they would consequently suffer less from the symptoms of asthma. To evaluate how much the patients had learned, they took a short test on asthma and its treatment before the series of talks and a similar test at the end of the course and their scores were recorded.

In this example, the two test scores for each patient are paired continuous outcome data (Table 7.3.4).

Table 7.3.4 Test scores of asthma patients before and after attending a series of talks about asthma and its treatment.

Subject	Test scores		Difference (A – B)
	Before talks (B)	After talks (A)	
1	51	55	4
2	43	49	6
3	48	52	4
4	19	32	13
5	57	62	5
6	39	44	5
7	37	40	3
8	46	46	0
9	43	39	–4
10	43	52	9
11	53	50	–3
12	58	61	3
13	49	54	5
Mean	45.08	48.92	3.85

We can see that this gives us a series of differences (A – B) for each person. The mean of these differences is 3.85, which is the difference between the mean of the after scores (A) and the mean of the before scores (B). When, in analysing paired data, we wish to carry out an hypothesis test, we need to use one that is based on the distribution of these differences. These are called **paired tests**. In this example we have continuous data (the scores), so we use a **paired t-test** (in contrast to the independent sample *t-test* we have already used). We will cover this and other paired tests in Chapter 11.

Paired data also arise if we **match** individuals in the control and intervention groups. The matching is with respect to factors that are likely to affect the outcome, such as age, sex, and area of residence. We have seen the use of this technique in Chapter 6 on case-control studies, and you have already encountered the hypothesis test used for matched case-control studies, known as **McNemar's test**. Here is an exercise to help you understand when it is appropriate to use a paired comparison.



Self-Assessment Exercise 7.3.5

Which of the following should use a paired approach to analysis?

1. Thirty male volunteers take a psychological test to determine reasoning. Before the test 15 volunteers drink two glasses of whisky. The test scores for the 15 who drank the whisky are then compared with scores for the 15 who did not.

2. Twenty male volunteers take two psychological tests to determine reasoning. Each volunteer has no alcohol before the first test, but drinks two glasses of whisky before the second test. Test scores obtained after whisky are then compared with those obtained prior to drinking.

Answers in Section 7.8

7.3.10 The Crossover Trial

The crossover trial is a particular type of RCT, which allows subjects to operate as their own controls. This design has been used most commonly for testing drugs. The structure is summarised in Figure 7.3.1.

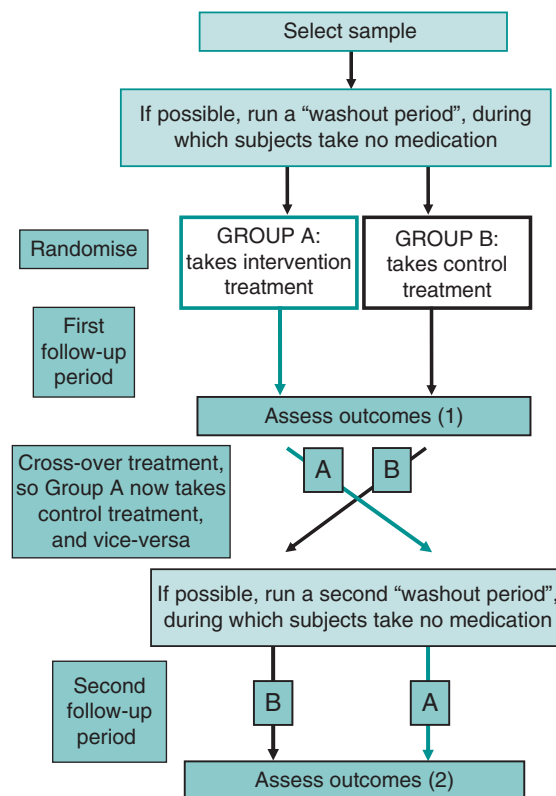


Figure 7.3.1 Schematic overview of a crossover trial.

During the course of the trial, each subject crosses over from receiving one treatment to receiving the other. This design would be suitable in, for example, an investigation of treatment for the control of symptoms in chronic disease, where the long-term condition of the subjects remains fairly stable.

There are a number of advantages to this study design. One of the principal attractions of the crossover trial is that subjects operate as their own controls, thus reducing the extent of confounding, although one must be alert to factors that may change over time during the course of the study (see below). Using subjects as their own controls also reduces the amount of random

error, implying that the study can be done with smaller numbers and is therefore cheaper and easier.

There are also some potential disadvantages, including that some factors influencing the outcome could change during the course of the study, as noted above. Another lies in possible **order effects**, such that taking one treatment first may have different effects from taking them in the reverse order. There could also be an effect lasting for an unknown duration following cessation of treatment, known as **carry-over**, and ideally a second washout period should follow the first round of treatment, as shown in Figure 7.3.1.

A crossover trial results in data in the form of matched pairs: There are two measurements for each subject, one for each treatment. Thus, we can analyse a crossover trial by methods for paired data, such as the paired *t*-test for a continuous outcome. However, there is a bit more to the analysis of the crossover trial. We need to take into account the order in which each subject received the two treatments and the possibility of the effect of one treatment carrying over to the next period in which the subject is receiving the alternative treatment referred to above. These methods are beyond the scope of this book, however, and are not described further.

Summary

- Randomisation is carried out to balance the distribution of confounding factors in the intervention and control groups. It is important to check how well this has been achieved.
- Loss to follow-up should be kept to a minimum. The methods for achieving this and the consequences of substantial and/or unrepresentative loss from the groups are essentially the same as for cohort studies.
- It is inevitable that some of the intervention and control subjects will stop or change the treatment that they were originally allocated to. The extent of the compliance with treatment should be assessed.
- To avoid bias arising from the unrepresentative nature of non-compliance, the appropriate method of analysis is by intention-to-treat.
- Per-protocol analysis may be used, with caution, where there is substantial deviation from compliance in randomised groups. Such analysis normally requires adjustment using regression techniques.
- Results should be presented with a 95% CI and a *p*-value from an appropriate hypothesis test.
- If, despite randomisation, there are judged to be substantial imbalances in important confounding factors, these can be addressed in analysis by multivariable regression.
- Paired data are generated in situations where subjects act as their own controls (the same people are assessed before and after an intervention) or when subjects are fairly closely matched. Paired data should be analysed by paired hypothesis tests.
- The crossover trial is a type of RCT that uses subjects as their own controls and thereby reduces random error and the required sample size. This strength has to be balanced against the problem of carry-over and order effects, so it can only be used for certain types of intervention.

7.4 Testing More-Complex Interventions

7.4.1 Introduction

The intervention we have considered so far was rather simple: taking a nicotine inhaler as needed. No other procedure or supporting advice was tested as part of the intervention. Although counselling was available, the effect of this was not being tested. RCTs of drugs are

extremely important; however, drug interventions are simple to research in comparison with the majority of treatments that are used in health care, health promotion, and public health. We will see that, when testing these more-complex treatments, it is often far more difficult to adhere to the **gold standard** of the blinded, randomised, placebo-control trial.

Complex interventions are those involving multiple, usually interacting, components, which may range from an individual treatment combined with advice or support at another level (e.g., the general practice or local government area) through to interventions with one or more actions at societal level including policy, awareness-raising through the media, regulation, and tax or other financial incentives. Evaluation of these types of interventions can be very difficult, randomisation is often not possible, and indeed experiment – beyond observing the impacts of initiatives such as new policies or laws – may also be impractical and/or unethical. The approach to carrying out such evaluation would typically involve a combination of research methods and paradigms (see Chapter 1). The UK Medical Research Council (MRC) has produced helpful guidance on this, ‘Developing and evaluating complex interventions: new guidance’ (Medical Research Council; Craig *et al.*, 2008).

For the current purpose, we restrict our discussion to a study that, although testing a complex intervention, adheres as closely as possible to the ideal of the RCT. It also introduces a **factorial design**, which involves testing more than one intervention at a time and allows us to determine whether an intervention is more effective in combination with another than when applied in isolation. Where a factor has more (or less) effect in combination than when administered alone, this is known as an **interaction**, or alternatively as an **effect modification**.

We now look at the key design elements of this through Exercise 7.4.1, based on Paper B.

7.4.2 Randomised Trial of Individuals for a Complex Intervention

The relevant text excerpts for this section are indicated in self-assessment exercise 7.4.1. When reading this paper, note that methods for calculating sample size are discussed in Section 7.6, and the main analytic method (survival analysis) is discussed in Chapter 8. You do not need to be familiar with these specific techniques in order to complete Exercise 7.4.1.



Self-Assessment Exercise 7.4.1

Using the information in the abstract and following excerpts (including Figure 7.4.1) from the methods section of Paper B:

1. List the criteria for inclusion and for exclusion of subjects for the trial.
2. How representative of people over 70 years of age was the sample of people who were randomised?
3. Briefly describe the intervention and control ‘treatments’. Don’t go into the details of the factorial design allocation, as we will deal with that later; just focus on the types of intervention treatment and control group management.
4. How generally applicable do you think the intervention treatments would be?
5. Does the treatment of controls conform to the ethical requirement we identified in Section 7.2 of this chapter?
6. Which of the three groups involved in the trial (subjects, research team, and health-care staff) were blinded to the intervention?

Answers in Section 7.8

Abstract

Objective: To test the effectiveness of, and explore interactions between, three interventions to prevent falls among older people.

Design: A randomised controlled trial with a full factorial design.

Setting: Urban community in Melbourne, Australia.

Participants: 1090 aged 70 years and over and living at home. Most were Australian born and rated their health as good to excellent; just over half lived alone.

Interventions: Three interventions (group based exercise, home hazard management and vision improvement) delivered to eight groups defined by the presence or absence of each intervention.

Main outcome measure: Time to first fall ascertained by an 18 month falls calendar and analysed with survival analysis techniques. Changes to targeted risk factors were assessed by using measures of quadriceps strength, balance, vision, and number of hazards in the home.

Results: The rate ratio for exercise was 0.82 (95% confidence interval 0.70 to 0.97, $p = 0.02$), and a significant effect $p < 0.05$ was observed for the combinations of interventions that involved exercise. Balance measures improved significantly among the exercise group. Neither home hazard management nor treatment of poor vision showed a significant effect. The strongest effect was observed for all three interventions combined (rate ratio 0.67 (0.51 to 0.88, $p = 0.004$)), producing an estimated 14.0% reduction in the annual fall rate. The number of people needed to be treated to prevent one fall a year ranged from 32 for home hazard management to 7 for all three interventions combined.

Conclusions: Group based exercise was the most potent single intervention tested and the reduction in falls among this group seems to have been associated with improved balance. Falls were further reduced by the addition of home hazard management or reduced vision management, or both of these. Cost effectiveness is yet to be examined. These findings are most applicable to Australian born adults aged 70–84 years living at home who rate their health as good.

Methods

Inclusion and Exclusion Criteria

Participants had to be living in their own home or apartment or leasing similar accommodation and allowed to make modifications. Potential participants were excluded if they did not expect to remain in the area for two years (except for short absences); had participated in regular to moderate physical activity with a balance improvement component in the previous two months; could not walk 10–20 metres without rest, help, or having angina; had severe respiratory or cardiac disease; had a psychiatric illness prohibiting participation; had dysphasia [difficulty with speech]; had had recent major home modifications; had an education and language adjusted score >4 on the short portable mental status questionnaire; or did not have the approval of their general practitioner.

Recruitment

When compared with data from the national census and health survey for Australians aged over 70 living at home, the study group differed as follows: a higher proportion (46.0% v 42.8%) were

aged 70–74 years and a lower proportion (7.3% v 9.8%) aged over 85 years old; a higher proportion (77.3% v 66.7%) were Australian born; a higher proportion (53.8% v 32.7%) were living alone; and a lower proportion (46.8% v 52.3%) were married. Study participants rated their health status considerably higher (very good to excellent, 62.6% v 30.7%), and a higher proportion (13.8% v 9.0%) reported taking antidepressant and hypnotic [sleep-inducing] medication.

Interventions

We sent all participants a letter outlining their assigned interventions advising of necessary actions. Strength and balance—Participants attended a weekly exercise class of one hour for 15 weeks, supplemented by daily home exercises. The exercises were designed by a physiotherapist to improve flexibility, leg strength, and balance and 30–35% of the total content was devoted to balance improvement. Exercises could be replaced by a less demanding routine, depending on the participant's capability. Transport was provided where necessary. Home hazards—Home hazards were removed or modified either by the participants themselves or via the City of Whitehorse's home maintenance programme. Home maintenance staff visited the home providing a quotation for the work, including free labour and materials up to the value of \$A100 (£37; \$54; €60).

Vision—If a participant's vision tested below predetermined criteria and if he or she was not already receiving treatment for the problem identified, the participant was referred to his or her usual eye care provider, general practitioner, or local optometrist, to whom the vision assessment results were given. Participants not receiving the vision intervention were provided with the Australian Optometrist Association's brochure on eye care for those aged over 40.

Assessment

Participants received a home visit by a trained assessor, who was initially blinded to group assignment. After informed consent was obtained, a baseline questionnaire was completed covering demographic characteristics; ability to perform basic activities and instrumental (more complex) activities of daily living; use of support services; social outings and interests; the modified falls efficacy scale; self-rated health; and falls and medical history. Current prescription and over the counter drugs were recorded from containers at the participants' homes. The targeted risk factors were assessed by using the methods outlined in table 1. Participants were then assigned (by computer generated randomisation) to an intervention group by an independent third party via telephone.

After 18 months the risk factor assessments were repeated in a proportion of participants $n = 442$ randomly selected by an assessor blinded to the intervention group (we used only a proportion of the participants because resources to reassess the whole study group were not available and this assessment was of secondary importance to the study's main goal).

Strength and balance were also measured at the final exercise class of the first 177 participants to complete the 15 week programme, 79 of whom were among the 442 subsequently selected for final reassessment.

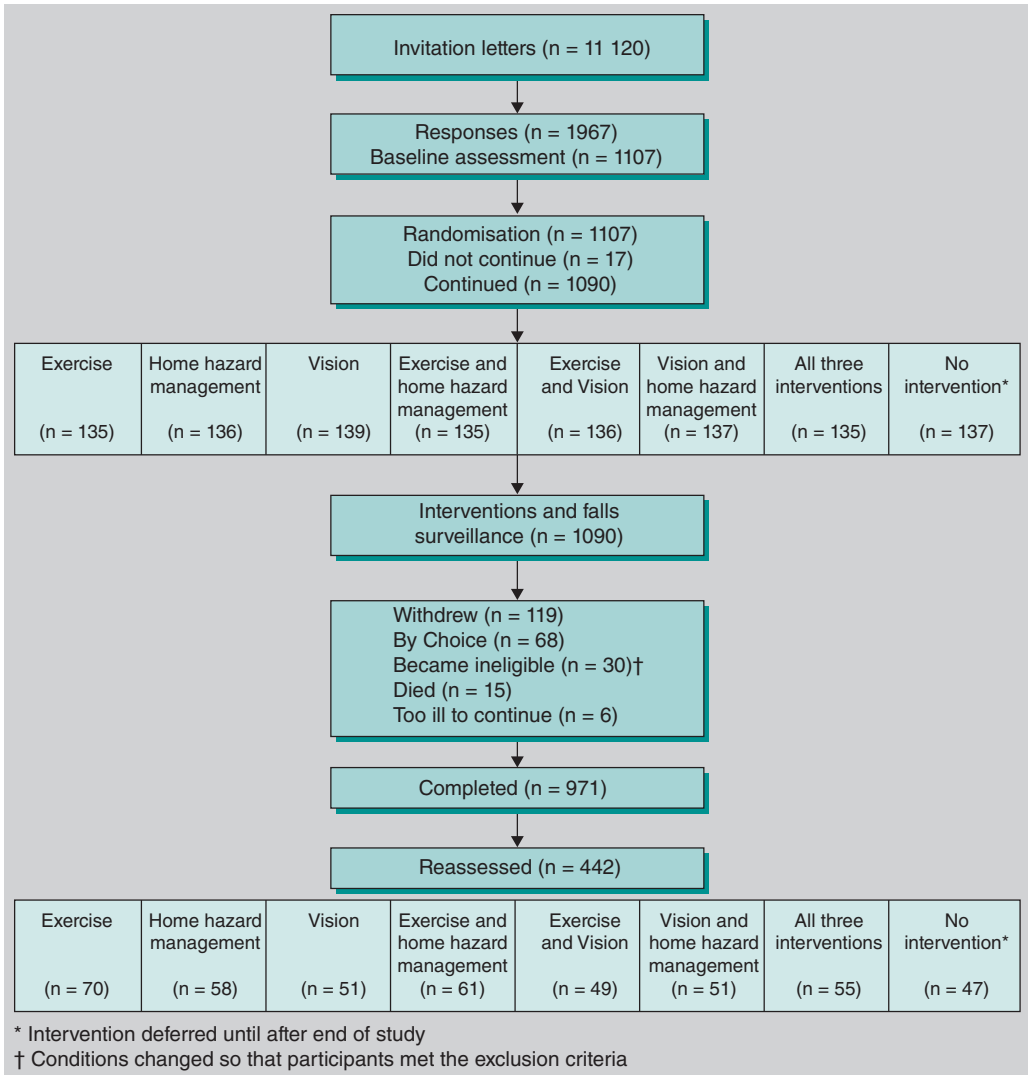


Figure 7.4.1 Flow chart showing stages in study protocol and numbers of participants (Figure 1 from Paper B). Source: Day 2002. Reproduced with permission of BMJ Publishing Group Ltd.

7.4.3 Factorial Design

A number of other features of intervention study design and analysis have been introduced in Paper B (falls prevention trial), and we now look at the most important aspects of these. In your further reading of this paper do not be concerned that you are unfamiliar at this stage with some of the methods and statistical procedures described, including factorial design, analysis of variance, Cox proportional hazards and Kaplan–Meier curves. We will work through and explain these at various points in this and subsequent chapters, beginning here with factorial design.

We saw in the falls-prevention study that, instead of simply comparing two groups (intervention and control) as in the nicotine-inhaler study, there were a total of eight groups, some with combinations of interventions and one with none. The allocations are set out in Table 7.4.1.

Table 7.4.1 Summary of intervention allocation in the falls prevention study (Paper B).

Intervention type			Number of subjects allocated	Number of interventions
Exercise programme	Home hazard removal	Vision test and referral		
Yes	No	No	135	One intervention
No	Yes	No	136	One intervention
No	No	Yes	139	One intervention
Yes	Yes	No	135	Two interventions
Yes	No	Yes	136	Two interventions
No	Yes	Yes	137	Two interventions
Yes	Yes	Yes	135	All three interventions
No	No	No	137	No intervention
Total			1090	

Source: Day 2002. Reproduced with permission of BMJ Publishing Group Ltd.

This design allows comparisons to be made in the following way:

- First of all, note that each type of intervention is experienced by half of the total sample. Thus, half the sample (135 + 135 + 136 + 135) had the exercise programme and half did not, so it is possible to compare exercise with no exercise in the whole sample. This comparison is the basis of the sample size calculation and what is meant by the *main effect* two-group comparison.
- Of course, only 135 subjects have the exercise programme alone; all the other subjects with exercise also have one or more other interventions. The factorial design, together with the way it is analysed using **analysis of variance** (ANOVA), allows the main effects (single intervention effects) and the effects in combination (*interactions*) to be identified. ANOVA is described further in Chapter 11.
- Overall, the design also allows comparisons to be made of
 - a single intervention versus another single intervention;
 - a single intervention versus no intervention;
 - two interventions together versus a single intervention, or no intervention;
 - three interventions together versus two interventions together, a single intervention, or no intervention.

7.4.4 Analysis and Interpretation

We now look at how the falls prevention study was analysed. You will find that some aspects of this are familiar, but some are new. Two familiar elements of the analysis include (a) characteristics of participants at baseline and assessing how well randomisation balanced variables across intervention and control groups, (Tables 2 and 3 from Paper B, reproduced in Table 7.4.2), and (b) analysis by intention-to-treat.

Baseline Characteristic and Balance of Randomisation

It may seem that Table 7.4.2 presents the baseline characteristics in an unfamiliar way. Since there are eight groups, rather than listing the mean and standard deviation (SD) (or per cent)

Table 7.4.2 Baseline values in the intervention and control groups (Tables 2 and 3 from Paper B).**Table 2:** Characteristics of participants at baseline.

Characteristic	All participants (n = 1,090)	Range across intervention groups (n = 1,090)*	Reassessed participants (n = 442)†
Mean (SD) age (years)	76.1 (5.0)	75.4–76.5 (4.7–5.5)	75.9 (4.9)
No (%) of women	652 (59.8)	77–93 (55.4–68.4)	261 (59.0)
No (%) of participants living alone	586 (53.8)	68–83 (50.0–61.0)	230 (52.0)
No (%) of participants who had a fall in past month	69 (6.3)	5–11 (3.7–8.1)	31 (7.0)
Mean (SD) score for activities of daily living‡	5.3 (1.1)	5.2–5.4 (0.92–1.2)	5.3 (1.1)
Mean (SD) no of medications	3.4 (2.6)	3.1–3.6 (2.4–2.9)	3.3 (2.6)

SD = standard deviation.

‡Score for instrumental activities of daily living, plus bathing.

Table 3: Targeted risk factor measures at baseline. Values are means (standard deviation).

Measure	All participants (n = 1,090)	Range across intervention groups (n = 1,090)*	Reassessed participants (n = 442)†
Quadriceps strength in stronger leg (kg)	22.6 (10.5)	21.3–24.0 (9.6–11.8)	23.2 (10.7)
Postural sway on foam pad (log)	2.8 (0.35)	2.7–2.8 (0.32–0.40)	2.7 (0.37)
Maximal balance range (cm)	13.3 (4.5)	13.0–13.7 (4.2–5.0)	13.5 (4.7)
Coordinated stability (sum of errors)	12.4 (8.4)	11.4–13.0 (7.6–9.2)	11.6 (8.0)
Timed 'up and go'(s)	11.7 (5.3)	10.9–12.3 (3.6–6.2)	11.6 (5.6)
High contrast acuity in best eye (logMAR)	0.08 (0.19)	0.05–0.11 (0.18–0.21)	0.06 (0.19)
Low contrast acuity in best eye (logMAR)	0.38 (0.19)	0.34–0.42 (0.17–0.20)	0.38 (0.19)
Dot pattern (No of patterns identified)	5.8 (3.2)	5.5–6.1 (3.1–3.4)	6.0 (3.1)
Field of view in best eye (No of correct identifications)	25.2 (3.0)	24.8–25.5 (1.9–4.5)	25.4 (2.7)
Home Hazards (No identified)	9.3 (4.8)	8.3–10.1 (4.3–5.4)	9.6 (4.7)

*Highest and lowest recorded among the eight groups. Measures for the remaining groups fall within the range.

†Participants randomly selected for reassessment at end of follow up.

Source: Day 2002. Reproduced with permission of BMJ Publishing Group Ltd.

values for each group, the range across the groups is given (column 2). Thus, for age, the lowest group mean was 75.4 years and the highest 76.5 years, with the other six groups distributed in between these values. This suggests that randomisation had balanced the groups quite well by age, but what about the other variables? Look, for example, at the balance of males and females in row 2 of the upper table (Table 2 from Paper B).

Compliance with Interventions

It is important to take a close look at compliance with interventions, as this will help us in interpreting the results. Also, since *intention-to-treat* analysis has been used, comparisons are made according to the group that subjects were randomised to, not according to whether they complied with the intervention offered in that group. Please now read the excerpt below on

intervention compliance, from the results section of Paper B. Exercise 7.4.2 explores the level of compliance and considers the implications for the analysis.

Intervention Compliance

Of the 541 participants receiving the exercise intervention, 401 started a class. The mean number of sessions attended was 10 (SD 3.8) and 328 participants attended more than 50% of their sessions. The mean number of additional home exercise sessions was nine a month. Of the 543 participants receiving the home hazard management intervention, 478 participants were advised to have modifications in their homes; 363 of these participants received help to do these modifications, which included hand rails fitted (275 participants), modifications to floor coverings (72), contrast edging fitted to steps (72) and maintenance to steps or ramps (66).

Of the 547 participants receiving the vision intervention, 287 were recommended for referral, of whom 186 had either recently visited or were about to visit their eye care practitioner. Of the remaining 101 participants, 97 took up the referral, resulting in 26 having some form of treatment – new or modified prescription glasses (20) or surgery (6).



Self-Assessment Exercise 7.4.2

1. Briefly summarise the information given on compliance.
2. Comment on compliance with each of the three interventions.

Answers in Section 7.8

Reassessment of Risk Factors During Follow-Up

Not all of the study subjects were reassessed for risk factors at 18 months. Exercise 7.4.3 explores this further, including implications for the interpretation of results.



Self-Assessment Exercise 7.4.3

According to the information from excerpts on assessment (Section 7.4.1) and Table 7.4.2 (Tables 2 and 3 from Paper B):

1. How many (what percentage) of the subjects were reassessed at 18 months?
2. Why did the investigators not reassess all subjects?
3. How did they select those who were reassessed?
4. What information is available to judge how representative the reassessed individuals were of all study subjects?
5. How representative do you think these individuals were?

Answers in Section 7.8

Analysis of Main Effects of the Intervention

The main results of the study are shown in Table 7.4.3 (Table 4 from Paper B). You may have seen from Paper B that the effect estimates, expressed as rate ratios, have been obtained by Cox proportional hazards regression (more generally these would be termed 'hazard ratios', where

Table 7.4.3 Effect on falls outcome, single and combined interventions (Table 4 from Paper B).

Intervention	No (%) having at least one fall	Rate ratio*		% estimated reduction in annual fall rate (95% CI)	Number needed to treat to prevent 1 fall
		Estimate (95% CI)	P value		
No intervention**	87/137 (63.5)	1.00 (Reference)			
Exercise	76/135 (56.3)	0.82 (0.70 to 0.97)	0.02	6.9 (1.1 to 12.8)	14
Vision	84/139 (60.4)	0.89 (0.75 to 1.04)	0.13	4.4 (-1.5 to 10.2)	23
Home hazard management	78/136 (57.4)	0.92 (0.78 to 1.08)	0.29	3.1 (-2.0 to 9.7)	32
Exercise plus vision	66/136 (48.5)	0.73 (0.58 to 0.91)	0.01	11.1 (2.2 to 18.5)	9
Exercise plus home hazard management	72/135 (53.3)	0.76 (0.60 to 0.95)	0.02	9.9 (2.4 to 17.9)	10
Vision plus home hazard management	78/137 (56.9)	0.81 (0.65 to 1.02)	0.07	7.4 (-0.9 to 15.2)	14
Exercise plus vision plus home hazard management	65/135 (48.1)	0.67 (0.51 to 0.88)	0.004	14.0 (3.7 to 22.6)	7

*See text: risk estimate from Cox regression is usually termed the 'hazard ratio'.

**No intervention until after the study had ended.

Source: Day *et al.*, p. 133.

this method has been used). Cox regression is the method of choice when using outcome data that include time until the event occurs, as is the case in this study. This is known as survival analysis and is described in Chapter 8. For now, we can interpret the results of Cox regression (the hazard ratio) in the way we have done for the odds ratio, in the sense that a value of 1.0 means no increased or decreased risk, <1.0 means a lower risk, and >1.0 a higher risk. A number of other techniques used in the analysis of this study will be covered later in the book. Figure 2 in Paper B (not reproduced here) shows the outcomes over time graphically in what are known as Kaplan–Meier curves, and these are also covered in Chapter 8. Analysis of variance (ANOVA), which is used to test main effects and interactions in a factorial design, is described in Chapter 11, together with paired *t*-tests and the Fisher's exact test.

First, look at Table 7.4.3. You are familiar with the interpretation of rate ratios (relative risk), 95 per cent CIs, and the *p*-value.

An important new concept is the **number needed to treat** (NNT), shown in the last column. This describes the number of subjects that need to be treated (given the intervention) to prevent one event (in this case, a fall). It is calculated as 1/difference between proportions experiencing events in the intervention and comparison groups (or 100/difference if this is calculated from percentages).

We now look at interpretation of the main effect results and one example of NNT in Exercise 7.4.4.



Self-Assessment Exercise 7.4.4

From the information in Table 7.4.3 (Table 4 from Paper B):

1. Interpret the rate ratios, 95 per cent CI, and *p*-values for receiving (a) exercise and (b) vision interventions.

2. Now interpret the rate ratio, 95 per cent CI, and *p*-value for receiving exercise and vision interventions in combination. How does this compare with the results for the interventions on their own?
3. The number needed to treat for all three interventions in combination is seven. See whether you can work out how this was derived. What does a NNT of seven for these interventions mean?

Answers in Section 7.8

7.4.5 Departure from the Ideal Blinded RCT Design

In the foregoing sections we have emphasised two important themes in respect of the design and application of intervention studies:

- The strength of the ideal RCT as a research design arises from the randomisation, controlled comparison, use of placebo, and application of blinding.
- For many interventions, however, it is not possible to adhere to all aspects of this ideal design. We have already discussed a number of interventions where blinding was not possible.

Table 7.4.4 summarises opportunities (or lack of them) for randomisation, use of placebo controls, and blinding for a range of intervention types, from drug and vaccine trials through to more-complex preventive and health-care interventions.

Table 7.4.4 Methodological opportunities and constraints in trials of various types of intervention.

Intervention	Randomisation	Control/placebo	Blinding
Drug or vaccine	Possible	Possible, often with placebo control	Blinding of all groups usually possible
Surgical operation	Possible	Control operation, but placebo not possible	Not possible for surgical team, or for patient in some circumstances. Some blinding of assessment should be done if possible
Medical treatment extending beyond medication alone, e.g., health education advice, management by a multidisciplinary team	Possible	Control treatment, but placebo not possible	Usually not possible; some blinding of assessment should be done if possible
Health promotion initiative at group level, e.g., whole (GP) practice	Can randomise clusters (groups, e.g., general practices) if a sufficient number are available	Control groups, but comparability may be difficult to achieve and maintain	Not possible; some blinding of assessment should be done if possible
Health promotion initiative at community, city, or regional level	Not possible	Control area, but comparability will be difficult to achieve and maintain	Not possible; blinding of assessment also very difficult with such a broadly applied intervention

7.4.6 The Cluster Randomised Trial

Reasons for Selecting a Cluster Randomised Design

In the two intervention studies discussed so far, individual people have been randomised to intervention and control groups. For some interventions it is more appropriate to randomise groups of people, such as schools or general practices. These groups are known as *clusters*, a term we have already encountered in sampling.

Group (cluster) randomisation is more suitable for interventions that involve, as part of their delivery, components that are not directed solely at individuals but at groups of service users – for example, health information posters in general practice or school-based health promotion activities. Although these may be adjuncts to interventions directed at individuals on a one-to-one basis, the group-based component means that randomisation of individual subjects within the practice or school would not be appropriate. When control subjects are exposed to influences intended to be solely for intervention subjects, this is known as *contamination*. Contamination may operate in rather subtle ways. For example, even if a practice-based intervention is applied to individuals, it may be very difficult to prevent information exchange between intervention and control subjects registered with the same practice, particularly for interventions that cannot be blinded and involve information and advice. An additional factor with a non-blinded intervention in settings such as general practice is that it may be difficult for the health-care staff to ensure equivalent treatment of intervention and control patients in the same practice. For all of these reasons it may be better to randomise the whole practice (or other type of cluster), rather than individuals within that cluster.

Most aspects of cluster randomised trials are similar to individual-based randomised studies, but there are some key differences. As an example we will look briefly at a study of the effectiveness of exercise advice and prescription (green prescription) for 40–59-year-olds habitually doing little physical activity, set in general practice in New Zealand (Paper C). The abstract for this study is reproduced below.

Abstract

Objective: To assess the long term effectiveness of the ‘green prescription’ programme, a clinician based initiative in general practice that provides counselling on physical activity.

Design: Cluster randomised controlled trial. Practices were randomised before systematic screening and recruitment of patients.

Setting: 42 rural and urban general practices in one region of New Zealand.

Subjects: All sedentary 40–79 year old patients visiting their general practitioner during the study’s recruitment period.

Intervention: General practitioners were prompted by the patient to give oral and written advice on physical activity during usual consultations. Exercise specialists continued support by telephone and post. Control patients received usual care.

Main outcome measures: Change in physical activity, quality of life (as measured by the ‘short form 36’ (SF-36) questionnaire), cardiovascular risk (Framingham and D’Agostino equations) and blood pressure over a 12 month period.

Results: 74% (117/159) of general practitioners and 66% (878/1322) of screened eligible patients participated in the study. The follow up rate was 85% (750/878). Mean total energy expenditure increased by 9.4 kcal/kg/week ($p = 0.001$) and leisure exercise by 2.7 kcal/kg/week ($p = 0.02$) or 34 minutes/week more in the intervention group than in the control group ($p = 0.04$). The proportion of the intervention group undertaking 2.5 hours/week of leisure exercise increased by

9.72% ($p = 0.003$) more than in the control group (number needed to treat = 10.3). SF-36 measures of self-rated 'general health,' 'role physical,' 'vitality,' and 'bodily pain' improved significantly more in the intervention group ($p < 0.05$). A trend towards decreasing blood pressure became apparent but no significant difference in four year risk of coronary heart disease.

Conclusion: Counselling patients in general practice on exercise is effective in increasing physical activity and improving quality of life over 12 months.

Please now read the following excerpt from Paper C describing the 'green prescription' used as the intervention in this cluster-randomised trial.

The 'Green Prescription' Intervention

- Primary care clinicians are offered four hours of training in how to use motivational interviewing techniques to give advice on physical activity and the green prescription.
- Patients who have been identified as 'less active' through screening at the reception desk and who agree to participate receive a prompt card, stating their stage of change, from the researcher, to give to the general practitioner during consultation.
- In the consultation, the primary care professional discusses increasing physical activity and decides on appropriate goals with the patient. These goals, usually home-based physical activity or walking, are written on a standard green prescription and given to the patient.
- A copy of the green prescription is faxed to the local sports foundation with the patient's consent. Relevant details such as age, weight, and particular health conditions are often included.
- Exercise specialists from the sports foundation make at least three telephone calls (lasting 10–20 minutes) to the patients over the next three months to encourage and support them. Motivational interviewing techniques are used. Specific advice about exercise or community groups is provided if appropriate.
- Quarterly newsletters from the sports foundations about physical activity initiatives in the community and motivational material are sent to participants. Other mailed materials, such as specific exercise programmes, are sent to interested participants.
- The staff of the general practice are encouraged to provide feedback to the participant on subsequent visits to the practice.



Self-Assessment Exercise 7.4.5

1. Why do you think the authors chose a cluster randomised design for this study, rather than individual-based randomisation?
2. What management do you think the control clusters would have received?

Answers in Section 7.8

Analysis of a Cluster RCT

The analysis of a cluster RCT requires that we take account of the fact that individuals (people or other 'units') within clusters may have factors more in common than when comparing such individuals across different clusters. We will look at this in more detail in Section 7.5, but in essence this means that for any given factor, including the outcome, variability may be less

within clusters as compared to *between clusters*. To deal with this in the analysis, we need to use a *multilevel model* (Section 7.5), that is, one that preserves the structure of the total sample as being made up of groups of people in clusters who may have factors in common. This issue also requires attention in sample size calculation for the same reason (discussed in Section 7.6).

7.4.7 The Community (Cluster) Randomised Trial

We began this chapter by looking at a pharmacological (drug-based) intervention to help people reduce and stop smoking: use of a nicotine replacement inhaler. A double-blinded, placebo-control, randomised trial was both feasible and appropriate. What approach would be appropriate, however, if we wanted to study the effectiveness of a range of family-based and community-based measures to prevent young people from taking up smoking?

We know that adoption of smoking by young people is influenced by many family, peer group, school, and wider societal factors. Interventions to address this complex, interrelated set of factors cannot realistically be delivered within tightly defined groups. This is a good example of a complex intervention with multiple interacting components delivered in a variety of ways. Despite this complexity, as we will see in our last example in this chapter (Paper D), an evaluation could still be carried out as a randomised trial, testing community approaches to preventing adolescent tobacco use. The units of randomisation were small communities in Oregon, USA. Comparison was made between a school-based intervention only (control communities) and school-plus-community intervention (intervention communities). The community intervention was carried out by a paid community co-ordinator and included

- media advocacy;
- anti-tobacco activities designed to be engaging and persuasive for young people;
- family communication about tobacco use;
- reducing youth access to tobacco in stores.

This study was published in 2000, and it preceded the now-widespread legislation on smoking in public places. This legislation is generally regarded as having had an important impact on smoking behaviour, and consequently it would need consideration in any study of tobacco control nowadays. Because that law is generally implemented at national or state level, however, it would not be possible to study the effect in this trial because it is based on a comparison between small communities.

In this section we only study the principal design issues involved in a community (cluster) randomised trial. Please now read the abstract and excerpt from the methods section (design) from Paper D, reproduced below. Exercise 7.4.6 will help you identify the key design points.

Abstract

Objective: Experimental evaluation of comprehensive community wide programme to prevent adolescent tobacco use.

Design: Eight pairs of small Oregon communities (population 1,700 to 13,500) were randomly assigned to receive a school based prevention programme or the school based programme plus a community programme. Effects were assessed through five annual surveys (time 1–5) of seventh and ninth grade (ages 12–15 years) students.

Intervention: The community programme included: (a) media advocacy, (b) youth anti-tobacco activities, (c) family communications about tobacco use and (d) reduction of youth access to tobacco.

Main outcome measure: The prevalence of self-reported smoking and smokeless tobacco use in the week before assessment.

Results: The community programme had significant effects on the prevalence of weekly cigarette use at times 2 and 5 and the effect approached significance at time 4. An effect on the slope of prevalence across time points was evident only when time 2 data points were eliminated from the analysis. The intervention affected the prevalence of smokeless tobacco among grade 9 boys at time 2. There were also significant effects on the slope of alcohol use among ninth graders and the quadratic slope of marijuana for all students.

Conclusion: The results suggest that comprehensive community wide interventions can improve on the preventive effect of school based tobacco prevention programmes and that effective tobacco prevention may prevent other substance use.

Design

The design of the current study is shown in fig 1. It was a randomised controlled trial in which small Oregon communities were assigned to one of two conditions (see below). The population of these communities ranged from 1,700 to 13,500. The principal economic activities are tourism, logging, fishing and farming. Communities were selected such that the possibility of contamination between communities was minimised. The communities share no common high schools and are at least 20 miles apart. In order to participate, school districts agreed to implement the school based intervention and to permit the in-school assessment sequence shown in fig 1.

Pairs of communities were matched on community socioeconomic status and population. One member of each pair was assigned at random (via the flip of a coin) to receive a school based tobacco and other substance use prevention programme (school based only (SBO) condition) in grades 6 through to 12. The other member received a community intervention programme in addition to the school based programme (CP condition). Communities in the two conditions did not differ statistically in size, per capita income, median household income, the percent of people below the poverty level, the proportion of minority students, the number of high school students per grade, or the proportion of high school graduates in the population.

Implementation of the design was carried out in three phases to allow refinement and streamlining of intervention procedures and to minimise demands on project personnel.



Self-Assessment Exercise 7.4.6

According to the information in the abstract and the excerpt on methods from Paper D:

1. How many communities were involved?
2. How large (population size) were the communities?
3. What factors helped to minimise contamination between communities?
4. How did the investigators try to minimise the effect of confounding factors in the comparison between control and intervention communities?
5. What evidence was there that they had succeeded in the objective described in (4) above?

Answers in Section 7.8

7.4.8 Non-Randomised Intervention Designs

In this chapter, all of the intervention designs we have looked at so far have used random allocation to intervention or control groups, whether these are of individuals or of groups (clusters). There are a number of other intervention study designs that, usually for pragmatic and/or ethical reasons, do not employ randomisation to generate the comparison between intervention and control groups. These are sometime called *quasi-experimental* designs.

Three of the most important types are described in Table 7.4.5, with an explanation of key design features and the main methodological limitations.

Table 7.4.5 Three of the most commonly used types of non-randomised intervention study designs.

Study type	Key design features	Methodological issues
Uncontrolled before-and-after study	After the study sample is identified, baseline measurements are made of the outcome of interest and of other relevant variables. The intervention is then applied. At some point following the intervention, the outcome and other measurements are repeated. This design generates paired data.	The strengths of this design are that it is relatively simple to carry out, and because subjects serve as their own controls, there is less random error, and confounding may also be reduced. The main weakness is that some important time-related variables affecting the outcome may change between baseline and follow-up.
Controlled before-and-after-study	A similar design to the uncontrolled before-and-after study, but with the addition of one or more control groups, for which baseline and follow-up measurements are made, but no intervention is applied. Efforts are usually made to ensure that the intervention and control groups are as similar as possible (often with some matching). This study design also generates paired data within the groups, as well as allowing comparison between groups.	This design is stronger than the uncontrolled before-and-after study, by virtue of the control group(s), which allow the effects of any time-varying factors to be assessed and adjusted for. A common approach to analysis is by 'difference in differences', that is, the change seen in the intervention vs control comparison between baseline and follow-up. The main weakness lies with the fact that the intervention is not randomly allocated, and it can be difficult to identify comparable control group(s).
Interrupted time-series	This type of study is used to assess whether an intervention has had an impact in circumstances where there is an underlying change in the rate or prevalence of the outcome over time (<i>secular trend</i>) due to other factors. A series of measurements are made prior to the intervention and again afterwards. The purpose of making the multiple measurements is to be able to define the underlying secular trend, and hence distinguish the additional effect of the intervention. Key to this design is the number and frequency of repeat measurements that can be made.	This design can provide good evidence of intervention impact, so long as there are sufficient repeat measurements before and after the intervention and appropriate statistical analysis is used, allowing for the greater similarity of values obtained closer in time, termed <i>autocorrelation</i> . The weaknesses are that, in practice, it is often difficult to obtain sufficient, well-standardised repeat measurements, and the design may not be able to distinguish between a true effect of the intervention and some other change occurring at the same time.

7.4.9 The Natural Experiment

One final type of intervention study design worthy of mention is the *natural experiment*. This is actually an observational study design, but one where the researchers take advantage of a natural (e.g., geological) social or economic event, to study the effect on one population group compared to another group not exposed to the change. This design is most useful if the event is fairly marked and rapid and if the population groups being compared are similar in other respects. The design may compare two independent groups, or the same population before and after the event. A key feature of a natural experiment that differentiates this study design from both the randomised and non-randomised intervention designs discussed above is that the researchers do not have control over the nature, timing, or allocation of the intervention. This means that natural experiments can be used for studying the effects of changes (in exposure, policy, etc.) for which deliberate implementation for research purposes would be impractical or unethical.

Perhaps the most famous historical example of a natural experiment is the work by John Snow in identifying the source of cholera in London in 1854. He was able to show that higher rates occurred among homes supplied by one company that drew water out of the Thames below the level at which sewage flowed into the river, while there were lower rates among homes supplied by another company drawing water from higher up the river. Snow had no control over who received which water, but he was able to observe the results of what he described as an ‘an experiment ... on the grandest scale’.

Other, more recent, examples include studies of the effects of radiation on cancer risk from weapons testing and use of the atomic bomb in Japan in 1945, and the beneficial health impacts of bans on smoking in public places. In neither case were these ‘events’ planned by the researchers, but they both offered opportunities to study the effects on health. In the former case, variation in the risk of cancer according to estimated dose of radiation could also be determined.

Summary

- The elements of RCT design that contribute to its scientific strength become more difficult to apply with more-complex preventative health-care and community-based interventions.
- Placebo control is especially difficult to achieve in these circumstances.
- Blinding may be difficult to achieve, but some degree of blinding of the assessment is usually possible and should be attempted.
- Cluster randomisation is appropriate for interventions that include components that can, or are intended to, affect groups and where there is a high chance of contamination (controls are exposed to influences intended only for the intervention group).
- Community-based trials, which may also be randomised, are appropriate where the intervention(s) seek to alter social determinants of behaviour and health and are most effectively addressed through action at the level of communities.
- Non-randomised intervention studies may be used when randomisation is not practical, but they require greater attention to the effects of confounding and other time-varying factors that can affect the outcome.
- The natural experiment, although an observational design, offers the opportunity to study the health impacts of important events or changes in social policy that could not be the subject of a deliberate research-led intervention.

- It is important to be aware of what is lost by deviating from the pure RCT design, and to preserve as much as possible. However, overzealous adherence to the RCT design ideal where this is inappropriate may well result in an artificial intervention and/or an atypical study sample, and it can ultimately compromise the usefulness of the results and the extent to which these can be generalised.

7.5 Analysis of Intervention Studies Using a Cluster Design

7.5.1 Why Does the Use of Clusters Make a Difference?

In describing the design of a cluster trial, we introduced the idea that subjects within clusters may be more alike in respect of certain characteristics – for example, body weight, behaviour such as smoking, or uptake of screening – than subjects compared across different clusters. In other words, variation between clusters is greater than within clusters because values for individuals within clusters are correlated.

We can think of clustered data as being on several levels. To explain this, we will consider the example of some (hypothetical) data for an index of adolescent mental health – the Strengths and Difficulties Score (SDQ) in a number of schools. Let's say we want to study the relationship between SDQ score and parental smoking at home (categorized simply as yes/no). So, in this example, we have two levels of data:

1. Level 1: individual child data: the SDQ score, parental smoking at home (yes/no), and other characteristics such as age and sex.
2. Level 2: The school that each child attends

We suspect that some features of each school (e.g., catchment area, the values promoted by the school, level of parental support) will influence SDQ score, and these factors make children in any given school more similar (correlated) in terms of SDQ score. As a result, in a sample of children selected as clusters from a small number of schools, the variation in SDQ score is less than if the same number of children had been selected randomly from a much larger number of schools. If the clustering is not taken into account in the analysis, the standard errors for mean SDQ scores for the two groups (i.e., smoking at home and not smoking at home) will be underestimated. Other ways of expressing this are that precision would be overestimated, confidence intervals would be artificially narrow, or the p -value when comparing the two groups would be artificially small. Thus, by not allowing for clustering, we could make a Type I error in our analysis of the relationship between smoking at home and SDQ score.

We have a very similar situation in the counselling and exercise trial (Paper C), in which the intervention was delivered via clusters (practices), which are Level 2 data, and the individual amounts of exercise and other characteristics are Level 1 data. Before looking at the methods the authors used to analyse these data allowing for clustering, we introduce the technique used to assess quantitatively how similar individuals within clusters are.

7.5.2 Summarising Clustering Effects: The Intra-Class Correlation Coefficient

The pattern of variation within and between clusters can be summarized quantitatively by the intra-class correlation coefficient (ICC). Using our example of children's SDQ scores, Figure 7.5.1 shows two scenarios for the distributions of scores within and between four schools. The Figure 7.5.1 (a) shows considerable variation between schools in mean SDQ

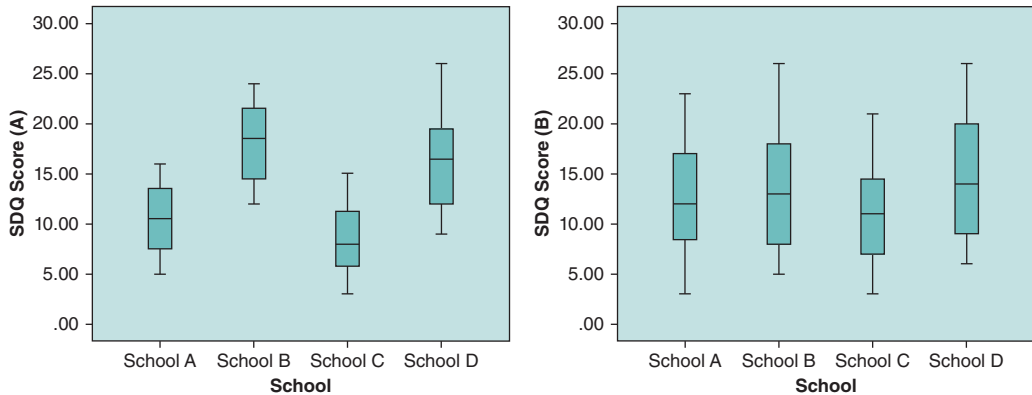


Figure 7.5.1 SDQ (performance) score for children in four schools; see text for further explanation.

score, with children in each one tending to be bunched towards higher or lower levels; the ICC for this scenario is 0.026. The second example, Figure 7.5.1 (b), shows a similar overall range in SDQ scores across all children, but the means for the schools do not vary as much, and they are not bunched in the same way towards higher or lower levels; the ICC for this scenario is lower at 0.016. In summary, the higher the ICC, the more similar individuals within clusters are compared to when looking across clusters.

7.5.3 Multi-Level Modelling

We can now consider the approach to analysing clustered data. In studying the relationship between parental smoking at home and SDQ score, we could ignore the clusters and simply use linear regression to obtain a coefficient to determine how much parental smoking affected the SDQ score in comparison with no parental smoking. The problem is that the precision will be over-estimated (artificially small standard error), and we may obtain a p -value of <0.05 , suggesting that there is a significant relationship when in fact this is not the case (Type I error).

A multi-level model can help with this by specifying coefficients for parental smoking at home (level 1) and school (level 2). The model uses a mix of fixed and random effects; typically, the intercept is random, which allows mean values of the outcome (SDQ score) to vary between level-2 units (schools) – which is what we suspect is needed. The model will also estimate the fixed effect, given by the slope of the regression equation (β -coefficient), which describes the average relationship between smoking and SDQ score for all schools.

These assumptions can be tested and the model revised as necessary. Looking back at Figure 7.5.1, we would expect that SDQ score does vary substantially between schools in scenario (a), but not in scenario (b); this is indeed the case, with the p -values for differences between schools being <0.0005 and 0.473, respectively, in analysis of variance (ANOVA) – see Chapter 11 for further explanation of this procedure.

7.5.4 Analysis of the Cluster RCT of Physical Activity

As referred to above, the trial of the effect of counselling in general practice on physical activity (Paper C) used a cluster randomised design in which the medical practices were the clusters. The main results, which are shown in Table 7.5.1 (Table 2 from Paper C), employed a multi-level model to allow for practice-related influences that could make individuals within practices

Table 7.5.1 Mean changes (with 95% confidence intervals) in physical activity, cardiovascular, and quality-of-life outcomes in the control and intervention groups at 12 months (Table 2 from Paper C).

Measure	Intervention* (n = 451)	Control*	Difference between groups† (n = 878)	P value
Primary outcomes:				
Total energy expenditure (kcal/kg/week)	9.76 (5.85 to 13.68)	0.37 (−3.39 to 4.14)	9.38 (3.96 to 14.81) (975 kcal/week)	0.001‡
Leisure physical activity (kcal/kg/week)	4.32 (3.26 to 5.38)	1.29 (0.11 to 2.47)	2.67 (0.48 to 4.86) (247 kcal/week)	0.02§
Leisure exercise (minutes/week)	54.6 (41.4 to 68.4)	16.8 (6.0 to 32.4)	33.6 (2.4 to 64.2)	0.04§
Systolic blood pressure (mm Hg)	−2.58 (−4.02 to −1.13)	−1.21 (−2.57 to 0.15)	−1.31 (−3.51 to 0.89)	0.2
Diastolic blood pressure (mm Hg)	−2.62 (−3.62 to −1.61)	−0.81 (−1.77 to 0.16)	−1.40 (−3.35 to 0.56)	0.2
4-year risk of coronary heart disease (%)	0.42 (0.23 to 0.60)	0.52 (0.32 to 0.72)	−0.10 (−0.43 to 0.23)	0.6
SF-36 quality of life scores:				
Physical functioning	3.16 (1.61 to 4.71)	1.63 (−0.04 to 3.31)	1.23 (−1.35 to 3.81)	0.3
Role physical	10.53 (6.8 to 14.3)	4.16 (0.63 to 7.68)	7.24 (0.16 to 14.31)	0.045§
Bodily pain	6.51 (4.28 to 8.74)	2.50 (0.15 to 4.86)	4.01 (0.78 to 7.24)	0.02§
General health	5.95 (4.43 to 7.47)	1.60 (0.22 to 2.99)	4.51 (2.07 to 6.95)	0.000‡
Vitality	5.36 (3.76 to 6.96)	3.06 (1.44 to 4.68)	2.30 (0.03 to 4.57)	0.047§
Social functioning	3.02 (0.68 to 5.36)	2.85 (0.57 to 5.13)	0.36 (−3.53 to 4.26)	0.9
Role emotional	5.32 (1.43 to 9.21)	5.70 (2.07 to 9.32)	−0.38 (−5.70 to 4.94)	0.9
Mental health	2.61 (1.17 to 4.04)	1.63 (0.28 to 2.98)	0.98 (−0.99 to 2.95)	0.3
Other variables:				
Body mass index (kg/m ²)	−0.11 (−0.25 to 0.02)	−0.05 (−0.18 to 0.07)	−0.06 (−0.24 to 0.12)	0.5
Cholesterol concentration (mmol/l)	−0.019 (−0.08 to 0.05)	0.01 (−0.05 to 0.06)	−0.02 (−0.12 to 0.09)	0.7

*Unadjusted for clustering.

†Adjusted for clustering by medical practice.

‡Significant at 0.01 level.

§Significant at 0.05 level.

Source: Elley 2003. Reproduced with permission of BMJ Publishing Group Ltd.

more alike in respect of physical activity. Please now read the short excerpt from the methods section from Paper C, reproduced below, which outlines the approach taken to analysis of the main outcomes.

We analysed differences between intervention and control groups in change of outcome variables by using random effects models in Stata 7.0 (generalised least squares) and SAS 8.2 (mixed model), to allow for clustering by practice. All outcome analyses were by intention to treat, according to random allocation.

We now examine and interpret the results through the following self-assessment exercise.



Self-Assessment Exercise 7.5.1

1. In Table 7.5.1, which sets of results (columns) provide information on primary and other outcomes unadjusted for clusters and adjusted for clusters?
2. Using the results that are unadjusted for clusters (that is, the whole sample of subjects taken as one group for intervention and one for control), what was found for 'Leisure exercise' in minutes per week (and the 95% CIs)?
3. What is the intervention effect (difference between groups) for the outcome 'Leisure exercise'? Although we are not provided with a p -value for this difference between groups, what do the 95% CIs tell us about the expected level of significance?
4. What was the intervention effect and p -value for this same outcome, following adjustment for clusters? Interpret your findings in comparison with the unadjusted effect.

Answers in Section 7.8

7.6 How Big Should the Intervention Study Be?

7.6.1 Introduction

Determining an appropriate sample size is as important for intervention studies as it is for surveys and other study designs such as case-control and cohort studies. Indeed, ethical considerations relating to the fact of carrying out an intervention (e.g., a new treatment; a disease-prevention strategy) on individuals or populations, if anything, demands additional attention being given to ensuring that the study has sufficient power to detect a useful benefit.

The approach usually taken to determine the sample size for an intervention study involves the comparison of two groups. Indeed, the methods described here apply to most situations in which you wish to calculate sample size for comparing two groups; for example, information on smoking (the variable of interest) in men and women (the two groups) obtained through a survey.

We have previously discussed the two main types of data for statistical analysis, *categorical* and *continuous*, and the fact that we use different formulae for calculating *standard error* for these two types of data. Since standard error forms the basis of the sample size calculation, we need to deal with categorical and continuous data outcomes separately.

7.6.2 Sample Size for a Trial with Categorical Data Outcomes

The following description from the falls prevention trial (Paper B) summarises the information required for sample size calculation with a categorical outcome.

To detect a 25% relative reduction (or more) in the annual fall rate, with 5% significance level and power of 80%, 914 individuals were needed. The calculation assumed a non-intervention annual rate of 35 per 100 people, and a main effects two-group comparison for each intervention. Allowing for a 20% dropout rate, 1143 subjects were needed (Paper B, p. 129).

The outcome is *categorical* because the subjects are classified as either experiencing a fall or not experiencing a fall. Thus, the information required is as follows:

- P_1 : The percentage (or proportion) of people with the characteristic in the control group. In this study the characteristic is falling (over a period of 1 year). The rate of falls in the non-intervention population is taken as 35 per 100 per year; hence, P_1 is 35 per cent (or 0.35, expressed as a proportion).
- P_2 : The percentage (or proportion) of people we expect to have the characteristic in the intervention group. Here the authors are designing the study to detect (at least) a 25 per cent reduction in the fall rate, so P_2 is 26.25 per cent (or 0.2625, expressed as a proportion). Note that this is a 25 per cent reduction on a rate of 35 per cent. In choosing these values, the authors are making a decision about the difference $P_1 - P_2$ (the benefit) that they regard as important to be able to detect.
- The significance level (α) chosen for the hypothesis test. In this study the chosen significance level is 5 per cent ($1 - \alpha = 95\%$). This is a two-sided significance level; one- and two-sided tests are explained further in Section 7.6.3.
- The power of the study ($1 - \beta$); for this study, 80 per cent power has been chosen, a value that is commonly used for calculating sample size.

With this information, we can calculate the number of people needed in each group (intervention and control) using OpenEpi software, as shown in Table 7.6.1.

Table 7.6.1 Sample size for detecting a difference in two group proportions.

Sample Size: X-Sectional, Cohert, & Randomized Clinical Trials	
Two-sided Significance Level(1-alpha):	95
Power(1-beta, % chance of detecting):	80
Ratio of Sample Size, Unexposed/Exposed:	1
Percent of Unexposed with Outcome:	35
Percent of Exposed with Outcome:	26*
Odds Ratio:	0.66
Risk/Prevalence Ratio:	0.75
Risk/Prevalence Difference:	-8.7

	Kelsey	Fleiss	Fleiss with CC
Sample Size-Exposed	433	432	454
Sample Size-Nonexposed	433	432	454
Total Sample Size:	866	864	908

References

Kelsey *et al.*, Methods in Observational Epidemiology 2nd Edition, Table 12-15

Fleiss *et al.*, Statistical Methods for Rates and Proportions, formulas 3.18 & 3.19

CC = continuity correction

Results are rounded up to the nearest integer.

Print from the browser menu or select, copy, and paste to other programs.

*The value of 26.25 was entered into OpenEpi, but it is automatically rounded in the output.

Using the non-continuity-corrected value (see Chapter 5, Section 5.6.3 for a description of the continuity correction), this shows that we need 433 subjects per group, a total of 866 to meet the criteria set. Using the continuity-corrected option provides a more-conservative estimate showing we require around 908 in total (454 in each group), close to the value of 914 subjects described in Paper B. The small difference between our calculation and that reported by the authors is likely to be due to software differences and possibly some rounding errors.

Allowing for Refusals and Drop-Outs

There is one final and very important step in sample size calculation, which was described in the excerpt from Paper B; namely, making some allowance for people who do not wish to take part (refusals) and others who do join the study but are lost to follow-up before the planned study completion (drop-outs).

In Paper B, allowance was made for 20% dropouts, resulting in a total of 1,143 subjects being required at recruitment. The value chosen for refusals and dropouts depends on the subjects (e.g., age, health), duration of follow-up (one can expect a higher dropout rate with a longer follow-up), and other factors relevant to the study setting. Ultimately, this requires judgment and researchers should err on the side of caution, but they may also be informed by empirical evidence where available – for example, from a prior study carried out in the same or a similar population.

7.6.3 One-Sided and Two-Sided Tests

In sample size calculations we may occasionally see that *one-sided* significance has been assumed. A *one-sided test* works on the assumption (perhaps based on other evidence) that the intervention can have an effect in only one direction. We therefore need fewer data to test our hypothesis and thus a smaller sample. Although using one-sided tests results in a smaller sample size requirement, this is a risky approach, and it is one that is likely to be frowned upon by statistical referees of research-funding applications and papers. In a trial of some new intervention, it is wise to allow for the possibility that the new treatment could actually be worse than the best existing treatment. We do not usually have reliable enough information about the direction of any difference, and it is strongly recommended that *two-sided* tests be used so that the study has sufficient power to detect either a better or worse outcome.

7.6.4 Sample Size for a Trial with Continuous Data Outcomes

Calculation of sample size for a continuous outcome variable follows the same principle as for the categorical outcome, although some of the information required is different.

As with the categorical outcome, we need to make a decision about the size of difference that we wish to be able to detect (demonstrate as statistically significant). If we label the means of the intervention and control populations μ_1 and μ_2 , respectively, we need to decide the difference ($\mu_1 - \mu_2$) to be detected. Since the standard error is calculated from the standard deviation, we also need an estimate of the standard deviation of the outcome variable. In summary, we require:

- $\mu_1 - \mu_2$: The difference between the population means of the outcome variable in the control group and intervention group.
- σ_1 and σ_2 : The standard deviation of each group. Note that we may have to assume this value is the same for each group, unless we have sufficient data to derive separate values for the two groups. Usually it is necessary to obtain estimates of standard deviation from other studies on

similar populations (e.g., from the literature), or preparatory studies that have been carried out on the same population.

- The significance level (α) required for the hypothesis test; e.g., 5 per cent.
- The level of power ($1 - \beta$) for the study; e.g., 80 per cent.

You can now calculate the sample size required for detecting a difference between two means in the following exercise. In this exercise, we also look at the effect that a change in the difference between means ($\mu_1 - \mu_2$) has on sample size.



Self-Assessment Exercise 7.5.1

1. In Paper B, Table 3 (not shown) gives data on a number of continuous variables. The 'balance range' was an important outcome for which the baseline mean was 13.3 cm with a standard deviation of 4.5 cm (we will assume the same value for each group, and that the ratio of sample sizes for the two groups is 1). Using the OpenEpi software (or a suitable alternative) calculate the sample size for detecting a 1.5-cm improvement in the balance range, with a significance level of 5 per cent and power of 80 per cent.
2. Comment on what you find in respect of the actual study size.
3. What is the effect on sample size of halving the balance range difference you wish to detect to 0.75 cm?

Answers in Section 7.8

7.6.5 Sample Size for an Intervention Study Using Cluster Design

The calculation of sample size for a cluster trial is similar to that described above for individual-based two-group comparisons, with one important difference. This is that we need to take account of the *intra-class correlation* described in Section 7.5.2, due to subjects within a cluster being more alike than subjects from different clusters, and the resulting loss of precision.

An adjustment therefore needs to be made to the sample size, to compensate for the effect of the cluster design on study power. This adjustment, which will increase the sample size, requires an estimated value for the intra-class correlation coefficient (ICC). Please now read the following excerpt on sample size calculation from Paper C, which describes how ICCs based on estimates from previous studies were used for the four main outcomes:

Sample Size Calculation

A sample size of 800 patients from 40 practices ($\alpha = 0.05$, power = 90%) was required to detect differences in change between the intervention and control groups of 1 hour of moderate physical activity per week, 4.5 mm Hg systolic blood pressure, 10% relative risk of cardiovascular events, and six points of SF-36 'vitality.' We assumed an attrition rate of 25%. To account for the effect of clustering, we adjusted the sample size calculations by using intra-class correlation coefficients of 0.05, 0.016, 0.0036, and 0.05 for physical activity, blood pressure, cardiovascular risk, and vitality, respectively, based on estimates from previous studies.

We will not go into the calculation of sample size for cluster trials in any more detail in this book. It is sufficient that you are aware of the need to allow for the effect of using clusters and to obtain a value for the ICC, and that this adjustment will increase the sample size required.

As previously mentioned, we recommend that statistical advice be sought for all but the most straightforward sample size calculations.

7.6.6 Estimation of Sample Size is not a Precise Science

We have now looked at sample size calculations for cohort, case-control, and intervention studies. These calculations involve the estimation of at least one of the parameters involved: A judgment will always have to be made about the size of the effect that you wish to detect, and often the prevalence, mean, or standard deviation has to be estimated from other studies. As a result, a calculated sample size should be seen as a guide and not as a precise or absolute value. In making a final decision about sample size it is also wise to err on the side of caution.

Finally, it is also necessary to make realistic allowances for non-response and refusals in recruiting the sample and for dropouts from the follow-up phase of the trial. These, in turn, cannot be predicted exactly, but good estimates can be obtained from other similar work in comparable circumstances.

Summary

- Sample size calculation is a vital part of the planning of an intervention study and is carried out to reduce the chances of making type I and type II errors in any given situation.
- It is wise to obtain advice from a statistician on the correct approach to calculating sample size for a given study.
- It is recommended that the α values (significance level of hypothesis test) associated with two-sided tests be used for sample size calculation.
- A decision is always required about the size of effect to be detected, and this in turn will depend on a judgement about what size of effect is important to demonstrate.
- Sample size for cluster study designs should allow for greater similarity of subjects within a cluster than between clusters; this involves use of the intra-class correlation coefficient (ICC) and results in an increase in the sample required compared to a non-clustered design.
- Sample size calculation provides a guide for study design and should not be seen as a precise estimate. If in doubt, err on the side of caution!
- Make realistic allowances for non-response, refusals, and dropouts.

7.7 Intervention Study Registration, Management, and Reporting

7.7.1 Introduction

Throughout this book, we have emphasised the importance of careful study design, attention to detail, and a thorough approach to research management in practice. In this chapter, the special nature of intervention studies has been highlighted, based on the fact that these involve the research team making changes in the treatment, exposures, or other aspects of people's lives. In addition to the ethical responsibilities this brings, there is also the expectation that certain standards of study management will be met covering trial registration, research management, subject safety, and reporting. Meeting these expectations is increasingly stated by journals as a condition of publication.

Key elements of the guidance and recommendations covering these aspects of intervention study management and reporting are provided in sections 7.7.2 to 7.7.4, together with some

Web links for further information and reference. Although this guidance has been designed primarily for randomised trials, some aspects should be considered in planning other types of (non-randomised) intervention study, again due to the fact that these involve introducing a change that has the potential to affect people's lives, health, and safety.

7.7.2 Registration

Before a trial is implemented, the protocol should be registered. The reasons for this are summarized in Box 7.1, drawn from the WHO's guidance on trial registration available at: http://www.who.int/ictrp/trial_reg/en/

Box 7.1: Reasons for Trial Registration

The registration of all interventional trials is considered to be a scientific, ethical, and moral responsibility because:

- There is a need to ensure that decisions about health care are informed by all of the available evidence.
- It is difficult to make informed decisions if publication bias and selective reporting are present.
- The Declaration of Helsinki states that 'Every clinical trial must be registered in a publicly accessible database before recruitment of the first subject'.
- Improving awareness of similar or identical trials will make it possible for researchers and funding agencies to avoid unnecessary duplication.
- Describing clinical trials in progress can make it easier to identify gaps in clinical trials research.
- Making researchers and potential participants aware of recruiting trials may facilitate recruitment.
- Enabling researchers and health care practitioners to identify trials in which they may have an interest could result in more effective collaboration among researchers. The type of collaboration may include prospective meta-analysis.
- Registries checking data as part of the registration process may lead to improvements in the quality of clinical trials by making it possible to identify potential problems (such as problematic randomization methods) early in the research process.

Source: WHO http://www.who.int/ictrp/trial_reg/en/

A number of registries are available for this purpose. Among these, the International Standard Randomised Controlled Trial Number (ISRCTN) Registry is a primary clinical trial registry recognised by WHO and the International Committee of Medical Journal Editors (ICMJE). It accepts all clinical research studies, whether proposed, ongoing, or completed. It provides content validation, preservation, and management and the unique study identification number necessary for subsequent publication of the trial. All of the study (trial) records in the ISRCTN database are freely accessible and searchable.

7.7.3 Trial Management

The UK Medical Research Council has produced guidance on the management of intervention trials, which can be found at <http://www.mrc.ac.uk/documents/pdf/good-clinical-practice-in-clinical-trials/>.

Although the management arrangements for different types of trial vary, it is considered good practice to ensure there is independent supervision. The two main mechanisms for securing this are through a trial steering committee (TSC), and a data monitoring and ethics committee (DMEC).

Trial Steering Committee (TSC)

The purpose of the TSC is to provide independent oversight of the study, to ensure that its conduct meets the rigorous standards expected of such intervention trials. The membership will vary according to the needs of the study, but it should include a chair who is independent of the principal investigators (PIs) and their research institution, one or more other independent members bringing specific expertise, and one or more of the PIs. The TSC usually reviews the study protocol before it is finalised (and registered – see Section 7.7.2), and meets at least once per year to review progress against the protocol (including the timeline), subject safety (including reports from the DMEC; see below), and any newly arising issues that may affect the trial. It would also normally be within the remit of the TSC to review and endorse annual (or other periodic) reports on trial progress prepared by the PI.

Data Monitoring and Ethics Committee (DMEC)

The purpose of the DMEC, which should be established by the TSC early in the process, is to carry out regular reviews of the data and conduct interim analyses. The DMEC should be wholly independent of the TSC, the PIs, and the PIs' host institutions. In reviewing the data at regular intervals, the DMEC will be looking for any important adverse events that may influence whether or not the trial should continue. Being independent, this group is able to conduct interim analysis (breaking the randomisation code) in order to detect adverse events and/or whether a large enough effect from the intervention can be reliably determined before the anticipated end of the study. In the latter case, it may be considered unethical to continue the trial if benefit has already been demonstrated, as it would be wrong to continue withholding the more-effective treatment from control subjects.

7.7.4 Reporting Standards (CONSORT)

In order to ensure higher standards of consistency and transparency in the reporting of trials, the CONSORT (Consolidated Standards of Reporting Trials) group has developed the CONSORT Statement. This statement is an evidence-based minimum set of recommendations for reporting randomised trials; it is available at: www.consort-statement.org. The statement comprises a 25-item checklist set out under the following headings:

- Title and Abstract
- Introduction
- Methods
- Results
- Discussion
- Other information (Registration, Protocol, Funding)

The CONSORT Explanation and Elaboration document explains and illustrates the principles underlying the CONSORT Statement, and the group strongly advises that this should be used in conjunction with the CONSORT Statement.

Accompanying the checklist is a skeleton flow chart that provides a useful guide to reporting numbers of subjects at enrolment, allocation (and randomisation), follow-up, and analysis, including recording the reasons for exclusions, dropouts, etc.

Extensions of the CONSORT Statement

A number of extensions to the statement have been developed to give additional guidance for RCTs with specific designs (including cluster randomised trials), intervention types, and data. This is available at: <http://www.consort-statement.org/extensions>, and the main headings are shown in Table 7.7.1.

Table 7.7.1 Extensions to the CONSORT statement for trial reporting.

Study designs	Intervention types	Data
Cluster trials	Herbal medicine interventions	CONSORT-Pro ³
Non-inferiority and equivalence trials ¹	Non-pharmacologic interventions	Harms
Pragmatic trials ²	Acupuncture interventions	Abstracts

¹Trials that seek to determine whether a treatment is no worse than another, or equivalent to another.

²Pragmatic trials are those that are designed to measure effectiveness, that is, whether an intervention works when used in typical conditions of medical care or other service delivery.

³Trials based on patient-reported outcomes.

7.8 Answers to Self-Assessment Exercises

Section 7.1

Exercise 7.1.1

If the research team had carried out an observational study (e.g., case–control or cohort), they would have to base their exposure assessment on measurement of existing patterns of nicotine inhaler use among people in a study sample. This was a relatively new product and may therefore not have been in use extensively, presenting practical difficulties for achieving the required sample size. Furthermore, people who were using the inhaler already would be unlikely to be typical of all ‘smokers unable or unwilling to quit’, so the research team would need to go to great lengths to try to characterise the features of users and non-users and adjust for these. It is unlikely they could avoid the bias resulting from this.

Furthermore, by using an observational approach, it would not be possible to blind the subjects or the research team to whether or not subjects were using an inhaler, thus introducing another source of bias.

Thus, to gain more-effective control of who used the inhaler and to minimise various other sources of bias, it was very appropriate to carry out an intervention study.

Exercise 7.1.2

1. Objectives:

- To determine the effectiveness of an oral nicotine inhaler in achieving long-term reduction in smoking.
 - To determine safety of nicotine inhaler use among people who are still smoking.
2. The population was people (male and female), aged 18 and older, living in (and around, depending on the readership of newspapers) Basle and Lausanne, Switzerland, who were smokers ‘unable or unwilling to give up, but interested in reducing their smoking’. When you look at the eligibility criteria you will see that this was further defined, and we consider this again in Exercise 7.2.1.

3. The sample was selected by using newspaper advertisements, asking for healthy smokers 'unable or unwilling to give up, but interested in reducing their smoking'. These were therefore volunteers. The question of whether this sample is representative of the population is considered further in Exercise 7.2.1, but even if demographic characteristics are similar between sample and population, one must always ask whether volunteers differ in other ways that are relevant to the study; for example, having a higher level of motivation or being prepared to try something new.
4. The intervention was an inhaler containing nicotine and menthol. The control was a placebo inhaler, containing only menthol. The use of placebos is considered further in Section 7.2.

Section 7.2

Exercise 7.2.1

1. The criteria that determined whether or not a person took part were as follows:

Criteria of eligibility to be included	Criteria for exclusion
<ul style="list-style-type: none"> ● Aged 18 and over. ● Both sexes. ● Smoke 15 or more cigarettes per day. ● Have a CO concentration in exhaled air of ≥ 10 ppm. ● Have smoked regularly for 3 or more years. ● Have failed in at least one serious attempt to quit smoking in the last 12 months. ● Want to reduce smoking as much as possible with the help of the nicotine inhaler. ● Prepared to adhere to the protocol. ● [Willing to provide informed consent.] 	<ul style="list-style-type: none"> ● Current use of nicotine replacement therapy or any other behavioural or pharmacological smoking cessation or reduction programme. ● Use of other nicotine-containing products. ● Have any condition that might interfere with the study.

It is very important that these criteria are clearly defined prior to the selection process and described (in the protocol and publications) so that others can see exactly who was included in the sample and who was excluded. The criteria specified do seem to be clear and easily understandable, although the last exclusion criterion does seem rather vague and is not further defined or discussed.

2. The selection process was not random, as volunteers were sought through newspaper advertisements. It is also useful to refer to the study flow diagram in Paper A (this is reproduced in Section 7.2.1). Diagrams such as this are an extremely useful way of summarising the recruitment, selection, randomisation, and follow-up process: You can see at a glance how many people got through each stage (a flow chart is also a condition of publication – see CONSORT statement in Section 7.7.4). At present we are only concerned with the initial stages of the study. Of 1,115 people answering the advert, less than 50 per cent (468) had telephone screening, of whom 407 were invited to a baseline visit. The paper does not explain why only 468 out of 1,115 were screened, nor does it indicate how representative those who progressed into the study were of all those responding to the advert. Of the 407 invited to baseline visit, only seven did not meet the 'admission' criteria (of course, many more were deemed ineligible at earlier stages of the process), and 400 were 'eligible and randomised'.
3. How representative is the sample of the Swiss population? As there is no sampling frame (list of names with identifying characteristics), we have no way of knowing whether the sample

was representative of the Basle and Lausanne populations of 18 years or older smokers unable or unwilling to give up. Based on the process review in the answer to point 2 above, there are some questions about why less than 50 per cent of those answering the advertisement progressed past the telephone screening, and we have no further information to judge how representative the sample is of the populations of the two cities. We would need additional information to assess whether findings on the response of these smokers to the nicotine inhaler is likely to be typical of how people throughout Switzerland would respond.

Exercise 7.2.2

1. The nicotine inhaler (which also contained 1 mg of menthol), termed the *active treatment*, was compared with a placebo inhaler that contained only menthol. In all respects, the placebo appeared identical to the nicotine inhaler.
2. The general principle for deciding what the appropriate comparison treatment should be is as follows. The control (comparison) group must be offered the best existing treatment in routine use. If there is no known effective treatment, then the controls can be given no active treatment (which, if practical, may be as placebo). In the case of this nicotine study, although it is well established that nicotine replacement can help smokers reduce or quit, it seems that 'successful abstinence is usually obtained in smokers with low to moderate nicotine dependence' (p. 329 of Paper A). Heavily dependent smokers have the highest relapse rates, and they are at highest risk for smoking-related disease, so smoking reduction can be seen as a useful (possibly intermediate) goal. As the inhaler was a relatively new product, it is implied (as this is not explicitly discussed) that it is not unethical to compare the new inhaler with no active treatment in this group of smokers.

This is a different situation from, say, testing the effect of a new surgical procedure that is thought might improve on an existing procedure. So long as there is evidence that the existing procedure has some beneficial effect, it would be unacceptable to compare the new operation with no operation at all. No ethical committee would allow a study to proceed that did not address this matter appropriately.

Exercise 7.2.3

1. The main outcome, or measure of success (the 'primary efficacy measure') was defined as 'self-reported reduction of daily cigarette smoking by at least 50 per cent compared with baseline from week six to month four.'
2. This outcome was validated by 'decreased breath carbon monoxide (CO) concentrations at week six and months three and four'. The authors do not appear to state how large a reduction in CO (or to what concentration) is required for validation.
3. The other outcome examined was smoking cessation, defined as not smoking from week 6 and a breath CO concentration of <10 ppm at all subsequent visits.

Exercise 7.2.4

1. Blinding in the nicotine inhaler study:

Group	Blinding applied
1. Subjects	Placebo should have ensured blinding, but it is likely that smokers who are highly nicotine dependent would notice that the active inhaler is having some effect on their craving, and vice versa for those with the placebo inhaler. By giving <i>informed consent</i> , all subjects would know they are being randomised to receive either active or inactive inhalers.

Group	Blinding applied
2. Research team	The research team would not know the allocation until completion of the study, but they could probably pick up some signals in the follow-up assessments, based on comments by subjects. Counselling was provided at each visit, although it is not clear by whom: This is one point at which – if subjects suspected which inhaler type they had and talked about it – some bias could have been introduced.
3. Health-care staff	The most important health-care staff are the independent pharmacists who dispensed the inhalers, and they should have remained blind. Other health-care staff could become involved, such as GPs. Although this was a healthy group of subjects (despite being smokers!), no doubt they sought care for the usual ailments, and in doing so they might mention their involvement in the study and what they felt about the type of inhaler they had been randomised to use; the advice from the GP – if she or he had the time to consider it – could have been influenced by this.

2. Reasons why blinding is important for these three groups:

Group	Reasons
1. Subjects	Could affect compliance if they felt they might be missing out on effective treatment. May influence reporting of outcomes and side effects.
2. Research team	May influence objectivity of data collection on outcomes and side effects.
3. Health-care staff	Knowledge of allocation may influence modifications to treatment for related or other health problems, with management potentially differing between treatment and control subjects.

3. Examples of interventions that cannot be blinded:

- A surgical procedure, the nature of which is clearly obvious to those carrying it out and caring for the patient postoperatively; such procedures may well be apparent to the patient, at least postoperatively.
- Counselling or other similar support and, indeed, the verbal advice and support component that relates to any treatment.
- Health education advice and materials.

It is possible that some of these could be blinded to the research team, or some members of it. The technique of blinding some people involved in the assessment of outcomes is very important. Thus, in a non-blinded study, the outcome should, if possible, be assessed by people or techniques that are not influenced by the open nature of the study. Using self-administered questionnaires (for the subjects themselves to complete) will also help. Although this will not remove *bias* introduced by the subjects' knowledge of their own treatment, the *instrument* is likely to be more objective than an interviewer (who is very likely to find out the allocation during the course of the interview).

Measurements, such as breath CO, are also very useful, as these are less likely to be affected by knowledge of the allocation, so long as such knowledge does not affect the way the observer conducts the measurement. Laboratory-based tests – for example, a blood sample stored and measured later in an independent laboratory – should be free of such bias.

Section 7.3

Exercise 7.3.1

1. Examples of variables that could affect the outcome (independent of the intervention) are age started smoking, number of cigarettes smoked per day, and the sex of subject (as social or other factors may affect men differently from women). The FTND score (nicotine dependence) may also affect the outcome, but this would not be independent of the intervention, as the nicotine inhaler will have an influence on this dependency.
2. Variability is expressed by the standard deviation (SD) and range, which is the correct approach for this purpose. The 95% CI should not be used, and the data are not being used for inference.
3. Variables that vary between the groups include the men-to-women ratio (men: 104/200 in placebo vs. 86/200 in intervention); age when started (about 1 year younger in placebo); number of cigarettes smoked per day (about two per day more in the placebo group). These last two variables might imply more difficulty cutting down in the placebo group, but the differences are not large. Age and weight were well balanced and, perhaps most importantly, so too were the exhaled CO concentration and the FTND score.
4. Beyond already knowing these are volunteers from two Swiss cities who are 'unwilling or unable to give up, but interested in reducing their smoking', these data tell us that there is a mix of men and women, of mean age around 46 years with a wide range (22–79 years), and who started smoking in their late teens. They are quite heavy smokers with an average of around 30 cigarettes per day and range of 15–70 per day: none are light smokers and some are very heavy. Overall, this gives us a fair idea of the characteristics of the sample and to whom the results may apply in other populations.

Exercise 7.3.2

1. Overall, inhaler use decreased over time from 60 per cent of participants using the inhaler at 6 weeks down to 10 per cent at 18 months.
2. Table 7.3.2 provides information on subjects in both groups who used the inhaler every day. The number of subjects using inhalers and the number of inhalers (cartridges) used per day decreased in both groups from the similar starting levels. The rate of decrease in numbers of subjects using inhalers appears to have been greater in the placebo group, but the number of inhalers used per day by these subjects was fairly similar.

Exercise 7.3.3

1. For primary efficacy measure at 24 months (Table 7.3.3)
 - (a) The 2×2 table for sustained reduction at 24 months is shown below. Note (from the percentages in Table 7.3.3), the total numbers are those that were randomised to each treatment ($n = 200$ per group), consistent with the *intention-to-treat* analysis.

	Intervention	Control	Total
Reduced	19	6	25
Did not reduce	181	194	375
Total	200	200	400

- (b) The Fisher's exact test would normally be used if one or more of the cells in the 2×2 table has an expected value of less than 5. The smallest expected value in this case is in the reduced/control cell: $200 \times 25/400 = 12.5$. Thus, although use of the Fisher's exact test is not strictly necessary (the chi-squared test with continuity correction could have been used), it may be that the authors were erring on the side of caution, or they wished to use the same test for all comparisons, some of which have smaller numbers.
- (c) The OR is 3.39, with a 95 per cent CI of 1.39 to 8.29. Hence, in the study (Paper A), those using the nicotine inhaler are around 3.4 times more likely to have a sustained reduction in smoking at 24 months, and we can be 95 per cent sure that the true effect (in the population) lies between 1.4 and 8.3 times. This interval does not include 1.0, which is the value for no effect. This is also consistent with a significant p -value result reported for the hypothesis test ($p < 0.05$).
2. An appropriate hypothesis test for comparing mean values of breath CO, a continuous variable, would be the independent samples t -test, so long as the assumptions are met. If the distributions were very skewed (if you look at the means, standard deviations, and ranges in Table 7.3.2, there is evidence of positive skewing), you could transform the data and use the t -test, or you could use a non-parametric test such as the Mann–Whitney U test. Transformation of data and non-parametric tests, such as the Mann–Whitney U test, are described in Chapter 11.

Exercise 7.3.4

1. This seems reasonable. Although we have identified and discussed some concerns, (for example whether the groups were really blind to whether or not their inhaler contained nicotine), the validation of reported smoking reductions with breath CO does add weight to the findings. It is always possible that subjects deliberately reduced their smoking for several days prior to their scheduled visit and CO test, but it seems unlikely that the nicotine group would be so much more aware of this that they would do so more consistently than the control group. We might, however, point out that the numbers of subjects with sustained reduction at 24 months was somewhat disappointing, with just 19 intervention subjects and six control subjects.
2. This is quite difficult to answer. It is not helped by knowing so little about how representative this sample is of all 'healthy smokers in Basle and Lausanne who are unable or unwilling to give up, but are interested in reducing'. Then we are faced with the question, if the intervention works in these smokers, would it work as well in other parts of Switzerland and in other countries? Perhaps the first step would be to look for other evidence, but one might reasonably assume that the response of adult smokers in Basle and Lausanne to this inhaler is probably quite typical of how other people living in Western Europe (at least) would respond.

Exercise 7.3.5

1. No, as these are two independent groups (whisky versus none).
2. Yes, since these are the same volunteers before and after.

Section 7.4

Exercise 7.4.1

1. Eligibility for inclusion and exclusion criteria.

Eligibility for inclusion	Exclusion
<ul style="list-style-type: none"> • City of Whitehorse, Melbourne. • 70 years and over. • Living in own home or leasing similar accommodation and allowed to make modifications. 	<ul style="list-style-type: none"> • Did not expect to remain in the area for 2 years (except for short absences). • Participated in regular to moderate physical activity with a balance improvement component in the previous 2 months. • Could not walk 10–20 metres without rest, help, or having angina. • Severe respiratory or cardiac disease. • Psychiatric illness prohibiting participation. • Dysphasia (difficulty with speech). • Recent home modifications. • Education and language adjusted score of >4 on the short portable mental status questionnaire. • Did not have approval of their general practitioner

- The sample was drawn from the electoral roll, and Figure 7.4.1 (Figure 1 in Paper B) provides numbers progressing through each stage. Response is difficult to judge as (the authors point out) they do not know how many of the total group contacted would have been eligible. Details are given in the text (recruitment) of how the study group compared with the general population (based on national census and health survey). Towards the end of page 133 in Paper B, the authors comment on the applicability of the findings, saying that these would be most relevant to ‘older adults living at home with similar characteristics – namely Australian born, aged 70–84 and rating their health as good to excellent’.
- The intervention and control treatments are as follows:

Intervention groups	Control
Offered as single interventions or in various combinations: <ul style="list-style-type: none"> • Strength and balance exercise programme, 1 hour/week for 15 weeks plus home exercises. • Home hazard identification and removal. • Vision assessment and referral for treatment as required. 	<ul style="list-style-type: none"> • No intervention (deferred until the end of the study).

- All of these interventions are, or could be, made available through routine services. So, depending on which turned out to be most effective, there is no reason why the interventions could not be made widely available, subject to costs and resource availability. One aspect that may be quite challenging to reproduce is the co-ordination of the interventions, particularly if the results show that two or more interventions should be combined for enhanced effectiveness.
- The comparison groups are of two types in this study:
 - Other types of interventions (and combinations of interventions), none of which have been shown to be more effective than others used in the study, either singly or in combination.
 - No intervention at all, although in effect these are deferred until the end of the study, which lasted 18 months. The introduction section states that there is evidence that a number of interventions (including those in the study) are effective, so it may be considered surprising that the ethics committee approved this ‘no intervention’ control group. We would need to know more about the case made to the committee to judge this further.

6. Blinding to the intervention:

Group	Blinding
Subjects	It was not possible to blind the subjects. Although we do not know exactly how aware they were of other options, it should be assumed that they would have a good general idea from the informed consent. The main outcome was self-reported by the subjects, although it was followed up by a blinded research assistant.
Research team	The team would be aware of the allocation, but some blinding was designed into the study. The assessor who carried out the initial home visits was blind to the allocation, as was the assessor who carried out risk factor assessments on a proportion of subjects ($n = 442$) at 18 months. Follow-up of self-reported falls and of subjects not returning the record card at the end of each month was carried out over the telephone by a blinded research assistant.
Health-care staff	Staff delivering the interventions could not have been blinded to the allocation. GPs (and presumably other primary care staff) were informed of the study (approval was sought from the GP) and probably would have known the allocation. Whether this knowledge would have affected their management in respect of falls is not known. This type of uncertainty is inevitable in a non-blinded (open) study, but there is really little alternative with this type of intervention.

Exercise 7.4.2. Compliance with interventions (Paper B)

Summary of compliance:

Group	Allocated	Initiated	Compliance
Exercise	541	401	401 (74%) out of 541 started sessions, and 328 (82%) of the 401 attended >50% sessions. Mean (SD) number of sessions = 10 (3.8).
Home hazard	543	478	Of the 478 advised, 363 (76%) received help to do modifications.
Vision	547	287	186 of 287 (65%) had recently visited or were about to visit for eye care; 97 of 101 others (96%) took up referral.

Some additional information is to be found in the discussion section; for example, with exercise (described as having achieved 'relatively poor compliance'), sessions were intended to be daily, but in fact they were performed twice weekly on average.

Commentary:

- Exercise: Around three quarters started the sessions, and 328 (61 per cent) of the total allocated attended 50 per cent of the sessions. Put another way, 40 per cent either did not attend any (26 per cent) or attended less than half of the sessions (14 per cent).
- Home hazard: Around three quarters of those thought to need modifications had the work done.
- Vision: Almost all of those needing a new referral took up this advice.

Overall, given the types of intervention and the fact that these were people living independently at home (as opposed to in an institutional setting), compliance was fairly good. However, the intention-to-treat analysis has to absorb the fact that around 40 per cent of the exercise groups

did no exercise sessions or did less than half of the sessions, and one quarter of those recommended to make changes in the home hazard group did not do so. Although virtually all of those needing attention to their vision took up this advice, the intervention only applied to 97 people (18 per cent of the total allocated), as all others were either not in need of eye care or had already arranged this.

Exercise 7.4.3 Reassessment at 18 months

1. A total of 442 out of 1,090 (40.6 per cent) were reassessed at 18 months. Figure 7.4.1 (Figure 1 from Paper B) shows the numbers in each of the intervention groups who were reassessed (from which you can calculate percentages).
2. The reasons given are that resources were not available to reassess the whole group and this assessment was of secondary importance to the study's main goal; this was falls, measured by a calendar completed by all subjects continuing in the study.
3. They were randomly selected by an assessor blind to the intervention group.
4. The percentage of each group reassessed can be determined from Figure 7.4.1. The baseline characteristics are shown in Table 7.4.2 (Tables 2 and 3 from Paper B).
5. Comparison of reassessed subject characteristics (column 4 in Table 7.4.2 [Table 2 and 3 from Paper B]) with those for all participants (column 1) shows that the reassessed individuals were similar in all respects studied.

Exercise 7.4.4

1. For (a) exercise: 18 per cent reduction in risk (95% CI, 3 to 30 per cent reduction), $p = 0.02$, which is statistically significant. For (b) vision: 11 per cent reduction in risk (95% CI, 25 per cent reduction to 4 per cent increase), $p = 0.13$, which is non-significant and consistent with a 95% CI that includes 1.0.
2. Exercise and vision in combination: 27 per cent reduction (95% CI, 9 to 42 per cent reduction), $p = 0.01$. This shows, more or less, that the effect of the two interventions in combination is roughly that of the individual effects added together (additive). This implies there is no *interaction*, whereby the effect of one (or both) is enhanced by the presence of the other. This is reported in the first paragraph of the discussion.
3. The NNT has been calculated by the per cent estimated reduction in the annual fall rate, which for the three interventions combined is 14 per cent (or 0.14 as a proportion). The NNT therefore is $100/14$ (or equivalently $1/0.14$) = 7. This means that seven people need to receive the intervention for 1 year in order to prevent one fall over the same time period. Note that this is based on the central estimate of 14 per cent and does not take account of the 95% CI (which can be done).

In interpreting the effectiveness of the interventions, it is important to remember that intention-to-treat analysis was used, and compliance was not perfect. Hence, a NNT of 7, which allows for the level of compliance with interventions, may be considered as quite favourable.

Exercise 7.4.5

1. Although the intervention was initiated on an individual basis mainly delivered by the GP through the consultation, there was some wider involvement of practice staff who were, for example, encouraged to give feedback. If within-practice individual randomisation had been used, there would have been a risk of contamination between intervention and control subjects (see discussion section, page 798 of Paper C). In addition, GPs (and other staff) would have found it extremely difficult to manage intervention and control patients strictly to protocol and to avoid bias.

- Control clusters received 'usual care', as described in the Abstract. No further information is given on details of this, but they were offered the intervention at the end of the study.

Exercise 7.4.6

- There were 16 communities: eight intervention and eight control.
- Populations ranged from 1,700 to 13,500.
- The communities shared no common high schools and were at least 20 miles apart.
- Pairs of communities were matched on socioeconomic status and population size, and randomisation was carried out within each pair.
- It was reported that communities 'in the two conditions' (this means intervention and controls) did not differ statistically in respect of a range of demographic, social, and economic factors, implying that a good balance had been achieved.

Section 7.5

Exercise 7.5.1

- Columns 2 and 3 are unadjusted for clusters; column 4 is adjusted for clusters, with the p -value for the adjusted result in column 5.
- The value for the intervention group is 54.6 (41.4 to 68.4) minutes per week; the value for the control group is 16.8 (6.0 to 32.4) minutes per week.
- The difference (intervention effects) is 37.8 minutes per week. We can see that the confidence intervals are in fact very far apart: The lower limit for the control group mean (41.4) is considerably higher than the upper limit for the intervention group mean (32.4), implying a highly statistically significant difference.
- The intervention effect for this same outcome, following adjustment for clusters, is a difference of 33.6 minutes per week (95% CI: 2.4 to 64.2), $p = 0.04$. The intervention effect adjusted for clusters is a little smaller than the unadjusted difference between intervention and control groups, but the p -value only just reaches statistical significance. The adjustment for clustering has reduced the precision of the effect estimate, as expected.

Section 7.6

Exercise 7.6.1

- With $(\mu_1 - \mu_2) = 1.5$, $\sigma = 4.5$, significance of 5%, and power of 80%, the output from OpenEpi (see below) shows that we need $n = 142$ per group, total 284, before allowing for non-response, dropouts, etc., Table 7.8.1(a).
- The total study sample size of 914 (before allowing for non-response, dropouts, etc.) is more than adequate to detect this difference.
- Decreasing the treatment effect to 0.75 cm results in a sample size requirement of 566 per group, Table 7.8.1(b). Thus, halving the size of the difference to be detected leads to an approximate fourfold increase in sample size.

Table 7.8.1(a) OpenEpi output for detecting a difference between means of 1.5 cm.

Sample Size For Comparing Two Means			
Input Data			
Confidence Interval (2-sided)	95%		
Power	80%		
Ratio of sample size (Group 2/Group 1)	1		
	Group 1	Group 2	Difference*
Mean	13.3	11.8	1.5
Standard deviation	4.5	4.5	
Variance	20.25	20.25	
Sample size of Group 1	142		
Sample size Group 2	142		
Total sample size	284		

*Difference between the means.

Results from OpenEpi, Version 3, open source calculator--SSMean.

Print from the browser with ctrl-P or select text to copy and paste to other programs.

Table 7.8.1(b) OpenEpi output for detecting a difference between means of 0.75 cm, with other parameters unchanged.

Sample Size For Comparing Two Means			
Input Data			
Confidence Interval (2-sided)	95%		
Power	80%		
Ratio of sample size (Group 2/Group 1)	1		
	Group 1	Group 2	Difference*
Mean	13.3	12.55	0.75
Standard deviation	4.5	4.5	
Variance	20.25	20.25	
Sample size of Group 1	566		
Sample size of Group 2	566		
Total sample size	1132		

*Difference between the means.

Results from OpenEpi, Version 3, open source calculator--SSMean.

Print from the browser with ctrl-P or select text to copy and paste to other programs.

8

Life Tables, Survival Analysis, and Cox Regression

Introduction and Learning Objectives

In the work we have done so far on cohort studies and trials, we have used incidence rates as the measure for outcomes, that is, the number of people dying (mortality) or becoming ill (morbidity) in a given time period. The simplest concept was *cumulative incidence*:

$$\frac{\text{Number of new cases arising from a defined population in a specified time period}}{\text{Number in defined at-risk population over the same time period}}$$

We also saw that a more-useful measure was *incidence density*, for which we used person-time (months, years, etc.) in the denominator.

With *survival analysis* we introduce a new but intuitively simple and important dimension. This is the time that people in the study survive until dying (mortality), falling ill (morbidity), or experiencing any other event of interest. Apart from that, what we are covering in this chapter builds on concepts we have already discussed.

We start by looking at the use of the Kaplan–Meier method (*Kaplan–Meier survival curves*) to study survival in cohort studies and trials where we are following up specific groups of people over time. We then discuss *Cox proportionate hazards regression*. This sounds complicated, but it is in fact just another form of regression analysis and is used for the purposes we have already covered in previous chapters (adjustment for confounding, prediction, etc.). However, in this new situation we have the added ingredient of ‘time until the event of interest’. Finally, we introduce the purpose and method of *current life tables*, which allow the calculation of life expectancy for a population using age-specific death rates.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe how survival analysis differs from comparisons of cumulative incidence or incidence density.
- Describe the circumstances when it is appropriate to use Kaplan–Meier survival curves and interpret examples of these, including using the curve to determine median survival and the proportion of a population surviving to a given time.
- Describe the use of the log-rank test for hypothesis testing with survival data, and interpret the results of the test.

- Describe the circumstances when it is appropriate to use Cox proportional hazards regression, determine that the assumption of proportional hazards has been met, and interpret results from such an analysis.
- Describe the purposes and principles of current life table analysis.

Resource Papers

Two papers have been selected for this chapter. The first (Paper A) investigates the risk of coronary heart disease (CHD) and stroke associated with passive smoking in the British Regional Heart Study, a cohort study that we introduced in Chapter 5. The second (Paper B) is a randomised control trial (RCT) comparing radiotherapy alone with radiotherapy plus drug treatment (cetuximab) for advanced head and neck cancer.

Paper A

Whincup, P.H., Gilg, J.A., Emberson, J.R., Jarvis, M.J., Feyerabend, C.F., *et al.* (2004). Passive smoking and risk of coronary heart disease and stroke: prospective study with cotinine measurement. *Br Med J* **329**, 200–205.

Paper B

Bonner, J.A., Harari, P.M., Giralt, J., Cohen, R.B., Jones, C.U., *et al.* (2010). Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival. *Lancet Oncol* **11**, 21–28.

8.1 Survival Analysis

8.1.1 Introduction

We begin by looking at the methods used for analysing data, typically derived from cohort studies or trials, that incorporates information on time until the event of interest for each person in the study. This is called *survival analysis*.

Thus, if we are looking at death as the event of interest, the analysis uses information on whether or not a person died within the study period and on how long the person survived from the time he or she entered the study.

We can apply survival analysis to many situations as well as survival time until death. For example, in a study of the duration of breastfeeding, we may record the length of time to completion of weaning. Completion of weaning is the *terminal event or end point*. We refer to any time-to-event situation by the common terminology of *survival time*. Thus, the event of interest could be recovery from disease, relapse, leaving intensive care, introduction of bottle-feeding, conception following fertility treatment, or any other discrete event. When we refer to ‘survival time’ we mean ‘time to the event of interest’, whatever that event may be.

8.1.2 Why Do We Need Survival Analysis?

In the analysis of cohort studies and trials, we are usually interested in one or more of the following questions:

- Describing the survival experience of a group of individuals; for example, what proportion of cancer patients are alive, or free of recurrent disease, 5 years after diagnosis and treatment?

- Comparing the survival experience of two or more groups of individuals; for example, is there a difference between the proportions of individuals taking treatments A and B who are still alive after 5 years? Or, alternatively, what are the average survival times for patients taking treatments A and B?
- Predicting the length of survival for an individual given a number of characteristics or prognostic factors; for example, what is the predicted time to recurrence of gallstones after dissolution (dissolving through treatment) given age, sex, history of previous gallstones, and other clinical information?

We already know some methods of estimating and comparing proportions and means, such as **confidence intervals (CIs)** and **hypothesis tests**. We have also seen how a **regression model** can be used for prediction. But the methods we have already met cannot usually be used to answer our questions about survival time. This is because data on survival time differ from the types of data we have studied so far in two important respects:

- Survival times are hardly ever normally distributed; instead, they are usually quite markedly skewed, as illustrated in Figure 8.1.1. These data, for which the mean is 67.8 weeks and the median 61.0, have been generated for this example and are presented in more detail in Table 8.1.1.
- Only some of the individuals studied actually experience the event of interest during the study period. The rest have what are known as **censored** observations. This is a critical element of survival analysis and is discussed next.

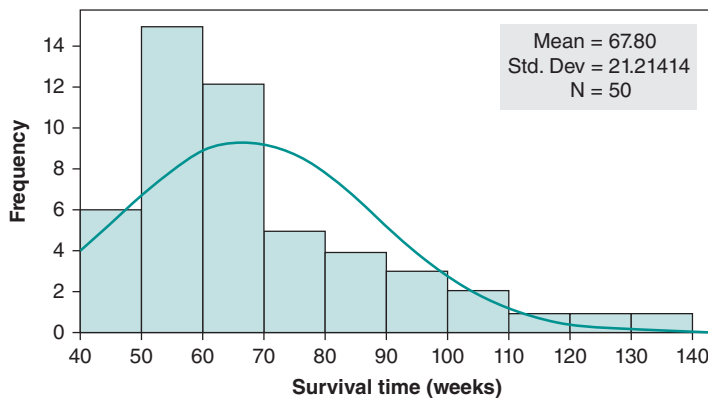


Figure 8.1.1 Survival data are usually positively skewed.

8.1.3 Censoring

In a cohort study we may observe a survival time of 5 years for each of two individuals. If one person died 5 years after entering the study and the other was still alive when the study ended 5 years after joining it, then these are clearly two different types of information. The fact that a person remained alive for the full 5 years is important, and we need a way of dealing with this difference in the analysis. This is called **censoring**. In this example, the time for the second person is censored; that is, the period of observation ended before the event (death) occurred. Figure 8.1.2 shows several features of survival data that are typically encountered in survival analysis.

As shown in Figure 8.1.2, it is often the case that recruitment of study participants does not always occur at the same time and they enter the study at different time periods. Some study

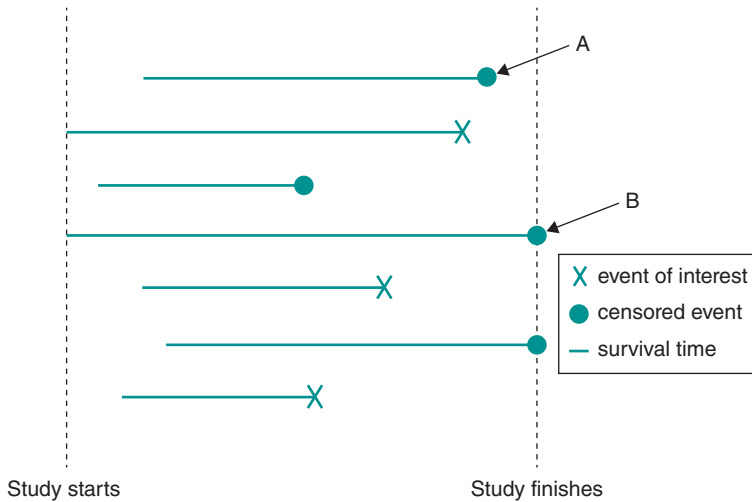


Figure 8.1.2 Illustration of survival data for a cohort study; see text for further explanation.

participants do not experience the event of interest, either dropping out of the study and becoming lost to follow-up or reaching the end of the study event free; these individuals provide censored data (●). Other study participants have the event of interest (X). All study participants provide event-free survival time (—).

Censoring involving individuals who become lost to follow-up or who withdraw from the study (A in Figure 8.1.2) or who reach the end of the study period without having the event (B in Figure 8.1.2), is termed **right censoring** because it occurs after (to the right on graphs of survival over time) recruitment and follow-up started, and it is the most common form of censoring.

Left censoring occurs when individuals have the event of interest before being enrolled in a study and are therefore lost from the cohort. This form of censoring is very rarely encountered.

In a set of survival data it is usual to indicate in some way which of the survival times are censored. Commonly, this is done by using an indicator variable (that is, a variable that can have the value 0 or 1 for each individual) called a **censoring indicator**. This indicator has the value 0 for each individual with a censored survival time and the value 1 for those whose survival times are not censored. There is an example of this in Section 8.1.4.

In the following sections we first look at a graphical display of survival data using the **Kaplan–Meier survival curve**, and then we consider the hypothesis test used to compare the survival of two or more groups (**log-rank test**) and lastly the predictive model used for survival data (**Cox regression**).

Summary – Survival Data

- We use survival analysis when we have data showing the times to occurrence of an event.
- Survival data are usually positively skewed, often highly skewed.
- Usually some observations are censored, and this information needs to be taken into account in the analysis.
- Censoring arises when an individual has not experienced the event of interest by the end of the study (e.g. is still alive), or the event had not occurred when the individual was last observed, but records stop before the end of the study (e.g. lost to follow-up).

8.1.4 Kaplan–Meier Survival Curves

We now start to look at how survival analysis is applied in practice through the example of the cohort study investigating passive smoking and heart disease (Paper A). Please now read the following excerpts, which cover the abstract and the introduction and methods. These excerpts do not include a description of the statistical methods, because we need to cover more on survival analysis in order for you to fully understand these.

Abstract

Objective

To examine the associations between a biomarker of overall passive exposure to tobacco smoke (serum cotinine concentration) and risk of coronary heart disease and stroke.

Design

Prospective population based study in general practice (the British Regional Heart Study).

Participants

4729 men in 18 towns who provided baseline blood samples (for cotinine assay) and a detailed smoking history in 1978–80.

Main Outcome Measure

Major coronary heart disease and stroke events (fatal and non-fatal) during 20 years of follow up.

Results

2105 men who said they did not smoke and who had cotinine concentrations <14.1 ng/ml were divided into four equal sized groups on the basis of cotinine concentrations. Relative hazards (95% confidence intervals) for coronary heart disease in the second (0.8–1.4 ng/ml), third (1.5–2.7 ng/ml), and fourth (2.8–14.0 ng/ml) quarters of cotinine concentration compared with the first ≤ 0.7 ng/ml were 1.45 (1.01 to 2.08), 1.49 (1.03 to 2.14), and 1.57 (1.08 to 2.28), respectively, after adjustment for established risk factors for coronary heart disease. Hazard ratios (for cotinine 0.8–14.0 v ≤ 0.7 ng/ml were particularly increased during the first (3.73, 1.32 to 10.58) and second five year follow up periods (1.95, 1.09 to 3.48) compared with later periods. There was no consistent association between cotinine concentration and risk of stroke.

Conclusion

Studies based on reports of smoking in a partner alone seem to underestimate the risks of exposure to passive smoking. Further prospective studies relating biomarkers of passive smoking to risk of coronary heart disease are needed.

Introduction

Active cigarette smoking is a well established major preventable risk factor for coronary heart disease (CHD). Many studies have reported that passive smoking is also associated with increased risk of CHD. Generally such studies have compared the risks of non-smokers who do or do not live with cigarette smokers, though a few have also considered occupational exposure.

Meta-analyses of case-control and cohort studies examining the effect of living with a cigarette smoker on risk among non-smokers have generally shown an overall increase in risk of about one quarter, after adjustment for potential confounding factors and with little evidence of publication bias. Passive smoking may also be related to risk of stroke. Although living with someone who smokes is an important component of exposure to passive smoking, it accounts for less than half of the variation in cotinine concentration among nonsmokers and does not take account of additional exposure in workplaces and in public places (particularly pubs and restaurants). Biomarkers of passive exposure to smoking, particularly cotinine (a nicotine metabolite), can provide a summary measure of exposure from all these sources. Although cotinine concentration in non-smokers has been related to prevalent CHD, there are no published reports of the prospective associations between serum cotinine concentration and risk of CHD and stroke in non-smokers. We have examined these associations in the British Regional Heart Study, a prospective study of cardiovascular disease in middle aged men, using retained baseline samples for retrospective measurement of cotinine.

Methods

The British Regional Heart Study is a prospective study of cardiovascular disease in 7735 men aged 40–59 years selected from the age and sex registers of one general practice in each of 24 towns in England, Wales, and Scotland (78% response rate).

Baseline Assessment

In 1978–80, research nurses administered a questionnaire on present and previous smoking habits (cigarettes, cigar, pipe), alcohol intake, physical activity, and medical history (including angina, myocardial infarction, stroke, and diabetes diagnosed by a doctor). Participants also completed a questionnaire on chest pain (Rose, World Health Organization). Two seated blood pressure measurements were taken with a London School of Hygiene and Tropical Medicine sphygmomanometer; the mean was adjusted for observer variation within each town. Non-fasting total serum cholesterol concentration was measured with a modified Liebermann-Burchard method on a Technicon SMA 12/60 analyser. High density lipoprotein (HDL) cholesterol was measured by the same procedure after precipitation with magnesium phosphotungstate. Serum samples were placed in long term storage at -20°C in the last 18 study towns. In 2001–2, these were thawed and cotinine concentration was measured with a gas–liquid chromatography method (detection limit 0.1 ng/ml).

Follow Up

All men were followed up for all cause mortality and cardiovascular morbidity. We collected information on deaths through the established ‘flagging’ procedures provided by the NHS central registers. We obtained information on non-fatal CHD events and strokes from general practitioners’ reports, supplemented by reviewing patients’ records every two years throughout follow up. Major CHD events included deaths with coronary heart disease as the underlying cause, including sudden death of presumed cardiac origin (international classification of diseases, ninth revision (ICD-9), codes 410–414) and non-fatal myocardial infarction, diagnosed in accordance with standard WHO criteria. Stroke events included deaths with cerebrovascular disease as the underlying cause (ICD-9 codes 430–438) and non-fatal stroke diagnosed in accordance with WHO criteria. The analyses presented are based on all first major CHD or stroke events during the follow up period to December 2000, with an average follow up of 18.5 years for men who had no myocardial infarction or stroke (range 0.2–20.0 years).

Definition of Baseline Smoking Status

Men were classified as 'current non-smokers' at baseline if they reported that they did not smoke cigarettes, cigars, or a pipe and had a serum cotinine concentration <14.1 ng/ml. Among these men, 'lifelong non-smokers' were those who reported never having smoked cigarettes, cigars, or a pipe. For comparison purposes, 'light active smokers' were men who reported smoking 1–9 cigarettes a day, irrespective of cotinine concentration.

Other Definitions

Pre-existing CHD included one or more of angina or possible myocardial infarction, or both, on the WHO Rose questionnaire; electrocardiographic evidence of definite or possible myocardial infarction or ischaemia; or participant's recall of myocardial infarction or angina diagnosed by a doctor. Study towns were considered to be in the south if they were to the south or east of a line joining the River Severn and the Wash. Physical activity and alcohol intake were categorised as in earlier reports.



Self-Assessment Exercise 8.1.1

1. You should be familiar with the methods of this cohort study. To make sure you are clear about the subjects being studied in this report, make brief notes on who was included.
2. What outcomes were studied?
3. How was passive smoking exposure determined? Comment on the likely validity of this assessment.
4. Can you see why survival analysis would be appropriate to this investigation? Does the research team have the necessary information for carrying out survival analysis?

Answers in Section 8.4

8.1.5 Kaplan–Meier Survival Curves

One of the most important methods in dealing with survival data from one or more groups is the graphical display, known as the *Kaplan–Meier* survival curve. To compare two or more groups, we show each group's survival curve on the same graph. We can then use the graph (or, more accurately, the calculations that produced it) to estimate the proportion of the population of such people who would survive a given length of time in the same circumstances.

Summary statistics such as the *median survival time*, or the proportion of individuals alive at a specified time (say, 1 month, or 1 year), can be read directly from the Kaplan–Meier graph. We look at an example and show you how to do this shortly. Confidence intervals should be calculated for these summary statistics. If there are no censored values we can use standard methods for calculating CIs for a population proportion, which is given by

$$p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$

where p is the proportion observed in the sample and n is the sample size. Generally, however, we need to use a modified formula to allow for the censoring.

We will not go into the detail of how to calculate the Kaplan–Meier survival curve here, which in any case is easily done on a computer. Essentially, however, for each unit of survival time, we estimate the proportion of the sample surviving for that period of those who have not experienced the outcome event during the immediately preceding time period. This results in the survival curve being presented as a *step function*; that is, the proportion surviving remains unchanged between events, even if there are some intermediate censored observations. The times of these censored observations are usually indicated by tick marks on the curve to show when they occurred. The median survival time is the time corresponding to a survival probability of 0.5; of course, if the curve does not fall below 0.5, we cannot estimate the median, but this is often possible and is a useful property of the graph. The following example will help you to understand these key elements of the Kaplan–Meier method.

Example

Table 8.1.1 shows times (in months) before recurrence for two groups of cancer patients in a trial comparing a new treatment against the existing standard treatment. These data are not real but have been created to illustrate key points for this section and are typical of such a trial. In this example, we are interested in *recurrence-free* survival; that is, the time that each person remains well, without recurrence of the cancer.

In Table 8.1.1, the *censoring indicator* shows whether the times are censored or not. A code of 0 means that the time observation is censored, and a code of 1 means that the event (recurrence) has occurred. The Kaplan–Meier survival curves for the two groups are displayed in Figure 8.1.3 (calculated with SPSS) and show that the treatment group has a higher recurrence-free survival rate than the control group.

In Figure 8.1.3, the cumulative survival (*y*-axis) gives the probability of surviving (not having a recurrence of the cancer) at any given time after treatment. The median survival is given by a probability of 0.5, but any other probability can also be read off the graph. Similarly, the probability of surviving until a specified time can be determined. For example, the probability of controls surviving to 10 months is (approximately) 0.65. Check that you can read this off the graph. The following exercise will help consolidate your understanding of this technique.



Self-Assessment Exercise 8.1.2

1. In the example data in Table 8.1.1, how many individuals in the intervention (new treatment) and control groups were censored? What are the possible reasons for this difference?
2. Use Figure 8.1.2 to estimate the median recurrence-free times for the two groups.
3. What proportions of controls and those treated with the new treatment are estimated to survive (be free from recurrence of their cancer) to 30 months?

Answers in Section 8.4

8.1.6 The Log-Rank Test

The log-rank test is a hypothesis test used to compare survival in two or more groups, such as in the cancer trial we have been discussing. This is the preferred test as it uses all of the available survival information, not just information at one point, such as when comparing the median

Table 8.1.1 Survival times and censoring information for intervention and control groups.

Intervention (new treatment)			Control (standard treatment)		
Subject ID	Survival (months)	Censoring indicator*	Subject ID	Survival (months)	Censoring indicator*
1	3.00	1	26	2.00	1
2	6.00	1	27	3.00	0
3	6.00	1	28	4.00	1
4	6.00	0	29	5.00	1
5	7.00	1	30	5.00	1
6	9.00	0	31	6.00	1
7	10.00	1	32	7.00	0
8	10.00	0	33	8.00	1
9	11.00	0	34	8.00	1
10	13.00	1	35	9.00	1
11	16.00	1	36	11.00	1
12	17.00	0	37	11.00	1
13	19.00	0	38	12.00	0
14	20.00	0	39	13.00	1
15	22.00	1	40	15.00	1
16	23.00	1	41	17.00	0
17	25.00	0	42	19.00	1
18	32.00	0	43	21.00	1
19	32.00	1	44	21.00	1
20	34.00	0	45	23.00	1
21	35.00	0	46	25.00	0
22	36.00	1	47	26.00	1
23	42.00	0	48	29.00	1
24	45.00	0	49	31.00	0
25	51.00	0	50	33.00	1

*Censoring indicator: 0 = censored; 1 = event (recurrence of cancer).

survival times or the proportions surviving at a given time. This test is available in standard computer programmes, and our discussion here is confined to looking at how it uses all of the available information.

The **log-rank test** is a form of chi-squared test and, as it is **nonparametric**, it does not require assumptions to be made about the distribution of survival times (nonparametric hypothesis testing is described in more detail in Chapter 11). It is used to test the null hypothesis H_0 that there is no difference between the populations in the probability of an event (e.g. death) at any time point. The analysis is based on the times of events. For each such time, the observed number of deaths in each group is calculated and compared to the number expected if there were in reality no difference between the groups. The log-rank test, in effect, compares the total

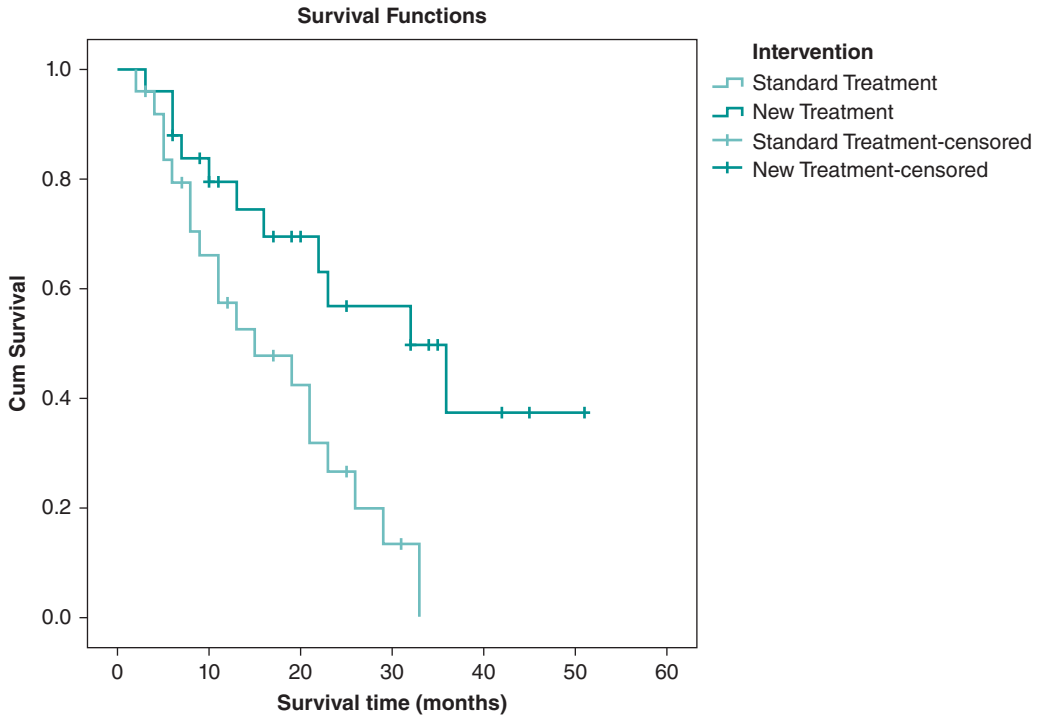


Figure 8.1.3 Kaplan–Meier survival curves for cancer trial example.

number of events in each group with the number of events expected if there was no treatment effect.



RS – Reference Section on Statistical Methods

The principle of the test is to divide the survival time scale into intervals according to the distinct observed times of death (or other event), but ignoring censored times. First, the survival times for all events are ranked in ascending order, combining data from both groups. For the purpose of this discussion, we call the people on the new treatment group A and those on the control treatment group B. Then, for each time period, we calculate the contribution from each group to the overall expected deaths for each group. These contributions to expected events at each time period are then summed to yield the total number of expected events for group A (notation is E_A). The total expected for group B (E_B) is also calculated, though this is simply the total number of observed events – E_A . This is how the log-rank test uses all of the available survival information.

The notation for the number of observed events in group A is O_A and for group B the notation is O_B . The log-rank test statistic has one degree of freedom and is calculated as

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}$$

The log-rank test can also be used to compare the survival of more than two groups.

8.1.7 Interpretation of the Kaplan–Meier Survival Curve

We have now covered enough on survival methods to start looking at what has been done in the passive smoking study, though we will skip the description of Cox regression until we have looked at this in Section 8.2.

Please now read the following excerpts (text, table, and figures) from Paper A and complete Exercise 8.1.3.

Statistical Methods

We used Cox proportional hazards models, stratified by town of residence, to examine the independent contribution of serum cotinine concentration to the risks of CHD and stroke. These produced relative hazards, adjusted for age and other risk factors, for each quarter of the distribution of serum cotinine concentration compared with the lowest. We carried out overall tests of the association, fitting the continuous relation between log cotinine concentration and risk of CHD. Relative hazard estimates for each five-year interval were calculated by fitting interaction terms with time (using three binary factors to separate the effects in the second, third, and fourth intervals from those in the first). Kaplan–Meier curves were used to display the differences in incidence of major CHD by cotinine exposure group; differences were assessed using the log rank test. All *P*-values were two sided.

We fitted age, body mass index, height, systolic blood pressure, diastolic blood pressure, serum total cholesterol, high density lipoprotein cholesterol, white cell count, lung function (forced expiratory volume in one second, FEV₁), and triglycerides as continuous variables. Physical activity was fitted as a factor with four levels (none, occasional, light, moderate or more), alcohol intake with three levels (none/occasional, light/moderate, heavy), and social class with seven levels (six registrar general categories and Armed Forces). History of cigarette smoking, pre-existing CHD, and diabetes were fitted as dichotomous variables.

Results

In the last 18 towns of the study, 5661 men took part (78% response rate). For 4729 of these we had detailed histories on smoking and blood samples for cotinine analysis. These men resembled the whole study population in reported smoking habits and risks of CHD and stroke. A total of 2158 men reported that they were current non-smokers, of whom 2105 (97.5%) had serum cotinine concentrations <14.1 ng/ml. Of these, 945 men were classified as lifelong non-smokers, the remaining 1160 as former smokers. The cotinine distributions of these two groups (fig 1)* were skewed, with a slightly higher geometric mean cotinine among former smokers (1.49 v 1.18 ng/ml). Few men in either group had cotinine concentrations close to the 14.1 ng/ml cut off.

Serum cotinine and cardiovascular risk factors—Among current non-smokers, cotinine concentrations were not consistently related to age, total cholesterol concentration, physical activity score, or prevalent CHD but showed graded positive associations with mean body mass index, systolic and diastolic blood pressure, high density lipoprotein cholesterol, white cell count, and triglycerides (weakly) and positive associations with the prevalence of former smoking, heavy drinking, and manual occupation (table 1).^{*} Cotinine concentrations were inversely associated with FEV₁, prevalence of low alcohol intake, and residence in southern England. These associations were generally little affected when we excluded former smokers. Light active smokers had lower mean body mass index, diastolic blood pressure, and FEV₁ and a higher mean white cell count than men who did not smoke.

*Figure 1 is reproduced in Figure 8.1.4 and Table 1 is reproduced in Table 8.1.2.

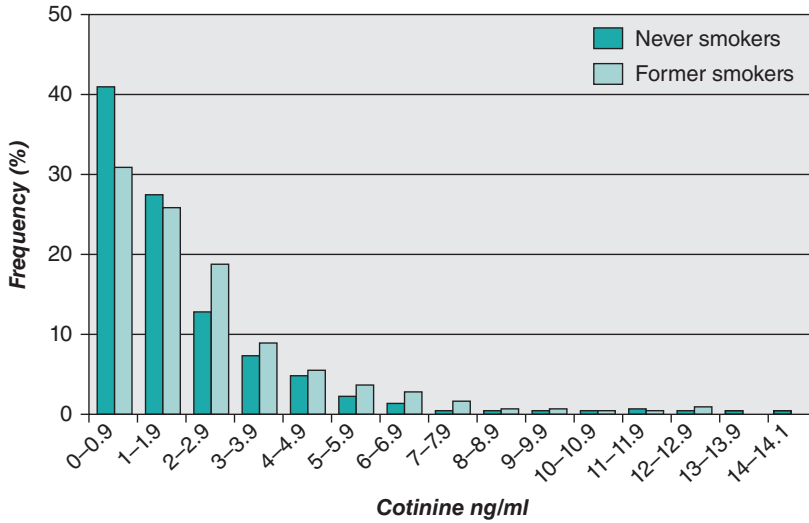


Figure 8.1.4 Distribution of serum cotinine concentrations among current non-smokers; lifelong non-smokers and former smokers are shown separately (Figure 1 from Paper A). *Source:* Whincup 2004. Reproduced with permission of BMJ Publishing Group Ltd.

Table 8.1.2 Means (SDs) and numbers (percentages) of cardiovascular risk factors by cotinine concentration: non-smokers and light active cigarette smokers (Table 1 from Paper A).

Risk factor	Passive smoke exposure (ng/ml cotinine)				Active smokers (1-9/day)	P value for trend*
	≤0.7	0.8-1.4	1.5-2.7	2.8-14.0		
Mean cotinine (ng/ml)	0.5	1.1	2.0	4.9	138.4	-
No. of men	575	508	506	516	192	-
Age (years)	50.5 (5.7)	49.9 (5.7)	50.2 (5.8)	50.5 (6.1)	50.7 (5.7)	0.96
BMI (kg/m) ²	25.5 (2.9)	25.4 (3.1)	26.3 (3.1)	26.5 (3.4)	25.0 (3.4)	<0.001
Height (cm)	174.0 (6.6)	173.5 (6.6)	173.5 (6.6)	172.4 (6.5)	172.8 (6.1)	0.03
Systolic blood pressure (mm Hg)	144 (22)	145 (21)	147 (20)	151 (22)	144 (22)	<0.001
Diastolic blood pressure (mm Hg)	82 (14)	83 (13)	85 (14)	87 (15)	83 (14)	<0.001
Total cholesterol (mmol/l)	6.3 (1.0)	6.3 (1.0)	6.3 (1.0)	6.3 (1.0)	6.3 (1.0)	0.50
HDL cholesterol (mmol/l)	1.14 (0.25)	1.16 (0.27)	1.14 (0.26)	1.20 (0.26)	1.15 (0.25)	<0.001
White cell count (10 ⁹ /l)	6.4 (1.4)	6.5 (1.4)	6.6 (2.3)	6.7 (1.4)	7.2 (1.6)	0.02
FEV ₁ (ml)	357 (68)	355 (72)	346 (74)	329 (80)	329 (78)	<0.001
Triglycerides [†] (mmol/l)	1.65	1.61	1.77	1.78	1.74	0.04
Evidence of CHD	134 (23)	126 (25)	117 (23)	142 (28)	48 (25)	0.19
Diabetes (diagnosed by doctor)	8 (1)	8 (2)	7 (1)	5 (1)	4 (2)	-
Physical activity: none or occasional	182 (32)	169 (33)	163 (32)	208 (40)	70 (36)	0.25

Table 8.1.2 (Continued)

Risk factor	Passive smoke exposure (ng/ml cotinine)				Active smokers (1–9/day)	P value for trend*
	≤0.7	0.8–1.4	1.5–2.7	2.8–14.0		
Alcohol intake: never or occasional	261 (45)	189 (37)	145 (29)	104 (20)	52 (27)	<0.001
Alcohol intake: heavy (>6 drinks/day)	10 (2)	22 (4)	38 (8)	85 (16)	22 (11)	<0.001
Former smokers	267 (46)	259 (51)	309 (61)	325 (63)	–	<0.001
Manual workers	246 (43)	258 (51)	287 (57)	346 (67)	120 (63)	<0.001
Live in south	303 (53)	196 (39)	158 (31)	113 (22)	64 (33)	<0.001 [‡]

*Across passive smoking groups only. Adjusted for age and town.

[†]Geometric means as log transformed (use of transformed data and geometric means for skewed data are discussed fully in Chapter 11 and explored in Exercise 8.1.3).

[‡]Adjusted for age only.

Source: Whincup 2004. Reproduced with permission of BMJ Publishing Group Ltd.



Self-Assessment Exercise 8.1.3

1. Why are geometric means quoted for the comparison of cotinine levels in ex-smokers and never-smokers?
2. Why do you think ex-smokers had higher cotinine levels than lifelong never-smokers? (see text and Figure 8.1.4).

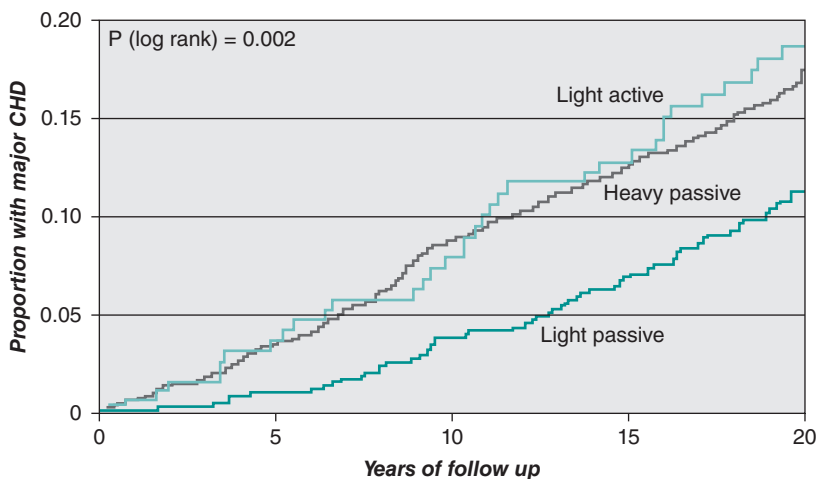


Figure 8.1.5 Proportion of men with major CHD by years of follow-up in each smoking group. 'Light passive' refers to lowest quarter of cotinine concentration among non-smokers (0–0.7 ng/ml), 'heavy passive' to upper three quarters of cotinine concentration combined (0.8–14.0 ng/ml), and 'light active' to men smoking 1–9 cigarettes a day (Figure 2 from Paper A). Source: Whincup 2004. Reproduced with permission of BMJ Publishing Group Ltd.

3. What type of graph is Figure 8.1.5?
4. What is the probability of having a CHD event after 10 years for light passive and heavy passive smoking? Can you determine the median probability of a CHD event?
5. Interpret the result of the log-rank test (result top left of Figure 8.1.5).
6. Why do the curves in Figure 8.1.5 start at a probability of 0.0, whereas those in our cancer trial example (Figure 8.1.3) start at a probability of 1.00?

Answers in Section 8.4

In the foregoing section we saw how survival analysis was used for a cohort study, as described in Paper A. The other common application of survival analysis is for intervention studies. We now briefly review Paper B, a randomised control trial carried out to test the benefit of adding an anti-cancer drug (cetuximab¹) to radiotherapy for head and neck cancer, to see how the method is applied in practice. Please now read the abstract, review Figures 8.1.6 and 8.1.7, and complete Exercise 8.1.4.

Abstract

Background

Previous results from our phase 3 randomised trial showed that adding cetuximab to primary radiotherapy increased overall survival in patients with locoregionally [restricted to a localized region of the body] advanced squamous-cell carcinoma of the head and neck (LASCCHN) at 3 years. Here we report the 5-year survival data, and investigate the relation between cetuximab induced rash and survival.

Methods

Patients with LASCCHN of the oropharynx, hypopharynx, or larynx with measurable disease were randomly allocated in a 1:1 ratio to receive either comprehensive head and neck radiotherapy alone for 6–7 weeks or radiotherapy plus weekly doses of cetuximab: 400 mg/m² initial dose, followed by seven weekly doses at 250 mg/m². Randomisation was done with an adaptive minimisation technique to balance assignments across stratification factors of Karnofsky performance score [a measure of functional impairment ranging from 0 to 100], T stage, N stage, and radiation fractionation. The trial was un-blinded. The primary endpoint was locoregional control, with a secondary endpoint of survival. Following discussions with the US Food and Drug Administration, the dataset was locked, except for queries to the sites about overall survival, before our previous report in 2006, so that an independent review could be done. Analyses were done on an intention-to-treat basis. Following completion of treatment, patients underwent physical examination and radiographic imaging every 4 months for 2 years, and then every 6 months thereafter. The trial is registered at www.ClinicalTrials.gov, number NCT00004227.

¹ Cetuximab is a monoclonal antibody that locks onto cancer receptors (epidermal growth factor receptors [EGFRs]) to inhibit cancer cells dividing and growing. It has been suggested that cetuximab might also make the cancer cells more sensitive to the effects of radiotherapy.

Findings

Patients were randomly assigned to receive radiotherapy with ($n = 211$) or without ($n = 213$) cetuximab, and all patients were followed for survival. Updated median overall survival for patients treated with cetuximab and radiotherapy was 49.0 months (95% CI 32.8–69.5) versus 29.3 months (20.6–41.4) in the radiotherapy-alone group (hazard ratio [HR] 0.73, 95% CI 0.56–0.95; $p = 0.018$). 5-year overall survival was 45.6% in the cetuximab-plus-radiotherapy group and 36.4% in the radiotherapy-alone group. Additionally, for the patients treated with cetuximab, overall survival was significantly improved in those who experienced an acneiform rash of at least grade 2 severity compared with patients with no rash or grade 1 rash (HR 0.49, 0.34–0.72; $p = 0.002$).

Interpretation

For patients with LASCCHN, cetuximab plus radiotherapy significantly improves overall survival at 5 years compared with radiotherapy alone, confirming cetuximab plus radiotherapy as an important treatment option in this group of patients. Cetuximab-treated patients with prominent cetuximab-induced rash (grade 2 or above) have better survival than patients with no or grade 1 rash.

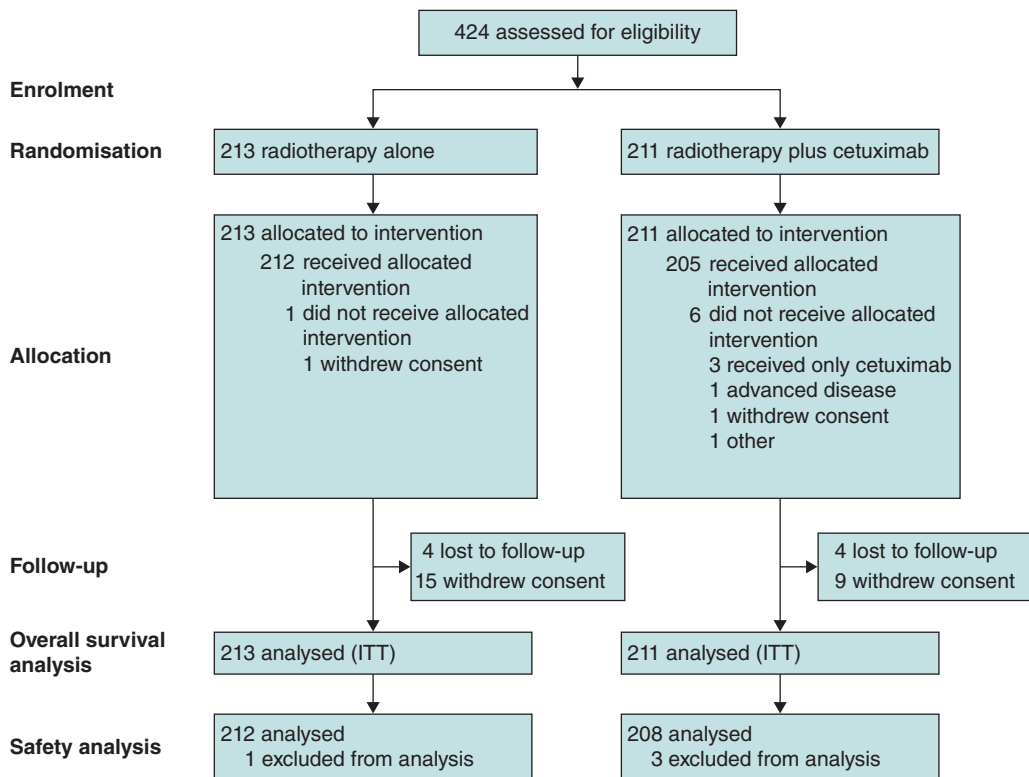


Figure 8.1.6 Trial profile (Figure 1 from Paper B). Source: Bonner 2010. Reproduced with permission of Elsevier.

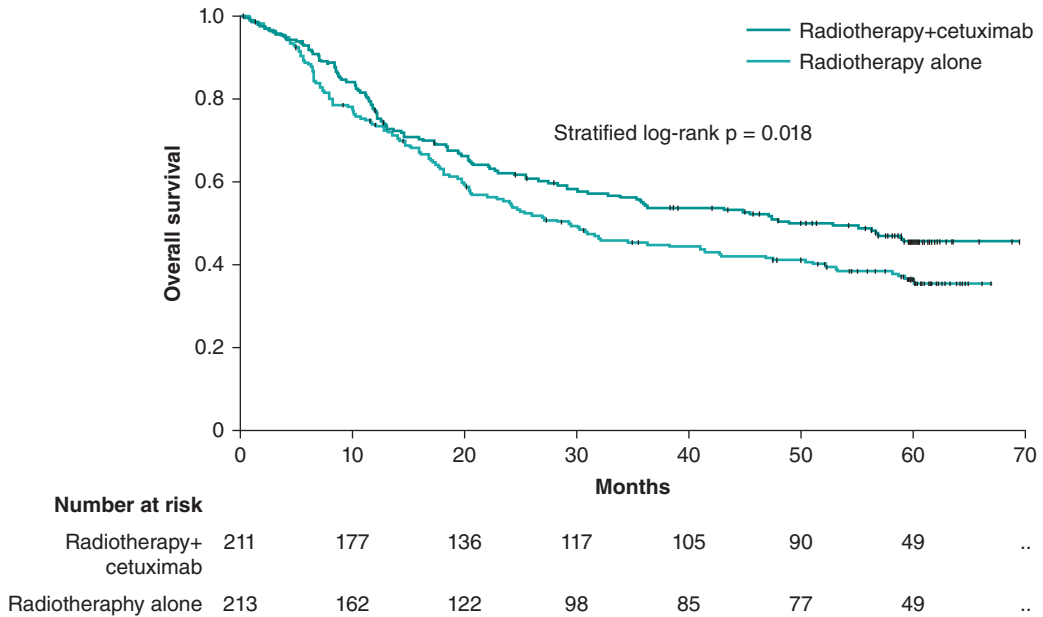


Figure 8.1.7 Percentage of patients surviving after date of randomization in control and intervention groups (Figure 2 from Paper B). *Source:* Bonner 2010. Reproduced with permission of Elsevier.



Self-Assessment Exercise 8.1.4

Based on the information provided in the Abstract and in Figures 8.1.6 and 8.1.7

1. List the criteria that defined the patients included in this study.
2. What were the primary and secondary events (outcomes) of interest?
3. From the Kaplan–Meier curve (Figure 8.1.7), what was the median ‘survival time’ and the probability of surviving to 60 months for the two groups? Do your answers agree with the figures for the two groups provided in the abstract?
4. The hazard ratio (see abstract and text in figure) is reported as 0.73, with a 95 per cent CI 0.56–0.95 and $p = 0.018$ for a log-rank test. Although you have not yet been introduced to hazard ratios (we do that in Section 8.2), you can regard this, for purposes of interpretation at this stage, as being similar to a relative risk. Have a go at interpreting this result.

Answers in Section 8.4

Summary: Survival Analysis

- We use survival analysis when we have data with times to occurrence of an event.
- These data require special treatment because
 - survival times are usually positively skewed, often highly so;
 - some observations are censored.
- We first picture the data as a Kaplan–Meier survival curve.

- Median survival and the proportion surviving at a given time can be read off the Kaplan–Meier survival curve. Although these are useful measures, they do not utilise all of the available data for comparison purposes.
- We use the log-rank test to test the hypothesis of no difference in survival between two or more groups; this test uses all of the available survival data.

8.2 Cox Regression

8.2.1 Introduction

We have seen how we can compare the average survival time (or proportions of individuals surviving to a given point in time) between groups using Kaplan–Meier methods and how to test the statistical significance of differences in survival times using the log-rank test. In addition we might want to calculate an overall estimate of the difference in survival times between groups. This is estimated by calculating the **hazard ratio** comparing the **hazard**, or probability, of events occurring in one group (e.g. treatment) as a ratio of the hazard of the events occurring in another group (e.g. control) at a particular time. Whilst the hazard ratio can be interpreted in a similar way to the risk ratio (relative risk), with which you are now familiar, the key distinction is the consideration of time. The hazard ratio gives an instantaneous estimate of risk at a particular time, whereas the relative risk gives a cumulative estimate of risk over a time span (the box below gives an example of this distinction).

Relative risk: At the end of the study, without considering time, the risk of dying with standard treatment is twice that compared with new treatment.

Hazard ratio: risk of dying with standard treatment is twice that compared with new treatment at any fixed point in time.

The hazard ratio, and its 95% confidence interval, can be calculated by using another type of regression analysis:– **Cox regression**.

We might also want to explore the effect of several variables on survival; for example, to compare the effects of two different treatments while allowing for the confounding effects of age and previous history. Indeed, you have seen from the statistical methods section of Paper A (British Regional Heart Study) that the research team intended to adjust for a large number of confounding variables. Thus, we want to be able to model how the risk of the event depends on a number of **explanatory variables**. As with the examples of multiple linear and logistic regression in Chapters 5 and 6, these explanatory variables may be a mixture of continuous variables (such as age) and categorical variables (such as smoking status). **Cox regression** is the regression method of choice for modelling the relationship between the risk of the event occurring at a particular time and a number of explanatory variables. It is also called **proportional hazards regression** – we will see why shortly. It is analogous to the multiple and logistic regression models we have met previously, but it is a regression model that takes into account time until the event occurs.

8.2.2 The Hazard Function

In survival analysis, the probability of an event occurring is called the **hazard**. This probability can vary with time, and the probability of individuals experiencing the event at a specific time

(which we call time t), given that they have survived up to that time, is denoted by $h(t)$ and is called the **hazard function**. Although the hazard may vary with time, the assumption in Cox regression (in calculating the hazard ratio between two groups) is that the hazard function in one group is a constant proportion of the hazard function in the other group. We look now at this assumption, and then briefly at the Cox regression model to help illustrate what it has in common with methods you are already familiar with.

8.2.3 Assumption of Proportional Hazards

In Cox regression the key assumption is known as **proportional hazards**: For any possible explanatory variable such as the type of treatment, we assume that if it affects the hazard, it does so by the same ratio at all times. Thus, although the hazard (risk of event occurring) on treatment A may vary over time, and similarly for B, it is assumed that the ratio of the hazards is the same at all times. So, if being on treatment B rather than A doubles the risk of dying at 1 week, it is assumed that it also doubles the risk of dying at 2 months, at 1 year, and so on.

8.2.4 The Cox Regression Model

Taking this example, we can write mathematically the statement (in Section 8.2.3) that the hazard ratio is 2 at all times as:

$$\frac{h_1(t)}{h_0(t)} = 2 \quad \text{for all times } t$$

where $h_0(t)$ is the hazard function for group A and $h_1(t)$ the hazard function for group B. The hazards are therefore in constant proportion. This is why the method is called **proportional hazards regression**. We are assuming that $h(t)/h_0(t)$ – this notation is used for the general form, explained below – depends only on the predictors, and not on time t , as it is constant over time.

$h(t)/h_0(t)$ is called the **hazard ratio**. It is the relative risk of an endpoint occurring at any given time and $h_0(t)$ is called the **baseline hazard function**. The **Cox regression model** generates the log hazard ratio:

$$\log \left(\frac{h(t)}{h_0(t)} \right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Where β_1 is the regression coefficient for the first explanatory variable x_1 and so on.

You can see the similarity between the Cox regression model and the multiple logistic regression model we studied earlier. Cox regression can be carried out in standard statistical software.

8.2.5 Checking the Assumption of Proportional Hazards

It is important that, before carrying out Cox regression analysis of survival data, we check whether the assumption of **proportional hazards** is reasonable. The Kaplan–Meier survival curves can be used for this purpose; the curves for each group should remain (at least approximately) proportional to each other and certainly not cross. We will look at this again in subsequent exercises. Another way of checking this assumption is to stratify the analysis at different times to see whether the hazard ratio is similar at different periods of follow-up. Indeed, this

what the authors of Paper A did. They summarised the effect of follow-up time on the CHD hazard associated with heavy passive smoking (cotinine >0.7 ng/ml) relative to no or light passive smoking (cotinine <0.7 ng/ml), by calculating the hazard ratio in 5-year periods of follow-up for the 20-year study. We return to this example later in the chapter when looking at the application of Cox regression in Section 8.2.8).

8.2.6 Interpreting the Cox Regression Model

Interpretation of the Cox regression model is similar to that for logistic regression, except the hazard ratio reflects the ratio of the hazard function for two groups (probability of an event at any fixed period of time) rather than the ratio of cumulative probabilities of having an event. As with odds ratios in logistics regression, in Cox regression the hazard ratios are found by exponentiating the estimated regression coefficients b_1, b_2 (that is, we raise the natural log e to the power of b , written as e^b). A hazard ratio

- Greater than 1 indicates an increased hazard relative to the baseline hazard.
- Less than 1 indicates a reduced hazard relative to the baseline hazard.
- Equal to 1 indicates that the hazard is equal to the baseline hazard.

Example

In another cancer drug trial (not the study reported in Paper B), 37 patients were randomised to the treatment group and 32 patients to the control group. Their survival times (until death, in this case) are measured in months, and some observations are censored. The main question of interest is whether survival is related to treatment, but survival time is thought to be related to the age and sex of the patient as well. As shown above, the Cox regression model is:

$$\log \left(\frac{h(t)}{h_0(t)} \right) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where the explanatory variables are:

Group $x_1 = 0$ for controls; 1 for treatment
 Sex $x_2 = 0$ for males; 1 for females
 Age x_3 in years

The results of the Cox regression are similar to the output from logistic regression and look like this (Table 8.2.1):

Table 8.2.1 Results of Cox regression for the cancer drug trial example (see text).

	Explanatory variable	Regression coefficient	Hazard ratio	95% CI	p-value
1	Group	-1.896	0.1052	0.086–0.262	<0.0001
2	Sex	-0.09135	0.9127	0.732–1.366	0.4342
3	Age	0.1196	1.127	1.103–1.152	0.002

Interpretation

1. The hazard ratio for the effect of treatment on outcome (death in this study) is 0.1052 (95% CI: 0.086, 0.262), indicating that risk of death is reduced by almost 90 per cent at any given

- time. The 95 per cent CI does not include 1.0 (no effect); indeed, the upper limit is still well below 1.0 at 0.262, equivalent to a 74 per cent reduction in risk. As expected, this is a highly significant result ($p < 0.0001$).
2. The hazard (probability of the outcome: death) for females does not significantly differ from that for males (the 95 per cent CI includes 1.0, the p -value is large): the sex of the patient does not affect survival.
 3. Each 1-year increase in age results in the hazard increasing by a factor of 1.127, with a 95 per cent CI that does not include 1.0 and a p -value of 0.002: age has a significant effect on survival.

Important Note on Interpretation of the Hazard Ratio for Age

It is important to note that a 5-year increase in age multiplies the death hazard by $1.127^5 = 1.818$. This calculation is $1.127 \times 1.127 \times 1.127 \times 1.127 \times 1.127$; this is *not* the same as 5×1.127 , which would give a factor of 5.635 – very different!

We saw in Chapter 6 that this also applies to the interpretation of odds ratios from logistic regression, which, as you recall, are also derived by exponentiating the *beta* coefficient.

8.2.7 Prediction

We have seen that one of the functions of regression is *prediction*. The quantity ($b_1x_1 + b_2x_2 + \dots + b_kx_k$) is known as the *risk score* or *prognostic index* and can be used to predict the survival of new cases by substituting their particular set of explanatory variable values into the equation.

The use of Cox regression to create a prognostic index in survival analysis is shown in the following example. In a study of patients who have advanced pancreatic cancer and are receiving palliative care, researchers wished to create a prognostic index model to predict differential survival (Xue *et al.*, 2015). They conducted a 7-year cohort study of 145 patients after collecting detailed data on 14 clinical variables (suspected to be important prognostic factors for cancer survival based on a literature review). Using Cox proportional hazards regression, they identified three of these variables to be independently associated with poor survival. These included

- Eastern Cooperative Oncology Score (ECOG) – This is a disability 5-point score based on activities of daily living, ranging from 0 ‘fully active’ to 5 ‘dead’.
- levels of carbohydrate antigen 19-9 (CA19-9) – This is a tumour marker for pancreatic cancer, with elevated levels indicating more-advanced disease and disease progression.
- C-reactive protein (CRP) levels – Elevated levels of this protein have been found to be associated with tumour progression and a poorer prognosis in a variety of cancers.

The authors found a better survival prognosis in patients with a low ECOG (0–1 vs 2: HR = 0.49 (95% CI = 0.26, 0.93), lower levels of CA19-9 (<1,000 vs $\geq 1,000$: HR = 0.48 (95% CI = 0.25, 0.92) and lower levels of CRP (<5 vs ≥ 5 : HR = 0.49 (95% CI = 0.25, 0.95). The authors then created a simple prognostic index score based on a positive response to these three clinical variables at presentation: ECOG >2, CA19-9 $\geq 1,000$, and CRP ≥ 5 . The index score ranged from 0 (no positive responses) to 3 (positive to all three). When they compared a high-risk group (score of 2–3) against a low-risk group (score of 0–1) they found the low-risk group had a significantly

longer survival time: 9.9 months vs 5.3 months; HR = 0.27 (95% CI = 0.14, 0.52). The low-risk group also had a better overall 1-year survival rate: 40.5% versus 5.9%, $p < 0.05$.

8.2.8 Application of Cox Regression

We have now covered enough on Cox regression to look at how this method was used in the analysis of Paper A. Please now read the following excerpt from the results of Paper A and review Table 8.2.2 (Table 3 from Paper A).

Results (continued)

Serum Cotinine Concentration and CHD Risk

We examined the association between quarters of the cotinine distribution and CHD hazard ratios among all 2105 current non-smokers using the complete follow up period (table 2). The risks in the upper three cotinine groups were markedly higher than the risk in the lowest group, with a relative hazard of 1.61 in the highest group in the simplest model (adjusted for town and age), a hazard estimate similar to that of light active smokers. The association between cotinine concentration and CHD seemed graded and was not markedly affected by adjustment for other cardiovascular risk factors. The results of analyses restricted to lifelong non-smokers were similar, though the confidence intervals were wider. Exclusion of men with pre-existing CHD had no effect on these findings (data not presented). When we examined the overall association between cotinine concentration and CHD, we found that a doubling of cotinine concentration was associated with a hazard increase of 16% (95% confidence interval 6% to 27%).

Influence of Follow Up Period

In a Kaplan–Meier plot showing the cumulative proportions of men with major CHD over time among three groups (light passive (lowest cotinine quarter), heavy passive (upper three cotinine quarters), and light active (1–9 cigarettes/day)) we found that the heavy passive and light active groups diverged rapidly from the light passive group during the first years of follow up but remained almost parallel during later years (fig 2). The corresponding hazard ratios for cotinine and risk of CHD in separate five year follow up periods were highest in the early years of follow up and declined with increasing duration of follow up (table 3). These patterns were little affected by adjustment for cardiovascular risk factors, and again the hazard ratios for the heavier passive smoking groups were comparable with those of light active smokers. Restriction of these analyses to lifelong non-smokers or to men with no evidence of pre-existing CHD had no material effect on the results.

Serum Cotinine Exposure and Stroke

There was no strong association between cotinine concentration and stroke among non-smokers, either before or after adjustment for major cardiovascular risk factors (table 4). Analyses based on lifelong non-smokers showed similar results. For stroke, there was no strong evidence that hazard ratios changed over time (data not presented).

Table 8.2.2 Cotinine group and risk of coronary heart disease (CHD): hazard ratios (95 per cent CIs) for specific 5-year follow-up periods (Table 3 from Paper A).

	Follow-up period (years)			
	0–4	5–9	10–14	15–20
Passive smokers*†				
Adjustment 1	3.45 (1.36 to 8.80)	1.90 (1.09 to 3.31)	1.27 (0.72 to 2.22)	1.09 (0.66 to 1.82)
Adjustment 2	3.14 (1.23 to 8.04)	1.93 (1.09 to 3.42)	1.10 (0.63 to 1.95)	1.00 (0.60 to 1.67)
Adjustment 3	3.73 (1.32 to 10.58)	1.95 (1.09 to 3.48)	1.13 (0.63 to 2.04)	1.04 (0.62 to 1.76)
Light active smokers‡				
Adjustment 1	3.44 (1.07 to 11.02)	1.50 (0.63 to 3.55)	1.59 (0.70 to 3.62)	1.41 (0.66 to 3.03)
Adjustment 2	2.99 (0.90 to 9.97)	1.58 (0.66 to 3.80)	1.43 (0.61 to 3.38)	1.47 (0.68 to 3.15)
Adjustment 3	3.32 (0.87 to 12.64)	1.66 (0.66 to 4.18)	1.71 (0.71 to 4.10)	1.34 (1.23 to 1.47)

*Hazard ratios for CHD events for passive smokers with cotinine above 0.7 versus passive smokers with cotinine below 0.7.

†For CHD: adjustment 1 stratified by town and adjusted for age; adjustment 2 additionally adjusted for systolic blood pressure, diastolic blood pressure, total cholesterol, HDL cholesterol, FEV₁, height, and pre-existing CHD; adjustment 3 additionally adjusted for BMI, triglycerides, white cell count, diabetes, physical activity (none, occasional, light, moderate or more), alcohol intake (none/occasional, light/moderate, heavy), and social class (I, II, III non-manual, III manual, IV, V, and Armed Forces).

‡Hazard ratios for CHD events for low active smokers (1–9 cigarettes/day) versus passive smokers with cotinine below 0.7.

Source: Whincup 2004. Reproduced with permission of BMJ Publishing Group Ltd.



Self-Assessment Exercise 8.2.1

1. Can you explain why Cox regression, rather than logistic regression, has been used for the analysis of this study?
2. What assumption should be satisfied before applying Cox regression? Was this requirement met? Refer to Figure 8.1.5 (Figure 2 from Paper A), reproduced in Section 8.1.7.
3. In Table 8.2.2 (Table 3 from Paper A), the result for passive smokers observed with adjustment model 2, for the follow-up period 10–14 years, was 1.10 (0.63–1.95). Interpret this result, noting which confounding factors had been included in the model.
4. Why do you think the hazard ratio is highest for the first 5-year follow-up period and declines thereafter?

Answers in Section 8.4

Summary: Cox Regression

- Cox regression allows estimation of the hazard ratio (together with its 95% CI and *p*-value) summarising differential event hazards between groups. The method assumes proportional hazards between the groups being compared.
- The Cox regression multivariable model allows investigation of the simultaneous effect of several explanatory variables on survival.

- As with other multivariable regression methods, Cox regression allows adjustment for confounding and obtains estimates of the independent effects (hazard ratios) of variables we are interested in.
- We must be able to assume that hazards are proportional over time. Examination of survival curves in the Kaplan–Meier graph/plot allows assessment of whether the difference in hazards is (approximately) proportional at all times, a key test being that the curves do not cross.
- The result of Cox regression is a relative hazard for the event and is calculated by exponentiating the regression coefficient for each explanatory variable in the same way as for logistic regression.

8.3 Current Life Tables

8.3.1 Introduction

An important indicator of the health of populations is how long people can expect to live: *life expectancy*. This can be calculated for different populations provided data are available on current rates of mortality and the size of the population. Life expectancy at birth – and at any age – can be calculated through the use of *current life tables*.

The life table is a method for studying the survival pattern of a population or group of people. There are essentially two ways the method can be used. A *cohort life table* is where groups of people are followed up over time and survival rates are calculated through the life table. This is an alternative to the Kaplan–Meier survival curve, though it is less widely favoured. A cohort life table approach can be used to answer research questions on, for example, the survival of patients with different forms of cancers, so that an average life expectancy from diagnosis can be estimated. Over time, this can also show how life prospects for cancer patients has changed, if the calculations are completed with different cohorts at different points in time. See, for example, Kirkwood (1988) for further discussion of this technique. Given the relatively infrequent application of this method, here we focus instead on the current life table method and how it can be used to answer different questions about population health for researchers and policymakers.

8.3.2 Current Life Tables and Life Expectancy at Birth

One of the most fundamental measures of population health is life expectancy at birth, which measures (in years) how long a child born at a certain point in time can expect to live as a population average. This is often presented for men and women separately. Life expectancy at birth can be used in many ways:

- Comparison between areas – such as differences in life expectancy at birth in the local authorities across England and Wales.
- Setting aspirational goals – for example, in the early 21st century the UK government set a target of reducing by 10% the gap in life expectancy between the fifth of areas with the worst life expectancy and the national average life expectancy between 2001–2010.
- Time-series analysis, to chart the fluctuations in life expectancy in a setting over time, as a measure of the efficacy of health care and health-care policy.

Current life tables can be generated using the entire population or using an arbitrary figure such as 100,000 (such as in the example here). For the former, mortality data would be the reported numbers, and for the latter a rate per 100,000 population would be used. This example uses a

full life table as opposed to an abridged life table, in which age groups are used rather than the single years of age used to populate a full life table.

Current Life Table – Worked Example

In Table 8.3.1 we see a worked example of a life table using hypothetical data.

Table 8.3.1 Structure of a full life table.

Age: for explanations here, each age is termed age _x	Number surviving at age _x	Number of deaths between age _x and age _{x+1}	Probability of surviving from age _x to age _{x+1}	Probability of dying between age _x and age _{x+1}	Expectation of life at age _x
X	l _x	d _x	p _x	q _x	e ^o _x
0	100 000 (<i>radix</i>)	647.06	0.9935294	0.0064706	
1	99 352.94	158.84	0.9984013	0.0015987	
2					
Etc.					
Etc.					
99					
100					

The first column contains each year of age, denoted as *x*. For age 0, the **number surviving at age_x** (*l_x*) is our arbitrary 100,000. For age 1 year, *l_x* would be 100,000 minus the number of deaths between 0 and 1 years (*d_x*). Therefore, *l_x* can be calculated by subtracting the number of deaths (rate per 100,000 population) in the row/age group above from the numbers surviving also in the row/age group above, (i.e. $l_x = [l_x - 1] - [d_x - 1]$).

However, such data are not always available, so we would need to derive the number of deaths from mortality rates to calculate *q_x*, the **probability of dying between age_x** and age_x+1. To do this for all but age_x = 0 (i.e. the first year of life), we use the following equation:

$$q_x = \frac{\text{Mortality rate (for age } x)}{1 + 0.5 (\text{mortality rate for age } x)}$$

As much infant mortality occurs in the first six months of life, an adjustment is made for age_x = 0, creating the following equation:

$$q_x = \frac{\text{Infant mortality rate}}{1 + 0.7 (\text{infant mortality rate})}$$

In this case, we have set infant mortality at 0.0065 (or 6.5 deaths per 1,000 live births). Therefore:

$$q_x = \frac{0.0065}{1 + 0.7(0.0065)}$$

$$q_x = 0.0064706$$

The number of deaths occurring during this year (d_x) is then calculated as $l_x \times q_x = 100,000 \times 0.0064706 = 647.06$.

Hence l_x (the numbers surviving to year 1) $= 100,000 - 647.06 = 99,352.94$.

And p_x (the probability of surviving from birth to 1 year) $= 1 - q_x = 0.9935294$.

For the cohort aged 1 year, the mortality rate is found to be 1.6 per 1,000 population, or 0.0016.

Using the general equation for q_x , we calculate $q_x = 0.0016/1 + 0.5(0.0016) = 0.0016/1.0008 = 0.0015987$.

Therefore, $d_x = 99352.94 \times 0.0015987 = 158.84$ deaths, and $p_x = 1 - 0.0015987 = 0.9984013$.

From this, l_x when $\text{age}_x = 2$ is $99352.94 - 158.84 = 99,194.1$.

Life expectancy at age x (e^o_x) is calculated once the other fields are completed for all ages (with the highest age usually being that of the oldest person in the population (e.g. 99 years), or the oldest ages being grouped (e.g. aged 85 years and older). At each age, l_x represents the person-years of life. So, l_x for age 0 = 100,000 person years; at age 1 = 99,352.94 person years, and so on. The sum of l_x across all ages equals the number of person-years lived in the population during the specified time period. To calculate life expectancy at birth, we add all of the figures for l_x from $\text{age}_0 + 1$ to the maximum age, then divide by l_x (for age 0 years, i.e. 100,000 in this example) and add 0.5. For life expectancy at birth, this calculation starts from $\text{age} = 0$.

8.3.3 Life Expectancy at Other Ages

Whereas life expectancy at birth is an important indicator of population health prospects, other variations of life expectancy can be utilised to answer different questions. Completing the steps in Section 8.3.2 will provide a full list of life expectancy at each year of life. For example, areas with high infant mortality generally have lower life expectancy than areas with low infant mortality. Calculating the life expectancy from age one year instead of zero is more informative for the health expectancy of those who survive beyond the first year of life. In this instance, life expectancy is calculated by summing the l_x figures from age 2 years to (for example) 99, instead of from age 1 (to age 99).

Another example might be for resource allocation when looking at how long someone currently approaching retirement age (for example, aged 65 years) can expect to live: l_x is added from age 66 to 99 years. With a full current life table, it is possible to calculate life expectancy from any year of age.

8.3.4 Healthy or Disability-Free Life Expectancy

When considering utilisation of health-care resources, life-expectancy estimates are informative, but alone they cannot describe the prospective burden of illness and life-limiting disability. For example, a population with high life expectancy may also equate to a population where a large proportion of elderly people are living for many years with chronic illnesses that require treatment, and thus expenditure on health care. Adding a measure of disability to the life table calculations is a means of addressing this.

Healthy life years are calculated for the European Union for the purposes of benchmarking and comparing the health of nations, with a view to reducing inequalities in healthy life years between countries. Data on disability is less complete than data on mortality, and the quality and completeness varies between countries and by illness. To address this, survey data on the presence of any disability or illness can be used as an approximation of overall population health

at a given age. The UK census in 2011 included a non-compulsory question on self-rated health that is commonly used in population health surveys:

How is your health in general?
Very good Good Fair Bad Very Bad

Such a question could be used as a proxy for healthy life. One way would be to place a value on each response, where very good = 1 and very bad = 0, with good = 0.75, fair = 0.5, and bad = 0.25, although there are other ways to categorise these answers. From this it would be possible to derive a population estimate for good health at each age. Multiplying this weighting proportion by the l_x for a given age would give the number of years being lived in good health at age $_x$. As with life expectancy, healthy life expectancy at age $_x$ is calculated by adding together the number of years in good health from age $_x + 1$ through to the maximum age, then dividing this figure by l_x and adding 0.5.

8.3.5 Abridged Life Tables

A full life table (as in Table 8.3.1) has the information entered for each year. In practice, however, age-specific death rates for each year of life may not be available, and an abridged life table may be used. This would typically have data for 5-year intervals, although each of the first 5 years of life may be included, as in Table 8.3.2.

Table 8.3.2 Structure of an abridged life table.

Age: for explanations here, each age is termed age $_x$	Interval (years)	Number surviving at age $_x$	Number of deaths between age $_x$ and age $_{x+1}$	Probability of surviving from age $_x$ to age $_{x+1}$	Probability of dying between age $_x$ and age $_{x+1}$	Expectation of life at age $_x$
X	n	$n l_x$	$n d_x$	$n p_x$	$n q_x$	e^o_x
0	1	100 000				
1	1					
2	1					
3	1					
4	1					
5	5					
10	5					
15	5					
Etc.	Etc.					

An additional column (interval) is included to show the size of the unequal intervals. The notation is also slightly different, with the n included for columns 3–6 to indicate the interval (in years) referred to.

8.3.6 Summary

The main points from this brief introduction to life tables are summarised below.

Summary: Current Life Tables

- The current life table is a means of summarising death (or other) rates in a given population at a specific time.
- Current life tables can be used to compare life expectancy at birth between populations (e.g. between regions) and over time (e.g. to measure the effectiveness of health-care policy).
- The calculation of a life table is based on age-specific (mortality) rates. A full life table includes every year; an abridged one has larger intervals.
- Since death rates for a specific time period are used, the life table does not indicate the true mortality experience of a cohort of people over time, as the death rates those people actually experience as they age will change over time (as social circumstances and health services improve, or maybe deteriorate).
- Information about each age (or age group in an abridged table) is preserved and available for comparison.
- The information from life tables can be compared across populations as age-specific rates are used.
- Current life tables can be used to calculate healthy or disability-free life expectancy.
- Life tables can be calculated in standard statistical packages.

8.4 Answers to Self-Assessment Exercises

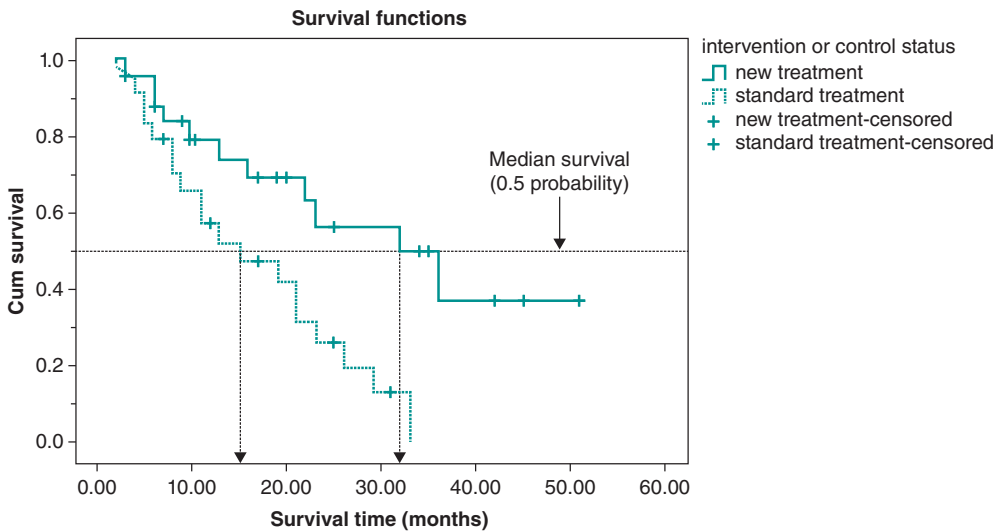
Section 8.1

Exercise 8.1.1

1. The subjects were a sub-sample of men from the British Regional Heart Study, a prospective investigation of men aged 40–59 years at recruitment drawn from 24 towns in Great Britain. The sub-sample for this analysis ($n = 4,729$) was drawn from the last 18 towns, which included serum samples (which could be subsequently analysed for cotinine). From the men in these towns, self-reported non-smokers ($n = 2,158$; this number was provided in a later excerpt – see Section 8.1.7) were selected, but only those with a serum cotinine of <14.1 ng/ml ($n = 2,105$) were included in the analysis.
2. Outcomes studied were major CHD events (deaths and non-fatal myocardial infarction) and stroke (deaths and non-fatal strokes).
3. Passive smoking was measured by serum cotinine levels. Cotinine is a metabolite of nicotine and is a ‘sensitive and specific’ indicator of exposure to tobacco smoke (the definition and interpretation of the terms *sensitivity* and *specificity* are covered in Chapter 10). Although it is a good indicator of recent history of exposure to tobacco smoke, the single baseline measurement becomes a progressively less accurate measure over time as the men’s exposure to tobacco changes.
4. The research team does have information on time between the baseline survey (recruitment) and the event; hence the ‘survival’ time. Therefore in these circumstances, survival analysis is an appropriate method.

Exercise 8.1.2

1. A total of 14 intervention group observations and six control group observations were censored. A possible reason for this difference might be that the better survival among intervention subjects has meant that these people are more likely to reach the end of their follow-up period without experiencing the outcome event (recurrence).
2. On the Kaplan–Meier curve (see annotated figure below), we have run a horizontal line across the graph from probability = 0.5. Where this line intersects the curves for (a) control and (b) intervention groups, we run a vertical line down to the x -axis (survival time). This gives values of approximately 15 months (control) and 32 months (intervention). These are confirmed when calculated in SPSS: median (with 95 per cent CI) on control treatment = 15 (3.9, 26.1); on new treatment, median = 32 (16.4, 47.6).



3. We can run a vertical line up from 30 months and see where this intersects with the two curves, and then run horizontal lines across to the y -axis, the survival probability. This is approximately at 0.13 (13 per cent) for the control group and 0.55 (55 per cent) for the intervention group.

Exercise 8.1.3

1. Geometric means are used because the distribution of cotinine levels is markedly right skewed, and this is normalised by log transformation. The use of transformation for skewed data and the interpretation of geometric means are discussed fully in Chapter 11.
2. Self-reported ex-smokers could have higher cotinine levels if they are still smoking some cigarettes, although, according to the interim surveys at 5 and 12 years, almost all (99 per cent) continued to report they were non-smokers. Alternatively (or in addition), ex-smokers may be more likely than never-smokers to associate at home, socially, and at work with current smokers.
3. Figure 8.1.4 is a Kaplan–Meier curve.
4. The probability of having a CHD event after 10 years for light passive smoking is approximately 0.04 (4 per cent); for heavy passive smoking, it is approximately 0.085 (8.5 per cent). It is not possible to read median probability of a CHD event from this Kaplan–Meier curve,

as the probability does not exceed 0.5: indeed, the probability of a CHD event after 20 years among the highest exposure group is just under 0.2.

5. The result of the log-rank test (result top left of Figure 8.1.4) is $p = 0.002$, a very significant finding. This indicates that we can reject the null hypothesis of no difference in survival between the groups studied (light passive, heavy passive, light active). Since the last two curves are close and cross several times, it is the difference between light passive and the other groups that is important.
6. The curves in Figure 8.1.4 start at a probability of 0.0 because this study looks at the probability of having an event, not avoiding it (i.e. survival without recurrence), as in the cancer trial example. At the start of follow-up, none of the men have had a new event; hence, the probability is 0.0.

Exercise 8.1.4

1. The abstract does not provide much detail about the inclusion criteria for selected patients. Essentially, eligible patients had locoregionally advanced squamous-cell carcinoma of the head and neck (LASCCHN) of the oropharynx, hypopharynx, or larynx with measurable disease.
2. The primary endpoint was described as locoregional control, with the secondary endpoint being survival.
3. The Kaplan–Meier graph is for survival (not dying), the secondary outcome. The graph is presented as percentage of patients surviving, where 100 per cent is equivalent to a probability of 1.0. As a result of the very high death rate (low survival), the probability of survival drops below 50 per cent during the course of the study, and it is therefore possible to read the median survival from the graph (unlike in Paper A). The results in the abstract are medians 29.3 and 49.0 months for radiotherapy and radiotherapy with cetuximab, respectively, and 36.4 and 45.6 per cent surviving to 5 years, respectively. These results are consistent with those you would read off the Kaplan–Meier curves.
4. A hazard ratio for survival of 0.73 with a 95 per cent CI 0.56, 0.95 ($p = 0.018$) implies that subjects receiving radiotherapy with cetuximab had a 27 per cent lower risk of dying, with a 95 per cent CI ranging from a decreased risk of dying between 5 per cent and 44 per cent. Since the 95 per cent confidence interval does not include 1.0, the result is statistically significant, confirmed by the p -value of 0.018.

Section 8.2

Exercise 8.2.1

1. Time to event data were available, and Cox regression makes the most of this information. Logistic regression could have been used for incidence density (CHD events/person-years), but this would not have been the most efficient use of the data. Cox regression is now commonly used in the analysis of prospective studies for this reason.
2. The Kaplan–Meier curves in Figure 8.1.5 do not cross (at least not those for light and heavy passive, which are the curves of principal interest for this analysis). However, we do know from the curves and particularly from the analysis of 5-year follow-up periods in Table 8.2.2, that the hazards do not remain proportional over the 20 years. Carrying out the Cox regression separately for each 5-year period, however, overcomes any concern about the assumption not being met across the whole 20-year period.
3. Interpretation: there is a 10 per cent increase in the relative hazard (risk of CHD if exposed to passive smoking with cotinine above 0.7 ng/ml compared to less than 0.7 ng/ml), but the 95 per cent CI is quite wide and includes 1.00, so this is a non-significant result. Adjustment

- 2, in addition to stratification for town and adjustment for age, included systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein cholesterol, FEV₁ (a measure of lung function), height, and pre-existing CHD.
4. Some comment on this is to be found in the paper, in the Discussion section. This considers the effect of relying on a single baseline measure of cotinine to characterise passive smoking exposure for (up to) the next 20 years. With changing smoking patterns (for example, spouses of subjects giving up, and growing restrictions on smoking at work, in public transport, and in public places) and generally greater awareness of the dangers of passive smoking, it is likely that many of these men have experienced considerable changes to their passive smoking exposure over the follow-up period.

9

Systematic Reviews and Meta-Analysis

Introduction and Learning Objectives

If you carry out a literature search on a defined question for almost any topic, you can expect to find several, if not many, published studies that in one way or another may contribute to answering the question. No two studies report identical results, in part because the methods and outcome measures may differ, but also for a number of other reasons including random variation (i.e. sampling error). Not infrequently, we find that studies appear to contradict each other. What are we to make of all this information? Do a large number of studies help us to arrive at a definitive answer, or do more studies lead to more confusion and lack of clarity? Finally, there are usually unpublished studies investigating the same question that may be hiding results that agree or disagree with the published studies.

As we will see over the course of this chapter, reviews of all available evidence on a particular question can be extremely useful. However, to assess the validity of the various studies, integrate the appropriate information, and arrive at an overall result, we need a rigorous, systematic approach. That is, we need to carry out a **systematic review** of all the evidence, both published and unpublished.

A **systematic review** is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review (Moher *et al.*, 2009).

As part of this process, we may want to obtain a quantitative summary (for example, of a treatment effect or an exposure risk) across comparable studies. This may be (most commonly) through combining the results of the individual studies or by analysing the raw data from the studies if they are available. The statistical methods required to carry out such analyses are known as **meta-analysis**.

Statistical methods (**meta-analysis**) may or may not be used to analyse and summarise the results of the included studies [from a systematic review]. **Meta-analysis** refers to the use of statistical techniques in a systematic review to integrate the results of included studies (Moher *et al.*, 2009).

Meta-analysis is commonly used to combine data from a number of randomised control trials (RCTs) of therapies or interventions. However, the techniques may also be used for observational epidemiological studies of risk factors.

Increasing Power by Combining Studies

The results of any one study may have too much random error to show any clear effect. That is, the study might not be powerful enough to demonstrate a statistically significant difference even where a true difference exists (that is, a type II error). In combining data from several studies, we increase the sample size and so increase power and obtain more-precise estimates. Doing this, however, requires that the studies we combine are sufficiently comparable, and we return to how this can be assessed later in the chapter.

Systematic reviews and associated meta-analyses are playing an increasingly important role in both research and practice. In this chapter we explore the rationale for carrying out systematic reviews and the methods for selecting, reviewing, and synthesising the results, and we look at some practical applications of the reviews. We then explore the principles and practicalities of carrying out a meta-analysis based on the results of a systematic review. This incorporates assessment of the suitability of studies to be included in a meta-analysis, methods of statistically pooling results from the selected studies, and approaches to investigating the effects of study quality on the results. Finally we introduce the Cochrane Collaboration, one of the most important current initiatives in the field of systematic reviews.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the purpose of systematic reviews and the contribution they can make to earlier introduction of effective practice.
- Describe the main steps in carrying out a systematic review, including how to minimise selection bias and assess the methodological quality of selected studies.
- Describe the most commonly used approaches to reporting the results of systematic reviews of RCTs and observational studies.
- Describe what is meant by publication bias and the commonly used techniques for determining whether publication bias is present and the implications for a meta-analysis.
- Describe what is meant by statistical heterogeneity between studies selected for a meta-analysis and the commonly used methods for determining the presence and extent of heterogeneity.
- Describe the two main statistical approaches used to combine results from studies in a meta-analysis, namely fixed-effect and random-effects models, and the implications of heterogeneity for the choice of method.
- Interpret the results of a meta-analysis as illustrated in a forest plot.
- Describe what is meant by sensitivity analysis and how this may be used to investigate the impact of specific aspects of methodological quality of studies on the results of a meta-analysis.
- Describe those aspects of methodology requiring special attention where a systematic review and meta-analysis of observational study designs is being carried out, rather than a review of randomised trials.

Resource Papers

Paper A

One resource paper has been selected for this chapter. It reviews a range of interventions aimed at reducing the incidence of diarrhoea in developing countries.

Fewtrell, L., Kaufmann, R.B., Kay, D. Enanoria, W., Haller, L., *et al.* (2005). Water, sanitation and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *Lancet Infect Dis* 5, 42–52.

9.1 The Why and How of Systematic Reviews

9.1.1 Why is it Important that Reviews be Systematic?

Conducting a systematic review involves a great deal of work, although it is generally quicker and less costly than carrying out a new study. Clearly, it is important to establish why all of this activity is so important.

An example of the importance of synthesising results from studies in a meta-analysis can be illustrated by studying the relationship over time between accruing evidence from meta-analyses with the recommendations made by experts in review articles and textbooks. In a now-classic study of thrombolytic drugs (drugs used to break down blood clots in heart attack patients) and recommendations for their routine implementation into current therapy, Antman *et al.* (1992) showed that such drugs were only first recommended for routine use in 1987, some 14 years after a statistically significant ($p < 0.01$) beneficial effect would have been identified if meta-analysis had been done at that time. The results from this study are illustrated in Figure 9.1.1.

Figure 9.1.1 is a graphical display of a *cumulative meta-analysis* of 70 RCTs of thrombolytic therapy. In the centre box, odds ratios (ORs) (blobs) and confidence intervals (CIs) (the horizontal lines through the blobs) are displayed, showing the pooled results of all studies up to a particular point in time. For example, at the top is an estimated OR of 0.5 (the OR is likely to have been less than 0.5, but scale has been truncated for the figure) and a wide CI (only the right half is shown), which is the result of a single, pre-1960, trial with only 23 patients. The wide CI includes 1, corresponding to no treatment effect, and reflects the lack of precision in this small trial. Between 1960 and 1965, a second study of 42 patients was completed. So the second line shows the pooled estimate from the two trials, with a total of 65 patients, and so on. With each additional study, more data are included in the meta-analysis, so the sample size increases and the CI for the pooled OR becomes narrower, reflecting increased statistical precision and hence certainty of the estimate.

By 1990, 70 trials had been completed with a total of 48,154 patients. The estimated OR and CI from all the trials clearly show the benefit of the treatment, although the evidence of this benefit was in fact available in the 1970s, if a systematic review had been carried out at that time. However, thrombolytic therapy was not recommended for routine treatment until some years after the available evidence actually confirmed the efficacy of the treatment. In fact, the entry in the second edition of the *Oxford Textbook of Medicine* (1987), more than 10 years after pooled results showed a statistically significant effect, stated:

The clinical benefits of thrombolysis whether expressed as improved patient survival or preservation of left ventricular function, remain to be established.

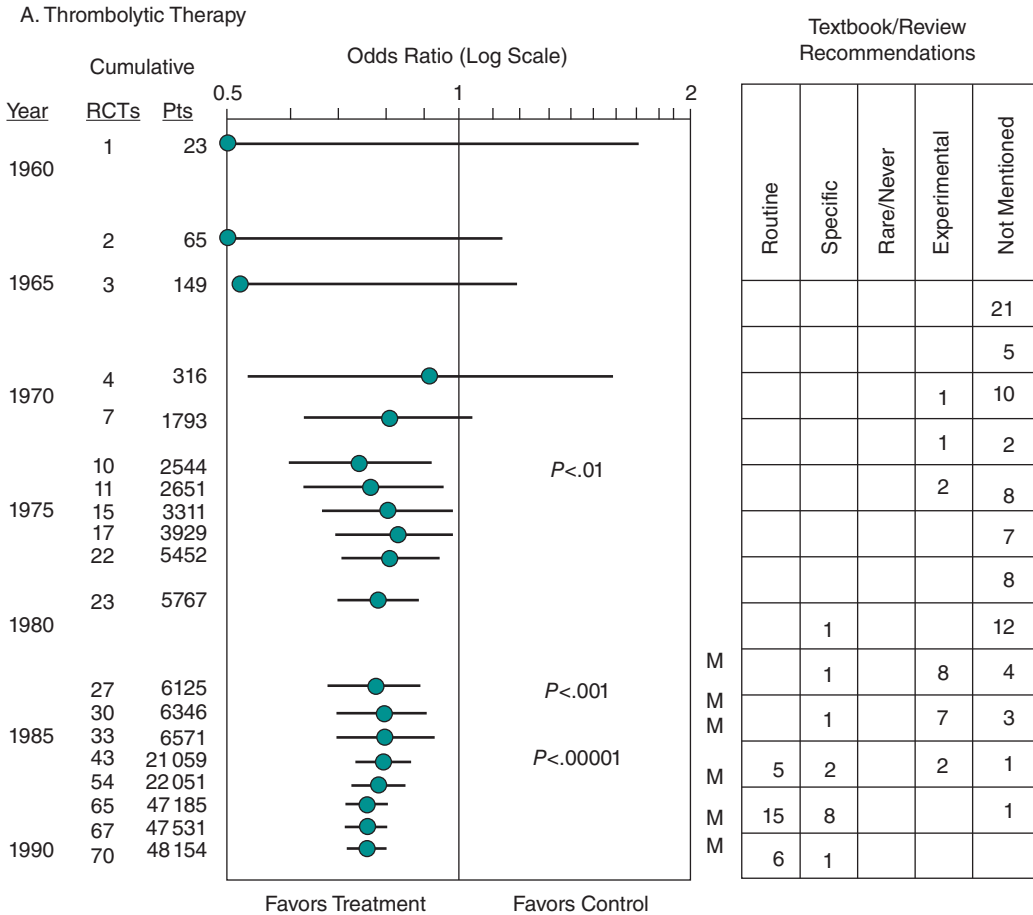


Figure 9.1.1 Recommendations to use thrombolytic therapy lagged well behind the evidence that would have been available from a systematic review and meta-analysis (adapted from Antman 1992). The letter M indicates that at least one meta-analysis was published that year.

Interestingly, this was published 2 years after a systematic review of the relevant RCTs had shown that the treatment reduced the risk of premature death after myocardial infarction (Yusuf *et al.*, 1985). Antman concluded,

Because reviewers have not used scientific methods, advice on some life-saving therapies has been delayed for more than a decade, while other treatments have been recommended long after controlled research has shown them to be harmful.

9.1.2 Method of Systematic Review – Overview and Developing a Protocol

In carrying out such reviews, we require systematic methods in order to establish whether research findings are consistent and generalisable. The process of carrying out a systematic review comprises the following steps:

1. Decide on the objectives of the review and develop the research question.
2. Define criteria for the inclusion and exclusion of studies for the review
3. Find studies which broadly address the topic being studied.

4. Select studies according to the eligibility (inclusion and exclusion) criteria.
5. Assess the methodological quality of the studies.
6. Extract data, that is, the main findings of each study.
7. Describe and compile the results of the review (synthesizing the evidence).
8. Report the results of the review.

Developing a structured review protocol detailing these steps is an essential step in the conduct of a systematic review. A protocol is the plan that reviewers will follow to complete the systematic review. It allows thinking around each step of the review to be focused and allocation of tasks to be determined. The methods to be used in a systematic review must be determined at the outset, so taking the time to prepare a clear protocol will reduce time spent during the systematic review process. We will see in Section 9.4 that there are published guidelines that set out crucial aspects of a systematic review that need to be considered in its reporting (**Preferred Reporting Items for Systematic reviews and Meta-Analysis – PRISMA**). One of these guidelines relates to the production of a structured protocol (and publication of this protocol) before conducting a systematic review. One reason this is such an important step is the iterative nature of systematic reviews: as new evidence (perhaps better-quality studies, larger studies, etc.) emerges, systematic reviewers may need to modify their original review protocol during its conduct. Another important reason, particularly in relation to publishing the protocol, is to let others know what you are planning to do.

We will now work through each of these steps to explain the process in more detail.

9.1.3 Deciding on the Research Question and Objectives for the Review

As we saw previously, with any study design, it is very important to state clearly the research question and study objectives prior to carrying out the study, and this is just as relevant to a systematic review. If the objectives are not explicitly stated, there is a risk that a reviewer will use the results to support a hypothesis that was not intended. For example, if after analysis of the results of individual studies, one or more are found to give unexpected or contrary findings in respect of the postulated objectives, there might be a temptation to revise the objectives of the review or to exclude some of the studies.

When developing the research question for the review, it is useful to frame it using **PICO** (as recommended by PROSPERO – an international prospective register of systematic reviews) to clearly define the components of the question. The review objectives should clearly state who the **Population** (study participants) is, what **Interventions** are to be investigated (or what **Indicators** or risk factors are to be measured for studies not involving interventions), what **Comparators** (alternative choices of action or alternative exposure) are available, and what **Outcomes** are to be assessed. In addition, the settings of studies for the review need to be stated and the criteria that will be used for inclusion and exclusion of subjects in the study should be clearly defined.

Please now read the following excerpts from the abstract and introduction of Paper A.

Abstract

Many studies have reported the results of interventions to reduce illness through improvements in drinking water, sanitation facilities, and hygiene practices in less developed countries. There has, however, been no formal systematic review and meta-analysis comparing the evidence of the relative effectiveness of these interventions. We developed a comprehensive search strategy designed to identify all peer-reviewed articles, in any language, that presented water, sanitation, or hygiene interventions. We examined only those articles with specific measurement of

diarrhoea morbidity as a health outcome in non-outbreak conditions. We screened the titles and, where necessary, the abstracts of 2120 publications. 46 studies were judged to contain relevant evidence and were reviewed in detail. Data were extracted from these studies and pooled by meta-analysis to provide summary estimates of the effectiveness of each type of intervention. All of the interventions studied were found to reduce significantly the risks of diarrhoeal illness. Most of the interventions had a similar degree of impact on diarrhoeal illness, with the relative risk estimates from the overall meta-analyses ranging between 0.63 and 0.75. The results generally agree with those from previous reviews, but water quality interventions (point-of-use water treatment) were found to be more effective than previously thought, and multiple interventions (consisting of combined water, sanitation, and hygiene measures) were not more effective than interventions with a single focus. There is some evidence of publication bias in the findings from the hygiene and water treatment interventions.

Introduction

Diarrhoeal disease is one of the leading causes of morbidity and mortality in less developed countries, especially among children aged under 5 years. Since the seminal reviews of Esrey and colleagues in 1985, 1986, and 1991, additional studies have been published on various water, hygiene, and sanitation related interventions aimed at population health improvements. The original reviews, and a study by Blum and Feachem, have led to a better understanding of methodological issues in this area. The reviews by Esrey and colleagues included studies that measured differences in health outcomes between groups that had different water or sanitation conditions. Since these original reviews, many studies have reported additional results of interventions to reduce illness through improvements in drinking water, sanitation facilities, and hygiene practices in the less developed world. There has, however, been no formal systematic review and meta-analysis comparing the relative evidence on the effectiveness of these interventions. We present a systematic review of all published studies and, where appropriate, meta-analysis of studies that reported interventions (planned or occurring as natural experiments) in water quality, water supply, hygiene, and sanitation in less developed countries. Less developed countries are defined here as any country not within a class A region under the WHO comparative risk assessment (class A countries have very low child and adult mortality).



Self-Assessment Exercise 9.1.1

According to the abstract and introduction of Paper A:

1. What were the stated objectives of this systematic review?
2. Was the context of the study clearly described (in terms of its PICO [population, intervention, comparator, and outcome])?
3. From the introduction, what types of studies were stated as eligible for the review?

Answers in Section 9.6

9.1.4 Defining Criteria for Inclusion and Exclusion of Studies

When we looked at intervention studies (Chapter 7), we saw the importance of clearly stating the subject inclusion and exclusion criteria. In a similar way, for a systematic review, it is

important to specify the criteria that studies need to meet before being eligible for inclusion and, by the same token, those that should not be included. Failure to do this can result in bias because study selection may be influenced by factors related to the reviewers' preconceptions, among other factors. The inclusion criteria should relate to the required study populations, treatments (which may be interventions or risk factors in observational studies), study outcomes, length of follow-up, and aspects of methodological quality. We will see later in the chapter how the methodological quality of studies can be assessed and how we can measure the influence that differing quality of studies can have on results of meta-analyses. The particular issues relating to reviews of observational studies are discussed in Section 9.3.

Please now read the following excerpt relating to the initial selection criteria of Paper A, taken from the methods section.

Methods

Initial Selection Criteria and Data Extraction

Two selection criteria were used to identify articles: (1) description of specific water, sanitation, or hygiene interventions, or some combination of such interventions; and (2) diarrhoea morbidity reported as the health outcome, measured under endemic (non-outbreak) conditions. In addition, only published studies were used, to maintain quality (via peer review) and transparency.

No study was excluded from the review or meta-analysis on the basis of quality criteria alone.



Self-Assessment Exercise 9.1.2

1. What were the main selection criteria used to identify relevant articles for the review?
2. Why do you think the authors did not include study quality in the selection criteria (we will be discussing the importance of assessing methodological quality of studies in systematic reviews in section 9.1.6)?

Answers in Section 9.6

9.1.5 Identifying Relevant Studies

Publication Bias

When carrying out a systematic review it is important to consider including relevant unpublished studies, as well as published, peer-reviewed studies. Thus a simple literature search using electronic databases is not sufficient (also because not all published studies are in the set of e-databases most people can get or are prepared to search). One of the most important reasons for this is that studies reporting statistically significant results are more likely to be published than those with non-significant results. This selective publication, or **publication bias**, means that we may reach over-optimistic or misleading conclusions if we include only published studies in our systematic review. This is particularly the case with small studies: we have seen that any study might occasionally produce a significant effect when no such effect really exists (type I error), but since small studies are more difficult to publish than large ones, there is a tendency for those with significant results to be offered (and accepted) for publication more frequently than small studies without 'interesting' results.

Statistical significance does not guarantee the quality, validity, or clinical significance of the research, and good studies that have conclusively demonstrated a lack of treatment effect or lack of association may never be published. Since 2005, in accordance with the International Committee of Medical Journal Editors (ICMJE), prior entry of clinical trials in a public registry is a condition for publication of results in ICMJE journals (including *The New England Journal of Medicine*, *Journal of the American Medical Association*, and *The Lancet*, among others). For UK clinical trials, the ICMJE recommends a website for public registration: www.controlled-trials.com, a UK site developed and maintained by Current Controlled Trials Ltd, part of the Current Science Group of biomedical publishing companies. Such registries make it easier to identify studies for a systematic review in order to reduce the risk of publication bias.

Publication bias is minimised by a comprehensive search strategy including unpublished work and foreign-language journals. However, it should be noted that the inclusion of data from unpublished studies can itself introduce bias. The studies that can be located might be an unrepresentative sample of all unpublished studies and, in general, unpublished trials may have a poorer methodological quality than those that are published.

As we shall see in Section 9.2 when we consider how to conduct a meta-analysis, it is possible to investigate both graphically (by a funnel plot) and quantitatively (by statistical methods) whether the selection of studies for a systematic review is likely to have been subject to publication bias.



Self-Assessment Exercise 9.1.3

Refer to the previous excerpt from Paper A (Section 9.1.4) that described the initial selection criteria and data extraction.

1. What reasons did the authors give for only including published studies in their review?
2. Do you think that these reasons were sufficient justification for excluding unpublished studies from their review?

Answers in Section 9.6

Specifying Search Terms

Before searching the literature it is important to develop an exhaustive list of terms that can be used to identify *all* potentially relevant studies that address the review question. These need to be clearly detailed in the study protocol for transparency. The best way of structuring this is to return to the review research question (specifically the **I**ntervention/**I**ndicator and **O**utcome components of the PICO). Taking the example from Paper A, the specified Intervention(s) were ‘water quality, water supply, hygiene, and sanitation interventions’ and the Outcome was ‘diarrhoeal disease’ – note that the authors state that this was diarrhoea ‘measured under endemic (non-outbreak) conditions’.



Self-Assessment Exercise 9.1.4

The authors of Paper A stated that they *paired aspects of ‘water’, ‘sanitation’, and ‘hygiene’ with ‘diarrhoea’ and separately with ‘intervention’*. How appropriate do you think this list of search terms was in identifying relevant studies for the systematic review?

Answers in Section 9.6

Searching the Literature

The next important step in ensuring that a review is systematic is to state what the search strategy will be prior to identifying relevant studies. The Cochrane website (www.cochrane.org; the Cochrane initiative is described further in Section 9.5) contains detailed information about how to develop a search strategy. In this section we briefly look at the main approaches to carrying out a thorough literature search.

The sources chosen to search for studies will be influenced by the types of study to be included in the review, that is, whether these are to be clinical trials, observational studies, qualitative studies, and so on, but generally the first step is to search the main health-related electronic databases. If the review relates to clinical trials, the *Cochrane Central Register of Controlled Trials (CENTRAL)* (see also Section 9.5) is the best single source of trials published in peer-reviewed journals. To ensure the literature is up to date, however, searches of *Medline* and *EMBASE (Excerpta Medica dataBASE)*, which is managed by the publisher Elsevier) should also be carried out. This would also be the starting point for systematic reviews of observational studies. The main attributes of Medline and EMBASE (as of December 2015) are shown in Table 9.1.1.

Table 9.1.1 Summary of the main attributes of the two main electronic journal databases for quantitative studies of health outcomes (Medline and EMBASE).

Medline	EMBASE
<ul style="list-style-type: none"> • approximately 18 million references • 5,200 journals indexed • Uses specific indexing (MeSH) • 1950 to present • Available through the Internet • 52% from the USA 	<ul style="list-style-type: none"> • approximately 24 million references • 7,500 journals indexed • Uses specific indexing (EMTREE) • 1947 to present • Available through the Internet • 33% from the USA

It is advisable to search both Medline and EMBASE because the overlap between the databases is only approximately 35 to 50 per cent. According to the Cochrane Handbook for Systematic Reviews of Interventions (accessed Dec 2015), of the 5,200 journals indexed by Medline, 1,800 were not indexed in EMBASE (35%) and of the 4,800 journals indexed by EMBASE, 1,800 were not indexed in Medline (38%). In addition, just over half the references on Medline were published in the USA compared to only a third of references on EMBASE, which has better coverage of European journals.

The next step is to supplement the literature search by utilising additional electronic databases of published health-related research. The most common of these databases are listed in the box below. It is advisable to consult a librarian to identify relevant databases in your research field.

Examples of Additional Electronic Databases of Published Studies in Health Fields (Accessed December 2015)

- **Allied and Alternative Medicine (AMED)** indexes articles relating to complimentary medicine, occupational therapy, palliative care, physiotherapy, podiatry, rehabilitation, and speech and language (records from nearly 600 journals).
- **Biological abstracts (BIOSIS)** indexes articles in the fields of biology, biochemistry, biotechnology, botany, pre-clinical and experimental medicine, pharmacology, zoology, agriculture, and veterinary medicine published since 1926.

- **CAB Direct** indexes articles in the field of applied life sciences, incorporating two types of bibliographic databases. One is CAB Abstracts, representing agriculture, animal and veterinary sciences, environmental sciences, human health, food, nutrition, microbiology, parasitology, and plant sciences. The other is Global Health, representing public health at the international and community level.
- **Cumulative Index to Nursing and Allied Health Literature (CINAHL)** indexes over 700 nursing and allied health journals from 1937 to the present.
- **Derwent Drug File** contains information on drugs and pharmaceutical sciences from over 40 countries, including journals and conference proceedings.
- **PsychInfo** indexes and abstracts peer-reviewed literature in the fields of psychology, behavioural sciences, and mental health.
- **Science Citation Index (SCI)** indexes in the field of science and technology, including articles from 6,500 journals, across 150 disciplines, from 1900 to the present.
- **Pascal Biomed** provides bibliographic indexing of scientific literature with multidisciplinary and multilingual coverage for science, technology, and medicine, with special emphasis on European content.
- **Latin American and Caribbean Health Sciences Information System (LILACS)** indexes more than 600 medical journals from Latin America and the Caribbean region. LILACS has interfaces in Portuguese, Spanish, and English and has a unique content, because most of these journals are not indexed in other databases.
- **China Knowledge Resource Integrated Database (CNKI)** was launched in 1998. The CNKI is an electronic platform reflecting a comprehensive bibliography of China-based information resources. It includes the China Academic Journals (CAJ) database incorporating over 9 million articles from 7,000 academic journals published in China since 1994.
- **African Index Medicus (AIM)** is an international index to African health literature and information sources operated by the World Health Organisation's Africa division.

After searching the electronic databases of published studies, it is also advisable to search the so-called *grey literature* for additional studies that have not been published in peer-reviewed journals. For example, a proportion of trials and observational studies are only published as meeting or conference abstracts. The main grey literature electronic databases are described in the box below.

Principal Electronic Databases of Grey Literature

- **British Library Conference Collections** is a comprehensive accessible database of International conference proceedings.
- **System for Information on Grey Literature in Europe (SIGLE)** provides open access to records of grey literature including technical or research reports, doctoral dissertations, conference papers, and some official publications (available through www.opengrey.eu).

To complete the search, checks should be made to ensure that the review has captured all relevant studies by checking the reference lists of published papers and reports, by hand searching the reference lists of key journals, by carrying out author-based searches, and by contacting experts in the field to identify any research they have conducted or know about that has not yet been published.

We now look at the search strategy used for Paper A, described in the following excerpt.

Methods

Search Strategy

Database searches of the Cochrane Library, Embase, LILACS, Medline, and Pascal Biomed were done with keyword searches that paired aspects of 'water', 'sanitation', and 'hygiene' with 'diarrhoea', and, separately, with 'intervention'. The Cochrane Central Register of Controlled Trials was particularly useful for identifying intervention studies; Embase and Medline provided very good coverage of English language papers; and LILACS and Pascal Biomed provided coverage of foreign language, Latin American, and Caribbean papers. Searches were limited to articles published before June 26, 2003 (when the search was done), and to articles about human beings. The reviews by Esrey and colleagues were used as an additional source to identify early studies, and author-based searches were used to identify subsequent work by the primary investigators, with additional information. All titles and abstracts (if available) from each of the searches were examined and then the relevant articles were obtained for review. Bibliographies of those articles were examined for additional references. No restrictions were put on study design, location, or language of publication.



Self-Assessment Exercise 9.1.5

1. List the databases and sources used for the search.
2. Do you feel these were sufficiently comprehensive?

Answers in Section 9.6

Selecting Relevant Studies for the Review

The process of searching the literature and selecting relevant studies (by referring to inclusion and exclusion criteria reported in the systematic review protocol) should ideally be conducted independently by two reviewers; if that is not possible, then a proportion should be reviewed independently. This ensures that key studies are not missed through reviewer error. When all sources of literature have been reviewed, a bibliographic database of titles and abstracts of potentially relevant studies should be created using bibliographic software (e.g. Endnote, Reference Manager, Mendeley). A three-stage filtering process should then be applied using an 'inclusive' approach (only excluding articles that are clearly not relevant in addressing the review question) reviewing the titles and then the abstracts before finally appraising the full text papers. This is a very important step in the conduct of a review, and a requirement specified in the PRISMA statement is to clearly document and provide numerical details of this selection process in the form of a flow chart. The main reasons for rejection of papers after screening abstracts and full-text articles should be briefly summarized. The structure of the flow chart of information through the different phases of a systematic review (adapted from the PRISMA statement (Moher *et al.*, 2009) is shown in Figure 9.1.2.

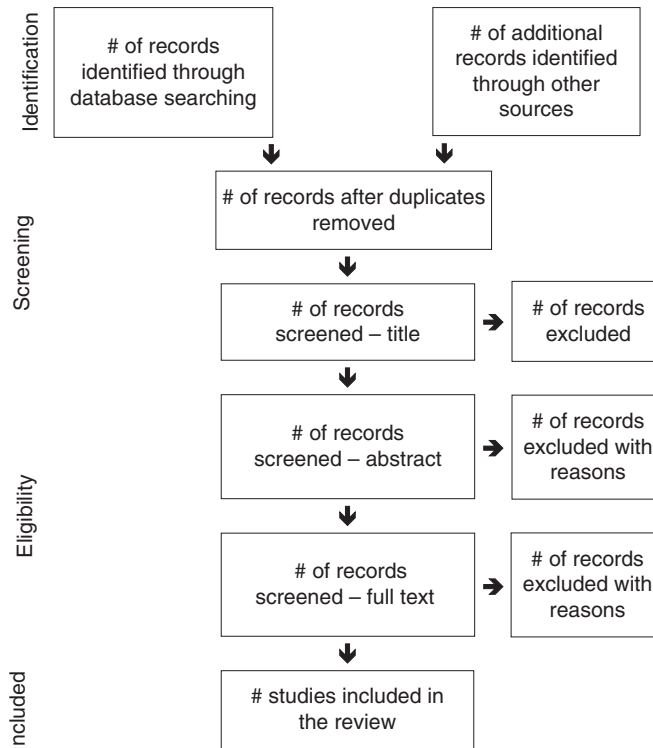


Figure 9.1.2 Flow of information through the different phases of a systematic review.

9.1.6 Assessment of Methodological Quality

What is Methodological Quality?

The concept of *methodological quality* is hard to define, but typically the term is used to describe the design, procedures, and conduct of a study, how the analysis has been carried out, and/or the quality of the reporting. For example, we might decide that ‘good’ quality studies for a systematic review of clinical trials would need:

- placebo controls (if possible);
- evidence of effective randomisation;
- blinding used, ideally at least double blinding; that is, of study participants and research staff assessing the outcome;
- near-complete follow-up of subjects;
- analysis by intention-to-treat.

Complete information is not always available, however, even in published journal articles, and it is sometimes necessary to define the methodological quality of studies in terms of only basic criteria. Alternatively, it may be necessary to contact the authors to obtain the missing information.

Why Measure Methodological Quality?

Even if we do exclude studies with poor methodological quality (noting this was not done by the authors of Paper A), it is likely that the remaining studies will still be of variable quality. It has

been demonstrated empirically that studies with poor quality can distort the results from systematic reviews and meta-analyses. For example, a study of 250 trials from 33 meta-analyses of a range of interventions relating to pregnancy and childbirth found that non-random treatment allocation and lack of double blinding of controlled intervention trials were associated with larger treatment effects. The main findings of this study, which compared the effect estimates of trials defined as having inadequate or unclear methodology with those from trials defined as having adequate methodology (an OR ratio of less than 1 indicates an exaggerated treatment effect), are presented in Table 9.1.2.

Table 9.1.2 Summary of the results of a study investigating methodological aspects of RCTs in relation to the size of treatment effect (Egger *et al.*, 2003).

Methodological quality item	Ratio of OR*	95% CI	Interpretation
Treatment hidden from subjects			
– adequate	1.0	(reference group)	
– unclear	0.67	0.60–0.75	Exaggerated effects
– inadequate	0.59	0.48–0.73	Exaggerated effects
Randomisation method			
– adequate	1.0	(reference group)	
– inadequate/unclear	0.95	0.81–1.12	Similar effects
Double blinding			
– yes	1.0	(reference group)	
– no	0.83	0.71–0.96	Exaggerated effects

*Comparison of ORs of studies with inadequate methodology and studies with adequate methodology.

Among this group of studies, the authors did not find that the randomisation method or exclusions after randomisation influenced the treatment effect. If the treatment was not adequately hidden (blinded) from study subjects, however, there appeared to be an exaggerated treatment effect of 41 per cent (95% CI: 27 to 52 per cent). In addition, a lack of double blinding was associated with an exaggeration of treatment effect by 17 per cent (95% CI: 4 to 29 per cent). These results highlight the importance of assessing the methodological quality of studies included in a systematic review, especially if the results are to be included in a meta-analysis.

How Should we Measure Methodological Quality?

Although it is possible to use study quality as an inclusion/exclusion criterion for a systematic review, this is rarely done in practice, as was the case in Paper A. This is due to the somewhat subjective nature of the decision process as to whether studies meet minimum inclusion criteria in relation to quality and the potential for relevant information being inappropriately excluded. Instead, as we shall see, the influence of study quality on the results derived from systematic reviews can be explored through *sensitivity analysis* (Section 9.2.6).

Given the subjectivity in deciding on a study's quality, it is good practice for two reviewers to check the eligibility of candidate studies independently, with disagreements being resolved through discussion with a third reviewer.

Methodological quality can be quantified by scoring the quality of studies on a pre-existing scale, such as that developed for randomised trials by Jadad *et al.* (1996). This scale has five items: two relate to blinding, two to randomisation, and one to the description of withdrawals or dropouts. When using the Jadad scale, each of the five items receives a 'yes' or a 'no',

resulting in an overall composite quality score that can range from 0 to 5; higher scores reflect better methodological quality. Creating quality scores can reduce the subjectivity involved in the quality decision-making process; however, it is not ideal to exclusively assess study quality on the basis of a quantitative summary.

As discussed in Chapters 4, 5, and 6, observational studies (cross-sectional surveys, cohort studies, and case–control studies) have specific methodological issues that can affect their results in terms of *bias* and *confounding*. Any assessment of the quality of observational studies needs to assess how well these issues have been controlled for. Examples of tools for assessing the quality of observational studies are the Newcastle–Ottawa Scale (for cohort and case–control study designs; best sourced via the Internet) and the Liverpool Quality Assessment Tools (LQATs) (for all observational study designs; available through the authors of this book). These provide both a quality score for included studies and a narrative for individual components of methodological quality. We look more closely at systematic reviews of observational studies in Section 9.3.

Although assessment of quality is a very important step, evidence suggests that the use of quality scales can be problematic. Comparisons of assessments of the quality of trials using different composite scales have found that the perception of the quality of a clinical trial, and hence (potentially) whether or not it is included in the review, varies according to which scale is used. The conclusions of a meta-analysis can therefore be affected by the choice of quality scale.

An alternative to using a quality score is to measure individual components of methodological quality. These can then be examined quantitatively for their influence on the results from a meta-analysis. An example of this was seen in Table 9.1.2, where the authors identified that inadequately hiding the treatment from subjects and not using double blinding were associated with exaggerated treatment effects. As we will see in Section 9.2.6, this can be investigated by performing a *sensitivity analysis*. Let's now look at the assessment of quality in Paper A, described in the following excerpt:

Methods

The quality of each study was examined on the basis of a set of methodological criteria for such studies previously suggested by Blum and Feachem. No study was excluded from the review or meta-analysis on the basis of quality criteria alone. If possible, issues of study quality were examined in the meta-analysis as a source of possible heterogeneity between results. Poor quality studies, for the purposes of this review, were defined as those that had any of the following design flaws: inadequate or inadequately described control groups, no clear measurement or control for confounders, no specific definition of diarrhoea or the particular diarrhoeal health outcome used, or a health indicator recall period (i.e. the maximum time between illness occurrence and the reporting of the illness) of more than 2 weeks. Studies without these flaws were categorised as being of good quality. Fewtrell and Colford have outlined further details on issues of study quality.



Self-Assessment Exercise 9.1.6

1. How were 'poor-quality' studies defined?
2. How did the authors deal with studies of poor quality?

Answers in Section 9.6

9.1.7 Extracting Data

We have now come to the stage of extracting data from articles selected for the systematic review to report in a narrative summary of the main results.

The process of data extraction should be carried out with as much care as was taken for assessing the methodological quality of studies. Again, it is important that two independent observers extract the data to ensure that errors are minimised. Data extraction requires that a form prepared for this purpose be used for all the studies selected for the review, and this should be carefully designed, pilot tested, and revised if required. Typically, information required includes the reference; the study design; the setting and sample of people being studied; the intervention, treatment, indicator, or exposure of interest and how this is measured; the health outcome and how this is measured; and a measure of the size of effect (e.g. odds ratio, relative risk, difference in means) with associated CIs.

9.1.8 Describing the Results

The descriptive presentation of results from a systematic review normally involves three stages. First, it is important to state clearly the numbers of studies included and rejected from the review. Second, the results of the articles included in the systematic review are presented in a structured table summarising the main attributes of the studies. The third stage is to provide a descriptive outline of the main results of the review. Finally, if there is enough information from studies included in the review, a quantitative summary of results can be provided by pooling the study data in the form of a *meta-analysis*.

Please now read the excerpt below and review Table 9.1.3 (Table 1 from Paper A), reproduced from the results section of Paper A.

The table and text relate to one objective of the systematic review (hygiene interventions) and show how the results of the review are presented as a descriptive summary of the studies. At this stage, you can ignore the information given at the bottom of Table 9.1.3, as this refers to the meta-analysis. We will return to this information in Section 9.2.

Results

Hygiene

15 articles, representing 13 distinct studies, were identified that examined hygiene interventions. 11 of these studies presented data that could be used for meta-analysis (Table 1). Hygiene interventions were typically of two types, those concentrating on health and hygiene education, and those that actively promoted handwashing (usually alongside education messages). The number of messages, the content of those messages, and the way in which they were delivered varied between studies. In general, education was aimed at the mothers, although the outcome was measured in children.



Self-Assessment Exercise 9.1.7

1. From the text excerpt, how did the authors summarise the types of hygiene intervention?
2. Referring to Table 9.1.3, briefly describe how information from each study included in the systematic review relating to hygiene is presented.

Answers in Section 9.6

Table 9.1.3 Studies of hygiene interventions and health effects (Table 1 from Paper A).

Reference	Intervention	Country (location)	Study Quality*	Health Outcome	Age Group	Measure	Estimate (95% CI)
Khan, 1982	Handwashing with soap	Bangladesh (unstated)	Good	Diarrhoea	All	RR [†]	0.62 (0.35–1.12) [‡]
Torún, 1982	Hygiene education	Guatemala (rural)	Poor	Diarrhoea	0–72 months	RR [†]	0.81 (0.75–0.87) [‡]
Sircar <i>et al.</i> , 1987	Handwashing with soap	India (urban)	Good	Diarrhoea	0–60 months >5 years	RR [†] RR [†]	1.13 (0.79–1.62) 1.08 (0.86–1.37)
				Dysentery	0–60 months >5 years	RR [†] RR [†]	0.67 (0.42–1.09) 0.59 (0.37–0.93)
				Combined outcome	Combined ages	RR [†]	0.97 (0.82–1.16) [‡]
Stanton <i>et al.</i> , 1988 Stanton and Clemens, 1987	Hygiene education	Bangladesh (urban)	Good	Diarrhoea	0–72 months	IDR [#]	0.78 (0.74–0.83) [‡]
Alam <i>et al.</i> , 1989	Hygiene education (and increased water supply)	Bangladesh (rural)	Good	Diarrhoea	6–23 months	OR	0.27 (0.11–0.66) [‡]
Han and Hlaing, 1989	Handwashing with soap	Burma Myanmar (urban)	Good	Diarrhoea	0–60 months 0–24 months 25–60 months	RR RR RR	0.70 (0.54–0.92) 0.69 (0.48–1.01) 0.67 (0.45–0.98)
				Dysentery	0–60 months 0–24 months 25–60 months	RR RR RR	0.93 (0.39–2.23) 0.59 (0.22–1.55) 1.21 (0.52–2.80)
				Combined outcome	0–60 months	RR [†]	0.75 (0.60–0.94) [‡]

Lee <i>et al.</i> , 1991	Hygiene education	Thailand (unstated)	Good	Diarrhoea	0–60 months	RR [†]	0.43 (0.32–0.56) [‡]
Wilson <i>et al.</i> , 1991	Handwashing with soap	Indonesia (rural)	Good	Diarrhoea	<11 years	RR [†]	0.21 (0.08–0.53) [‡]
Haggerty <i>et al.</i> , 1994	Hygiene education	Zaire (rural)	Poor	Diarrhoea	3–35 months	RR [†]	0.89 (0.80–0.98) [‡]
Pinfold and Horan, 1996	Hygiene education	Thailand (rural)	Poor	Diarrhoea	0–60 months	RR [†]	0.61 (0.37–1.00) [‡]
Shahid <i>et al.</i> , 1996	Handwashing with soap	Bangladesh (periurban)	Good	Diarrhoea	All	IDR [#]	0.38 (0.33–0.43) [‡]
					0–11 months	IDR [#]	0.39 (0.29–0.54)
					12–23 months	IDR [#]	0.53 (0.37–0.77)
					24–59 months	IDR [#]	0.44 (0.34–0.59)
					5–9 years	IDR [#]	0.27 (0.19–0.37)
					10–14 years	IDR [#]	0.28 (0.16–0.49)
					>15 years	IDR [#]	0.38 (0.30–0.49)

Results of the meta-analyses: fixed-effect estimate of relative risk (RR) 0.75 (95% CI 0.72–0.78); heterogeneity $p < 0.01$; random-effects estimate of RR 0.63 (95% CI 0.52–0.77); Begg's test $p = 0.19$.

*For definition of quality, see main text.

[†] Calculated.

[‡] Result used for the overall meta-analysis, which provided a pooled estimate of relative risk.

[#] IDR = Incidence density ratio (interpreted as RR but is the ratio of incidence density rates rather than incidence rates).

Adapted from Fewtrell 2005.

Summary

- A systematic review is a review of the methods and results of all individual studies that are designed to answer the same research question and that conform to a set of pre-agreed criteria.
- Systematic reviews can provide essential information for the early introduction of effective practice.
- It is important that reviews employ systematic methods in order to establish whether research findings are consistent and generalisable.
- The process of carrying out a systematic review comprises the following stages:
 - (i) decide on the research question and objectives of the review, which should be clear and explicit;
 - (ii) define the inclusion and exclusion criteria, which should also be clear and explicit, and may include key methodological features or quality;
 - (iii) identify relevant studies by carrying out an effective and wide-ranging literature search that will minimise the likelihood of publication bias;
 - (iv) filter studies (based on titles, abstracts and full text review) to identify the final set of studies to include in the review; ideally, at least a proportion should be carried out by two reviewers. Record results of filtering in a review flow chart;
 - (v) assess the methodological quality of studies to be included in the review, using (at least) two independent assessors;
 - (vi) extract data using a form developed for the purpose of the review, with (at least) two independent reviewers;
 - (vii) present results from the review in the form of a structured table of the main study attributes (including quality and main methodological limitations), with a descriptive narrative summary.

9.2 The Methodology of Meta-Analysis

After investing a substantial amount of time and effort in conducting a careful and thorough systematic review, the icing on the cake is to be able to combine the results from studies selected into the review quantitatively by a *meta-analysis*. Meta-analysis is the name given to the statistical analysis of data from studies included in a systematic review.

The importance of this final step was illustrated in the previous example by Antman *et al.* (1992) in Section 9.1.1, where we saw that conducting a meta-analysis would have demonstrated a statistically significant treatment effect for thrombolytic drugs some 14 years before the drugs started to be used routinely.

9.2.1 Method of Meta-Analysis – Overview

Essentially there are four main steps in carrying out a meta-analysis. These are:

1. An assessment of publication bias using a funnel plot (or a statistical analogue of the funnel plot) to look for asymmetry.
2. A statistical test for heterogeneity (difference) of the intervention effect between the selected studies.

3. A pooled estimate (e.g. RR, OR, difference in means) and 95% CI for the intervention effect after combining all the trials, the statistical approach used depending on whether or not statistical heterogeneity has been identified between the selected studies.
4. An hypothesis test for whether the intervention effect is statistically significant or not.

We discuss each of these steps in the remainder of this section.

9.2.2 Assessment of Publication Bias – the Funnel Plot

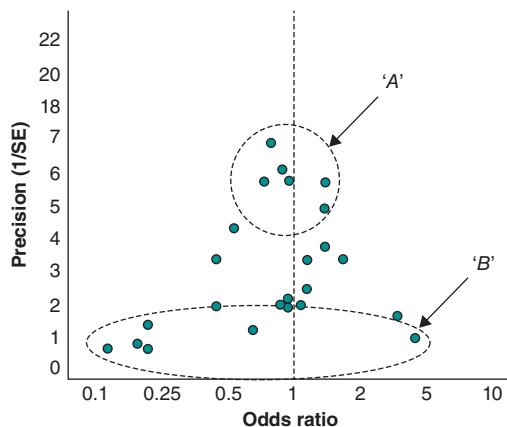
Graphical Presentation

Funnel plots are used to identify whether a systematic review might have been subject to publication bias in the selection of studies for the review (see Section 9.1.5 for an overview of publication bias). If substantial publication bias is identified, the results should not be pooled in a meta-analysis without firstly considering potential explanations for this bias by revisiting the original search strategy. Secondly, it is possible to statistically adjust for the effects of the bias prior to pooling data (e.g. through trim-and-fill methods; this approach simulates inclusion of the missing studies in the analysis and has the effect of attenuating the *biased* pooled effect estimate).

A funnel plot is a scatterplot showing the spread of results from studies selected for a review. Each point on the graph corresponds to one study and shows the relevant effect estimate (e.g. an OR) and its precision (i.e. how precisely it is estimated). The precision of the estimate is measured by its estimated standard error (SE), which depends on the variability in the study and the sample size: larger samples provide more-precise estimates.

The examples illustrated in Figure 9.2.1 clarify this and show why this type of graph is called a funnel plot. The plots show the estimated OR for each study on the x -axis, with a vertical line through the value for no effect (OR = 1.0), and the precision (expressed as the inverse of the standard error; $1/SE$) is shown on a log scale on the y -axis. Moving up the scale on the y -axis corresponds to a more precise estimate. In general, larger studies have smaller standard errors and hence provide more-precise estimates with higher values on the y -axis. In the absence of

(a) Symmetrical funnel plot



(b) Asymmetrical funnel plot

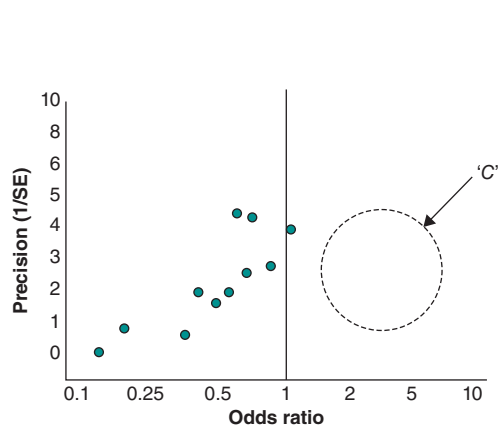


Figure 9.2.1 Funnel plots from two different systematic reviews.

publication bias, we expect to see a symmetrical funnel plot, as illustrated in Figure 9.2.1(a). We will consider this first and then look at an asymmetrical funnel plot illustrated in Figure 9.2.1(b), which is suggestive of publication bias.

In Figure 9.2.1(a), larger studies with greater precision (denoted by 'A') have a narrow spread, whereas smaller studies (denoted by 'B') scatter widely at the bottom of the graph due to the lack of power to estimate effects precisely. In the absence of publication bias, the plot resembles a symmetrical inverted funnel because the large (precise) studies all have estimates fairly close to the true effect (the OR is a little below 1.0 in this example), while the less-precise, smaller studies have greater variability, but these should still be scattered symmetrically either side of the true effect. We can see that this is indeed the case, and there are more or less the same number of studies (whether large or small) scattered on either side of the true effect.

In Figure 9.2.1(b) we have an asymmetrical funnel plot because studies that have found against a beneficial treatment effect (i.e. studies with an OR greater than 1.0, denoted by the area of the plot labelled 'C', have not been included in the systematic review, and this is particularly the case for small and medium-sized studies (those with lower values on the y -axis). We expect to see a number of medium-sized and smaller studies to be in this area, to balance those on the side with ORs below 1.0; the absence of the former group of studies is evidence of possible publication bias.

If the funnel plot is not symmetrical, as in Figure 9.2.1(b), we should not calculate an overall estimate from the combined studies. The asymmetry may be due to publication bias: For example, studies demonstrating no effect or the 'opposite' effect have not been published and/or small studies have not been published unless significant results or sizeable risks or protective associations have been found. If the search did not cover unpublished studies, these missing effect estimates will not have been included in the systematic review.

We mentioned that an asymmetric funnel plot is evidence of *possible* publication bias, as there are other explanations for this finding. For example, smaller studies are often conducted and analysed with less methodological rigour than larger studies. An asymmetrical funnel plot could therefore also be the result of exaggerated treatment effects of smaller studies of lower methodological quality.

Statistical Methods for Detecting Funnel Plot Asymmetry

Statistical methods are commonly used to detect whether funnel plot asymmetry is present. This translates the graphical approach given by the funnel plot into a statistical model. The two main statistical approaches include a rank correlation method proposed by Begg and Mazumdar (1994) and a linear regression method proposed by Egger *et al.* (1997).

The rank correlation method examines the association between the effect estimates and their variances (or their standard errors), whereas the linear regression approach is equivalent to a weighted regression of effect estimate (e.g. log OR) on its standard error, with weights inversely proportional to the study variance. Since both approaches look for an association between the study's treatment effect and its standard error, they can be seen as statistical analogues of funnel plots.

For both methods, evidence of possible publication bias (corresponding to an asymmetrical funnel plot) is indicated by a p -value of less than 0.05. However the sensitivity of both methods has been found to be low in meta-analyses based on fewer than 20 studies, and this can result in evidence of publication bias being missed, a false-negative test result (type II error). For this reason, a higher significance level (e.g. $p < 0.1$ or $p < 0.2$) is often taken to judge statistical significance. Of course, increasing the significance level causes a greater likelihood of detecting a false-positive result (type I error), that is, falsely identifying publication bias.

Please now read the following excerpt taken from the methods section of Paper A, describing the approach used to carry out a meta-analysis. For now, concentrate on how the publication bias was assessed, and we will return to the issues of heterogeneity and random and fixed-effect models later in this section.

Methods

Meta-Analysis

Risk estimates from the selected studies for each category of intervention were pooled in meta-analyses by use of STATA software (version 8; STATA Corporation, College Station, TX, USA). Random-effects models and fixed-effect models (which both use a form of inverse variance weighting) were generated for each analysis. Random effects models were used to summarise the relative risk estimates if the test of heterogeneity for a group of study results was significant (defined conservatively as $p < 0.20$). In the absence of heterogeneity, fixed-effects models were used. Publication bias was explored through the use of Begg's test, and a result with a p -value less than 0.20 was defined, a priori, to indicate the possible presence of bias.



Self-Assessment Exercise 9.2.1

1. What approach was used to ascertain whether publication bias might have occurred?
2. How do you think the authors' choice of a p -value of <0.2 would have affected their judgement of whether publication bias had occurred?
3. Refer back to Table 9.1.3 (Table 1 from Paper A, in Section 9.1.8), and in particular the paragraph at the bottom of the table. Did the authors identify whether there was any publication bias in their review of hygiene interventions?

Answers in Section 9.6

9.2.3 Heterogeneity

By **heterogeneity** we mean 'not of the same type'. In the previous section we discussed how studies with a similar design should be selected by set criteria.

By doing this, we aim to avoid including studies in the review that differ in important aspects of design, exposure or intervention, outcome, and so on, although we still record and comment on any such differences in the study summary table and accompanying text of the review (Section 9.1.8).

Having selected studies for the review, you need to decide whether or not it is appropriate to continue to meta-analysis. In addition to examining the funnel plot for asymmetry, it is also important to assess the extent to which the actual results of the various studies differ, known as **statistical heterogeneity**. If these differ too much, pooling the results is likely to be misleading, since the studies might actually be measuring different effects. Thus, statistical heterogeneity of results across studies means that the estimates from individual studies have different magnitudes, or even different directions (e.g. some showing increased risk, others showing reduced risk).

Statistical heterogeneity may be caused by recognisable differences in treatment or subjects in a trial or by methodological differences, or it may be related to unknown or unrecorded study characteristics. Whether or not we can identify the reasons, we can assess the heterogeneity of study estimates of effect by

- Looking to see whether the 95% CIs for each of the studies included in a systematic review overlap. This can be ascertained by looking at individual study results on a **forest plot** (discussed in Section 9.2.5, presentation of results).
- Carrying out an hypothesis test to assess whether there is evidence of statistically significant heterogeneity in the results of the studies. As we shall see, it is possible to use the **critical value** of the hypothesis test to quantify statistically the amount of heterogeneity between studies included in a meta-analysis.

The hypothesis test for assessing whether there is statistically significant heterogeneity between studies to be included in a meta-analysis is a version of the **chi-squared test** known as **Cochran's Q** (producing the **Q statistic**). The test examines whether the observed variability in effect sizes of included studies is within the range that can be expected if all the studies shared a common population effect size. Because the value of the Q statistic has a chi-squared distribution, results of meta-analyses frequently quote the value of the Q statistic from a test of heterogeneity as a chi-squared statistic, χ^2 .

A *p*-value for the Q statistic is often quoted as an indication of the extent of between-study variability. However, as with the statistical test for funnel plot asymmetry, the sensitivity of the Q statistic is low when only a few studies (i.e. $n < 20$) are included in the meta-analysis, so that the test could fail to detect even a moderate degree of heterogeneity. To compensate for this poor sensitivity, a higher significance level is usually taken (e.g. $p < 0.1$ or $p < 0.2$) for statistical significance. In most published meta-analyses, the test for heterogeneity is non-significant, but this cannot be interpreted as evidence for **homogeneity** of all the results in the selected studies: Recall that we can never say that we accept the null hypothesis, but only that there is insufficient evidence to reject it.

It is, however, possible to quantify the amount of statistical heterogeneity between studies by using the Q statistic to calculate the proportion of total variability between studies explained by the heterogeneity, over that occurring by chance. This is done by calculating the **I² statistic**, calculated by the formula

$$I^2 = 100 \times \frac{Q - df}{Q}$$

In the formula, Q represents the Q statistic for the hypothesis test for heterogeneity, and *df* (degrees of freedom) represents the number of studies minus 1 included in the meta-analysis. Negative values of the I² statistic (which arise if the degrees of freedom are larger than the value of Q) are put to zero so that the I² statistic lies between 0 and 100 per cent. A value of 0 per cent indicates no observed heterogeneity, and larger values show increasing heterogeneity. As a general rule, low, moderate, and high values of the I² statistic are assigned to 25, 50, and 75 per cent, respectively, to aid interpretation of this test result.

To some extent we must make a judgment about whether the differences between the studies described in the review are acceptable and of practical importance. The I² statistic helps us make this decision, but we must also try to understand what methodological and other features of the studies have led to the observed heterogeneity. Not surprisingly, there is often room for disagreement about whether results should be pooled.



Self-Assessment Exercise 9.2.2

1. Referring back to the previous excerpt describing the methods of meta-analysis of Paper A (Section 9.2.2), how did the authors investigate whether there was statistical heterogeneity between the studies included in their meta-analysis?
2. Look again at the last paragraph of Table 9.1.3 (Table 1 of Paper A) in Section 9.1.8. Did the authors identify significant heterogeneity between studies included in their meta-analysis of hygiene interventions?

Answers in Section 9.6

9.2.4 Calculating the Pooled Estimate

The quantitative summary of study results is generally considered the most important conclusion from a meta-analysis. The choice of method for calculating this pooled estimate depends mainly on whether or not significant heterogeneity has been identified.

The Fixed-Effect Model

If we are confident that the studies in our review are all providing similar estimates (that there is no significant heterogeneity between the studies), then we can calculate a common estimate from all the data. This is known as a *fixed-effect* meta-analysis. The summary measure used in a fixed-effect meta-analysis depends on the outcome of interest, such as a RR, an OR, or a difference in means.

The pooled estimate is essentially a summary measure of the results of the selected studies in the review. It is not simply an average, however, as different studies do not all provide information of equal value. It is therefore a *weighted* combination where the weight given to each study is related to the inverse of the variability within the study. This means that more account is taken of the more-precise studies (larger sample sizes, more information) than the studies with low precision (smaller samples, less information).

There are a number of different methods for calculating a pooled estimate by a fixed-effect model. Specific details are not given here and are beyond the scope of this book, but methods you may commonly see referred to are the Mantel–Haenszel, Peto, and general inverse variance-weighted methods. These are further described by the Cochrane Collaboration’s Statistical Methods Group (Deeks and Higgins, 2010), including their use with Review Manager Software 5 (see Section 9.5.4).

The Random-Effects Model

If we suspect that there is moderate or substantial heterogeneity between the studies, an alternative approach is required to obtain a pooled-effect estimate. This alternative is the *random-effects* meta-analysis. This approach assumes a different underlying effect for each study included in the meta-analysis and takes this into consideration as an additional source of variation. Effects are assumed to be randomly distributed, and the central point of this distribution is the focus of the combined effect estimate. In other words, the effects for the individual studies are assumed to vary around some overall average effect.

Under the random-effects model, individual effect sizes are assumed to have a normal distribution with mean and variance (this variance is denoted by the Greek letter τ^2 , pronounced ‘tor-squared’). This variance (τ^2) is taken into account when weighting each individual study, with the result that weights are smaller and more similar to each other than the weights in

fixed-effect models that are based on the sample sizes of the individual studies. This means that random-effects meta-analyses are more conservative (the CIs are wider) than fixed-effect meta-analyses. It also means that random-effects models give relatively more weight to smaller studies than is the case with fixed-effect models.

The most commonly used approach for conducting a random-effects meta-analysis is the method of DerSimonian and Laird (1986). Again, specific details of this approach are beyond the scope of this book.

If we identify significant heterogeneity between studies selected for a meta-analysis, we should always exercise caution before using random-effects models. Quite often the heterogeneity between studies is overlooked when using a random-effects model. In fact, rather than ignoring heterogeneity, we should look carefully at the possible reasons why it occurred.

Please now read the following excerpt from the results section of Paper A, describing the meta-analysis of hygiene interventions.

Results

Hygiene

Although the studies show a wide range of effectiveness, the summary meta-analysis suggests that hygiene interventions act to reduce diarrhoeal illness levels (random-effects model pooled estimate of relative risk 0.63, 95 per cent CI 0.52 to 0.77), although there is some evidence of publication bias (Begg's test $p < 0.20$). Reanalysis of the data after exclusion of studies thought to be of poor quality resulted in a pooled estimate of the relative risk of 0.55 (95% CI 0.40 to 0.75).



Self-Assessment Exercise 9.2.3

1. Referring to the text excerpt above and the last paragraph of Table 9.1.3 (Table 1 of Paper A, Section 9.1.8), what method of pooling the study results was used for the analysis of the hygiene interventions? Why was this approach adopted?
2. What did the authors' main meta-analysis for the hygiene interventions reveal?

Answers in Section 9.6

9.2.5 Presentation of Results: Forest Plot

As we saw in Table 9.1.3 (Table 1 of Paper A) for the hygiene interventions, the results of a meta-analysis can be displayed in tables showing the results for each individual study, the pooled summary estimate and CI, and the results of the tests for heterogeneity and publication bias.

The estimates and CIs for each study and the pooled results should also be illustrated pictorially in a *forest plot*. Figure 9.2.2 (Figure 2 from Paper A) is a forest plot illustrating the results of the meta-analysis of studies investigating one of the other objectives of the review, namely the effect of household treatment water quality interventions on diarrhoea. We now look in more detail at the information provided and how this should be interpreted.

Interpretation of a Forest Plot

The vertical solid line represents no difference between the two groups, as, for example, when the relative risk equals one. Each rectangle represents the results of one trial in terms of its effect estimate, the rectangle having an area that reflects the weighting given to each study. The horizontal line through each of the rectangles is the 95% CI for each effect estimate (note

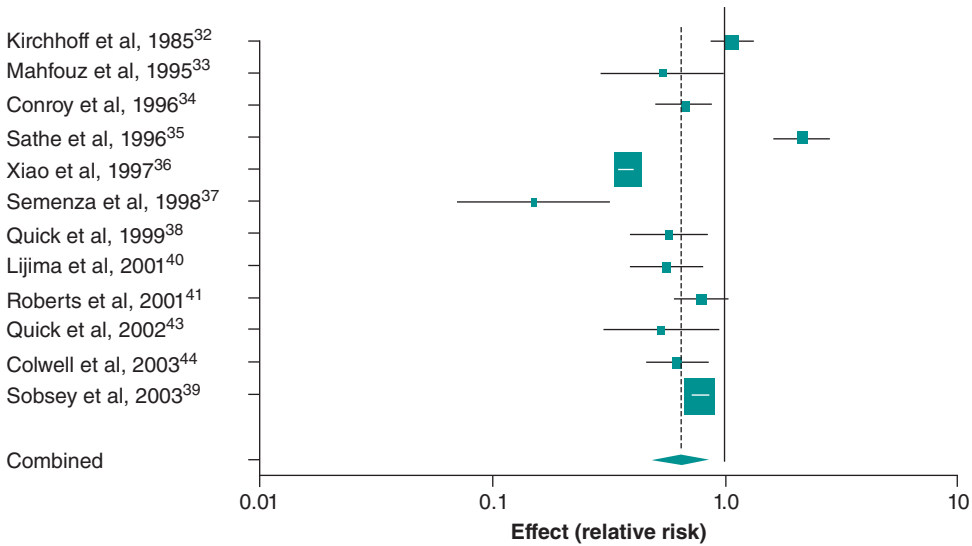


Figure 9.2.2 Random-effect meta-analysis of household treatment water quality interventions (Figure 2 from Paper A). *Source:* Fewtrell 2005. Reproduced with permission of Elsevier.

that for the two largest studies by Xiao *et al.* and Sobsey *et al.*, the 95 per cent CI is shown within the borders of each study rectangle). Estimates to the left of the vertical solid line indicate a protective effect of the intervention with (in this case), a relative risk (RR) of less than 1.0. Estimates to the right of the vertical line represent results that are associated with an increase in risk ($RR > 1$).

The final result illustrated with a diamond shape illustrates the combined result. The central points of the diamond indicate the pooled estimate, and the lateral points indicate the 95 per cent CI around that estimate. In the figure, the dashed vertical line aids in locating the pooled estimate relative to the individual study estimates. The combined result generally has a narrow CI relative to any one study, indicating the increased power of the combined analysis, and is usually placed at the bottom of the forest plot. Note that in a random-effects meta-analysis, the confidence interval for the pooled estimate might be wider than for large individual studies that are not given as much relative weight as they would be in a fixed-effects meta-analysis, which is the case in Figure 9.2.2.

Figure 9.2.2 illustrates well the presentation of results for a random-effects model. We can see that the 95% CIs for the studies conducted by Sathe *et al.* (1996) and Semenza *et al.* (1998) do not overlap with the other studies included in the meta-analysis. This is visual confirmation that the results of studies included in this meta-analysis are heterogeneous. All of the studies, except the outlier conducted by Sathe *et al.* (1996), are consistent with a protective effect of the household treatment water quality interventions. The pooled result (diamond) indicates a statistically significant protective effect for the interventions.

9.2.6 Sensitivity Analysis

A final stage in carrying out a meta-analysis is to investigate how sensitive the results of the meta-analysis are to the inclusion of studies of differing size, quality, and other specified methodological differences. This is known as *sensitivity analysis*. A sensitivity analysis can involve repeating the analysis on *subsets* of the original data as well as determining how any one study (or group of studies with similar attributes) might influence the overall summary

statistics. One very large study might have a profound, and perhaps misleading, influence on the overall result if, for example, it has serious methodological limitations. Sensitivity analysis can provide insight into the individual study factors that can affect the results and that will be important to consider in future studies.

We saw earlier how poor methodological quality can distort the findings of a meta-analysis (Table 9.1.2), and assessment of the influence of methodological quality on pooled estimates should be considered an integral part of the process. Sensitivity analysis can be conducted on separate aspects of study quality, as was done in Table 9.1.2. It is also possible to carry out an analysis to study the effects of multiple aspects of quality on treatment effects, for example, by using a composite quality score, or by meta-regression, although the methods for this are beyond the scope of this book.

Let us now return to how methodological quality and the impact on the meta-analysis were assessed in the water and sanitation review. Please now read the following excerpt taken from the discussion section of Paper A, which considers the quality of studies included in the review.

Discussion

In general, estimates calculated after the removal of poor quality studies from the meta-analyses indicated a stronger effect of the intervention. Although some of the studies identified in this review pre-dated the methodological critiques provided by Blum and Feachem and Esrey and Habicht, poorly done or poorly reported studies still make up a substantial part of the literature with 32% (12 of 38) of the identified studies classified as poor, with 50% (six of 12) of these being published after 1990. In addition to the studies classified as being of poor quality, data could not be extracted from 17% (eight from 46) of the studies identified. It seems clear that this research agenda would benefit from further guidance in terms of issues to be examined, reiteration of quality considerations, and guidelines in terms of reporting and presentation of results.



Self-Assessment Exercise 9.2.4

According to the excerpt from the discussion (above) and the previous excerpt describing the results of Paper A (Section 9.2.4):

1. How did the authors address the possible influence of methodological quality on pooled estimates derived from meta-analyses?
2. Was there evidence that methodological quality had influenced the pooled estimates?

Answers in Section 9.6

9.2.7 Statistical Software for the Conduct of Meta-Analysis

The Cochrane Collaboration has developed software called *Review Manager* (RevMan) that is free to download and is used for preparing and maintaining Cochrane Reviews (see Section 9.5.4); the latest version at the time of writing is *RevMan 5.3* (released on 13 June 2014). This bespoke software can be used to conduct both fixed-effect and random-effects meta-analyses for a range of outcomes (e.g. dichotomous, continuous, and ratio). The software will analyse statistical heterogeneity between studies (e.g. I^2 statistic), will present forest plots for the meta-analyses (overall and for specified subgroups and sensitivity analysis), and can produce a funnel plot of studies included in the meta-analyses. At the time of writing, *RevMan* cannot be used to

calculate statistical asymmetry of funnel plots (i.e. to assess publication bias), and other statistical software is required to conduct Begg's and Egger's tests of funnel plot asymmetry (for example *Stata* (StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP), or *R* (R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>).

9.2.8 Another Example of the Value of Meta-Analysis – Identifying a Dangerous Treatment

The following example is taken from a 1998 paper on the treatment of critically injured patients and contains a salutary lesson on the value of meta-analysis (Cochrane Injuries Group Albumin Reviewers, 1998). This is a systematic review of randomised trials comparing the administration of albumin with no albumin in critically ill patients. Patients in the studies had one of three types of problem following severe injury: hypovolaemia (reduced blood volume), burns, or hypoalbuminaemia (reduced blood albumin level). The outcome measure was mortality from all causes. Thirty-two trials met the criteria for inclusion in the review. Deaths occurred in 24 of the studies, and a meta-analysis of these studies was carried out. A funnel plot of the results of the included studies is shown in Figure 9.2.3, and Exercise 9.2.5 provides an opportunity to interpret this.

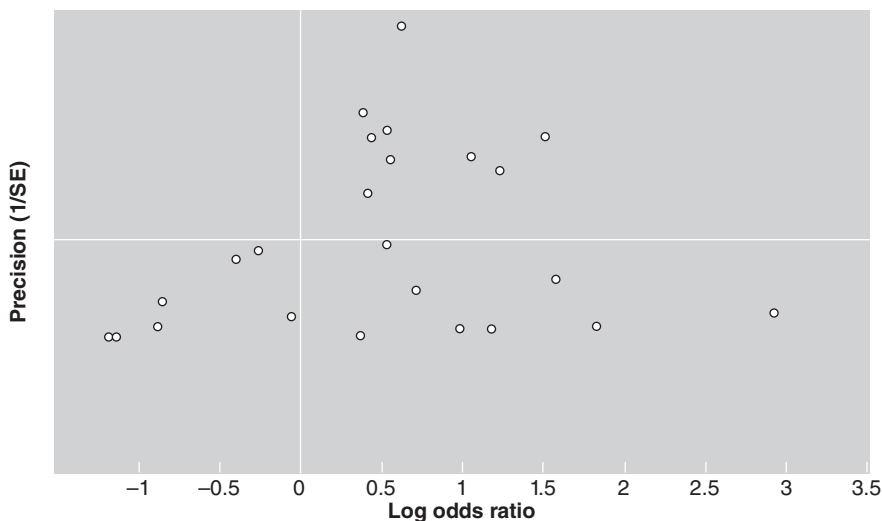


Figure 9.2.3 Funnel plot of the 24 trials in which deaths occurred. *Source:* Cochrane Injuries Group Albumin Reviewers 1998. Reproduced with permission of BMJ Publishing Group Ltd.



Self-Assessment Exercise 9.2.5

1. Describe the appearance of the funnel plot in Figure 9.2.3 (note the x-axis shows study effect estimates as log odds ratios, and therefore a log odds ratio of 0 is no effect (OR = 1).
2. What are the likely implications of this pattern for publication bias?

Answers in Section 9.6

Figure 9.2.4 shows how the results of the meta-analysis are presented in a forest plot. The results from each trial are presented separately, grouped according to type of injury.

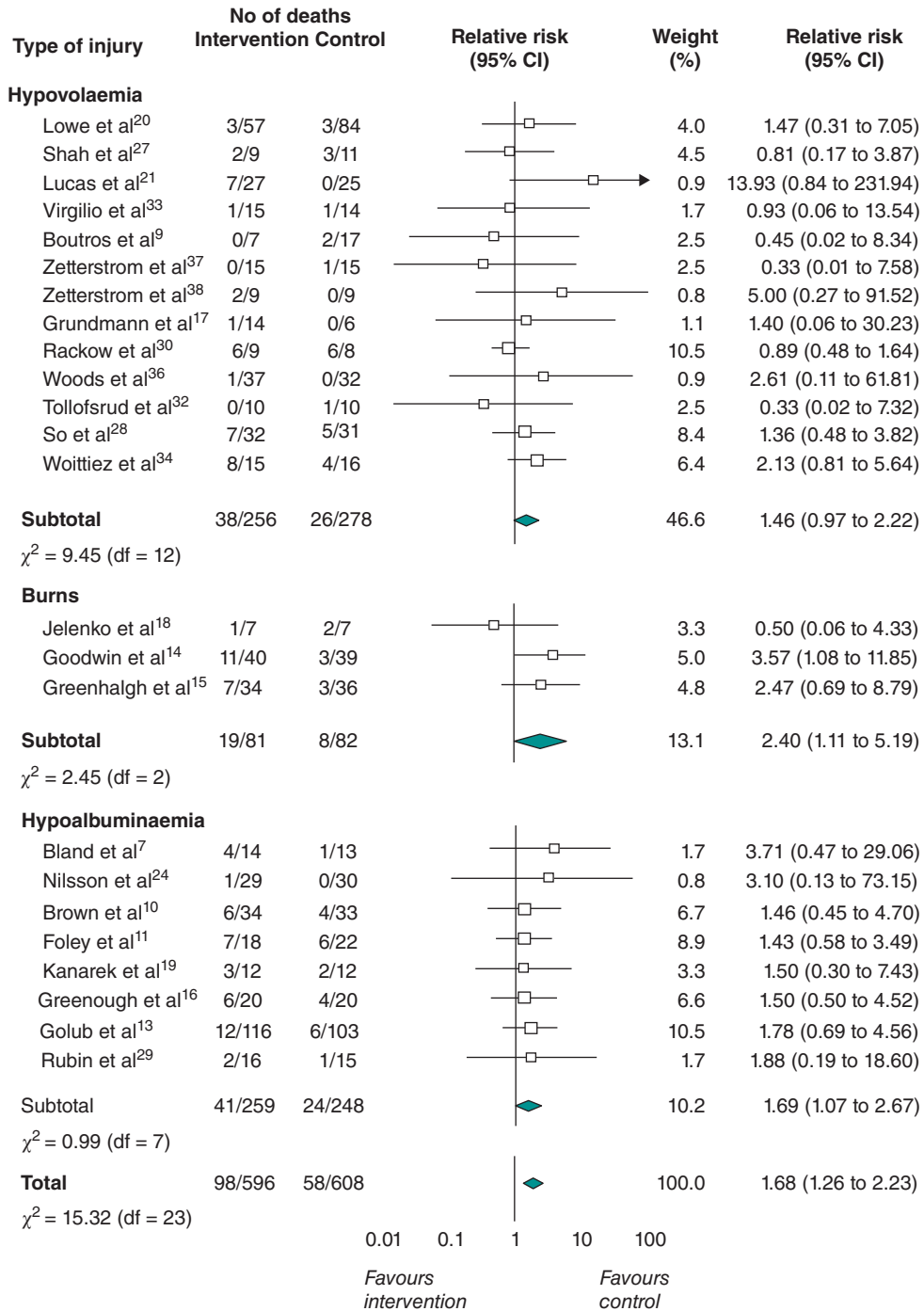


Figure 9.2.4 Meta-analysis of relative risk of death associated with intervention (albumin) compared with control (no albumin) in critically ill patients. *Source:* Cochrane Injuries Group Albumin Reviewers 1998. Reproduced with permission of BMJ Publishing Group Ltd.

For example, the first trial is by Lowe *et al.* (1977), and the results were as follows (Table 9.2.1):

Table 9.2.1 Initial albumin meta-analysis.

	Deaths	Survivals	Total
<i>Albumin</i>	3	54	57
Control	3	81	84
Total	6	135	141

The estimated relative risk (RR) of death for those given albumin compared with controls is

$$RR = \frac{3/57}{3/84} = 1.47$$

and the 95 per cent CI is (0.31, 7.05). This interval is very wide and includes 1, corresponding to no treatment effect.

We saw that in a forest plot, the square shows the estimated treatment effect (i.e. the relative risk) for each trial, the horizontal line through the square is the 95 per cent CI, and the area of each square represents the amount of information (number of subjects) contributed by that trial to the pooled estimate. The solid diamonds are pooled estimates. There is one for each subgroup and an overall estimate for all types of injury, labelled 'total'. As we identified previously, the centre points of each diamond indicate the pooled estimate of the relative risk, and the lateral points of the diamond show the 95% CI. The following exercise is based on this example and will help consolidate your understanding of meta-analysis, including the interpretation of the forest plot.



Self-Assessment Exercise 9.2.6

1. In the forest plot (Figure 9.2.4), do any of the 95 per cent CIs for individual studies not overlap? If so, which studies? What do you conclude from this?
2. The chi-squared statistics at the bottom of each section of the forest plot (below the 'subtotal' lines) and for the whole set of studies (below 'total') are derived from the Cochran's Q test for heterogeneity. *P*-values are not given, but the values (with the stated degrees of freedom) all have a *p*-value of >0.1. What do you conclude from this? Are your conclusions consistent with those you arrived at from looking at the 95% CIs in question 1?
3. The amount of heterogeneity can be estimated with the I^2 statistic, which was described in Section 9.2.3. Calculate this statistic from the information on the test for overall heterogeneity ($Q = 15.32$, $df = 23$), and interpret the result.
4. Given the findings so far, what type of statistical model should be used for meta-analysis?
5. Interpret the pooled estimates for hypovolaemia, burns, hypoalbuminaemia, and total. What do you conclude?

Answers in Section 9.6

Unfortunately, a *sensitivity analysis* was not conducted as part of this meta-analysis, so it is not possible to investigate how different aspects of study quality might have influenced the pooled estimate.

Conclusion – Albumin Administration was Increasing the Risk of Death

The clinical implications from this review bear some emphasis. The meta-analysis found that the administration of albumin in critically ill patients was associated with a significant increase in mortality, which was not identified in any single previous study. When interpreting these findings, it is necessary to appreciate whether this association is evidence of a *causal* relationship between the intervention (albumin administration) and the outcome (risk of death). The meta-analysis comprised randomized control trials and so would not have been affected by confounding and, if the trials were conducted appropriately (e.g. blinding, intention-to-treat analysis), would not have been subject to significant bias (we would need additional information on the quality of these RCTs to make this judgment). Therefore it is not unreasonable to conclude that, given it is likely that albumin had been given routinely to patients for some time prior to this meta-analysis, earlier recognition of the harmful effects of albumin administration might have avoided unnecessary deaths.

Summary: Meta-Analysis

- A meta-analysis is the statistical analysis of the data from studies included in a systematic review.
- It is an efficient way of analysing and interpreting the results from many studies.
- The main steps and techniques of a meta-analysis are as follows:
 1. generate and review a funnel plot and/or statistically assess whether there is funnel plot asymmetry;
 2. carry out and review tests of heterogeneity (Cochran's Q-statistic and I^2 statistic);
 3. if there is no significant heterogeneity, derive a pooled estimate and the 95% CI using fixed-effect meta-analysis;
 4. if there is significant heterogeneity, derive a pooled estimate and 95% CI using random-effects meta-analysis; an attempt should be made to explain any heterogeneity;
 5. present the results in the conventional format, namely the forest plot;
 6. carry out a sensitivity analysis to investigate the effects of study quality and/or different characteristics of interventions or risk factors being studied.
- Meta-analyses help to formulate hypotheses for future work.
- Meta-analyses help to derive consensus for clinical treatments and other interventions.

9.3 Systematic Reviews and Meta-Analyses of Observational Studies

9.3.1 Introduction

Traditionally, systematic review and meta-analysis methods have been used for pooling the results of RCTs of, typically, medical interventions. However, of growing importance – and becoming increasingly more common – are systematic reviews and meta-analyses of observational studies.

9.3.2 Why Conduct a Systematic Review of Observational Studies?

As we have seen in previous chapters, there are many research questions for which randomised intervention studies would be inappropriate. For example, we saw how a cohort study design was utilised to investigate the protective effect of vigorous physical activity in relation to cancer

(Chapter 5). We also saw how a case–control study design was adopted to investigate how sleeping position might be associated with late stillbirth (Chapter 6). Clearly, for both of these research questions, and for many others, trials would be at best difficult and in most cases impractical and/or unethical.

For these other research questions and potential health interventions, systematic review and meta-analysis methods can offer the same benefits as we have seen for data derived from intervention studies, by summarising all of the available evidence and increasing the statistical power by pooling estimates from multiple studies. In principle, these methods can be applied to observational studies, but there are important issues arising from key differences in study design that we need to take into account.

9.3.3 Approach to Meta-Analysis of Observational Studies

Observational studies of risk factors and potential interventions often investigate relationships with relatively small underlying risks and not uncommonly report conflicting results. It is therefore very desirable to combine the results of such studies statistically, using meta-analysis, with the intention of obtaining more-precise and definitive answers. However, extra care needs to be taken when combining the results of observational studies in this way due to methodological problems inherent in this type of study design.

As we saw in earlier chapters, observational study designs are susceptible to *bias* and *confounding*. Bias is particularly associated with case–control studies, although bias can also affect trials; confounding exclusively affects non-randomised studies. Although good study design and appropriate analysis can reduce these problems in observational studies, we need to be aware that it is generally difficult to avoid them altogether. This is particularly the case with confounding, and even after careful design (e.g. matching) and thorough adjustment in analysis, there may well be some *residual confounding*.

We should keep in mind that meta-analyses are carried out with the assumption that the studies being combined to produce the pooled effect estimate are free from bias and confounding, and that any differences between the results of the studies are due to random variation – essentially sampling error and the effects of differing sample sizes. Bias and confounding result in effect estimates that may deviate from the true underlying relationships beyond the effects of chance. Thus, a meta-analysis incorporating biased studies or studies suffering from residual confounding leads to an incorrect effect estimate. Furthermore, if this pooled estimate is based on a reasonable number of studies of moderate size, it appears to be precise with relatively narrow CIs. This appearance of precision may lead to unjustified confidence in the result, and in turn it can lead to misleading conclusions and potentially inappropriate decisions about health care or prevention policy. Reinforcement of a biased result in this way is more likely if most or all of the studies on a given issue are biased in the same direction due to common difficulties in avoiding such bias; for example, if a similar set of confounders exists that are difficult to avoid and fully adjust for.

An example of this is given by Egger *et al.* (2003), who carried out a meta-analysis of cohort studies that had investigated the association between cigarette smoking and suicide. By pooling the adjusted effect estimates from the four available studies, they found a significant *dose–response* relationship between the number of cigarettes smoked and the risk of suicide (Table 9.3.1); the more cigarettes smoked, the greater was the risk of suicide.

A causal relationship between smoking and suicide was considered improbable, and it is likely that the social and mental states that make an individual susceptible to suicide are also related to smoking behaviour; such factors confound the relationship between smoking and suicide. The authors pointed out that even though the cohort studies had adjusted for a number of known confounders, residual confounding of factors that cannot be precisely measured is likely

Table 9.3.1 Relationship between daily cigarette consumption and suicide: pooled relative risk based on four cohort studies (Egger *et al.*, 2003).

Cigarette consumption	Relative risk	95% CI
Non-smokers	1.00	
1–14 cigarettes	1.43	1.06, 1.93
15–24 cigarettes	1.88	1.53, 2.32
25 or more cigarettes	2.18	1.82, 2.61

Adapted from Egger 2003.

to have occurred. This is unfortunately often the case for observational epidemiological studies of complex social and environmental issues.

This cautious introduction and the example from Egger *et al.* might lead us to question whether we should ever carry out a meta-analysis of observational studies. However, there are still important benefits in summarising the evidence from a systematic review with a pooled quantitative estimate of effect, so long as this is done carefully. We now look in more detail at these requirements.

9.3.4 Method of Systematic Review of Observational Studies

The methods of conducting a systematic review of observational studies are in many respects the same as described for trials in Sections 9.1 and 9.2. However, there are some key differences in approach and emphasis, and these are summarised in Table 9.3.2.

The most important additional considerations for a systematic review of observational studies relates to the assessment of methodological design issues, specifically in relation to bias and confounding. For this reason, it is necessary to provide detailed information about the types of study design selected for review. We must also consider methodological issues of quality relevant to each type of study design. For example, both cohort studies and case–control studies are susceptible to confounding; however, due to the typically retrospective nature of the collection of exposure data, case–control studies tend to be more susceptible to bias. Pre-established quality-assessment instruments, such as the Newcastle–Ottawa Scales (case–control and cohort studies) and the Liverpool Quality Assessment Tools (LQATs) (all observational study designs), allow the overall quality of different observational study designs to be summarised, whilst clearly documenting specific aspects of methodological quality relevant to each study design, for example, bias in exposure assessment due to the retrospective nature of case–control studies. This allows different features of methodological quality to be explored in *sensitivity analyses* in meta-analysis of observational studies.

9.3.5 Method of Meta-Analysis of Observational Studies

The stages in conducting a meta-analysis of observational studies are also much the same as for trials, with the main differences again relating to the methodological limitations of observational studies. The key issues are described in Table 9.3.3.

The most important additional issues for a meta-analysis of observational studies relates to the consideration of possible sources of heterogeneity between observational study results, particularly where it is planned to use a random-effects model. *Sensitivity analyses* must be conducted to investigate the influence that aspects of methodological quality, especially in relation to bias and confounding, have on the effect estimate derived from the meta-analysis.

Table 9.3.2 Stages of a systematic review: intervention studies versus observational studies.

Review stage	Section described for randomised control trials	Particular issues to consider when considering review for observational studies
1. Decide on the objectives of the review	Section 9.1.3	Same process.
2. Define inclusion/exclusion criteria for studies	Section 9.1.4	Same process, but should detail types of observational studies to be included in the review (e.g. cohort, case-control, cross-sectional).
3. Search the literature (published/unpublished)	Section 9.1.5	Same process. Cochrane Controlled Trials Register (CCTR) not relevant.
4. Assess methodological quality	Section 9.1.6	Some elements are the same (e.g. minimising the potential for error and possible bias in sampling, measurement, and follow-up). Other types of bias also need to be considered in observational studies (e.g. recall/interviewer bias in case-control studies). In addition, it is important to assess how well confounding has been dealt with at the design stage or adjusted for in analysis. Composite quality assessment scales are available, tailored to individual observational designs (e.g. Newcastle-Ottawa Scale, LQATs).
5. Extract data from studies for the review	Section 9.1.7	Same process. The data extraction sheet should include information about study design (and features of the design) together with details of all confounders measured, matched (if relevant), and adjusted for.
6. Describe/present results from the review	Section 9.1.8	Same process. Summary table should be presented by type of study design if more than one observational study design is included. Quality scores should be indicated according to scale used.

Table 9.3.3 Stages of a systematic review: intervention studies versus observational studies.

Review stage	Section described for intervention studies	Issues to consider when considering review for observational studies
1. Check for funnel plot asymmetry as an indicator of publication bias.	Section 9.2.2	Same process.
2. Carry out and interpret a test for heterogeneity (Q-statistic and I^2 statistic).	Section 9.2.3	Same process. Heterogeneity is especially important for observational studies for reasons given in Section 9.3.3.
3. Calculate a pooled effect estimate using fixed-effect or random-effects meta-analysis.	Section 9.2.4	Same process, but paying particular attention to explaining heterogeneity if using a random-effects meta-analysis.
4. Present results (individual studies and pooled estimate) in a forest plot.	Section 9.2.5	Same process. The forest plot should present results separately for the different study designs included in the meta-analysis. Interpretation of the overall pooled result should be treated with caution.
5. Carry out a sensitivity analysis to investigate the influence of study attributes on effect estimate.	Section 9.2.6	Sensitivity analyses based on aspects of methodological quality is especially important for a meta-analysis of observational studies.

A good example of the importance of carrying out such a sensitivity analysis was demonstrated in a systematic review of the effect of household air pollution from using biomass fuel (e.g. wood, dung, charcoal) on lung cancer by Bruce *et al.* (2015). The authors summarized the evidence base for this relationship based on all identified observational studies (14 case–control studies). Pooling the results from all the studies identified a slight increase in the risk of lung cancer from the use of biomass fuel for cooking and heating (OR = 1.17; 95% CI = 1.01 to 1.37) and for cooking only (OR = 1.15; 95% CI = 0.97 to 1.37). Sensitivity analyses were carried out to consider studies with adequate adjustment for confounding, to address the fact that women were more likely to be exposed to household air pollution from biomass fuel used for cooking, and to account for potential bias involved in studies using a polluting reference group in their comparisons (e.g. excluding those studies comparing biomass fuel with kerosene fuel because kerosene also produces high levels of pollution). When analysis was restricted to studies with a clean fuel (e.g. gas or electricity) as a reference group, the authors found a significant 21% increase in the risk of lung cancer in men (OR = 1.21; 95% CI = 1.05 to 1.39) and more than a two-fold increase in risk of lung cancer in women (OR = 2.33; 95% CI = 1.23 to 4.42). The authors evaluated the results of the systematic review against the Hill viewpoints for causal inference (you were introduced to these viewpoints in Chapter 5, Section 5.7.1). This sensitivity analysis contributed to the authors' conclusion that 'for women, sub-analyses of higher quality studies with clean reference groups report moderately strong effects [and] support causal inference.'

Summary: Systematic Reviews and Meta-Analysis of Observational Studies

- Systematic reviews of observational studies are important as they provide a means of synthesising information about aetiological (causal) hypotheses and about interventions that cannot be tested by a trial design for ethical and/or practical reasons.
- A meta-analysis based on a systematic review of observational studies provides a useful summary of the evidence from the review as a pooled effect estimate.
- Caution should be exercised in both the conduct and interpretation of meta-analyses because observational studies are generally more susceptible than randomised control trials to bias and in particular to residual confounding.
- The methods of a systematic review of observational studies require additional information to be collected about methodological quality in terms of bias and confounding.
- The methods of a meta-analysis of observational studies require additional attention in relation to explaining heterogeneity where this exists and to sensitivity analysis to investigate the impact of methodological differences between studies. Sensitivity analysis should pay particular attention to possible effects of bias, confounding, and type of study design.

9.4 Reporting and Publishing Systematic Reviews and Meta-Analyses

The importance of systematic reviews and meta-analyses in summarizing evidence accurately and reliably is now well recognized. The utility of systematic reviews and meta-analyses is multi-factorial:

- They help inform a wide audience (e.g. practitioners, clinicians, and policy makers).
- They provide evidence for policy makers to judge risks, benefits, and harms associated with health care behaviours and interventions.
- They gather together and summarize related research for patients and their carers.

- They provide a starting point for the development of clinical practice and other guidelines.
- They provide summaries of previous research for funders wishing to support new research.
- They help editors judge the merits of publishing reports of new studies.

However, the value of a systematic review depends on how well it has been conducted and the clarity of how the review is reported. Unfortunately, there is much evidence that key information is often poorly reported in systematic reviews, and this limits their potential usefulness. As is true for all research, systematic reviews should be fully and transparently reported to allow the reader to assess the strengths and weaknesses of the investigation. To aid systematic reviewers in this process, the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement was developed with the aim of ensuring clear and transparent reporting of the planning, conduct, and results of systematic reviews (Moher *et al.*, 2009).

The PRISMA Statement comprises a checklist of 27 items to include when reporting a systematic review and meta-analysis representing all components from title, abstract, and introduction to methods, results, and discussion, with disclosure of any funding source. This is reproduced in Table 9.4.1 (overleaf), as it provides a valuable and comprehensive overview of all of the key components of a systematic review and meta-analysis, and it is useful in planning the review and in preparing the report.

In recognition of the importance of transparent and systematic reporting of the methods and results of systematic reviews and meta-analyses for policy and practice, a completed PRISMA checklist is required under the instructions to authors by many of the major peer-reviewed journals.

9.5 The Cochrane Collaboration

9.5.1 Introduction

In 1979 Archie Cochrane, a British epidemiologist, wrote:

It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials.
(Cochrane, 1979)

He recognised that people who wanted to make more-informed decisions about health care did not have ready access to reliable reviews of the available evidence. He had already suggested the establishment of a central international register of clinical trials in an earlier book, *Effectiveness and Efficiency*, published in 1972, which caused little reaction at the time. Unfortunately he never lived to see his vision realised. Neither did he see the creation of the central register he had proposed become a reality in the form of the **Cochrane Controlled Trials Register**, (now **Cochrane Central Register of Controlled Trials – CENTRAL**) which was set up some years later (see below).

Shortly before his death in 1988, Cochrane referred to a systematic review of RCTs of care during pregnancy and childbirth as ‘a real milestone in the history of randomised trials and in the evaluation of care’, and he suggested that other specialities copy the methods used. The NHS Research and Development Programme took this up, and funds were provided to establish the **UK Cochrane Centre**, which opened in Oxford in 1992. From the outset it was hoped that an international response might be established, and a year later at a meeting in Oxford in 1993, 77 people from nine countries co-founded the **Cochrane Collaboration**. At the time of writing there were 15 Cochrane Centres around the world.

Table 9.4.1 The PRISMA checklist for reporting of systematic reviews and meta-analyses.

Selection/Topic	#	Checklist Item	Reported on Page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g. Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g. PICOS, length of follow-up) and report characteristics (e.g. years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g. databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e. screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g. piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g. PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g. risk ratio, difference in means).	

Table 9.4.1 (Continued)

Selection/Topic	#	Checklist Item	Reported on Page #
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g. I^2) for each meta-analysis.	
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g. publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g. sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g. study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome-level assessment (see Item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group and (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analyses	23	Give results of additional analyses, if done (e.g. sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g. health care providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g. risk of bias), and at review level (e.g. incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g. supply of data); role of funders for the systematic review.	

9.5.2 Cochrane Collaboration Logo

The adopted logo of the Collaboration is the forest plot of the results of the systematic review referred to by Cochrane in 1988 of care during pregnancy and childbirth. Only two trials showed statistically significant effects, but the pooled data from all seven studies strongly indicated that corticosteroids reduced the risk of babies dying from the complications of immaturity. The treatment was shown to reduce the odds of the babies dying by 30–50 per cent. No systematic review of these trials was published until 1989, and so most clinicians had not realised that the treatment was so effective.



9.5.3 Collaborative Review Groups

The Cochrane Collaboration is an international organisation, consisting of a network of individuals and institutions, which aims to help people make well-informed decisions about health care and policy by preparing, maintaining, and ensuring the accessibility of systematic reviews of the effects of health interventions.

It is organised around *collaborative review groups*, each of which focuses on a particular health problem. Each group uses methods designed to minimise bias, to identify relevant trials, and to prepare and maintain reviews that are then published on the *Cochrane Database of Systematic Reviews*. At the time of writing there were 53 review groups, which covered most of the important areas of health. Each group has a co-ordinating editor and a supporting editorial team.

The many achievements of the Collaboration in such a short space of time reflect the goodwill and efforts of the individuals who contribute their time and effort, often without specific funding to do so. The organisations that do provide support tend to be public institutions such as government agencies and universities.

9.5.4 Cochrane Library

The Cochrane library consists of several databases together with a handbook on the science of reviewing research, a glossary of methodological terms, and Cochrane terminology and contact details for review groups and other groupings in the Collaboration. The entire Cochrane library is available on the Cochrane Library website (www.thecochranelibrary.com).

The databases are listed below.

Cochrane Database of Systematic Reviews

At the time of writing there were 9,201 published reviews on the database. Reviewers contribute their completed reviews and protocols to this database. Each review consists of a cover sheet giving details of the title, authors, and so on; an abstract; a structured report of the review;

discussion of the results; implications for research and practice; and a citation list and summary table of the studies included in the review, together with details of any eligible studies that were excluded, and tables of results.

Cochrane Central Register of Controlled Trials (CENTRAL)

CENTRAL was introduced in Section 9.2. It is a bibliographic database of all controlled trials identified by contributors to the Collaboration as part of the international effort to search systematically the world's health-care journals and other sources of information to create an unbiased data source for systematic reviews.

Database of Abstracts of Reviews of Effectiveness (DARE)

DARE includes structured abstracts of systematic reviews that have been critically appraised by reviewers at the NHS Centre for Reviews and Dissemination in York (UK) and by others, such as members of the American College of Physicians' Journal Club and the journal *Evidence-Based Medicine*.

Cochrane Methodology Register (CMR)

The Cochrane Methodology Register is a bibliography of articles on the science of research synthesis.

NHS Economic Evaluation Database (EED)

The NHS Economic Evaluation Database is a bibliography of published economic evaluations of health-care interventions.

Cochrane Collaboration Review Software

As discussed in Section 9.2.7, the Cochrane Collaboration has developed software to assist in preparing and maintaining reviews, called *Review Manager (RevMan)*. This allows entry of the review protocol, text commentary, characteristics of studies, comparison table, and study data. It can then be used to carry out all stages of the review, including meta-analysis, and to present the results graphically as forest plots (Section 9.2.5). RevMan is developed through a continuing process of consultation with its users. The software can be obtained (free of charge) from the Cochrane Collaboration's website (www.cochrane.org).

You can use **RevMan** for protocols and full reviews. It is most useful when you have formulated the question for the review, and it allows you to prepare the text, build tables showing the characteristics of studies and the comparisons in the review, and add study data.

RevMan is a major component of the Cochrane Information Management System (IMS), which is designed to enable contributors to the Cochrane Collaboration to meet the demands of producing high-quality systematic reviews of the evidence of the effects of health care and deliver these for publication in **The Cochrane Library** and elsewhere.

9.6 Answers to Self-Assessment Exercises

Section 9.1

Exercise 9.1.1

1. The objectives of the study are not clearly described, although the authors do state that the paper presents a systematic review of all published studies and a meta-analysis (where appropriate) looking at interventions in water quality, water supply, hygiene, and sanitation in less-developed countries. The meta-analyses were designed to provide summary estimates of the effectiveness of each type of intervention in relation to diarrhoeal disease.

- The context of the review in terms of its PICO is not fully described. Whilst not specifically describing the demographics of the target Population (age, sex, location, e.g. rural or urban), the authors state the study population incorporates 'less developed countries'. This is further defined as 'any country not within a class A region under the WHO comparative risk assessment' identified as not having very low mortality rates in both adults and children. The Interventions are identified as those involving water quality, water supply, hygiene, and sanitation, planned or occurring as natural experiments, but there is no indication what these will be Compared to (e.g. a control group with no intervention or before the intervention in before-and-after studies, or both). It is assumed that these will be existing traditional practices and facilities, but the comparison groups are not presented in the descriptive table of the included studies (Table 9.1.3). The Outcome is 'diarrhoeal disease occurring in non-outbreak conditions'.

Population	The population is not clearly stated in the abstract or introduction; however, the setting is less-developed countries (defined as any country not within a class A region under the WHO comparative risk assessment; class A countries have very low child and adult mortality).
Intervention	Interventions (planned or occurring as natural experiments) in water quality, water supply, hygiene, and sanitation (from Introduction).
Comparator	Again, this is not clearly stated, but it can be assumed that this will be absence of the intervention.
Outcome	Diarrhoea morbidity as a health outcome in non-outbreak conditions (from abstract).

- The authors mention that they will consider 'all published studies' involving interventions for inclusion in their review. This, as we shall see when we consider the implications of publication bias, might not have been the most appropriate choice for sourcing all relevant data.

Exercise 9.1.2

- There were two selection criteria. First, the studies were required to have described a specific water, sanitation, or hygiene intervention (or combination of these interventions). Second, studies had to report diarrhoea morbidity as a health outcome, measured under endemic (non-outbreak) conditions. Also, as mentioned previously, the authors required the studies to have been published 'to maintain quality (via peer review) and transparency'. We will return to this when we consider publication bias in Section 9.2.2.
- The authors state that 'no studies were excluded from the review on the basis of quality criteria alone'. They therefore included studies with potentially 'poor' quality (we will see in Section 9.1.6 that study quality relates to how well the study addresses the potential for bias and confounding that influence the study findings). Studies of poor quality (e.g. with biased results) that are included in the synthesis of findings from a systematic review (i.e. meta-analysis) have the potential to distort pooled results from the review. We will see in Section 9.2.6 how we can use *sensitivity analysis* to explore the impact of study quality on the results of systematic reviews and meta-analyses.

Exercise 9.1.3

- The authors state that only published studies were included in the review because this ensured quality via peer review and transparency.

2. There is no guarantee that studies published in peer-reviewed journals are of good quality; as we saw above, published studies identified as 'poor quality' were still included in this systematic review. On the other hand, many studies of sufficient or 'good' methodological quality are not published in peer-reviewed journals. For example, studies with non-significant or contradictory findings might be rejected for publication or not even put forward for publication.

Exercise 9.1.4

These search terms are rather limited, and there is a chance that relevant studies might be missed that did not include these key terms. For example, studies that focussed on a specific intervention like 'hand washing' without mentioning 'hygiene' in the title or abstract could be missed. The authors also include the paired term 'intervention' in their searches (meaning that studies were required to have this term in addition to the other terms). This is rather restrictive and might mean the search missed relevant studies that did not specifically define water, sanitation, or hygiene practices as an intervention. Typically, systematic reviews should create separate exhaustive lists of all terms and variants for the intervention or indicator and for the outcome before combining them in a search. This is a very important step in the review process, and these lists should be developed by the review team and piloted in at least one bibliographic database (see 'searching the literature'). Whilst the researchers were interested in endemic (non-outbreak) cases of diarrhoea, the inclusion of search terms to identify this outcome is likely to have been both impractical (finding suitable terms) and too restrictive (e.g. relevant studies might not make this distinction in their titles).

Exercise 9.1.5

1. The authors searched the main English-language databases, including Medline and EMBASE and the Cochrane Central Register of Controlled Trials. They also used previous reviews (by Esrey and colleagues) to identify early studies and carried out author-based searches to identify subsequent work by the primary investigators. In addition, they utilised foreign language journal databases (LILACS and Pascal Biomed). Hand searching the extracted journal articles for further references is also described.
2. The search is fairly comprehensive and should identify most relevant published articles. Although 'author-based' searches were carried out to identify subsequent work by the primary investigators, we are not told whether they were contacted directly to obtain more information.

Exercise 9.1.6

1. Previously established criteria were used to examine the quality of studies in the review (Blum and Feachem). However, details of these criteria and the procedure used to describe the quality of the studies, including the number of reviewers, are not given in the paper. The cited reference for Blum and Feachem describes seven methodological problems associated with studies in a previous review of water supply and sanitation investments on the outcome of diarrhoeal disease in 1983. However, it is not clear how these issues have been assessed in relation to the current review, and there is no indication in the paper how these methodological problems can be used to assess quality in other systematic reviews. Poor-quality studies were identified as having any of the following design flaws: inadequately defined control groups, no clear measurement or control of confounding factors, no specific definition of diarrhoea or the particular diarrhoeal health outcome being used, and a health indicator recall period of more than 2 weeks.

2. Although an attempt was made by the authors to assess the methodological quality of studies selected for the review, they stated that ‘no study was excluded from the review or meta-analysis on the basis of quality criteria alone’. They indicated that where possible, they would examine issues of study quality as ‘a source of possible heterogeneity between results (variation between studies)’. We will discuss the assessment of heterogeneity and sensitivity analysis when we look at meta-analysis in Section 9.2 and will return to the question of how the authors dealt with poor-quality studies at that point.

Exercise 9.1.7

1. Two types of hygiene intervention were identified: those relating to hygiene and health education (six studies) and those relating to hand washing, usually in combination with hygiene education (five studies). The authors reported that this education varied between studies, and although education on hygiene was aimed at the mothers, the health outcome was measured in the children.
2. The information for each study was tabulated by intervention, country, study quality, health outcome, age group, effect estimate used, and 95 per cent CI. This is a good example of a summary table, making it relatively straightforward to compare descriptors of the individual findings for each study from such a summary table.

Section 9.2

Exercise 9.2.1

1. Publication bias was evaluated by applying Begg’s test, and a result with a p -value less than 0.2 was defined, a priori, to indicate the possible presence of bias.
2. Begg’s test has a rather low sensitivity for meta-analyses based on fewer than 20 trials, so typically a more-conservative p -value of $p < 0.1$ (or <0.2) is used as opposed to the $p < 0.05$ level commonly used for most statistical tests. In this study the authors chose a p -value of $p < 0.20$. It is likely that they were erring on the side of caution in choosing this conservative cut-off for statistical significance, as they could be fairly certain that publication bias had not occurred with $p > 0.20$. The authors did not present a funnel plot (plotting the effect estimates against their variance) for visually appraising funnel plot asymmetry.
3. The sub-analysis on hygiene interventions did indicate the presence of publication bias, since the Begg’s test p -value was 0.19 ($p < 0.20$). However, interpretation of this result is not straightforward, as we need to consider the generally low sensitivity of the test, the fact that this sub-analysis was based on 13 studies (relatively few), and that the p -value is very close to the very conservative cut-off for statistical significance the authors used ($p < 0.20$). A cautious interpretation would be to suspect that some publication bias does exist – the visual appraisal of a funnel plot figure might have helped here.

Exercise 9.2.2

1. They carried out a test for heterogeneity (although the Cochran’s Q statistic has not been specified), setting the p -value for statistical significance to $p < 0.2$, again a conservative value to allow for the low power of the test with fewer than 20 studies (there were 11). The authors did not attempt to quantify the statistical heterogeneity between studies (using the I^2 statistic).
2. Yes, the subanalyses for hygiene interventions recorded a p -value of $p < 0.01$.

Exercise 9.2.3

1. Although the authors present results from both fixed-effect and random-effects models at the bottom of Table 9.1.3 (Table 1 from Paper A), the pooled relative risk for the random-effects model is more appropriate given the significant heterogeneity identified between the study results. From the footnote on Table 9.1.3 we can see there was evidence of significant heterogeneity between the studies ($p < 0.01$).
2. The pooled estimate of the relative risk by the random-effects model was 0.63 (95% CI 0.52 to 0.77), indicating a significant reduction in risk of diarrhoea by hygiene interventions: the intervention group were 37 per cent less likely to experience diarrhoea, with a 95 per cent CI of 23 to 48 per cent lower risk of diarrhoea. However, any conclusions should be treated with caution, because of the tentative evidence of publication bias (Begg's test, $p < 0.20$), discussed in Exercise 9.2.1. The implication might be that publication bias has occurred from the exclusion of unpublished studies from the review and/or that studies finding no effect or an opposite effect were not found and hence not included in the analysis. The authors also state that they carried out a 'reanalysis of the data after exclusion of studies thought to be of poor quality'. This is sensitivity analysis and is discussed in more detail in Section 9.2.6.

Exercise 9.2.4

1. Although the authors do not explicitly state that they carried out a sensitivity analysis, they do – in effect – conduct this by testing the effect of removing poor-quality studies on the pooled estimates for each of the intervention types.
2. For hygiene interventions, removal of the poor-quality studies led to a 13 per cent increase in the observed treatment effect, with the relative risk reduced to 0.55, compared with the 0.63 observed when all studies were included. Inclusion of poor-quality studies in the meta-analysis appears to dilute the effect of the hygiene interventions on reducing diarrhoea risk.

Exercise 9.2.5

1. The funnel plot shows the desired symmetric shape with an average effect of (log OR) between 0.5 and 1.0, equivalent to an exponentiated OR between 1.65 and 2.72.
2. The symmetrical shape is consistent with no or minimal publication bias.

Exercise 9.2.6

1. All of the 95 per cent CIs overlap, suggesting no statistical heterogeneity.
2. As the p -values are all $p > 0.1$, we can conclude that there is minimal statistical heterogeneity, particularly for the overall test, since, given the relatively large number of studies, the test has good sensitivity at this level of significance ($p = 0.1$). This conclusion is consistent with our observation from the 95% CIs in question 1.
3. The value of the I^2 statistic is -50.13 , which (see Section 9.2.3) is set to zero. This value of the statistic is interpreted as showing no heterogeneity.
4. Given the results so far, it is reasonable to use a fixed-effect model.
5. For hypovolaemia, the pooled RR = 1.46 (95% CI: 0.97, 2.22), indicating a 46 per cent increase in risk. The increased risk is not quite statistically significant, with a lower 95% CI just under 1.0. For burns, the pooled RR = 2.40 (95% CI: 1.11, 5.19), indicating that the risk is more than doubled and, despite the wide 95% CI, this increase in risk is statistically significant. For hypoalbuminaemia, the pooled OR is 1.69 (95% CI: 1.07 to 2.67), indicating a 69 per cent increase in risk. Again, the increased risk is statistically significant. Overall, the total pooled RR = 1.68 (95% CI: 1.26, 2.23), indicating a 68 per cent increase in risk, and a result that is clearly statistically significant.

10

Prevention Strategies and Evaluation of Screening

Introduction and Learning Objectives

In this chapter we look at applying some important epidemiological concepts to prevention strategies, including screening programmes, and we carry out some further exploration of methods for descriptive analysis over time – termed cohort and period effects – that can help inform prevention.

We begin by exploring concepts of risk and measures of disease burden, and we consider the implications of these for strategies of prevention. We then look at how screening programmes are evaluated, the commonly used measures of validity of screening tests, and the particular types of bias that arise in epidemiological studies of the effectiveness of screening. Finally, we examine how variations in patterns of mortality rates according to when people are born, or when they die, called cohort and period effects, can provide important information on disease causation. For the purposes of this chapter, we use the terms coronary heart disease (CHD) and ischaemic heart disease (IHD) interchangeably.

Although we will introduce some new ideas and terminology, this material mainly builds on ideas discussed in previous chapters.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Define and compare, with examples, relative risk (RR) and attributable risk (AR).
- Calculate attributable risk (you are already familiar with the calculation of RR).
- Describe what is meant by attributable fraction (AF), for exposed groups and for the population, and how these are calculated.
- Describe what is meant by years of life lost (YLL) and years lived with disability (YLD), and how these are calculated.
- Explain how YLL and YLD are used in burden of disease assessment, including the meaning and use of disability-adjusted life years (DALYs).
- Describe, with examples, the epidemiological background to high-risk and population approaches to prevention.
- Describe the advantages and disadvantages of high-risk and population approaches to prevention and the implications of these for our growing knowledge of the human genome.
- Explain why the evaluation of screening programmes is important and list, with examples, the criteria commonly used for this purpose: the Wilson and for Jungner criteria.

- Define and calculate commonly used validation criteria (sensitivity, specificity, predictive values, likelihood ratio, accuracy) and interpret information provided by the receiver-operator characteristic (ROC) curve.
- Describe the most common types of bias that arise in epidemiological studies of the effectiveness of screening programmes.
- Describe what is meant by cohort and period effects, with examples. Describe the graphical methods used to identify these effects and explain how they can contribute to identifying disease causation and the effect of preventive measures.

Resource Papers

There are two resource papers for this chapter:

Paper A

Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *Br Med J* **282**, 1847–1851 (no abstract available).

Paper B

Gunnell, D., Middleton, E., Whitley, E. Dorling, D., and Frankel, S. (2003). Influence of cohort effects on patterns of suicide in England and Wales, 1950–1999. *Br J Psychiatry* **182**, 164–170.

10.1 Concepts of Risk

10.1.1 Relative and Attributable Risk

We begin by comparing and contrasting two measures of risk, **relative risk (RR)** and **attributable risk (AR)**. We discussed RR in Chapter 5, and this was defined as incidence in the exposed group divided by incidence in the unexposed group.

We can see that RR is a ratio. For example, the RR for smoking and lung cancer in a study might be reported as 10, meaning that the risk of lung cancer among smokers is 10 times that in non-smokers. As a measure of risk, RR gives us a good idea of the increased risk, on average, that an individual faces as a result of being exposed to the factor of interest. AR tells us something equally important, but rather different. We will explore this alternative perspective on risk through the following exercise.



Self-Assessment Exercise 10.1.1

In the following table, three activities are listed, together with injury and disease outcomes that can result from these activities. Two blank columns are provided for you to complete. In the first blank column (risk to individual), make some notes on the level of risk that engagement in each of the activities poses for the individual, that is, the RR. Make this assessment in everyday language such as very low, low, high, etc.; there is no need to guess actual values of the RR. For the second column (burden to society), think about the burden that each activity poses for society as a whole (a country, for example).

Activity	Outcome	Risk to individual	Burden to society
Base jumping (from buildings, bridges, etc., with a parachute)	Death from injury due to failure of parachute to open in time or hitting buildings or cliffs, etc.		
Driving a car, or being a passenger	Death or serious injury in crash		
Smoking tobacco	Death from lung cancer or cardiovascular disease		

Answers in Section 10.5

In this exercise we have identified that:

- Some activities (e.g. base jumping) are dangerous for the people who engage in them, but they lead to a very small burden of ill health for society because they are pursued by a tiny minority of people.
- Others, such as car occupancy, can result in a substantial burden to health services and society, even though the risk to individuals on any given journey (or even over a year) is relatively low. The reason for this is that very large numbers of people travel in cars and very many journeys are made, at least in the high- and middle-income countries (and also in more-urban areas of many developing countries).
- Cardiovascular disease (CVD) is very common, so that if smoking increases risk by two- to threefold, the impact on society, in terms of the numbers of cases of, and deaths from CVD (burden of disease), will be very great.

The key elements here are the risk to the individual associated with exposure, the prevalence of the exposure in the population, and how common the disease or outcome is in the population. As we have said, the risk to the individual is measured by the RR. Another measure, AR, takes account of how common the disease outcome is and represents the incidence that can be attributed to the exposure. The AR therefore gives us a better idea of the public health impact (or burden) of the exposure.

10.1.2 Calculation of AR

The calculation of AR is quite simple. It is the difference between the incidence rate for people who are exposed and the incidence for those who are unexposed. This calculation could also have been made for mortality rates. Synonyms you may come across are the *risk difference* and the *causal risk difference*. The AR is defined as follows:

$$\begin{aligned} \text{Attributable risk (AR)} &= \text{incidence in exposed group} - \text{incidence in unexposed group} \\ &= I_{(\text{exposed})} - I_{(\text{unexposed})} \end{aligned}$$

Note that, whereas RR is a ratio of rates and therefore expressed as a number, AR is the difference between rates and is therefore expressed as a rate. You will have an opportunity to calculate and compare these two measures of risk in the next exercise.

If the association between a risk factor and a disease is causal, the AR indicates the number of cases per year (or whatever time period is used for the rate) among those exposed that can be attributed to the risk factor (cause). This is also the number that could (theoretically) be prevented if the exposure was eliminated. This notion is of course rather simplistic, since it assumes that any causes other than the risk factor being investigated have equal effects on the exposed and unexposed groups, but this is nevertheless a useful way to start thinking about what AR tells us. One can think of some examples where this is fairly clear. For example, the difference in incidence rates of liver cirrhosis among heavy drinkers and non-drinkers would, in a European country, give a good indication of the rate of cirrhosis due to heavy alcohol consumption because heavy drinking is the most important cause. We could apply a similar argument to smoking and lung cancer.



Self-Assessment Exercise 10.1.2

The following table shows incidence rates for people exposed and unexposed over a lifetime to factor A (which increases the risk of a rare cancer), and factor B (which increases the risk of ischaemic heart disease (IHD)). IHD is far more common in the population than the rare cancer.

Risk factor	Disease risk	Rate/10 000 per year		Relative risk	Attributable risk
		Exposed	Unexposed		
Factor A	Cancer	1.2	0.4		
Factor B	IHD	660	220		

1. Calculate the RR and the AR for the rare cancer (factor A).
2. Calculate the RR and the AR for IHD (factor B).
3. Comment on what you find.

Answers in Section 10.5

10.1.3 Attributable Fraction (AF) for a Dichotomous Exposure

Most disease conditions are caused or influenced by a variety of factors, even if – as in the case of lung cancer – one cause, namely smoking, is very dominant. An important question that arises in the assessment of the role of disease risk factors is the proportion of that disease that can be attributed to a specific cause. This proportion is given by the **attributable fraction** (AF), and it can be calculated for the exposed group only or for the population as a whole. The difference lies in the fact that while all people in the exposed group have the exposure (by definition), only some of the population will be exposed. The higher the percentage of the population that is exposed, the more closely will the population attributable fraction resemble that for the exposed group.

In the following explanations of AF for exposed groups and whole populations, we will begin with a simple dichotomous categorisation (yes/no) of exposure, using the example of smoking and lung cancer. Exposure is therefore stated in terms of smokers and non-smokers, for now ignoring the effects of different levels of smoking, ex-smokers, and exposure to second-hand smoking. Following these explanations and the self-assessment exercise, we will look at how to

derive AF for an exposure with more than two categories and also for a continuous exposure-risk function.

Exposed Attributable Fraction

With a dichotomous exposure, the fraction of disease (lung cancer) that can be attributed to the exposure (smoking), among those exposed (smokers) is given by the following equations (which are equivalent), numbered (1) and (2):

$$AF_{\text{exposed}} = \frac{I_{\text{exposed}} - I_{\text{unexposed}}}{I_{\text{exposed}}} \quad (1) \quad \text{or} \quad AF_{\text{exposed}} = \frac{RR-1}{RR} \quad (2)$$

The US Surgeon General's report (2004) states that the RR for lung cancer associated with average smoking consumption is 23 for men and 13 for women. Hence, using formula (2) with these RR data, the AF(exposed) for men is $23 - 1/23 = 0.957$, or 95.7%. Thus, among male smokers, more than 95% of lung cancer cases can be attributed to smoking. For women, the AF(exposed) = $12/13 = 0.923$ or 92.3%.

Population Attributable Fraction (PAF)

If we want to determine the proportion of all lung cancer in the population (for example the country) that is caused by smoking, we need to take into account that only a proportion of the population smoke, and that lung cancers in non-smokers (as well as some among smokers) are caused by other factors, for example, radon exposure in homes, air pollution, and diet lacking in fruit and vegetables. Again, assuming the dichotomous exposure categorisation, the alternative equations for the population attributable fraction numbered (3) and (4) are:

$$AF_{\text{population}} = \frac{I_{\text{population}} - I_{\text{unexposed}}}{I_{\text{population}}} \quad (3) \quad \text{or} \quad AF_{\text{population}} = \frac{P_{\text{exposed}}(RR-1)}{1 + P_{\text{exposed}}(RR-1)} \quad (4)$$

Where P_{exposed} is the proportion (prevalence) of the population with the exposure. Returning to our example and using equation (4) as $I_{\text{unexposed}}$ may not be easily available, if the prevalence of smoking is 25% and the RR for men is 23, then the PAF = $0.25 \times 22/1 + (0.25 \times 22) = 0.846$, or 84.6%. Thus, even though 'only' 25% of the male population smoke, almost 85% of all lung cancers in men are caused by smoking. This is an important conclusion, not only for smoking but also because a significant minority (and quite large numbers across a population) are due to other causes that also need attention.

Although equation (4) is most commonly used for PAF, it can lead to a biased estimate in the presence of confounding, even if the RR has been adjusted (Steenland and Armstrong, 2006). The relationship between smoking and lung cancer is so strong, that the effect of confounding in this example is likely to be small, but this would not be the case for disease conditions with multiple important and associated causes, for example if we are looking at the impact of smoking on ischaemic heart disease (the subject of self-assessment exercise 10.1.3).

The following equations, numbered (5) and (6), for PAR are not biased by confounding:

$$PAF = \frac{P_c(\text{Incidence}_{\text{exposed}} - \text{Incidence}_{\text{unexposed}})}{\text{Incidence}_{\text{exposed}}} \quad (5) \quad \text{or} \quad PAF = \frac{P_c(RR-1)}{RR} \quad (6)$$

where P_c is the proportion of cases that are exposed. Using the lung cancer example among men, if 86% of all cases of lung cancer in the population are smokers ($P_c = 0.86$) (Doll *et al.*, 2005), then using equation (6) the PAF = $0.86 \times (23 - 1)/23 = 0.823$, or 82.3%.

Another way to avoid the bias from confounding is to use a weighted sum method. In principle, this involves applying equation (4) for PAF to each stratum of the confounder using RR values specific to that stratum, weighting the resulting AF values by the proportion of cases in each stratum, and taking the sum of these. Further discussion of this topic, and alternative methods, are provided by Steenland and Armstrong (2006).



Self-Assessment Exercise 10.1.3

For this exercise, we look at the same exposure (smoking), but we look at a disease for which the RR is much lower and for which confounding is more of an issue. Let's say that, in a European country, the prevalence of smoking among men is 25 per cent and coronary heart disease (CHD) mortality for all men is 200 per 100,000 per year. From various studies, estimates of the mortality rates among non-smoking men is 150 per 100,000/year and among smoking men is 317 per 100,000/year. It has also been reported that around 30% of CHD cases in men are smokers.

1. Calculate the RR and the AR for smoking and death from CHD.
2. Calculate and interpret the PAF using equation (4).
3. If the adult male population of the country is 20 million, how many deaths could theoretically be prevented in one year if the smoking prevalence among men was reduced to 15 per cent? You will need to recalculate the PAF; use equation (4) again for this purpose.
4. It is likely that the relationship between smoking and CHD is confounded by other risk factors, including cholesterol and exercise. Calculate and interpret the PAF using equation (6), and interpret your findings.

Answers in Section 10.5

10.1.4 Attributable Fraction for Continuous and Multiple Category Exposures

In the earlier example (Section 10.1.3) we mentioned that, in using a simple yes/no classification for smoking, we ignored the important fact that risk varies with amount smoked and for ex-smokers versus never-smokers, and that second-hand smoking also carries a significant albeit lower risk for the same diseases as active smoking. We may also have a situation where data on risk is available across a continuous distribution of exposure, for example for BMI or air pollution. How do we calculate AR in these cases?

For the multiple categories, this is done by calculating the AF in each category of exposure, and then adding these values up. For the continuous exposure, we use RR values for each level of exposure – you can think of this as slicing the exposure distribution into very thin slices, each with its own RR and proportion of the population exposed to that level. Mathematically, this is done through integration.

We will not look at the formulae here, but a useful presentation of these methods, and how they have been applied, is provided in the GBD-2010 study of comparative risk assessment paper by Lim *et al.* (2012), to which we return in Section 10.1.5.

10.1.5 Years of Life Lost (YLL) and Years Lived with Disability (YLD)

We have made several references so far to the calculation of *burden of disease*. Since the 1990s, this methodology has become increasingly important for health systems and public health

agencies in understanding the extent and causes of diseases, how these are changing over time, and as a basis for prioritising resources. We have referred already to one key publication from the Global Burden of Disease (GBD)–2010 study (Lim, 2012), and we will look further at the methods used in this study and some of the findings shortly.

The concepts and measures we discussed in Sections 10.1.1 to 10.1.4 are central to the GBD study methods, and we now look at two additional measures that form the basis of disease burden summary statistics, namely *years of life lost* and *years lived with disability*.

Years of Life Lost (YLL)

YLL is a measure of the amount of life lost due to premature mortality. Although there are multiple ways of calculating YLL, the basic concept is to multiply the number of deaths (in a specified sex and age group, for a given population) by the reference standard life expectancy for that same group. For example, among women aged 50–54 in England and Wales there were 4,736 deaths in 2010; if the life expectancy (see Chapter 8 for calculation of life expectancy) for this group (average age, 52.5 years) was 30 years, a total of $30 \times 4736 = 142,080$ years of life have been lost. This has been calculated for a 5-year age group, but it can be done for each year or for any other interval so long as the data required are available.

The most important issue in calculating YLL is what to take as the reference standard expectation of life. Historically, WHO has used life expectancy from Japan, this being the highest in the world. The GBD-2010 study constructed standard life expectancy using the lowest observed death rate in each age and sex group across countries with a population of more than 5 million. This methodology assumes that, all other factors being equal, all people could reach these life expectancies if the causes of premature death were at the levels seen in those countries with the highest values for each age group.

Years Lived with Disability (YLD)

Whereas YLL quantify the life lost to premature mortality, YLD provide a measure of life lived in a state of less than ideal health through pain, other symptoms, and restrictions resulting from disease and injuries and their sequelae. An example of a sequela is renal (kidney) disease as a result of diabetes mellitus. This impaired health is collectively termed 'disability'. Although the importance of this measure of ill-health can be appreciated, so too can the complexity and reliance on social constructions of what constitutes disability, and how severe this is in both absolute and relative terms.

YLD are calculated for each disease or injury by determining the average duration (in years) of disability of the condition and its sequelae, and multiplying this by a 'disability weight'. This takes a value lying between 0 (no disability) and 1 (a state equivalent to death), and the derivation of these weights is described further below. The term sequelae is used for conditions that result from having a disease, for example the type of visual impairment that occurs in people with diabetes mellitus.

The major issues for calculating YLD, especially for countries where data on disease incidence are limited, are:

- obtaining reliable incidence rates for diseases and injuries;
- determining the sequelae for each disease and injury and not double-counting the latter; that is, if multiple diseases cause a sequela, the total amount of that sequela should not exceed what is observed, and ;
- obtaining valid disability weights, which are meaningful for the populations concerned, and globally for the GBD studies.

Among the key aspects of the methodology employed for YLDs are use of the *DisMod-MR model* to derive estimates of incidence, and a survey-based method used to obtain disability weights. A brief summary of these is provided below.

DisMod-MR

We will not look at this in detail, and you can find out more in Murray *et al.* (2012). For GBD-2010, DisMod-MR used Bayesian meta-regression methods (an introduction to Bayesian methods is provided in Chapter 11) incorporating epidemiological data (i.e. incidence, prevalence, remission, excess mortality, and cause-specific mortality) to estimate disease rates, and variables that can help explain true variation between studies of these rates and variation arising due to study design, case definitions, or diagnostic technology. This model is evolving with new iterations of the GBD study, so you may wish to review later publications as these become available.

Disability Weights

The method for developing disability weights for the GBD-2010 study are described by Salomon *et al.*, 2012. The information used was obtained in two ways. One way was through interviews ($n = 13,902$ respondents) in four developing countries (face-to-face, in home) and one developed country (by telephone); the other was by an open access Web survey available for around ten months ($n = 16,328$ participants). The method involved asking subjects to judge between hypothetical paired comparisons of selected health states, to provide an opinion on which was more healthy. This provided the basis for assigning weights. A very interesting finding was the relatively high degree of concordance between weights obtained from the various surveys and across different cultural environments. A majority of the weights reflected an assessment of mild disability with 26% below 0.05; overall, they ranged from very mild at 0.01 (mild anaemia) to the most severe at 0.76 (schizophrenia).

10.1.6 Disability-Adjusted Life Years (DALYs)

DALYs provide a single measure of disease burden that combines premature mortality with life experienced with morbidity. They are calculated by adding the disease-, age-, sex-, and country-specific values for YLL and YLD.

The following two exercises will help to consolidate what we have discussed about YLL and YLD and the application of DALYs as a summary measure of disease burden.



Self-Assessment Exercise 10.1.4

Figure 10.1.1 presents results for YLL and YLD for 1990 and 2010, for the 22 sub-regions of the world that were used for presenting the results of the GBD-2010 study.

1. Describe the patterns of variation in proportions of YLL and YLD making up the DALY totals across the sub-regions and (broadly) how these have changed between 1990 and 2010.
2. Making reference to specific sub-regions, suggest some of the possible reasons for the greatest differences in proportions of YLL and YLD making up the sub-regional DALY totals in 2010 and the pattern of change between 1990 and 2010.

Answers in Section 10.5

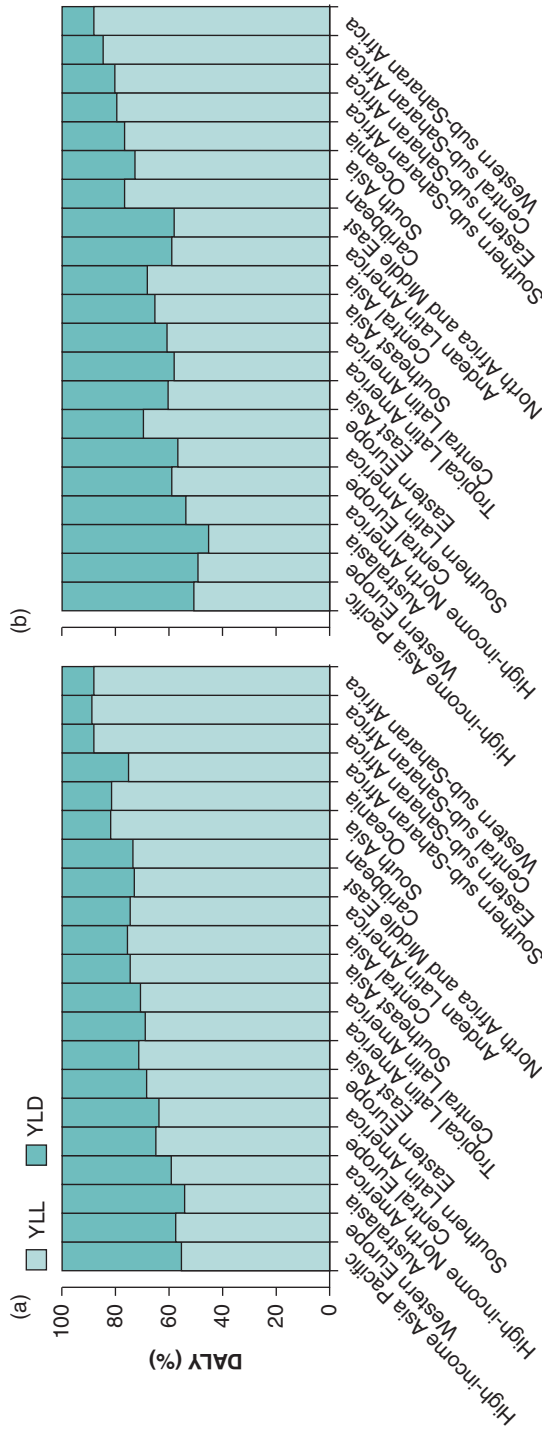


Figure 10.1.1 YLL and YLD for the 22 sub-regions used in the GBD-2010 study; data are for 1990 (a) and 2010 (b). Source: Murray 2012. Reproduced with permission of Elsevier.



Self-Assessment Exercise 10.1.5

Figure 10.1.2 shows on the following page the rank order (with 95% uncertainty intervals – UIs; these are analogous to 95% confidence intervals with which you are familiar, but they are calculated using different methods) of the 25 most important disease and injury causes in 1990 and 2010, based on the disease burden they contributed measured in DALYs. It also shows the changes in ranking, and the percentage change in number of global DALYs (with 95% UI), over the period. Causes that moved into, or dropped out of, the top 25 are shown at the bottom of the listings: for example, HIV/AIDS went from 33rd in 1990 to 5th in 2010, and measles went from 16th in 1990 to 56th in 2010).

1. Briefly highlight:
 - a. the important findings in respect of the five most important causes in 1990 and 2010;
 - b. three causes that have notably increased between 1990 and 2010
 - c. three causes that have notably decreased between 1990 and 2010
2. Using a few examples, describe the findings for causes in 2010 that have high morbidity (YLD) but relatively low premature mortality (YLL). Because YLL and YLD are not shown in Figure 10.2, you will need to make your own judgments about which causes are dominated by morbidity (YLD) rather than premature mortality (YLL).

Answers in Section 10.5

10.1.7 Burden Attributable to Specific Risk Factors

We are now in a position to see how the disease burden resulting from a specific risk factor – for example serum cholesterol or air pollution – can be calculated. For this, we need to know three key items of information:

- The disease conditions that are causally linked to the risk factor
- The total disease burden for these conditions and their sequelae
- The population attributable fractions (PAFs) for each disease resulting from exposure to the risk factor

The attributable burden for each disease is calculated by multiplying the risk factor–specific PAF by the total burden of that disease. For example, earlier we saw that the PAF for lung cancer from active smoking in men was 84.6% in a population with 25% of the population smoking. In the UK in 2010, the total (men and women combined) burden from lung cancer was 612,000 DALYs (95% UI: 498,000 to 767,000) (Murray, 2013). Using these data, the attributable burden for lung cancer caused by smoking in 2010 would be 517,750 DALYs (95% CI: 421,300 to 648,880).

Tobacco smoking is of course responsible for other diseases, including cardiovascular disease, chronic respiratory disease, and a number of other cancers. The total attributable burden is obtained by adding up the disease-specific attributable burdens. For the UK in 2010, the total attributable burden for active tobacco smoking was 1,965,000 DALYs (95% UI: 1,728,000 to 2,244,000) for men and women combined, with 1,145,000 (95% UI: 1,020 to 1356) for men alone (Murray *et al.*, 2013).

Many diseases are caused by more than one risk factor, with CVD being a good example as it is caused by smoking, high cholesterol, lack of exercise, raised blood pressure, and other factors. The sum of the attributable burdens calculated for each of the contributing risk factors typically adds up to more than the total burden of the disease (CVD) observed in the population. This

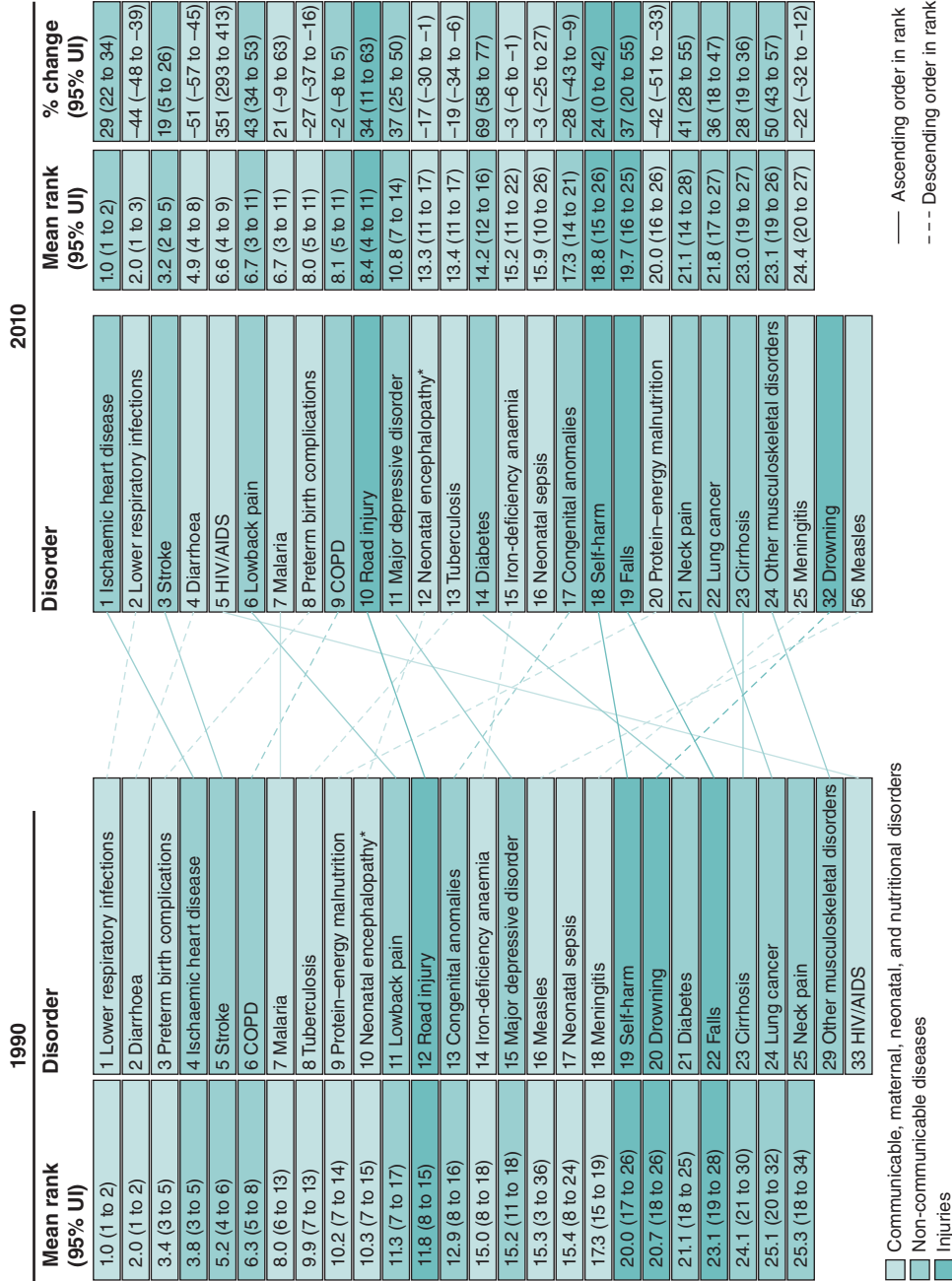


Figure 10.1.2 Global disability-adjusted life year ranks with 95% uncertainty intervals (UIs) for the top 25 causes in 1990 and 2010, and the percentage change with 95% UIs between 1990 and 2010. Source: Murray 2012. Reproduced with permission of Elsevier.

Communicable, maternal, neonatal, and nutritional disorders
 Non-communicable diseases
 Injuries
 — Ascending order in rank
 - - - Descending order in rank

does not mean that the methodology is flawed, but it does need to be kept in mind. It should also be kept in mind that attributable burdens for risk factors that are not independent in their action – that is, where the effect of one is mediated at least in part through another – they should not be added together.

Summary: Measures of Risk and Disease Burden

- RR: the ratio of the incidence rate of exposed and the incidence rate of unexposed groups (or 'more' and 'less' exposed, respectively, if we are dealing with degrees of exposure). The RR is a ratio and provides a measure of risk of being exposed for the individual.
- AR: the difference between the incidence rate among people who are exposed (or more exposed) and the incidence rate among those who are not exposed (or less exposed). AR is a rate and is a measure of the burden of disease attributable to a risk factor and that could be prevented if the risk factor was removed.
- Attributable fraction (AF): the proportion of cases or deaths that would not have occurred in the absence of a specific risk factor. This can be determined for the exposed group or for the whole population, which includes exposed and non-exposed individuals. AF can be calculated from the incidence rates in exposed and unexposed groups or from relative risks (or odds ratios if the disease is rare). Additionally, for PAF, information is required on the proportion of the population exposed or the proportion of all cases exposed.
- With the most commonly used equation for population attributable fraction (PAF), that is, equation (4) above, that uses the RR and proportion of the population exposed, the estimate may be biased by confounding. Equation (6) may be used, but this relies on an estimate of the proportion of cases that are exposed, information that may not be easily available. Alternative methods can also be used that add up weighted PAF values calculated for strata of the confounding variable.
- Calculation of the AF (for exposed groups or the population) where the exposure has more than two categories, or is continuous, employs methods that obtain the AF for each category and adds these together (for multiple categories) or integrates mathematically (for continuous exposure-risk functions).
- Years of life lost (YLL) are calculated by multiplying the mortality rate at a given age (group) by the expectation of life (in years) at that age, and adding these up across age groups. This provides a summary of life lost to premature mortality.
- Years lived with disability (YLD) are calculated by multiplying, for each age group, the average duration lived with a disease and its sequelae by the respective 'disability weights', then adding across age groups. This provides a summary of the number of years lived with symptoms, pain, or restrictions (collectively termed 'disability') resulting from a disease.
- Disability-adjusted life years (DALYs) are calculated by adding YLL and YLD, and they provide a useful measure of overall burden as this combines both premature mortality and duration of time lived with morbidity (disability).

10.2 Strategies of Prevention

10.2.1 The Distribution of Risk in Populations

We now look at why the concepts of relative and attributable risk discussed in Section 10.1 (and in particular the way that risk is distributed in the population) have important implications for prevention strategy. Much of this discussion is based on paper A. Although published in 1981,

this paper was groundbreaking in terms of understanding risk in a population and remains worth studying because it illustrates very clearly how conclusions about prevention strategies can be derived from epidemiological principles and research findings. Following our discussion of this paper, we will briefly review how these ideas have stood the test of time (Section 10.2.4) and also the implications that new knowledge about the role of the human genome in disease have for prevention strategies (Section 10.2.5).

Exercise 10.2.1 and the subsequent discussion will help your understanding of these concepts. Please now read the excerpt from Rose (1981) (paper A) headed ‘Absolute and Relative Risk’, review Figure 10.2.1 (Figure 2 from paper A), and complete Exercise 10.2.1. Absolute risk is just another way of saying this is the actual rate (mortality in this case); you can see this from the scale on the left-hand y -axis in Figure 10.2.1, which is deaths per 1,000 per year.

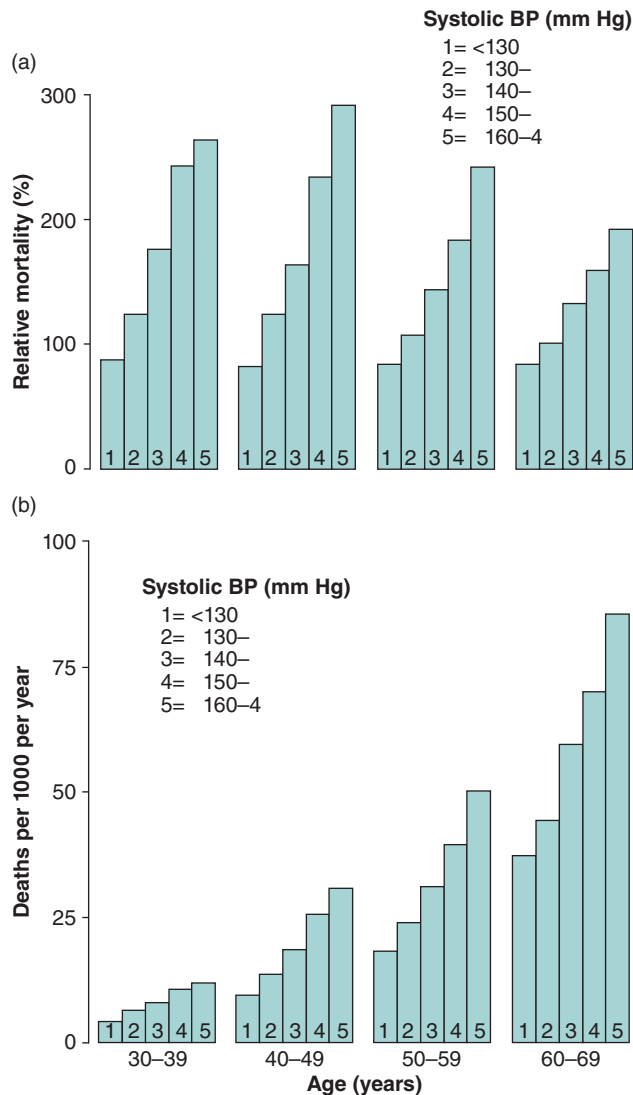


Figure 10.2.1 Age-specific mortality in men according to blood pressure and age, from life insurance data: (a) relative risk, and (b) absolute risk (Figure 2 from Paper A). Adapted from Rose 1981.

Absolute and Relative Risk

Life insurance experts concerned with charging the right premiums taught us that “high risk” meant “high relative risk,” and in this until recently they have been abetted by the epidemiologists. Figure 2(a), taken from life insurance data, shows for each of four age groups the relation of blood pressure to the relative risk of death, taking the risk for the whole of each age group as 100. The relative risk is seen to increase with increasing pressure, but the gradient gets a little less steep as age advances. That is perhaps not surprising, because a systolic pressure of 160 mm Hg is common in older men, and we would not expect it to be so unpleasant as at younger ages, when it is rare.

In figure 2(b) the same data are shown but with a scale of absolute instead of relative risk. The pattern now appears quite different. In particular, the absolute excess risk associated with raised pressure is far greater in the older men. A systolic pressure of 160 mm Hg may be common at these ages, but common does not mean good. To identify risk in relative units rather than absolute units may be misleading.



Self-Assessment Exercise 10.2.1

1. From Figure 2(b), the lower chart, estimate (as the actual values of the mortality are not given) the attributable risk of group 5 (systolic blood pressure 160–164) versus group 1 (systolic blood pressure <130), for age 30–39 years. Recall, as described in Section 10.1, AR is calculated as the difference between (mortality) rates.
2. Also from Figure 2(b) estimate the AR for group 5 versus group 1 at age 60–69 years.
3. Comment on what you find.

Answers in Section 10.5

We now look at the implications of these findings for prevention strategy. Figure 10.2.2 (Figure 3 from paper A) is central to this discussion.

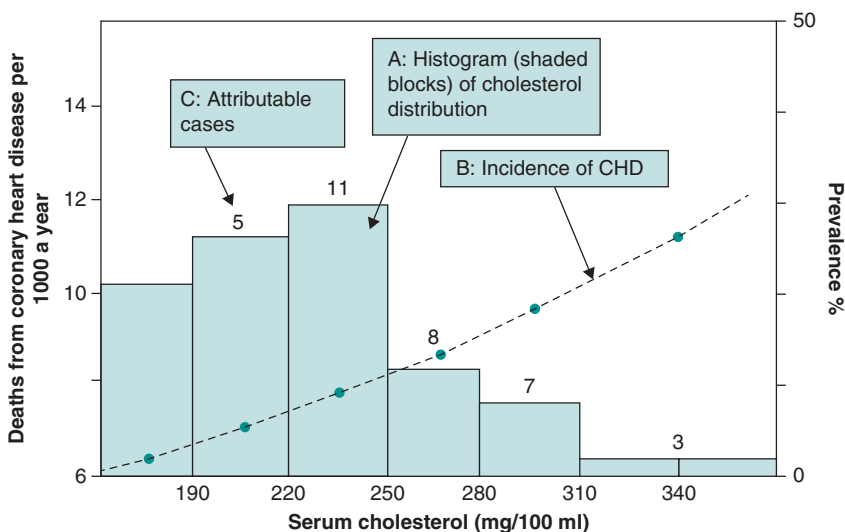


Figure 10.2.2 Prevalence distribution of serum cholesterol concentration related to coronary heart disease mortality in men aged 55–64 years. Number above each bar represents estimate of attributable deaths per 1,000 population per 10 years. Based on Framingham Study (Figure 3 in paper A).

Information in the Diagram

Figure 10.2.2 presents three distinct pieces of information, each of which is highlighted by a box and arrow and is discussed further here. First (A), there is a histogram (shaded columns) illustrating the distribution of serum cholesterol levels in the Framingham study: The scale (prevalence per cent) is on the right-hand y -axis, and you will see that the distribution is slightly positively skewed. Second (B), the dashed line rising across the histogram is the incidence (mortality) rate, with the scale on the left-hand y -axis. The rate for the highest category of cholesterol is about twice that of the lowest. Hence the RR for the highest category versus the lowest is about 2.0. Third (C), numbers are shown on top of the histogram: These are the estimated numbers of cases (deaths from CHD) that can be attributed to the respective cholesterol levels, and we discuss these further below.

Distribution of Risk

From the histogram we can see that there are very few people with cholesterol levels that are associated with the highest RR. However, the majority of the population, although not experiencing a RR as high as 2, does nevertheless have an elevated RR compared to the lowest cholesterol group, and this RR for the majority is 1.2 to 1.5 times that in the lowest cholesterol group.

Attributable Risk

When we look at the cases attributable to raised cholesterol (the numbers on top of the histogram) there is an important, and perhaps surprising, finding. These numbers have been obtained by applying the mortality rates (deaths per 1,000 population per 10 years) at each level of cholesterol to the numbers of people with those levels. We find that $5 + 11 + 8 = 24$ out of the total of 34 (70 per cent); that is, the majority of all the cases arise from the many people with 'typical' levels of cholesterol and moderately raised RR and not from the relatively few people with the high levels of cholesterol and high RR. This is a very important conclusion. In this population, 'typical' levels of cholesterol were not healthy. The implications of these observations for prevention strategy are now becoming apparent, and they are discussed further in the next section.

10.2.2 High-Risk and Population Approaches to Prevention

In paper A, Geoffrey Rose discusses two approaches to prevention, termed *high risk* and *population* (or mass). The following sections and diagrams help to explain what is involved for both approaches in terms of the distribution of risk factors and associated RRs.

The High-Risk Approach

The high-risk approach is illustrated in Figure 10.2.3. The diagram is a schematic summary based on Figure 10.2.2 (Figure 3 in paper A), and shows

- The distribution of cholesterol levels as a skewed curve, including the mean (dotted line).
- The RR as a rising dashed line.
- A cut-off level of cholesterol above which an individual is considered 'high-risk' (solid arrow with shaded portion of curve for values above the cut-off).

In the *high-risk approach*, individuals with cholesterol levels above a value associated with high risk ('level indicating high risk' in Figure 10.2.3) would be identified through screening and then managed with advice on exercise and diet and, if necessary, cholesterol-lowering drugs. These people are represented by the shaded area under the distribution to the right of the cut-off arrow.

Although this strategy is appropriate for those individuals with a high RR, you can see that it will have no impact on the majority of people with moderately raised RR. We now look at an

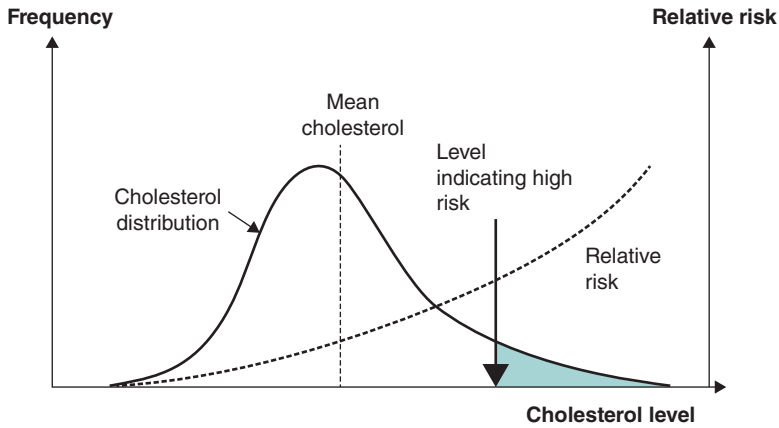


Figure 10.2.3 Schematic representation of distribution of cholesterol in the population, the relative risk (RR) across the range of cholesterol, and a cut-off level of cholesterol indicating high risk (based on Figure 3 from paper A).

alternative strategy, the population approach, which is aimed primarily at the majority of the population who have moderately increased risk.

The Population (or Mass) Approach

The population approach is illustrated in Figure 10.2.4. The diagram is also based on Figure 10.2.2 (Figure 3 in paper A), and it builds on the information shown in Figure 10.2.3. This new figure shows

- The original distribution of cholesterol levels as a skewed curve (A), including the mean (dotted line).
- As a new feature, a downward (left) shifted distribution of cholesterol levels as a skewed curve (B), including the mean cholesterol after the shift.

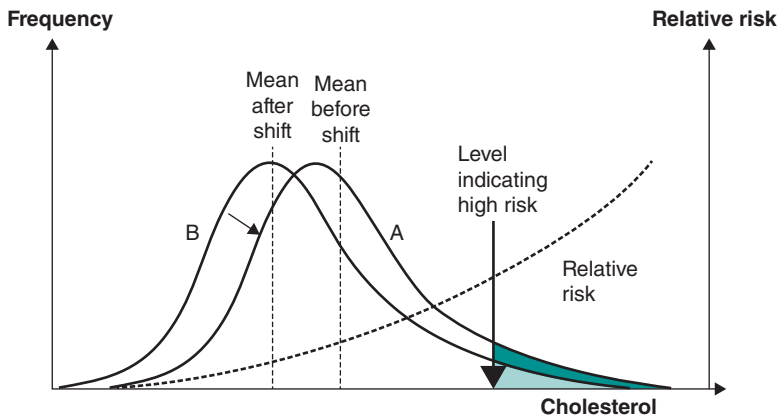


Figure 10.2.4 Schematic representation of distribution building on Figure 10.2.3 of cholesterol in the population, the relative risk (RR) across the range of cholesterol, a cut-off level of cholesterol indicating high risk, and a shift of the distribution to the left.

- The RR, as before, as a rising dashed line. Note that this does not change position or slope because the relationship between cholesterol level and RR has not changed.
- The same cut-off level of cholesterol above which an individual is considered to be at high risk.

With a population approach, measures (that we will consider shortly) are taken to move the whole distribution downwards (to the left). This has two effects:

- The RR of the majority of the population is reduced, since the central part of the distribution of cholesterol is now located over a lower portion of the RR curve. From our earlier discussion in which we found that most cases arise from the majority of the population towards the centre of the distribution, this should have a substantial impact on the overall disease burden.
- The proportion of the population that is now above the high-risk cut-off has also been reduced, as shown by the reduced shaded area under the shifted distribution curve (B). Although this proportion is likely to be reduced somewhat with the population approach, it will not be eliminated.

Both Strategies are Required

From this examination of the distribution of risk we can see that neither strategy is sufficient on its own. The high-risk approach will not address the underlying problem of moderately raised risk among the majority of the population, and the population approach will leave a tail of individuals at high risk who, although fewer, still need to be identified and offered advice and treatment as appropriate.

The following exercise will help you to think about how these two approaches would be applied in practice to prevent heart disease by action on cholesterol levels. It also provides an opportunity to start thinking about some of the practical, political, and social issues involved in implementing these strategies.



Self-Assessment Exercise 10.2.2

Make a list of the ways society might attempt to reduce the risk of IHD arising from elevated blood cholesterol through a high-risk strategy and a population strategy.

Answers in Section 10.5

This exercise perhaps raised more questions than it answered, such as

- What measures are likely to be most effective (and cost-effective)?
- Which professional groups, institutions, and agencies are best placed to implement these measures, and how?
- Are the many people with only moderately raised RR likely to take the advice being offered?
- What are the political and social implications of debate about individual freedom of choice?

Some of these issues are discussed in paper A. For a more-detailed discussion, including the relevance to other health and social issues, see Rose (1992). For a historical example of the application of the concepts of high-risk and population prevention to a range of international public health issues, see the *World Health Report 2002* (WHO, 2002). We will look at another, more recent example, shortly.

10.2.3 Safety and the Population Strategy

We have seen that the population strategy involves applying measures to large numbers of people, which may be the whole population or an important subgroup of the population. Please now read the excerpt 'Safety is paramount' from paper A, on the consequences of applying one of the early cholesterol-lowering drugs (clofibrate) to large numbers of people, and complete Exercise 10.2.3.

Safety is Paramount

The recent World Health Organisation controlled trial of clofibrate produced disturbing results. In the treated group non-fatal myocardial infarction was reduced by 26% (about the effect predicted from the fall in cholesterol concentrations). Mortality from non-cardiac causes, however, increased by one-third, an effect rather unlikely to be due to chance. This finding is important to the strategy of prevention. Clofibrate has been in use for many years and has been given to enormous numbers of patients. Until the results of this trial appeared there was no suspicion that it might kill. Indeed, by clinical standards it can still be called a relatively safe drug, since the estimate of excess mortality works out at only about one death per 1000 patient-years. In patients with severe hyperlipoproteinaemia we would be prepared to take such a risk if it was thought that the drug might reduce their very high death rate.

Intervention for prevention where the risk is low is totally different. I suggested earlier that a large number of people exposed to a small risk might yield more cases in the community than a small number exposed to a big risk. There is a counterpart to that in regard to intervention. If a preventive measure exposes many people to a small risk, then the harm it does may readily – as in the case of clofibrate – outweigh the benefits, since these are received by relatively few. Unfortunately we cannot have many trials as large as the clofibrate study, nor are we able to keep such trials going for longer than a few years, usually five at the most. We may thus be unable to identify that small level of harm to individuals from long-term intervention that would be sufficient to make that line of prevention unprofitable or even harmful. Consequently we cannot accept long-term mass preventive medication.



Self-Assessment Exercise 10.2.3

1. One strength of the population approach is that small reductions in risk achieved by many people can have a substantial impact on disease burden. How can this line of argument be applied to the safety of population prevention measures?
2. Why might it be difficult to identify the risks of a population prevention measure?
3. The above excerpt from paper A concludes by stating 'Consequently we cannot accept long-term mass preventive medication'. With around 7 million people in the UK taking statins (to lower blood cholesterol level), and the National Institute for Health and Care Excellence (NICE) recommending (as of 2014) their wider use, this conclusion from Geoffrey Rose appears to be being ignored. You may wish to carry out some enquiry of your own as to the balance of benefits and risks from the use of statins (which are certainly safer than clofibrate) and consider whether 'long-term mass preventive medication' is an appropriate policy. In the answers, we have provided an excerpt from the *Guardian* newspaper from February 2014 that raises some of the issues in this debate, and we encourage you to explore how this develops.

Answers in Section 10.5

The final exercise in this section will help to consolidate ideas about prevention strategy. Questions 1 and 3 require reading all of paper A, and reference to Rose (1992) is recommended if you are interested in more in-depth discussion.



Self-Assessment Exercise 10.2.4

1. In paper A, Geoffrey Rose refers to the 'prevention paradox'. What does this mean?
2. To which prevention strategy (high-risk or population) do the following statements best apply?
 - a. shifting the mean of the distribution of a risk factor;
 - b. screening for people with levels of a risk factor at or above a given point, and treating as necessary;
 - c. screening for people with average levels of a risk factor, and treating as necessary;
 - d. increasing tax on alcoholic drinks;
 - e. water fluoridation (fluoride at around 1 ppm prevents tooth decay).
3. Prepare a summary list of the advantages and disadvantages of the high-risk approach and of the population approach.

Answers in Section 10.5

10.2.4 The High-Risk and Population Strategies Revisited

We have seen that these ideas on prevention strategy were first proposed by Geoffrey Rose in the early 1980s, so how well have they stood the test of time? It is fair to say that they are still widely accepted and applied, although it is always valuable to look critically at established ideas (see 'paradigm shifts' in Section 1.1.1 of Chapter 1).

One such reassessment of the Rose hypotheses was made by Brown *et al.* (2007) in a study of BMI and non-communicable disease risk among Australian women. Using self-reported data on height, weight, and doctor-diagnosed hypertension (HT) or diabetes mellitus (DM), from 13,716 women aged 45–50 years (in 1996) from the Australian Longitudinal Study of Women's Health, the authors modelled the impact on HT and DM of three approaches to prevention:

1. A 1-unit reduction in BMI for the whole population (that is, a shift of the distribution to the left, by one unit of BMI, an example of the population approach)
2. A 3-unit reduction in BMI for the top 20% of the distribution (an example of the high-risk approach)
3. A 2-unit reduction in BMI for the top 50% of the distribution (the authors term this a 'middle road' strategy).

The authors do not discuss how they arrived at these strategies, the somewhat precise definitions that are unlikely to be achieved so neatly in practice, or their expected equivalence. Neither do they discuss the practicalities of how these would be implemented, especially (c), which requires a substantial change in one half of the population, and one must assume that any societal-level policy and other changes required to achieve this would affect the other 50% to some degree. Setting these points aside, the study does identify some useful perspectives on prevention.

The results differed for HT and DM, principally due to the shape of the exposure-risk functions. For HT, risk increased steadily from low to high BMI, whereas for DM, risk increased

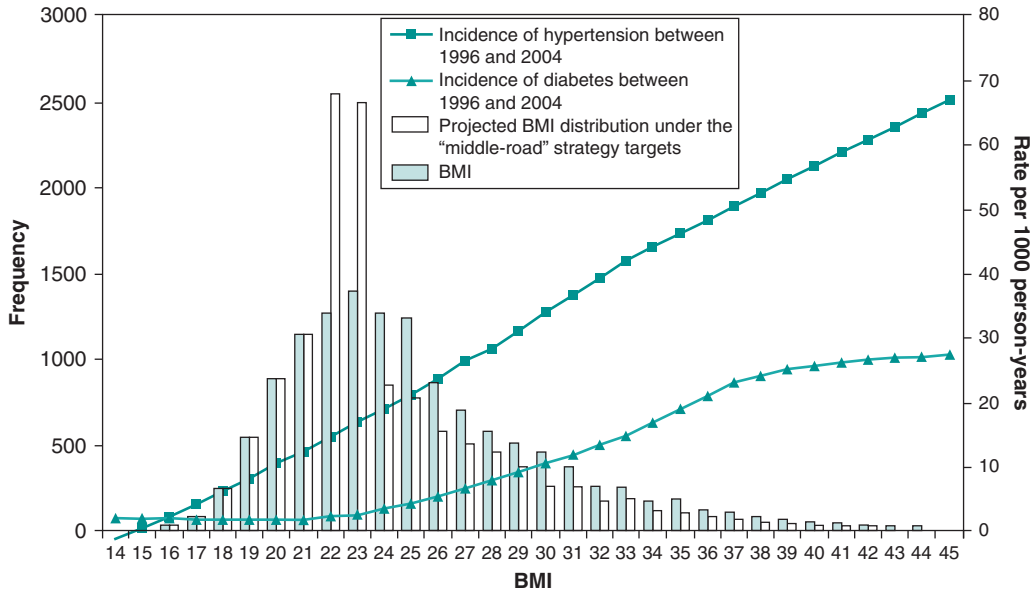


Figure 10.2.5 Distribution of BMI in 13,716 Australian women aged 45–50 years in 1996 (Figure 1 from *Brown et al.*).

more steeply at higher levels of BMI (between 25 and around 40), as shown in Figure 10.2.5. The risk appears to flatten off at BMI above 40, but there are relatively few people with such high values.

The more steeply the risk of a disease outcome rises at high levels of exposure to a risk factor, the greater will be the impact of a high-risk approach relative to a population approach. Conversely, if risk flattens off at higher levels of exposure, the high-risk approach will be relatively less effective compared to the population approach.

The principles underlying Rose’s proposals remain just as relevant today, and this study serves to remind us of the importance of assessing and understanding the distribution of risk factors in the population and how risk varies by level of exposure.

10.2.5 Implications of Genomic Research for Disease Prevention

Recent years have seen enormous strides being made in our knowledge of the human genome, and the influence of genetic makeup on disease, so it is important to ask whether this new information affects the way we should view prevention strategy.

This question was considered by Burton *et al.* (2012), using breast cancer as an example. This is a disease that, like so many others, has a complex aetiology involving lifestyle risk factors (e.g. obesity, alcohol consumption, and smoking), reproductive influences (e.g. early menarche, late menopause, higher age of first childbirth), medication (e.g. oral contraceptive pill, hormone replacement therapy), and genetics. Genes do play a part, but most breast cancer occurs in women without a family history. Although highly penetrant susceptibility genes carry a very high lifetime risk of the disease (between 30 per cent and 80 per cent, depending on type), these

are responsible for less than 5 per cent of cases. In discussing the wider genetic contribution to breast cancer, Burton *et al.* (2012) state:

For the common complex forms of breast cancer, genome-wide association studies have identified many risk alleles although the increase in risk conferred by each is small (usually a per-allele relative risk of less than 1.5). The most current estimate by Michailidou *et al.* (13) is that the polygenic risk based on 67 common genetic susceptibility variants explains approximately 14% of the genetic component of breast cancer risk.

The main issue the authors go on to discuss is the extent to which knowledge of genetic make-up could be used to inform prevention strategies and, more specifically, to stratify population sub-groups that may benefit from different approaches. For example, those in a higher risk stratum may be offered a different screening programme from those in a lower risk stratum.

While arguing that this could increase efficiency and benefit-to-harm ratio, they also acknowledge that this raises multiple ethical, legal, and social issues, including discrimination in insurance and employment. It is also not clear whether and how such stratification could apply to reducing risk factor exposures (in addition to potentially different screening strategies). Nevertheless, such knowledge about genetic influences is now available and is likely to become more so, and it is useful to read and give some thought to the issues raised by this paper.

Summary: High-Risk and Population Approaches to Prevention

- The distribution of risk in a population has very important implications for planning prevention. For common diseases, where risk factors are widely distributed in the population, most of the cases may arise from the majority of people with only moderately raised RR, rather than the few with very high RR.
- The high-risk approach involves identifying and managing individuals with high RR, the tail of the distribution. This approach would typically be delivered through the health system.
- The population approach involves measures that can shift the whole exposure distribution, thereby reducing the RR for the majority. This approach typically requires social, economic, and political action.
- It is very important that preventive measures applied to a population are safe – even a small risk can have disastrous results when applied to a very large number of people.
- Both high-risk and population approaches have advantages and disadvantages.
- Although initially proposed in the 1980s, these concepts have found wide acceptance and application and have stood the test of time.
- The shape of the relationship between level of exposure and relative risk has an influence on the relative impacts of population and high-risk approaches.
- A comprehensive prevention strategy requires both high-risk and population approaches to achieve the combined benefits of action to reduce population risk and to identify and manage smaller numbers at high individual risk.
- The knowledge gained in recent years about the human genome, and the influence of this knowledge on disease, have some implications for prevention strategy. While use of genetic markers to define alternative prevention strategies for population sub-groups that differ in terms of susceptibility may have a place, this also raises tricky social and ethical issues. This is therefore an important topic to watch.

10.3 Evaluation of Screening Programmes

Cancer Smear Test Errors Killed 14 Women, Thursday May 3, 2001

Fourteen women have died of cervical cancer despite being given the all-clear after smear tests, an audit into cervical screening services in Leicestershire revealed today.

The study of 403 women diagnosed with cervical cancer in a seven-year period showed that smear tests produced inaccurate results in nearly a third of cases.

Sixty-four more women eventually had to undergo radical treatment, including hysterectomies, after wrongly being given the all-clear at a time when their condition could have been treated with a simple operation, the Department of Health said.

The audit of women diagnosed with the disease between January 1993 and August 2000 found 122 discrepancies, of which 84 patients had a 'false negative' smear. Of those, 14 died and 64 underwent more radical treatment than they might otherwise have needed. The remaining women were not believed to have suffered any adverse consequences.

Speaking at a press conference at the Department of Health in central London, the government's cancer 'tsar', Mike Richards, said it was tragic that women had died and offered his condolences to bereaved relatives but stressed that Leicestershire was not failing.

He admitted the NHS had made mistakes but denied it had been negligent.

He said: 'That is why the audit was undertaken in the first place, to learn and then improve the service.'

Mr Richards said that the relatives of the 14 women who died would be told about the exact circumstances by Leicestershire health officials later today. He admitted that some could take legal action.

'No screening test can ever be 100% accurate, it is not an exact science'

Women who have been diagnosed with cervical cancer in Leicestershire will be offered an appointment with their gynaecologist and counselling.

Mr Richards stressed that the cervical cancer screening programme in Leicestershire had to be understood in context.

'Leicester is not failing, it performs well on all qualitative assurance measures and death rates from cervical cancer are falling in line with national trends,' he said.

'We recognise this may and will cause distress to some of the patients. The NHS say sorry when they get things wrong. We want to put the audit into the public domain so that learning is not just in Leicestershire but all around the country.'

Government health figures show that the NHS Cervical Screening Programme saves around 1,300 lives a year. It is directly responsible for a 42 per cent drop in incidents of cervical cancer between 1988 and 1997. It is estimated that the programme has saved more than 8,000 lives in that time.

Mr Richards said it was unlikely that the problem of wrongly interpreted smears was restricted to Leicestershire. 'We can assume that around the rest of the country there will be other cases where diagnosis of cervical cancer or abnormality on the smear were not picked up.'

Philip Sturman, chief officer of Leicestershire Community Health Council, which represents patients, urged the public not to panic.

'It is important not to scare people with these results – screening saves lives,' he said.

The NHS cervical screening programme was set up in 1988 and is based in Sheffield. It has screened almost four million women in England each year with around 3,450 new cases of cervical cancer – the most common form of cancer in women aged under 35 – diagnosed.

Figures show that 83.7 per cent of women of screening age have had a smear in the last five years with 3.3m women aged 25 to 64 screened.

The smear test spots abnormal cells, but does not test for cancer itself.

From the *Guardian*, 3 May 2001

10.3.1 Purpose of Screening

This article illustrates well the aims and benefits of screening (i.e. it has saved thousands of lives in the UK alone), as well as the complexities and challenges of running a programme (i.e. the ‘discrepancies’ that may falsely reassure some and miss cases among others).

A screening programme is designed to assess people in the population who are at risk for a disease, and classify them as having either a high or low chance of actually having the disease, at an early stage. Those people with a high chance of having the disease can then be referred for a definitive assessment (for example, a biopsy) and treatment as necessary, and those with a low chance can be reassured.

To achieve this, a programme is required to carry out repeated examinations of large numbers of well people (who do not have, or are unaware of, symptoms), using a test that combines accuracy with simplicity, low cost, safety, and acceptability. This is a substantial and demanding task. As the newspaper article illustrates, screening programmes are complex operations that have limitations. It emphasises that screening tests are not perfect and that evaluation is very important.

10.3.2 Criteria for Programme Evaluation

A number of criteria are commonly used in the evaluation of screening programmes, known as the Wilson–Jungner criteria. These were first described in 1968 in a WHO-commissioned report and are summarised in Table 10.3.1. Once we have discussed these and worked through some examples, we will consider how well they – like the ideas of Geoffrey Rose – have stood the test of time, and the implications of the rapidly increasing knowledge of the human genome and the opportunities this is bringing for genetic-based screening.

Table 10.3.1 Criteria for the evaluation of screening programmes.

Criterion	Explanation and discussion
The condition should be an important health problem.	There may be little value in developing a screening programme if the condition is not important, either in terms of severity, or incidence/prevalence, or both.
There should be an accepted treatment for patients with recognised disease.	It would be unacceptable to screen a population, detect a possibly serious disease, and then not be able to treat that disease effectively.
Facilities for diagnosis and treatment should be available.	In addition to there being an effective treatment, it is also very important that the health system has the resources needed to provide that treatment to all people with confirmed disease following screening, in a timely and effective manner.
There should be a recognizable latent or early symptomatic phase.	This relates to the natural history and also to opportunities for treatment. For screening to have any chance of being effective, there must be a phase during which the disease can be detected but before it has become too advanced to treat successfully.
There should be a suitable test or examination.	A number of measures are used to assess how suitable the screening test is, summarised by its validity. These validity checks compare the screening test to a definitive diagnostic test (the gold standard). These measures are sensitivity, specificity, predictive value, accuracy, and likelihood ratio. The discrimination of the test can also be analysed and displayed with a receiver-operator characteristic (ROC) curve. These are all explained in more detail later in this section.

(continued)

Table 10.3.1 (Continued)

Criterion	Explanation and discussion
The test should be acceptable to the population.	The test must also be acceptable (not unduly painful or anxiety provoking) and safe. For example, it has been necessary to show that the radiation dose used in mammography (screening for breast cancer by radiography) is safe and will not of itself cause breast cancer when used repeatedly during a woman's life.
The natural history of the condition, including development from latent to declared disease, should be adequately understood.	It is important to know about the natural development of a disease that is being considered for screening; that is, how it develops without any human interference or treatment. This is called the <i>natural history</i> . For example, does it always progress to a more-serious stage requiring treatment? If not, in what percentage of people does it progress, from what stage does it progress, and what factors determine this progression?
There should be an agreed policy on whom to treat as patients.	For subjects with results that may be termed 'borderline' (that is, neither clearly normal nor actual disease), there needs to be a clear policy (including communication) regarding the circumstances in which these people should be treated, as opposed to being followed up and reviewed (through repeat screening or other diagnostic tests).
The cost of case-finding (including diagnosis and treatment of patients diagnosed) should be economically balanced in relation to possible expenditure on medical care as a whole.	The costs of preventing a case (incidence and/or mortality) should be calculated, taking into account the full costs of screening and treatment. These costs should then be assessed in comparison with other procedures competing for resources.
Case-finding should be a continuing process and not a "once and for all" project.	<p>Screening on a large scale, such as for breast and cervical cancer, is a major operation that must be maintained over time, unless subsequent evaluation studies show it is not actually sufficiently cost-effective or otherwise unacceptable. Carrying out mammography, for example, requires expensive equipment, well-trained radiographers (to take X-rays), and radiologists (to read X-rays), and a good system for contacting and recalling women who require treatment or routine re-examination. This system has to be offered at intervals (e.g. every 5 years) to those already being screened while within the eligible age range and to those becoming eligible. In addition, advice and support are necessary for people in various stages and categories of the screening process:</p> <ul style="list-style-type: none"> ● people who may be afraid of attending for screening; ● people with a positive screening test, who turn out not to have the disease (false positive); ● people with a positive screening test, who turn out to have the disease (true positive); ● people with a negative screening test, who turn out to have the disease (false negative).

10.3.3 Assessing Validity of a Screening Test

Chapter 4 looks at validity in relation to questionnaire design and measurement in surveys. Validity is a measure of accuracy and can be defined as the capacity of a test or question to give

the true result. These concepts apply equally to screening tests, which, after all, are measurement tools used to establish whether or not an individual is likely or not to have a disease. We now introduce an example that helps to illustrate these various measures of validity.

Example

Blood cholesterol testing is increasingly common, and many self-test machines are now available. Self-testing is a form of screening (albeit mainly for a selected population), and it is of interest to assess how *valid* the results are. In the following (hypothetical) example, results on 100 people from a self-test machine have been compared (using the same blood sample) with a well-calibrated laboratory analyser, which can be treated as a *gold standard*. The results for each machine have been expressed in terms of the number (per cent) of people tested who had moderately or very high cholesterol (termed 'raised'), which, for the purpose of this example, is taken to be above (\geq) 6.5 mmol/l (Table 10.3.2).

Table 10.3.2 Validity of test results.

		Laboratory test		Total
		≥ 6.5 mmol/l (Raised)	< 6.5 mmol/l	
Self-test machine	≥ 6.5 mmol/l (Raised)	24	10	34
	< 6.5 mmol/l	3	63	66
Total		27	73	100

Measures of Validity

The following are commonly used measures of test validity, and we will examine each in turn by reference to the above example.

Sensitivity: Ability of the Screening Test to Recognise People with Disease

We consider first how well the self-test machine recognises people who really have an elevated cholesterol, that is, the 27 people testing high (≥ 6.5 mmol/l) on the laboratory analyser, which is as close as we are going to get to the truth. Of these 27 people, the self-test recognised 24. Expressed as a percentage, this is 88.9 per cent, which we might judge to be good. This measure of test performance is called *sensitivity*:

The *sensitivity* of a test is the probability that it will correctly identify people who have the characteristic that is being measured.

Sensitivity should be presented with the 95 per cent confidence interval (CI). It is a proportion, so the formula for the standard error (SE) and 95 per cent CI are as for a proportion (see Chapter 4). In this case, the SE is 6.05 (%) as shown below and the 95 per cent CI is $88.9 \pm (1.96 \times 6.05) = 77.1\text{--}100\%$.

$$SE (\text{sensitivity}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{88.9(100-88.9)}{27}} = 6.05$$

Specificity: Ability of the Screening Test to Recognise People Without Disease

There were 73 people whose test result on the laboratory analyser was < 6.5 mmol/l and who therefore were not cases of raised cholesterol. How well did the self-test do? Of these 73 people,

the self-test reported 63 as having a cholesterol level of <6.5 mol/l, which is 86.3 per cent, and is also quite good. This measure of test performance is called **specificity**:

The **specificity** of a test is the probability that it will correctly identify people who do not have the characteristic that is being measured.

Specificity should also be presented with the 95 per cent CI, which is 78.4–94.2 per cent. This is narrower than the 95 per cent CI for sensitivity because there are more cases ($n = 73$ versus 27) from which to derive the specificity estimate.

Predictive Value: How Trustworthy is the Result Predicted by the Screening Test?

By predictive value we mean, if the test predicts that the person is (or is not) a case, how likely is it that the test is correct? We can examine this aspect of performance for a positive (case) and a negative (not a case) result.

- **Positive result:** Of the 34 people with a high result on the self-test machine, only 24 really had a high level according to the laboratory test, which is not so good. This is 70.6 per cent and is termed the **positive predictive value**. A good way to think about the implications of this is to imagine that you were one of the 34, thinking from the self-test that you had a raised cholesterol, only to discover later that it was not true. In practice, this could cause people a lot of anxiety (which might be a lot worse if you were dealing with a test for cancer), and in research it would lead to serious bias in, for example, a **prevalence** estimate.
- **Negative result:** Of the 66 people testing negative on the self-test, 63 results really were negative. This is 95.4 per cent, which is good. If anything, the problems caused by a poor **negative predictive value** would be worse than for a poor positive predictive value: We saw the effects of this in cervical cancer screening in the newspaper article with cases of cancer being missed.

Accuracy: How Likely is it that Any Result from the Screening Test is Correct?

Accuracy considers all tests and asks what is the probability that any test, whether positive or negative, has provided the correct result. This is 24 (correct positive results) + 63 (correct negative results) divided by the total (100): this is 87 per cent, which is fairly good.

Likelihood Ratio

As with predictive value, likelihood ratio can be determined for both a positive and a negative test.

- **Likelihood ratio for positive test** looks at how much more likely it is that a positive screening test result will be found in a person with the condition, as opposed to a person without the condition. This is calculated (with sensitivity and specificity as ratios rather than percentages) as:

$$\text{Likelihood ratio (positive test)} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

In this case, it is 6.49. We can interpret this as meaning that it is 6.5 times more likely that a positive test will be found in someone with the condition than someone without it.

- **Likelihood ratio for negative test** looks at how much more likely it is that a negative screening test result will be found in a person with the condition, as opposed to a person without the condition.

$$\text{Likelihood ratio (negative test)} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

In this case, it is 0.13.

Summary

A good way to help understand what these measures of test performance show, and to calculate them, is to present the situation as a general 2×2 table (Table 10.3.3):

Table 10.3.3 Validity of test results.

		True situation		
		Positive	Negative	Total
Test	Positive	a (true test positives)	b (false test positives)	a+b
	Negative	c (false test negatives)	d (true test negatives)	c+d
	Total	a + c (true positives)	b+d (true negatives)	a+b+c+d

This table can now be used to help in calculating the test measures (Table 10.3.4):

Table 10.3.4 Calculation of test measures.

Measure	Summary (performance of test)	Calculation
Sensitivity	Percentage of true positives detected	a/(a+c)
Specificity	Percentage of true negatives detected	d/ b+d)
Positive predictive value	Percentage of test positives that are correct	a/(a+b)
Negative predictive value	Percentage of test negatives that are correct	d/(c+d)
Accuracy	Percentage of test results that are correct	(a+d)/(a+b+c+d)



Self-Assessment Exercise 10.3.1

In this exercise, the screening test is for a cancer that is much less common than the prevalence of raised cholesterol (27 per cent) in our earlier example. Comparison with the gold standard of biopsy provided the following results:

		Gold standard (biopsy)		
		Have disease	Disease free	Total
Screening test	+ve on test	325	450	775
	-ve on test	25	4,800	4,825
	Total	350	5,250	5,600

1. What is the prevalence of cancer in this example?

2. Calculate (with 95 per cent CI) and interpret the sensitivity of the test.
3. Calculate and interpret the specificity.
4. Calculate and comment on the positive and negative predictive values.
5. Calculate and comment on the accuracy of the test.
6. Calculate and interpret the likelihood ratios for a positive test and for a negative test.

Answers in Section 10.5

The Receiver-Operator Characteristic (ROC) Curve

We have seen that a screening test is designed to classify individuals as having a high or a low risk of a disease. For most tests, a decision has to be made about the value (which may be a questionnaire score, a physiological measurement, or physical findings such as cells or an X-ray) that, if exceeded, will be taken as indicating high risk of having the disease. For any given test, if a higher value (or more definite finding) is used, it will be more certain that the person is at high risk; that is, the test will give a better specificity. The other side of this coin, however, is that there will also be an increased probability that people with the disease will be missed; that is, there is lower sensitivity. Conversely, if a lower (less definite) 'cut-off' level is taken, sensitivity will be increased at the expense of specificity.

These characteristics of a screening test, including the trade-off between sensitivity and specificity, can be described by the *receiver-operator characteristic* (ROC) curve. This also allows identification of the cut-off value, giving the best discrimination optimising sensitivity and specificity. The term 'ROC curve' came from their use with radar during the 1939–1945 war. In that situation, they were used to identify settings that would best assist (radar) receiver operators in distinguishing between reflected signals from an object of interest (ship, aircraft) and other background objects or conditions that were not a threat. If the equipment was set to be too sensitive, there would be many false positives; for example, fighter planes being sent to intercept flocks of birds. If it was set to be too specific, there would be more risk that some incoming enemy aircraft would not be detected.

Example

Chapter 4 discusses validity in respect of designing a back pain disability questionnaire. Using this example, let's assume we have now scored the questionnaire responses on a scale of 0–30 obtained from 235 subjects in an occupational setting. We wish to determine the value of the disability score that best discriminates between the presence or absence of persistent back pain. The presence of back pain has been assessed by self-reports of the nature and duration of symptoms. In this example, we are using the disability score as the screening test and the presence or absence of back pain as the gold-standard assessment of the outcome we are interested in. Table 10.3.5 and Figure 10.3.1 provide some information on the distributions of the scores among those with and without back pain.

Table 10.3.5 Distributions of disability scores among cases and non-cases of back pain (output from SPSS).

Persistent back pain	Mean	N	SD	Median	Inter-quartile range	Minimum	Maximum	Range
No	5.66	116	3.580	5.00	3.0–7.0	0	19	19
Yes	10.23	119	4.938	9.00	7.0–13.0	1	30	29
Total	7.97	235	4.883	7.00	5.0–10.0	0	30	30

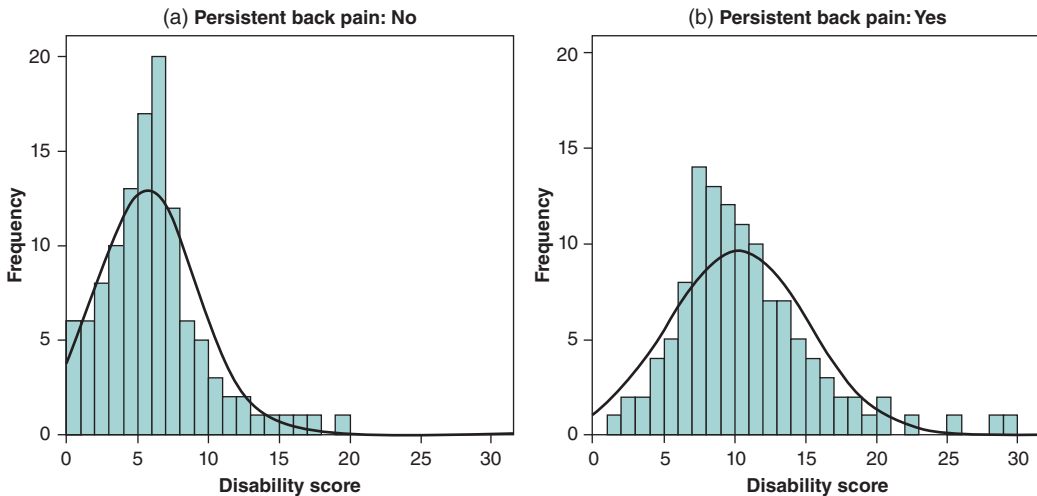


Figure 10.3.1 Distributions of disability score for (a) subjects without persistent back pain and (b) subjects with persistent back pain.



Self-Assessment Exercise 10.3.2

Using the information in Table 10.3.5 and Figure 10.3.1

1. Describe the distributions for those with and those without back pain, including how the distributions differ.
2. What cut-off score would be needed to ensure that test positive results included only people with back pain? Approximately what proportion of people with back pain would be wrongly classified as a result?
3. What cut-off score would be needed to ensure that test negative results included only people without back pain; that is, all people with back pain had a test positive? Approximately what proportion of people without back pain would be wrongly classified as a result?
4. Hazard a guess at the score you think will discriminate best between those with and without back pain.

Answers in Section 10.5

ROC Curve Analysis

We are now ready to look in a bit more detail at the analysis and interpretation of the ROC curve. Table 10.3.6 (obtained from SPSS ROC curve analysis output) shows the disability score values (range, -1 to 31 ; see footnote 3 in table), the sensitivity, and $1 - \text{specificity}$ associated with values of the questionnaire disability score. To aid interpretation, we have re-calculated the specificity in column 4 of this table.

The extremes of the disability score represent perfect sensitivity and worst specificity (shown by a disability score value of -1) and perfect specificity with worst sensitivity (shown by a disability score value of 31). Figure 10.3.2 shows the ROC curve based on the data

Table 10.3.6 SPSS output showing sensitivity and 1 – specificity values associated with scores from the disability questionnaire.

Output from SPSS ¹			
Positive (back pain present) if greater than or equal to ³ :	Sensitivity ⁴	1 – Specificity ⁴	Re-calculated ² Specificity
–1.00	1.000	1.000	0.000
0.50	1.000	.948	0.052
1.50	.992	.897	0.103
2.50	.975	.828	0.172
3.50	.958	.741	0.259
4.50	.924	.629	0.371
5.50	.882	.483	0.517
6.50	.815	.310	0.690
7.50	.697	.207	0.793
8.50	.588	.155	0.845
9.50	.487	.112	0.888
10.50	.395	.086	0.914
11.50	.311	.069	0.931
12.50	.252	.052	0.948
13.50	.193	.043	0.957
14.50	.151	.034	0.966
15.50	.118	.026	0.974
16.50	.092	.017	0.983
17.50	.076	.009	0.991
18.50	.059	.009	0.991
19.50	.050	.000	1.000
21.00	.034	.000	1.000
23.50	.025	.000	1.000
26.50	.017	.000	1.000
29.00	.008	.000	1.000
31.00	.000	.000	1.000

Notes:

¹These three columns are output from the SPSS ROC curve analysis.

²We have re-calculated the specificity [1 – column 3] to make it easier to see how the point of maximal discrimination optimises both sensitivity and specificity.

³The smallest cut-off value is the minimum observed test value minus 1, and the largest cut-off value is the maximum observed test value plus 1. All the other cut-off values are the averages of two consecutive ordered observed test values.

⁴Expressed as a proportion.

from the 235 respondents (SPSS output). We now discuss what these results mean for the screening test.

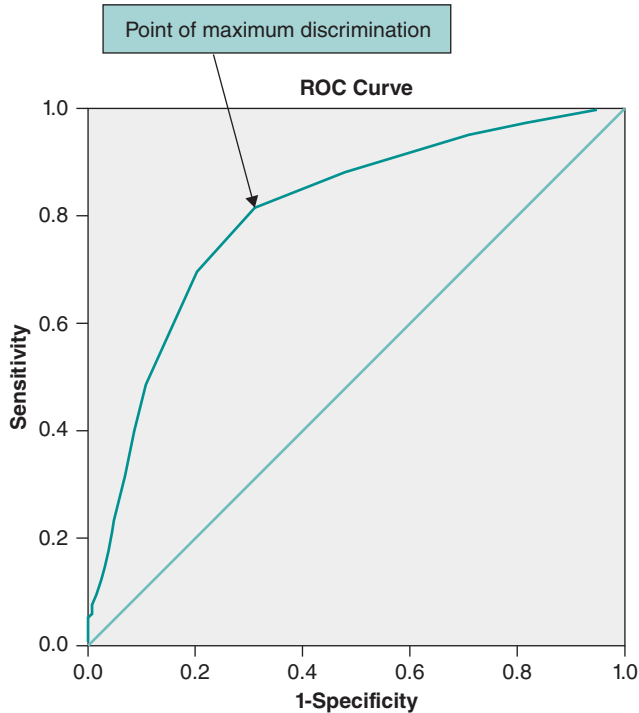


Figure 10.3.2 Receiver-operator characteristic (ROC) curve describing the discrimination of the disability score for persistent back pain.

The important results from ROC curve analysis are as follows:

1. The **area under the curve**, which quantifies the overall capacity of the test to discriminate between those who have the disease (back pain) and those who do not. This analysis yielded an estimate of the area at 0.802, with a standard error of 0.029 and 95 per cent CI 0.745 to 0.859, $p < 0.0001$ (SPSS output). A test with no power whatever to discriminate has an area of 0.5 and is illustrated by the diagonal line in Figure 10.3.2, so the better the test, the larger the area to the left of the diagonal line. Our test is very much better than the zero discrimination line, and the 95 per cent CI shows that we have a fairly precise estimate of how much better. We interpret this result as follows: For a test where higher values are associated with greater likelihood of disease (as in this case), the area represents the probability that a randomly selected person with the disease will have a higher test result than a randomly selected person who does not have the disease. In this case the probability is just over 80 per cent, or 8 out of 10.
2. The **point of maximal discrimination** (marked on Figure 10.3.2), is the point farthest from the diagonal line of no effect. This also indicates the sensitivity and 1 – specificity at the points of maximal discrimination, which are approximately 80 per cent and 30 per cent, respectively (and therefore the specificity is 70 per cent).

3. The **sensitivity** and **specificity** associated with the full range of scores, including the point of maximal discrimination, are shown in Table 10.3.6. In this table it is very clear how sensitivity falls as specificity rises and vice versa (remember that column 3 is $1 - \text{specificity}$). The optimal values of both sensitivity and specificity identified from the curve are associated with a score of 6.5 (bold typeface in table), so an integer score of 7 would be taken as the cut-off: This is the answer to question 4 in Exercise 10.3.2. You can also see the results for the questions about cut-offs (2 and 3) in Exercise 10.3.2. Question 2 requires a specificity of 100 per cent (true negatives must be correctly identified), which is achieved at a score of 20 (only integer scores are possible). Question 3 requires a sensitivity of 100 per cent (all cases are to be included), which is achieved at a score of 1.

Which Measures of Validity are Most Useful?

All of these measures are useful, but there are some pointers to which measures are most informative in any given set of circumstances.

Sensitivity and specificity are particularly useful in assessing the overall performance of a test, as for a screening programme or in epidemiological studies. For example, if we are trying to identify cases in a survey or cohort study, or of course in screening, it is valuable to know how good the test is at finding genuine cases. A test with a low sensitivity but very high specificity would not find all the cases but would be unlikely to identify people as cases if they did not have the disease. Conversely (and not uncommonly in epidemiological studies), it may be useful to use an initial test with a high sensitivity and low specificity to identify cases and possible cases (thus missing very few real cases), and then apply a more-rigorous test for the final case definition.

Positive and negative predictive values are particularly relevant to the health-care setting, where the results of an individual test need to be discussed with a patient. These measures help with answering questions such as ‘I have a negative test result – does that mean I definitely don’t have cancer?’ If the negative predictive value were 95 per cent, the response would be, ‘It is unlikely, but there is still a small chance, about 1 in 20, that you could have the disease.’

10.3.4 Methodological Issues in Studies of Screening Programme Effectiveness

Of the various epidemiological study designs available, RCTs and case-control studies are used most commonly to study the effectiveness of screening programmes. The design of these studies raises some quite complex issues, in particular the avoidance and/or interpretation of potential sources of bias that arise because it is a screening programme that is being studied. The three most important types of bias in epidemiological studies of screening programmes are summarised here.

Selection (Volunteer) Bias

People taking part in screening programmes may differ from those who do not. If this difference is associated in some way with their general health, severity of disease (for which they are being screened), and likely compliance with advice and treatment, then these are all ways the screened group may be biased towards a better outcome. This source of bias is more likely to arise with observational studies (e.g. case-control) than with RCTs, since randomisation should balance the characteristics of people screened and not screened.

Length-Based Sampling (Prognostic) Bias

Length-based sampling bias results from the observation that screening may detect people with less rapidly progressing (less-severe or less-aggressive) disease than the usual route of

presenting to a doctor with symptoms or other problems. This occurs because the less rapidly advancing cases have a longer preclinical phase than the more severe cases. Thus, when screening is carried out, people with the less severe form are more likely to be picked up; there are in effect more person-years of preclinical less severe disease around in the population than person-years of more severe disease. If this bias occurs, the screened group tends, on average, to have less severe disease and may consequently be observed to have better outcomes.

Lead-Time Bias

Lead-time bias occurs because screening can pick up cases earlier in the natural history than when they present through the usual route; that is, to a doctor when symptoms arise. Of course, an effective screening and treatment programme might actually achieve a better outcome (longer survival; cure) as a result of earlier detection, but that is *effectiveness*, not *bias*. What we are talking about here is the perception that survival is longer, simply because the diagnosis has been made through screening, say, 6 months or a year earlier than it would otherwise have been if the diagnosis had been made when symptoms arose and the person visited a doctor. This is summarised in Figure 10.3.3.

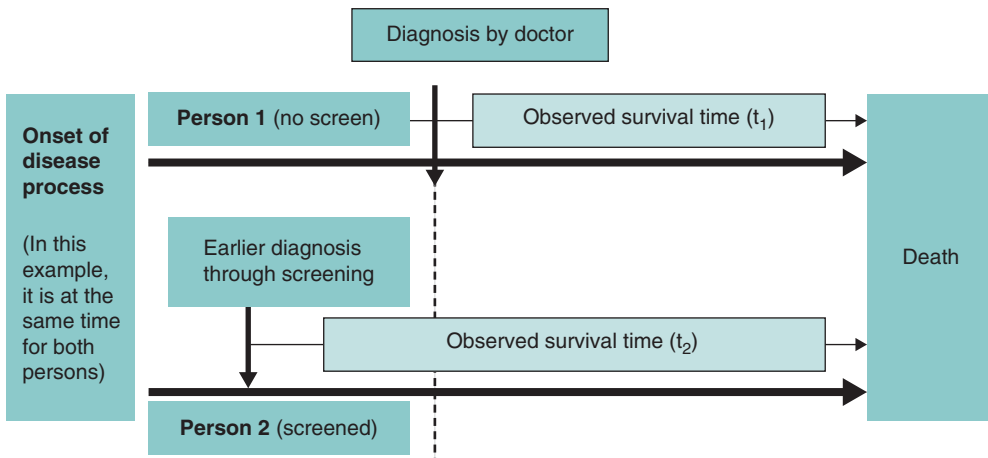


Figure 10.3.3 Illustration of lead-time bias that may occur in evaluation studies of screening programmes.

In this example, two people develop a disease that in fact runs exactly the same course from onset to death. Person 1, who is not screened, becomes ill and visits the doctor, and the disease is diagnosed and treated, but after a period of time (t_1), the person dies. Person 2, on the other hand, is screened, and as a result, the diagnosis is made some time earlier for person 2 than for person 1: this person's disease is also treated, but the person subsequently dies after a period of time (t_2) that is longer than t_1 . This gives the impression that person 2 survived longer, but in fact, treatment following detection of the disease at screening is no more effective than treatment of disease diagnosed when the person became clinically ill. This situation could lead to a mistaken conclusion that screening and early treatment resulted in longer survival.

10.3.5 Are the Wilson–Jungner Criteria Relevant Today?

Since these criteria were published in 1968, it is important to look at whether they are still relevant and the extent to which developments in screening tests and programmes mean that some revisions are needed.

A brief discussion of current relevance, and in particular in the context of the growing area of genetic screening, is provided by Andermann *et al.* (2008). Based on an extensive review of literature on developments with the Wilson–Jungner criteria and consultation with experts, they propose a set of criteria that represent a synthesis of perspectives that have emerged since 1968 (Box 10.3.1)

Synthesis of Emerging Screening Criteria Proposed Since 1968 (Source: Andermann *et al.* 2008)

- The screening programme should respond to a recognized need.
- The objectives of screening should be defined at the outset.
- There should be a defined target population.
- There should be scientific evidence of screening programme effectiveness.
- The programme should integrate education, testing, clinical services, and programme management.
- There should be quality assurance, with mechanisms to minimize potential risks of screening.
- The programme should ensure informed choice, confidentiality, and respect for autonomy.
- The programme should promote equity and access to screening for the entire target population.
- Programme evaluation should be planned from the outset.
- The overall benefits of screening should outweigh the harm.

These criteria are intended for guiding genetic screening in particular, but they can usefully be seen as encompassing a broader and more current perspective on screening than was previously available. They also incorporate experience from the delivery and evaluation of screening programmes over this period.

Andermann *et al.* conclude that the Wilson–Jungner criteria have indeed stood the test of time, and they remain the basis for planning and evaluation of screening. They also point out that the authors of the 1968 report encouraged debate and development of the original criteria, so while the criteria remain a sound foundation, it is important to also consider new perspectives as screening tests and programmes evolve.

Summary: Evaluation of Screening Programmes

- Although the validity of the screening test is very important, evaluation must include all aspects of a screening programme.
- The criteria and specific characteristics discussed in Sections 10.3.2 and 10.3.3 summarise the attributes that must be considered when planning a screening programme or evaluating an existing one.
- The Wilson–Jungner criteria have stood the test of time, although you may encounter multiple versions developed to address different settings and types of screening. The advent and expansion of genetic screening is one such example.
- Epidemiological studies designed to assess the effectiveness of screening programmes are subject to particular forms of bias, namely volunteer, length-based, and lead-time bias.

10.4 Cohort and Period Effects

10.4.1 Analysis of Change in Risk Over Time

In Chapter 2 we consider how routine data collected on populations could be used to generate hypotheses about possible associations between risk factors (exposures) and a disease or outcome of interest. One quite powerful way of doing this is through analysis of how risk (e.g. mortality) changes over time in groups of people of different ages or born at different times. The two most important influences on risk that can be identified in this way are known as *cohort* and *period effects*, and these are now described.

Cohort Effect

A group of people born at a particular time or over a relatively short period of years (known as a *birth cohort*) will experience a unique set of social and environmental conditions, health services, etc., over the course of their lifetimes. These conditions at times present exposures that carry some increased (or decreased) risk of disease. Such exposures might occur while their mothers were pregnant, at or around birth, or at a particular age in their lives such as teenage or young adult years. The increased risk of disease relating to one or more of these exposures will be identifiable among people born at a particular time in history. This is known as a *cohort effect* and is related in some way to the time these people were born, even though the exposure may occur at some later stage in their lives.

Cohort effects may be apparent in the incidence over time of diseases with a long interval between the cause and the onset of the disease. For example, increasing mortality rates from lung cancer among women have been attributed to smoking trends among different (birth) cohorts of women. In the UK, women in their late teens and early twenties at the time of the Second World War and during the years immediately afterwards, became part of a growing trend of smoking among women. This trend was related to smoking being perceived as more acceptable and fashionable and also because the war years had resulted in many more women working in industry and offices. As a result it can be observed that women born in the 1920s and 1930s, who were therefore at a vulnerable age (around 16–25 years old) during a period when smoking was being rapidly taken up by young women, have higher rates of lung cancer mortality than women born before the 1920s.

Period Effect

A contrasting situation occurs when people (of all ages or any age) in a given population, experience a change in exposure at a particular time (or period) in history. This is known as a *period effect*. This is often seen for health outcomes with more-immediate causes, or those with a short latency period that affect all age groups (or at least a wide age range) at the same time. An example of this was the introduction in the UK of the law enforcing the wearing of front seat belts in cars, which resulted in a rapid and sustained reduction in mortality from road traffic accidents. When front seat belt wearing was made compulsory in the UK, compliance with the law was very good and the rate of wearing belts increased rapidly from less than half to around 90 per cent of all front-seat passengers. The effect was a dramatic reduction in deaths and serious injuries from road accidents (among car occupants, not pedestrians!). This is a *period* effect, and it was not restricted to any particular age group.

It can be appreciated that the identification of cohort and period effects can be very useful in relating observed rates of mortality and morbidity to the patterns of events and circumstances that have occurred over time in order to understand the aetiology of disease and the impact of prevention measures. Thus, specific studies can be designed to investigate these effects,

incorporating *cohort* or *period analyses*, to help investigate possible causal associations. In the following example based on analysis of suicide trends in the UK, we will see how information about social trends can be related to observed cohort and period effects and how valuable this is in thinking about prevention and public health policy.

10.4.2 Example: Suicide Trends in UK Men and Women

In Chapter 2, we investigate some of the changing trends in suicide rates over the last 35 years by age and sex. We are now going to look at these trends in more detail. The following examples are taken from paper B (Gunnell *et al.*, 2003). In this study, trends in suicide were examined to see whether changes in the UK over time could be explained by cohort or period effects. We will begin by looking at trends in age-specific suicide rates.

Trends in Age-Specific Suicide Rates

Mortality rates for virtually all causes of death increase with age. Plotting death rates by age group over time is probably the most common way of illustrating trends in mortality, and it is a method we have used previously. Figure 10.4.1, taken from paper B, illustrates trends in male suicide since the 1950s for defined age groups.

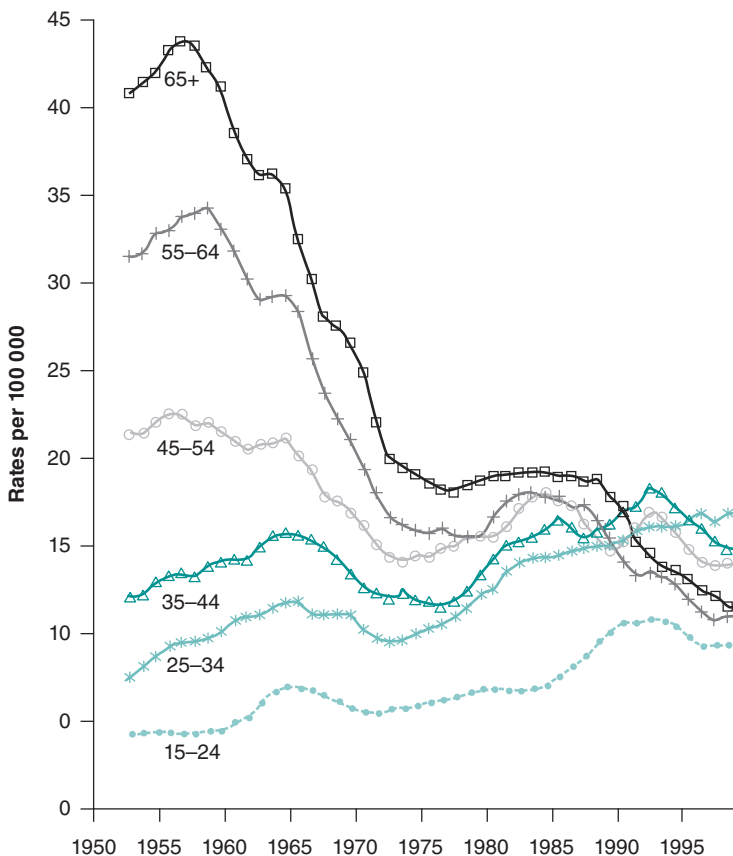


Figure 10.4.1 Age-standardised suicide rates: 1950–1999, England and Wales (3-year moving averages) in males (Figure 1(a) in paper B).¹

¹Note that the y-axis on this graph and some of those that follow have been altered to 100,000, as this was incorrectly labelled as 100,00 in the published paper.

Three-Year Moving Averages

The presentation of data in Figure 10.4.1 involves the use of 3-year moving averages. This technique is used to smooth year-on-year fluctuations, which are mainly random. The most common method of calculation is as follows. For year X, the smoothed rate = $[(0.25 \times \text{rate for year } X - 1) + (0.5 \times \text{rate for year } X) + (0.25 \times \text{rate for year } X + 1)]$.



Self-Assessment Exercise 10.4.1

1. From Figure 10.4.1 what can you say about trends in suicide deaths among men in the age groups illustrated?
2. Notice that suicide rates fell across all age groups during the late 1960s. What kind of effect describes this pattern of decline in death rates?
3. Can you think of any possible explanations for the decline in suicide rates identified in question 2?

Answers in Section 10.5

Analysis of Period Effects

In Figure 10.4.2, reproduced from paper B, suicide rates in six age groups are shown for five periods of death for males and females. This method of graphical analysis is used to identify whether there is any evidence of a period effect, since each line illustrates the death rates for all age groups and death rates at a particular time. If a period effect is occurring, this should be reflected in the suicide rates for deaths of all (or a wide range of) age groups at a particular period in history. On the graph, where a period effect is evident, we would therefore expect to see one or more of the lines with higher rates and a tendency for the lines to be separated in

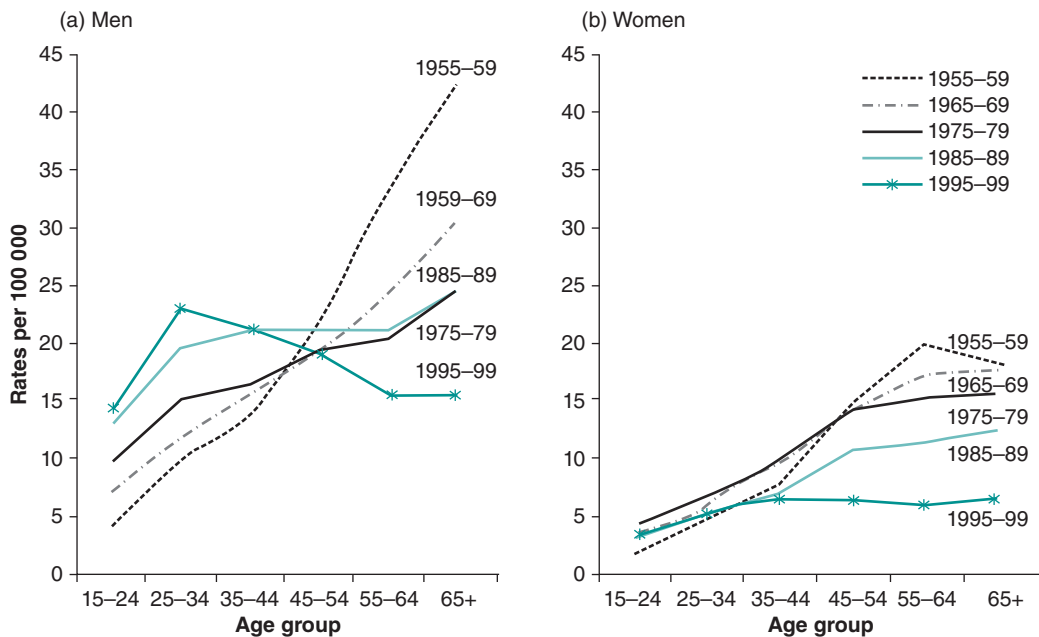


Figure 10.4.2 Suicide and undetermined death rates by time period (of death) and by age group in (a) men and (b) women in the UK (Figure 2 in paper B).

a parallel way. The following exercise explores whether there is evidence of period effects for suicide among men and women.



Self-Assessment Exercise 10.4.2

1. The x-axis of Figure 10.4.2 shows age groups: What are these ages?
2. What, approximately, is the suicide rate for men, dying at age 30 years, in 1977?
3. Does Figure 10.4.2 (a) provide evidence of a period effect among men?
4. Is there evidence in Figure 10.4.2 (b) of a period effect for women?

Answers in Section 10.5

Analysis of Cohort Effects

Graphical methods are also commonly used for analysis of cohort effects, and this is illustrated in Figure 10.4.3, reproduced from paper B. The way this graph is plotted may seem unfamiliar at first, but if you work through the interpretation systematically, it is not difficult to understand.

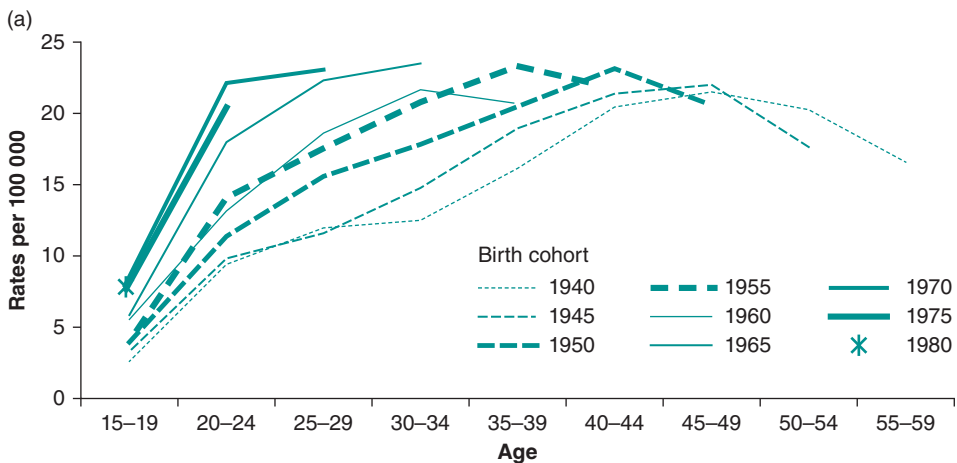


Figure 10.4.3 Rates of male suicide and undetermined death in successive 5-year birth cohorts by age group. The x-axis (age-groups) shows age at death (Figure 3(a) in paper B).

To demonstrate cohort effects, we plot *age-specific (suicide) rates* (that is, age at death) for different birth cohorts. For example, the first group shown by the fine dotted line (lowest on graph) shows male suicide and undetermined death rates for the birth cohort born between 1940 and 1944 and dying at different ages. The data range from the suicide rate among 15- to 19-year-olds (born 1940 to 1944) on the far left, through to the suicide rate among 55- to 59-year-olds (born 1940 to 1944) on the far right end of the dotted line.

This graph shows that for each successive birth cohort, suicide rates in young men (that is, up to the age of 30–34 years; we do not have data on older suicide deaths after the 1960–1964 birth cohort) have increased across all age groups. This indicates that since 1940, we can identify an effect such that the more recently males are born, the greater is the risk of suicide up to age 34 years. This indicates a probable birth cohort effect. The question then is, what might be progressively increasing the risk of young male suicide in this way?

A second question arises from the observation (from data in Figure 10.4.3) that, as we look across from earlier-born cohorts (e.g. 1940–, 1945–, etc.) to later-born cohorts (1970–, 1975–,

etc.), the suicide rates appear to peak at progressively younger ages. This is the case at least up to the 1960–64 birth cohort; we don't yet observe peaks for more recently born cohorts. The peak for the birth cohort 1945–49 is age 45–49 years, which must therefore have occurred in the period of death 1990– (calculated by adding the age at death to the birth year). Similarly, the peak for the 1950–54 birth cohort is age 40–44 years, and the period of death is also 1990–. In fact, the peaks in death rates occurred around 1990 for all birth cohorts for which we see such peaks, indicating a *period effect*.

The final exercise looks at the data for males from paper B in more detail, by examining data on suicides that exclude those due to overdose and gassing (Figure 10.4.4). This helps us draw some conclusions about what might be going on.

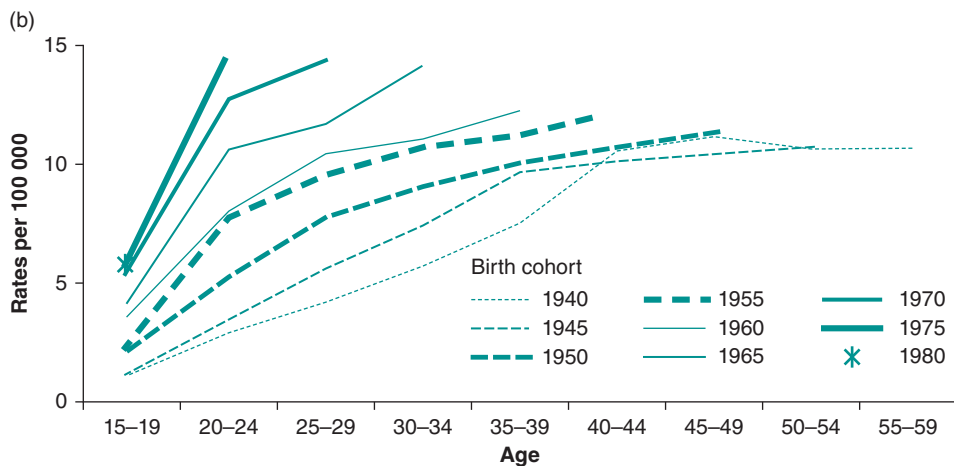


Figure 10.4.4 Rates of male suicide and undetermined death in successive 5-year birth cohorts by age group, excluding overdose and gassing. The x-axis (age-groups) shows age at death (Figure 3(b) from paper B).



Self-Assessment Exercise 10.4.3

1. From Figure 10.4.4, what is the death rate (suicide and undetermined, excluding overdose and gassing) for men born in 1963 and dying at age 32 years?
2. How do the patterns of death rates for birth cohorts in Figure 10.4.4 differ from those in Figure 10.4.3, which was based on all methods of suicide?
3. What might be the explanation for the finding identified in question 2?
4. What can we conclude so far from the analyses of male suicides we have discussed in this section?

Answers in Section 10.5

Summary: Cohort and Period Effects

- Cohort effects result from the experience of a group (cohort) born at a particular time and can be studied by examining death rates in relation to the year of birth.
- Period effects result from circumstances occurring at a particular time or period and can be studied by examining death rates in relation to the year of death.

- Cohort and period analysis can play a useful part in the study of disease causation and can help shed light on the associations between historical events; social, political, and environmental trends and prevention measures; and patterns of disease incidence.

10.5 Answers to Self-Assessment Exercises

Section 10.1

Exercise 10.1.1

Activity	Outcome	Risk to individual	Burden to society
Base jumping (from buildings, bridges, etc., with a parachute)	Death from injury due to failure of parachute to open in time or hitting buildings or cliffs, etc.	High: this is a dangerous activity	Very low, because relatively few people do base jumping
Driving a car, or being a passenger	Death or serious injury in crash	Relatively low: the chances of an individual being injured over a year or even a lifetime are not great*	Fairly substantial, because even though individual risk is not high, it is a very common activity
Smoking tobacco	Death from lung cancer or cardiovascular disease (CVD)	High: it is estimated that one in three smokers will die of a condition related to smoking	Very high, because not only is the individual risk high, but at least one third of the (UK) population smoke, or have smoked, and CVD is very common

*For some population subgroups, such as young males, there are substantially higher risks to individuals.

Exercise 10.1.2

Risk factor	Disease risk	Rate/10,000 per year		Relative risk	Attributable risk
		Exposed	Unexposed		
Factor A	Cancer	1.2	0.4	3.0	0.8 per 10,000 per year
Factor B	IHD	660	220	3.0	440 per 10,000 per year

The RR is the same for both factors, but the AR is very different. Thus, although exposure to factor A carries a RR of 3.0, the exposure and cancer are relatively rare. As a result, there are only 0.8 cases per 10,000 population per year *attributable* to exposure to this factor. In contrast, although factor B has the same RR of 3.0 for IHD, 440 cases per 10,000 per year can

be attributed to this exposure. This is because the exposure is very much more common, as is the disease. This emphasises two key points:

- Clearly the common exposure, factor B, has a far greater public health impact. This is very important in terms of reducing the burden of a common disease.
- For individuals exposed, both factors are of similar concern, although the context can be expected to differ. For example, although factor B may be an exposure of concern to the general public, factor A might be restricted to a specific occupational setting. In this exercise we found the attributable risk of IHD to be 440 per 10,000 per year. For every 10,000 people exposed, 440 per year will suffer IHD attributable to factor B. Thus, if exposure is eliminated for 10,000 people, 440 cases of IHD per year could be prevented (subject to related risk factors being unchanged).

Exercise 10.1.3

1. The $RR = 317/150 = 2.113$; the $AR = 317 \text{ per } 100,000/\text{year} - 150 \text{ per } 100,000/\text{year} = 167 \text{ per } 100,000 \text{ per year}$.
2. Using equation (4), $PAF = (0.25 \times 1.113)/[1 + (0.25 \times 1.113)] = 0.218$, or 21.8%
3. With 25% of the population smoking, 21.8% of the 40,000 total deaths from CHD (200 per 100,000 \times 20 million) in the year would be due to smoking, which is 8,720 deaths. When the prevalence of smoking is reduced to 15%, the PAF [again using equation (4)] is 14.3%, resulting in 5,720 attributable deaths. The difference of 3,000 is the number of deaths prevented.
4. Using equation (6), the PAF is $0.3 \times (2.113 - 1)/2.113 = 15.8\%$, somewhat lower than the 21.8% obtained using equation (4). This finding suggests that some confounding is occurring, and consequently the percentage of CHD that can be attributed to smoking is reduced when this is taken account of.

Exercise 10.1.4

1. Note that the order in which the sub-regions are presented has not been changed between the two graphs. For 1990, the proportions of YLL (green) and YLD (orange) vary from about 0.55/0.45 (high-income Asia Pacific) to around 0.85/0.15 (western sub-Saharan Africa). A similar range in variation is seen for 2010, although generally the proportion of YLD has increased over the period and, for example, the YLD are the larger component for Australasia in 2010. The ranking of sub-regions is not easy to discern from this graph (and the order of presentation is unchanged), but broadly, those with the higher proportions of YLL in 1990 also show this in 2010.
2. In 2010, the most contrasting sub-regions are Western Europe and Australasia (with the highest proportions of YLD) and central and western sub-Saharan Africa (with the highest proportions of YLL). As we have seen, YLL represent years of life lost, and this will be particularly high where there is a high rate of child mortality, as the expectation of life lost is greatest. This, along with high adult mortality (e.g. with HIV/AIDS) explains why YLL still dominates DALYs in sub-Saharan Africa. In contrast, in more-developed countries, the majority of people survive into adulthood, and the morbidity associated with non-communicable diseases and other chronic conditions is reflected in the much higher proportions of DALYs being made up by YLD. We noted above that the period 1990 to 2010 has seen a general increase in the proportion of DALYs made up by YLD, and this reflects the epidemiological transition from communicable to non-communicable diseases (the latter with more YLD), and the fact that while life expectancy is generally increasing, the amount of time that people live with illness is not decreasing.

Exercise 10.1.5

1. a. In 1990, the top five causes globally were lower respiratory infections (e.g. pneumonia), diarrhoea, pre-term birth complications, IHD and stroke. In 2010, four of these (except pre-term birth complications) were in the top five, although the order had changed with cardiovascular causes being more important, and HIV/AIDS now being placed in 5th rank. Apart from the prominence of HIV/AIDS, reflecting the historical growth in that disease over the period, these findings emphasise the continuing importance of respiratory infections as a cause of child death globally (mainly in developing countries) and the steady transition to non-communicable diseases throughout the world.
- b. Based on the changes in DALYs (last column), three causes that have increased substantially are HIV/AIDS (351 per cent) as discussed above, low back pain (43 per cent), and diabetes (69 per cent).
- c. Based on the changes in DALYs, three causes that have decreased substantially are lower respiratory infections (−44%), diarrhoea (−51%) and protein-energy malnutrition (−42%), all major causes of death and disease in developing countries.
2. Such causes include conditions such as low back pain (rank 6), major depressive disorders (rank 11), and neck pain (rank 21). Although most of these do carry some increased risk of premature death, by far the majority of the disease burden arises from the disability experienced by sufferers, reflected in the YLD. For example, the disability weights in GBD-2010 for back pain vary between 0.27 and 0.37, depending on whether the back pain is acute or chronic and is with or without leg pain. Weights for major depressive disorder range between 0.16 for mild to 0.66 for severe episodes. Their high rankings in the 2010 listing reflect the fact that they are common, these conditions are often chronic in relatively young people (contributing many years of morbidity), and the disability weights are relatively high.

Section 10.2**Exercise 10.2.1**

1. At age 30–39 years, the AR is approximately $12/1,000/\text{year} - 3/1,000/\text{year} = 9/1,000/\text{year}$.
2. At age 60–69 years, the AR is approximately $87/1,000/\text{year} - 37/1,000/\text{year} = 50/1,000/\text{year}$.
3. We can see that, although the RR is larger in the younger age group, the AR is far larger in the older age group, since the absolute rates are much greater in the 60–69 year age group. As a result, raised blood pressure has a far greater public health impact at older ages than it does among young people.

Exercise 10.2.2

Strategies for reducing IHD risk associated with cholesterol

High-risk approach	Population approach
<p>Screening based mainly in primary care, which would include</p> <ul style="list-style-type: none"> ● Detection of people with raised levels of cholesterol and of other related risk factors ● Management by dietary advice, medical treatment (drugs), exercise, etc. ● Monitoring 	<ul style="list-style-type: none"> ● Dietary advice to the population (what is effective?) ● Labelling of fat content and type on food ● Agricultural practices that reduce the fat content of meat, etc. ● Policies to address issues of accessibility and cost of healthier foods, especially for poorer people ● Policy on exercise facilities and transport, etc., that can encourage greater physical activity

Exercise 10.2.3

1. The argument about the benefits of the population strategy is essentially the same, but turned around, when considering safety. A small risk (associated with a prevention measure) applied to a large number of people may result in a disturbingly large number of adverse events (non-cardiac deaths in the case of clofibrate). For people at high risk of a disease, it may be reasonable to accept the small risk associated with the intervention. This would not be acceptable for the rest of the population made up of people with low individual risk.
2. It is rare for studies (ideally trials) of prevention measures to be carried out that are large enough to detect the small risks we are concerned with. Even a study with (for example) 20,000 to 30,000 subjects is small in comparison with the application of a prevention measure to a population of tens or hundreds of millions of people. In addition, such trials can rarely be continued for more than a few years, so problems emerging after 5 to 10 years may well be missed.
3. Excerpt from *The Guardian* (newspaper), 12 February 2014:

Statins, the cholesterol-busting drugs already taken by 7 million people in England, should be offered to millions more who have only a low risk of heart disease or stroke, new NHS guidance says.

The National Institute for Health and Care Excellence (NICE) says in draft guidance which now goes out to consultation that the threshold for GPs to prescribe statins to their patients should be cut in half. At the moment they are given to those with a 20% risk of cardiovascular disease, but NICE says that should be reduced to 10%.

The plan is applauded by some doctors and heart disease charities but criticised by other experts, who warn that the drugs have side-effects – which can include muscle pains, memory loss and erectile dysfunction – and may not be as effective as is claimed.

They are also concerned that GPs will hand out pills instead of tackling the root causes of heart attacks and strokes by encouraging people to stop smoking, reduce their drinking, eat more healthy food and take more exercise.

Exercise 10.2.4

1. The 'prevention paradox' is described as the situation in which a measure that brings large benefits to the community offers little to each participating individual. Thus, a preventive action that reduces the risk of the majority (and therefore substantially reduces the population burden of the disease) does little for the individual because the RR for most people is only moderately raised; as individuals, they are not very likely to suffer from the disease anyway. In contrast, a preventive measure targeted at individuals with high RR can bring a substantial benefit to them but may do little to reduce the population burden of disease.
2. The statements apply as follows:
 - a. Population approach.
 - b. High-risk approach.
 - c. This statement does not really apply to either approach. Screening and treating individuals with average levels of the risk factor (around the mean) is not suitable for the population approach and is likely to be ineffective. It is very time-consuming, as there are so many people in this category, and adopting an individual approach to reducing only moderately raised risk is difficult for both the professional(s) and the individual concerned.
 - d. An example of government action that can reduce general alcohol consumption and that is consistent with a population approach.
 - e. Another example of a population approach and also of the 'safety is paramount' issue that is seen with the controversy surrounding the safety of adding fluoride to drinking

water for prevention of tooth decay. Let us consider, for the purposes of illustration, the consequences of adding to drinking water a substance with a small RR of serious disease (e.g. cancer). Since virtually everyone in the supplied area would use that water, the PAR would be substantial. The view that fluoridation may have serious adverse health effects is not accepted by those with responsibility for dental public health. However, concern about such effects (e.g. from lobbying groups) combined with freedom-of-choice arguments has effectively prevented the authorities in many parts of the country from proceeding with what is known to be a highly effective means of preventing tooth decay.

3. Summary of advantages and disadvantages of high-risk and population strategies

a. High-risk strategy	
Advantages	Disadvantages
Appropriate to the individual who is at high risk. Avoids interference with those who are not at special risk. Prevention becomes medicalised, which may be beneficial, but it can also have disadvantages if prevention requires action on factors (e.g. lifestyle, economic and fiscal policy) best addressed outside of the health system.	Success is only palliative and temporary from a policy perspective (although this may not be true for the individual concerned). It does little to alter the situations that determine exposure, or the underlying causes of the health problem. See also discussion of medicalisation of prevention in column 1.
Readily accommodated within the ethos and organisation of medical care. Easy for doctors to see high-risk people as 'almost patients', and treat them accordingly.	The strategy is problematic because so much behaviour is determined by social norms and what peers do. The high-risk approach requires that individuals identified as at high risk should subsequently behave differently from most of the rest of society.
To be effective, interventions can be quite resource intensive. Focusing on those at highest risk may offer more cost-effective use of resources.	Limited by the poor ability to predict the future outcome of individuals, even if in high-risk group.
Since interventions have costs (adverse consequences) as well as benefits, and assuming costs are similar for all, focusing on those at higher risk will improve the benefit-to-cost ratio.	Problems of feasibility and costs. Costs of screening, including human resources, and of treatment and monitoring may be high.
	The contribution to overall control of a disease may be disappointingly small.
b. Population strategy	
Advantages	Disadvantages
The strategy is radical, as it offers the chance to address the underlying causes.	May not be accepted easily, especially by the medical profession, although this is changing. Also it is difficult to see (and establish) the links between action and results, especially for individuals.
It is potentially powerful, as shifting the distribution of common risk factors can have substantial effects on the incidence of disease.	May not be feasible where political interests do not lie with the social, economic, and environmental well-being of the population.

It is behaviourally more appropriate, as the social, cultural, and economic determinants of behaviour are addressed. This means that individuals are not being asked to change their ways against the grain of society.

There may be substantial costs, with benefits being seen only in the long term, and they may be difficult to attribute to the investment.

Safety is paramount and is a potential concern for population approaches. On p. 1850 of paper A, the potential hazards of applying an intervention to a substantial proportion of the population are discussed. The example discussed is of a drug, clofibrate, which was used to lower cholesterol. The conclusion was that if (as in this case) a drug has even a slightly elevated RR and is given to very large numbers of people, the resulting numbers of serious adverse effects are disturbingly large.

Section 10.3

Exercise 10.3.1

1. High-risk strategy. The prevalence of cancer is $350/5,600 = 6.25\%$.
2. The sensitivity of the test = $325/350 = 92.9\%$ (95% CI = 90.2 to 95.6%). This sensitivity is high, and it is estimated fairly precisely due to the numbers of people in the study with disease ($n = 350$).
3. The specificity = $4,800/5,250 = 91.4\%$: this is also high. We have not calculated the 95 per cent CI, but this would be even more precise than for the sensitivity, as there are many more disease-free people in the sample.
4. The positive predictive value (PPV) = $325/775 = 41.9\%$, which is very low. As a consequence, subjects with a positive test (only 42 per cent of whom will turn out to have the disease) would need to be managed carefully and considerably. This is a result of the low prevalence of the disease, despite high values for sensitivity and specificity. Note that sensitivity and specificity are fixed characteristics of a given test, but the predictive values depend on the prevalence of the disease or characteristics being screened for. The (negative predictive value) NPV = $4,800/4,825 = 99.5$ per cent, so subjects with a negative test could be reassured with a high level of confidence.
5. The accuracy of the test = $(325 + 4,800)/5,600 = 91.5\%$, which is good for a screening test.
6. The likelihood ratios for (a) a positive test = $92.9/(100 - 91.4) = 10.8$, so it is nearly 11 times more likely that a positive test will be found in a person with the disease than in a person who is disease free. For (b) a negative test = $(100 - 92.9)/91.4 = 0.08$, so there is only an 8 per cent chance that a negative test result will be found in a person with the disease compared to a person who does not have the disease.

Exercise 10.3.2

1. Both distributions are unimodal and positively skewed. The distribution for respondents without back pain has a mean of 5.66 (SD 3.580), a median of 5 (IQR 3.0–7.0), and a range of 0–19. The distribution for those with back pain has a mean of 10.23 (SD 4.938), a median of 9.0 (IQR 7.0–13.0), and a range of 1–30. Thus, the distribution for those with back pain has a larger range and spread, with higher measures of central tendency, but there is considerable overlap between the two distributions. (For hypothesis tests of no difference between the two distributions, $p < 0.0005$ for both t -test and Mann–Whitney test.)

2. The cut-off score needed to ensure that test positive results included only people with back pain is 20, as the people without back pain have scores up to, but not in excess of, 19. At this cut-off, the test would therefore have a specificity of 100 per cent. However, the great majority of cases of back pain would be wrongly classified as false negatives.
3. The cut-off score needed to ensure that test negative results include only people without back pain is 1, as only scores less than 1 (zero) include no people with back pain. At this cut-off, the sensitivity is 100 per cent, as all people with back pain are included in positive tests. The great majority of people without back pain are wrongly classified as false positives.
4. We discuss the score with the best discrimination after introducing the receiver-operator characteristic (ROC) curve.

Section 10.4

Exercise 10.4.1

1. Suicide rates among older age groups have reduced markedly for men older than 45 years since the 1950s. However, for younger age groups there has been an increase. In fact, for men aged 15–44 years, rates doubled, although for males 15–24 and 35–44 years rates have decreased since the early 1990s.
2. This shows a period effect. A decline was experienced by all age groups at a particular point in time (late 1960s), albeit more marked for older age groups.
3. The reason for the period effect in the late 1960s is probably related to the introduction of natural gas for domestic use. Studies examining suicides by cause of death during this time noted that the reduction was associated with a reduction in the rate of suicides by domestic gas asphyxiation.

Exercise 10.4.2

1. These age groups are for the age at death by suicide.
2. For men aged 30 years (age group 25–34 years) dying in 1977 (period of death 1975–79), the suicide rate is (approximately) 15 per 100,000.
3. For men, the lines do not show a consistent parallel separation. In fact, while the older age groups do show some evidence of progressive (and almost parallel) decrease across age at death periods, the youngest age groups show the opposite – an increase across age at death periods. Thus, we might conclude that there is no evidence of a consistent period effect across all age groups, but one explanation could be different period effects affecting suicide in older and younger men in quite different ways. We did notice that the reduction in age-specific mortality seen in the 1960s, which we suggested in Exercise 10.4.1 may be evidence of a period effect resulting from gas detoxification, was much greater among older men. These findings from period effect analysis for men should also be considered in light of the cohort effect analyses, which are studied next.
4. For women in the age groups 35–44 years and older there is consistent evidence of decreasing suicide rates across the time periods, with lines more or less parallel. This indicates a period effect. At younger ages, there has been very little change in the rates.

Exercise 10.4.3

1. The suicide rate at age 32 years is (approximately) 11 per 100,000.
2. In contrast to Figure 10.4.3, there is no evidence in Figure 10.4.4 that the suicide rates were peaking progressively earlier as we look from 'earlier-born' to 'later-born' cohorts.
3. Since all birth cohorts (with overdose and gassing excluded) now show similar patterns of rates without progressively earlier peaking, we can assume that the period effect identified

in Figure 10.4.3 applies to suicide by overdose and/or gassing, but not to other methods. This fits the explanation we considered earlier, namely the detoxification of domestic gas. To this can be added the increasing use of catalysers in cars, which effectively remove carbon monoxide from the exhaust. To reiterate, these would result in period effects because they are introduced at a particular time in history, and they will tend to affect all (or at least a wide range of) age groups.

4. In conclusion, from the period and cohort analyses of male suicides, we can say that there does appear to be a period effect for suicide by overdose and gassing, and we have some plausible explanations for this (domestic gas and catalysers). Figure 10.4.3 showed evidence of a birth cohort effect, and the exclusion of overdose and gassing in Figure 10.4.4 illustrated this very clearly once the period effect was removed. A number of potential social, economic, and method (of suicide) explanations for this progressive rise in suicide rates with later-born cohorts of males since 1940 are considered in the discussion of paper B.

11

Probability Distributions, Hypothesis Testing, and Bayesian Methods

Introduction and Learning Objectives

This final chapter has two main aims. The first is to extend and consolidate your understanding of the principles behind hypothesis testing. The second is to serve as a reference and guide to selecting the correct hypothesis test for a range of data types and comparisons.

You are by now familiar with the concept of an hypothesis test, and you have used a number of the most common ones, including the chi-squared and the two-sample t -test (and z -test). We have also introduced somewhat more specialised tests, such as McNemar's test, which is used for matched categorical data (for example, in matched case-control studies – Chapter 5).

We have described and emphasised the assumptions for each test, and we have pointed out that these assumptions should be assessed prior to applying any given test. One such assumption concerns the distribution of the data and how this can be related to a theoretical probability distribution. For example, you will recall that the t -test requires that the distribution of any characteristic under investigation is (approximately) normal among the population from which the sample is drawn and also that the sample standard deviations for the two groups are similar. We begin this chapter by describing the main theoretical probability distributions to help your understanding of the theoretical basis of hypothesis testing. We also introduce some important probability distributions that have been referred to briefly but not covered in detail in previous chapters, such as the Poisson distribution, which can be used to make assumptions about count data for rare events, and the binomial distribution, which is useful in quantifying the accuracy of estimates of disease prevalence.

The rest of this chapter is aimed at extending your knowledge of hypothesis testing in a number of specific ways:

- We introduce you to other commonly used tests for particular situations, including the paired t -test, which is used for continuous paired (or matched) data; analysis of variance (ANOVA), which is used for continuous data with more than two groups; and the chi-squared test for trend, which is used for ordered categorical data to assess the significance of a trend in the relationship between an exposure and an outcome.
- We discuss the use of transformation to convert skewed data into normally distributed data, so that the assumptions of the hypothesis test can be met for continuous data.
- We discuss the use of non-parametric tests, which do not rely on the distribution of the data: These tests offer a very useful alternative to transformation for use with continuous data or with ordered categorical data.

In the final section, we discuss Bayesian methods. Bayesian methods are gaining increasing recognition and application, and they provide a different approach to determining the probability of an outcome.

Before introducing these new hypothesis tests, we review the key ideas about probability and the nature of probability distributions, and we see how these distributions apply to the common tests we have already used, such as the t -test and chi-squared test.

Learning Objectives

By the end of this chapter, you should be able to do the following:

- Describe the main theoretical probability distributions and their uses.
- Describe when it is appropriate to use a paired test (as opposed to an independent sample test) and carry out and interpret this test.
- Describe what is meant by transformation of data and how to select the appropriate power of the transformation.
- Describe when it is appropriate to use ANOVA and interpret the results of this test.
- Describe the key features that distinguish between parametric and non-parametric hypothesis tests.
- Describe, carry out, and interpret non-parametric hypothesis tests for continuous or ordered categorical data in different situations, that is, comparing two or more sets of paired (matched) and independent group data, and measuring associations.
- Describe what factors should be considered when choosing an appropriate hypothesis test.
- Select an appropriate hypothesis test for use with different types of data and analytic requirements (e.g. comparison, association).
- Determine whether a set of data complies with the assumptions of the selected test.
- Describe how the problem of multiple significance testing arises, and utilise a suitable correction method.
- Describe the main features of Bayesian methods and ways these can be applied in health research.

Resource Papers

This chapter has one resource paper that provides an example of the application of Bayesian methods.

Paper A

Powell, J., Geddes, J., Deeks, J., Goldacre, M., Hawton, K. (2000). Suicide in psychiatric hospital in-patients: risk factors and their predictive power. *Br J Psychiatry* **176**, 266–272.

11.1 Probability Distributions

11.1.1 Probability – A Brief Review

We previously discussed the importance of using data from a sample to draw conclusions about the population from which it was selected. For example, let's say we have identified an

improvement in survival from a new treatment compared to an old treatment in a sample of patients taking part in a trial. What we really want to know is whether this improvement would be seen in the whole population of patients, or whether the finding from the sample could be due to chance (sampling error). We introduced probability theory in Chapter 4, Section 4.2, and we considered how we could use it to relate samples to populations and hence to draw conclusions about populations.

Before discussing this application of probability distributions further, it is useful to review the main properties of probability, as summarised in the box below.

The Main Features of Probability

- The probability of an event can be defined as the proportion of times that the event would occur if the experiment or observation were repeated a large number of times.
- Probability lies between 0 (the event never happens) and 1 (the event always happens).
- The probability of the complementary event (the event does not occur) is 1 minus the probability of the event occurring.
- If two events are mutually exclusive (that is, when one happens the other cannot happen), the probability that one or the other happens is the sum of their probabilities. For example, a die may show a five or a six, but not both. The probability that it shows a five or six = $1/6 + 1/6 = 2/6$ ($1/3$).
- If two events are independent (that is, knowing when one has happened will not tell us anything about whether the other will happen), the probability that both will happen is the product of their probabilities. For example, if a coin is tossed twice, the probability of two heads occurring is $1/2 \times 1/2 = 1/4$. We will return to this example when we look at **probability distributions** in the following section.

11.1.2 Introduction to Probability Distributions

When collected together, data from a sample (observed data) form an **empirical or frequency distribution**. In introducing descriptive data in Chapter 2, Section 2.4.3, we graphically represented data from a continuous variable (particulate matter in air pollution, PM_{10}) in a histogram showing the frequency distribution of the data (Figure 11.1.1). We described this variable as having an approximately normal distribution.

Frequency distributions display the actual data for a variable taken from a sample. In contrast, a **probability distribution** is theoretical and shows how the total probability (which equals 1) is distributed among the different possible values of a variable. As with frequency distributions, probability distributions can be illustrated as a histogram. An important example of a probability distribution, the sampling distribution, was introduced in Chapter 4. Recall that this was a theoretical distribution, showing the probability of different values of the mean of a variable assessed through repeated samples of a given size. The concept of a probability distribution can probably be most easily understood through a simple example: Figure 11.1.2 illustrates the probability distribution for obtaining heads after two tosses of a coin.

The y -axis measures the probability for each value of the variable. We can see that the probability for obtaining zero (or two) heads is $1/4$ (0.25), and it is the probability of obtaining a head after the first toss of the coin ($1/2$) multiplied by the probability of obtaining a head after the second toss of the coin ($1/2$). The probability distribution represents the probability of all possible events (no heads, one head, and two heads), and therefore the total probability is 1.

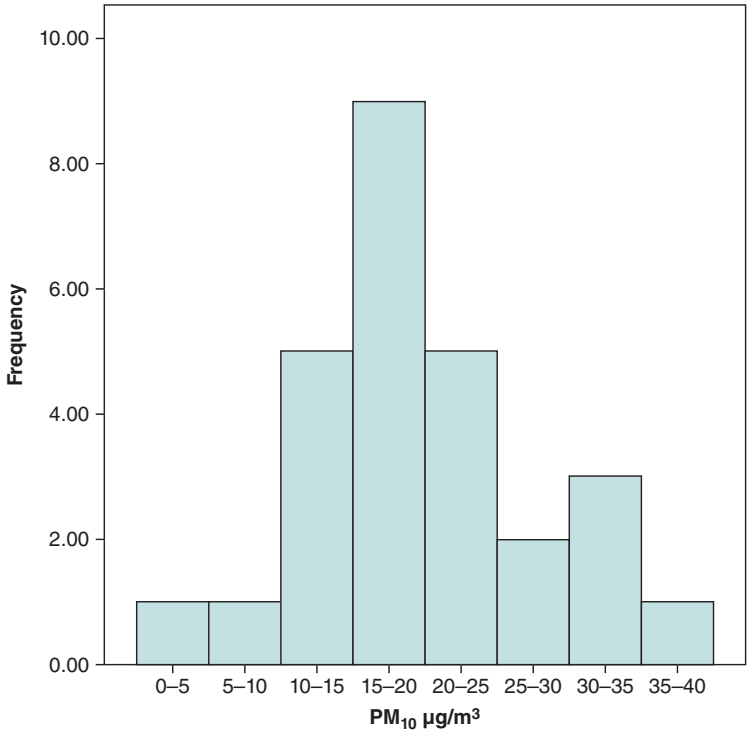


Figure 11.1.1 Frequency distributions of PM₁₀ concentrations (same data as Chapter 2).

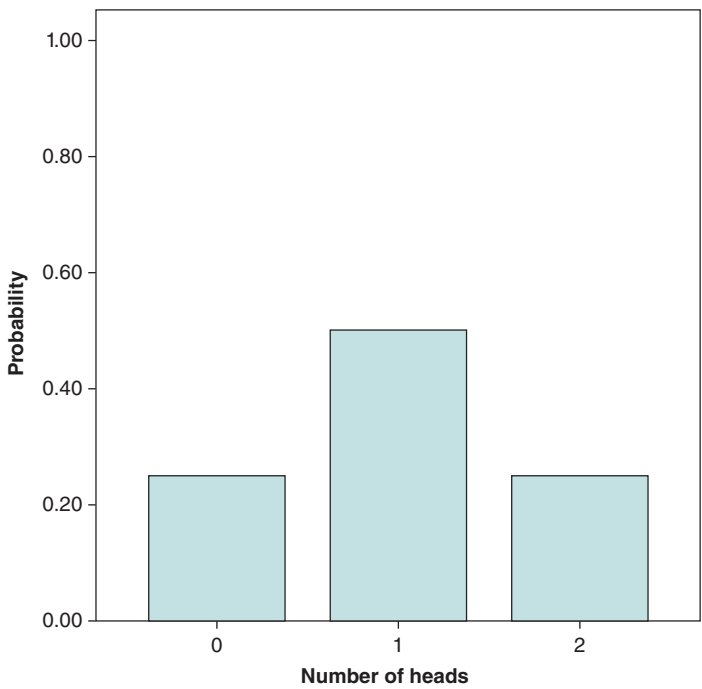


Figure 11.1.2 Probability distribution for obtaining heads after two tosses of a coin.

This example only applies to discrete (categorical) variables; for example, tossing a coin can only result in two outcomes: a head or a tail. For continuous variables, the probability of any particular value is zero as the number of possible values is infinite, so the **probability density**, rather than the actual probability, is plotted on the y -axis. Figure 11.1.3 illustrates the probability distribution representing the probability density of a continuous variable such as PM_{10} that we studied in Chapter 2.

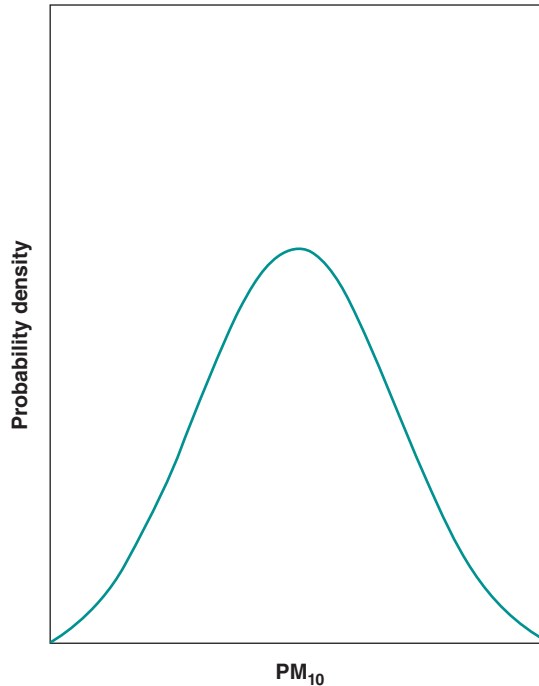


Figure 11.1.3 Probability distribution for PM_{10} data showing probability density.

As with discrete variables, the probability distribution represents the probability of all possible values of continuous variables. Therefore, the total area under the probability density curve is 1. The shapes of probability distributions are characterised by the same features as frequency distributions. These features include the number and position of modes that indicate the values at which the probability (or probability density for continuous variables) reaches a maximum, and the skewness of the distribution.

When observed data form a frequency distribution that approximates a particular probability distribution, theoretical knowledge of that probability distribution can be used to answer questions about the data. The most common example is the normal distribution that was introduced in Chapter 4. We will now look at this in more detail, as well as other important probability distributions.

11.1.3 Types of Probability Distribution

We already came across some of the more common probability distributions when we introduced hypothesis tests for continuous variables with a sample size of more than 30 (z -test) and a sample size of less than 30 (t -test, although applicable for all sample sizes) and for categorical variables (chi-squared test) (Table 11.1.1).

Table 11.1.1 Hypothesis tests and probability distributions described previously.

Probability distribution	Hypothesis test	Where test is described
Normal distribution	z -test	Chapter 5, Section 5.5.4
t -Distribution	t -test	Chapter 5, Section 5.5.4
Chi-squared distribution	Chi-squared test	Chapter 5, Section 5.5.3

There are a number of other probability distributions that are commonly used for data with various properties. Visual inspection of a set of observed data can often indicate which probability distribution is most suitable. Depending on whether the variable is continuous or discrete, the probability distribution can be continuous or discrete, examples of which are shown in Table 11.1.2.

Table 11.1.2 Example of continuous and discrete probability distributions.

Continuous	Types of data
Normal (Gaussian) distribution	Continuous
Log-normal distribution	Continuous
t -Distribution	Continuous
Chi-squared distribution	Continuous
F -Distribution	Continuous
Binomial distribution	Categorical (dichotomous)
Poisson distribution	Discrete (counts)

Before we describe the practical application of probability distributions in making inferences from samples to populations, it is useful to summarise briefly each distribution and the *parameters* used to describe the distribution. Parameters are the mathematical properties that determine the shape and location of the distribution, allow the determination of the probability associated with specified sections of the distribution, and so on. The mathematics involved in describing these probability distributions can be rather complex and is beyond the scope of this book. It is, however, useful to be familiar with the different types of probability distribution, how these differ for different types of variable, and how each distribution is described by its parameters. This will help you understand the theory behind *parametric* hypothesis testing and why we also sometimes need to use other procedures including *transformation* (Section 11.2) and *non-parametric* tests (Section 11.3).

Continuous Probability Distributions

Continuous probability distributions typically have no upper limit, and some have no lower limit. However, there are three other general points applicable to all continuous probability distributions, some of which have been touched on earlier. First, because a continuous variable has an infinite number of possible values, the probability of any particular value is zero. It is therefore only possible to calculate the probability that the variable will take a value within a specified range.

Second, as we saw in Figure 11.1.3, if all possible values of a continuous variable are plotted on a horizontal axis, we can draw a probability density curve using the equation of the probability distribution. The total area under this curve is 1, equal to the total probability.

Third, to use a probability distribution, the area corresponding to a particular range of values is typically considered. We saw this when we introduced confidence intervals (CIs) in relation to the sampling distribution (Chapter 4). Because the sampling distribution has a **normal** probability distribution, we know that approximately 95 per cent of values lie between ± 1.96 times the standard deviation (σ) of the population mean (μ). Therefore we can say that there is a 95 per cent chance that we choose a sample whose mean lies in the interval $\mu + / - 1.96 \sigma / \sqrt{n}$.

The Normal Distribution

The normal distribution is the most commonly used probability distribution. We described the normal distribution in detail in Chapter 4, Section 4.4. A summary of what we already know about the normal distribution is given in the box below.

The Normal Distribution

- is described by two **parameters** (the **mean** (μ) and the **standard deviation** (σ))
- is **unimodal** and bell-shaped, symmetric about its mean
- usually has no upper or lower limits
- has a curve that is shifted to the right if the mean is increased and shifted to the left if the mean is decreased
- has a curve that is flattened if the standard deviation is increased and becomes more peaked if it is decreased
- is used to analyse continuous data from one or two large ($n > 30$) samples
- can be used to define **confidence intervals**.

As we shall see later in this chapter, other probability distributions (including discrete distributions) approximate the normal distribution under certain circumstances, and therefore the normal distribution can often be substituted for these other distributions.

It is possible to test data statistically to determine whether they follow a normal distribution by either the Shapiro–Wilk W test or the Kolmogorov–Smirnov test. Details of these tests can be found in standard statistical reference books.

There are an infinite number of normal distributions, depending on the parameters (mean [μ] and standard deviation [σ]) of the distribution. One frequently used normal distribution is the **standard normal distribution**. This distribution has a mean of 0 and a standard deviation of 1. The standard normal distribution is particularly useful because the probabilities relating to the distribution are widely available in tables, although computer software incorporates these for all probability distributions used for hypothesis testing.

The t -Distribution

We used Student's t -distribution previously when carrying out an hypothesis test on samples with fewer than 30 subjects (the t -test) (Chapter 5, Section 5.5.4). The distribution has only one parameter, and this is the degrees of freedom, equal to the number in the sample minus 1 ($n - 1$). The t -distribution is similar in shape to the normal distribution but is more spread out with longer *tails* due to the typically small sample size. As the sample size increases, the shape of the t -distribution becomes increasingly like the normal distribution. This is one reason the t -test is appropriate for both small and large samples.

The t -distribution is used for a number of hypothesis tests, including the two-sample (or independent sample) t -test and the **paired t -test** (introduced in Section 11.3.6 of this chapter). The distribution can also be used to calculate CIs for sample means.

The Chi-Squared Distribution

As we have seen, the chi-squared distribution is a particularly useful probability distribution for analysing categorical data (Chapter 5, Section 5.3.3), although it is still a **continuous probability distribution**. This is because the critical value of the chi-squared test (the value obtained from the test relating to a specific probability, e.g. 0.05 or 0.01) forms a chi-squared distribution. When we introduced this hypothesis test, we saw that the chi-squared probability distributions are characterised by their degrees of freedom. Typically, the chi-squared distribution is highly positively skewed; however, as the degrees of freedom increase, the shape of the distribution approaches normality.

In Chapter 5, we saw how the chi-squared distribution could be used in the chi-squared hypothesis test to compare two or more proportions. Later in this chapter, we look at other applications of this distribution, as in comparing three or more groups of continuous or ordered categorical data when using the Kruskal–Wallis hypothesis test.

The F-Distribution

If two normally distributed populations have equal variances (e.g. the square of the standard deviation), the ratio of the variances of samples drawn from each should follow an F -distribution. The **F-test** is carried out using an **F probability distribution**, which is positively skewed. One important and common application is the use of the F -test prior to carrying out a t -test, in order to investigate whether the variances of the two groups to be compared are similar – one of the key assumptions that needs to be met for the t -test. Another common use is in comparing three or more means in one-way analysis of variance (ANOVA), to which we will return in Section 11.3.8.

Discrete Probability Distributions

As illustrated in Figure 11.1.2, for categorical variables it is possible to derive the probability of every possible value, and the sum of all these probabilities is 1. The two most common probability distributions for data with these properties are the binomial and Poisson distributions, both of which were briefly introduced in Chapter 5, Section 5.9.5.

Binomial Distribution

If there are a number (n) of what are often termed ‘trials’ (a clear example being tosses of a coin) that are independent of each other, and in which the outcome is either ‘success’ (heads) or ‘failure’ (tails), the number of observed successes follows a **binomial distribution**. In our example in Figure 11.1.2, we had a variable describing the number of heads (successes) in two tosses of a coin, taking values of 0, 1, and 2. This was an example of the binomial distribution.

Let us now look at an example of the binomial distribution in a health context. If we carried out a survey investigating the prevalence of smoking in a population (for which we use the symbol π) and took a random sample, the probability of any subject chosen being a smoker is p . In effect, as we assess each person, we are carrying out a series of independent assessments (referred to as ‘trials’ above) where the person either does or does not smoke. In effect, we can think of this as having a series of independent trials, each with the probability of ‘success’ (being a smoker) of p .

If we were to repeat the study with a series of randomly selected samples, the number of successes (smokers) in these repeated samples will follow the binomial distribution. The properties

of the binomial distribution enable us to say how accurate the estimate of prevalence obtained is. We can use the binomial distribution whenever we have a series of independent assessments (trials) with two possible outcomes. If we treat a group of patients, the number who recover has a binomial distribution. If we measure the blood pressure of a group of people, the number classified as hypertensive has a binomial distribution, and so on.

The binomial distribution has two *parameters*, the number of individuals in a sample (or repetitions of the individual assessments), n , and the true probability of success in each individual assessment, π . Because the number of different probabilities in a binomial distribution can be very large, we usually need to summarise these probabilities in some way. Just as a frequency distribution can be described by its mean and standard deviation (or variance, which is the square of the standard deviation), so can a probability distribution and its associated variable. The mean, which is the value for the variable we expect if we look at n individuals, is $n\pi$. The variance is $n\pi(1 - \pi)$. When n is small, the distribution is skewed to the right if $\pi < 0.5$ and to the left if $\pi > 0.5$. As n increases, the distribution becomes more symmetric and approximates the normal distribution if both πn and $n(1 - \pi)$ exceed 5.

In health studies, the binomial distribution can be used if researchers are interested in whether or not an event has occurred rather than the magnitude of the event. For example, when looking at a smoking-cessation intervention, we might be more interested in whether individuals successfully stop smoking than in how much they have reduced their smoking in terms of numbers of cigarettes smoked. The binomial distribution is the most commonly used distribution to describe discrete data.

Poisson Distribution

Another important discrete probability distribution useful in health research is the *Poisson distribution*. This describes variation in the rate at which usually fairly uncommon events occur over time, or spatially, provided that these are:

- Independent of each other, meaning that the timing of an event (or its location) does not depend on another event that has already taken place;
- Occur randomly over time or in geographical space (Figure 11.1.4).

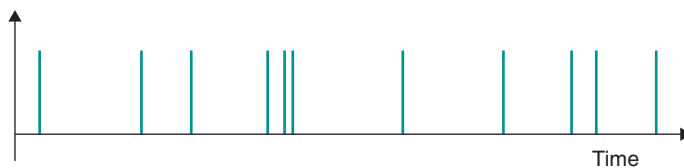


Figure 11.1.4 Events such as these that occur randomly in time and are not dependent on the timing of other events are described by the Poisson distribution.

Examples of the Poisson distribution A classic example of the Poisson distribution is radioactive emissions. You may be familiar with the Geiger counter that clicks as it is brought close to a radioactive source. These clicks speed up if either the sensor is brought closer or there is a stronger source. Each click represents detection of an emission, that is, an event.

This is an example of a Poisson distribution because the timing of each detected emission (click) is independent of when others occur, and over time the pattern is random. Although a clear example, it is not particularly relevant to most health research. The same ideas can be applied to rates of disease (or death), where, for example, each new (incident) case is an event.

Again, these must be independent and occur randomly in time or space. Some types of disease incidence, such as infectious disease, are not independent. Generally, however, so long as there is no strong evidence of clustering of cases in time or space, it is useful to apply the Poisson distribution.

Key Features of the Poisson Distribution The shape of the Poisson distribution depends on just one parameter, the *mean number of events* occurring over periods of the same length (or over equal regions of space). This is called μ . Figure 11.1.5 shows the shapes of Poisson distribution for four different values of μ .

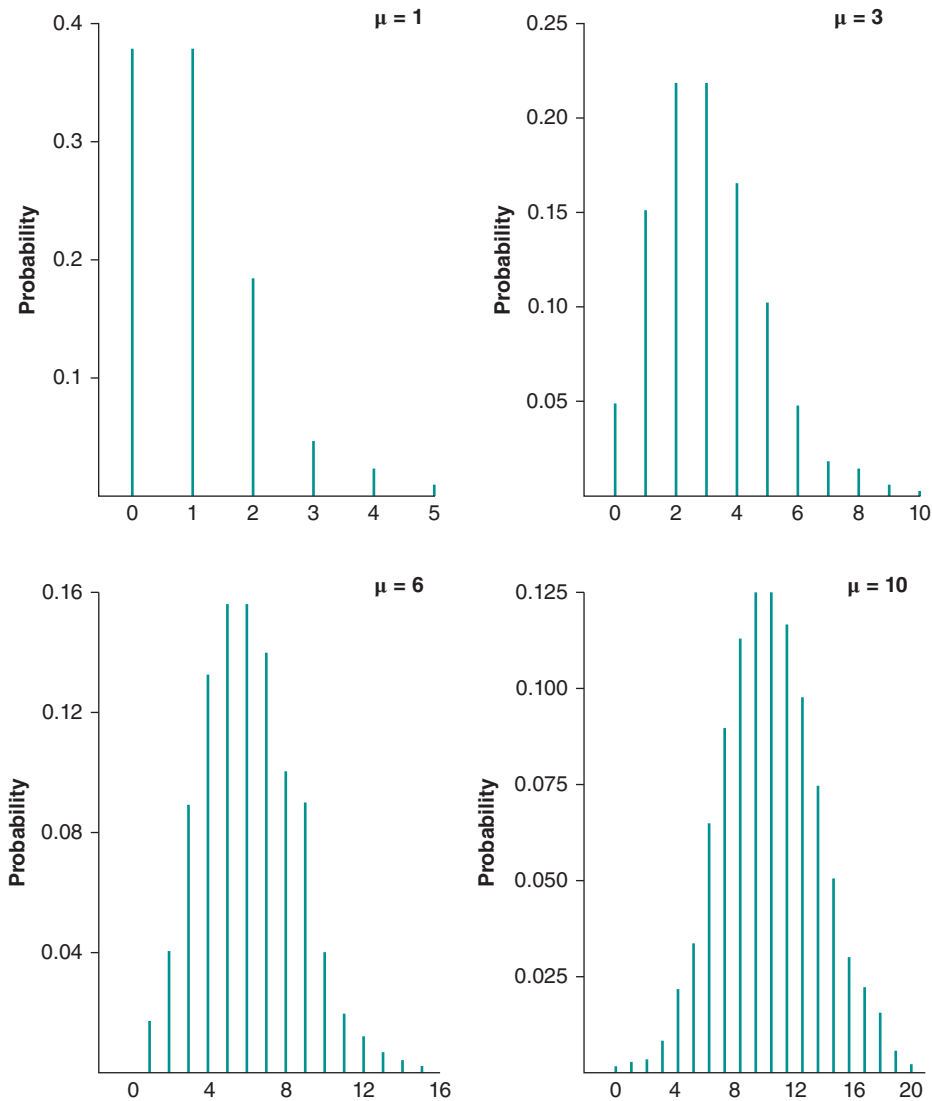


Figure 11.1.5 Poisson distribution for increasing values of μ . *Source:* Kirkwood 2010. Reproduced with permission of John Wiley & Sons.

When events are very rare, there is a good chance of there being none or very few over the time period, and the distribution is therefore very positively skewed. However, as soon as we reach a mean of 10 or more, the distribution approximates closely to the normal distribution.

Summary: Probability Distributions

- A probability distribution shows the probabilities of all possible values of a variable.
- It is used to calculate the theoretical probability of different values occurring.
- All probability distributions are described by one or more parameters (e.g. the mean and standard deviation, the degrees of freedom, etc.).
- The most appropriate probability distribution to use for hypothesis testing, calculating 95% confidence intervals, etc., will depend on the nature of the data being analysed.

11.1.4 Probability Distributions: Implications for Statistical Methods

Many statistical methods are based on the assumption that observed data are a sample from a population and that the sample, if repeated, has a distribution with a known theoretical form. It is not possible to know whether this assumption is true, but only whether it is reasonable. If the assumption is reasonable, we can use methods making distributional assumptions (known as *parametric methods*) that rely on probability distributions to calculate CIs and carry out hypothesis tests. We will look in more detail at the application of probability distributions to determining the standard error and precision of estimates and to comparison of groups using an hypothesis test later in the chapter.

If the assumption appears to be unreasonable, we need either to use statistical methods (e.g. hypothesis tests) that do not make assumptions about distributions (*non-parametric methods*) or to *transform* our variable so that it does meet a distributional assumption. These techniques are described in more detail in the following sections.

The following exercise will help to consolidate the ideas about probability and probability distributions we have covered in this section.



Self-Assessment Exercise 11.1.1

Probability Distributions

1. If the probability of a man aged 50 of having high blood pressure is 0.15 and the probability of his having a cold is 0.10, which of the following statements is correct?
 - (a) The probability of his having both conditions is 0.015.
 - (b) The probability of his having both conditions is 0.25.
 - (c) If the man has high blood pressure, the probability of his also having a cold is 0.15.
2. Which of the following variables follow a binomial distribution?
 - (a) The height of a sample of schoolchildren.
 - (b) The proportion of men who smoke.
 - (c) The number of admissions to hospital for a rare disease per month.
 - (d) The number of people with back pain (assessed as yes or no) in a random sample of farmers.
3. Which of these statements is correct about the normal distribution?
 - (a) It has a symmetric distribution about its mean.
 - (b) It is followed by all continuous variables measured in humans.

- (c) It is the distribution towards which the binomial distribution will approximate as the number of trials (or the sample size) increases.
- (d) It is the distribution towards which the Poisson distribution will approximate as the number of events becomes rarer.
- (e) It is a common probability distribution followed by many variables.

Answers in Section 11.6

11.2 Data That Do Not Fit a Probability Distribution

As discussed in Section 11.1, there are often occasions when our data do not meet the assumptions required by statistical parametric methods using probability distributions. Parametric hypothesis tests are based on probability distributions and thus rely on more assumptions than do non-parametric tests. We must decide whether our data meet the requirements of a particular test by exploring the data (pictures and summaries) and using our own experience and knowledge of the situation. We may find, for example, that

- the data are skewed and do not appear to be from a normally distributed population; or
- the standard deviations of two samples are very different; or
- the relationship between two variables in a scatterplot looks curved, not linear.

When the evidence shows that the assumptions of a parametric test are not fulfilled, we may approach the analysis in one of three ways:

1. Rely on the **robustness** of the test we want to use (Section 11.2.1).
2. **Transform** the data into a new set of data that are consistent with the assumptions, and carry out the parametric test on the new data (Section 11.2.2).
3. Use a **non-parametric** hypothesis test (Section 11.2.3).

11.2.1 Robustness of an Hypothesis Test

We say that a test is **robust** to departures from the assumptions if we can still obtain valid results when the assumptions are not strictly met. For example, if we have large samples, the t -test can still be used even if the data are not from normally distributed populations: t -tests are robust to departures from the normal distribution for large samples. Also, as a general rule, the two-sample t -test is fairly robust against unequal population standard deviations: If the ratio of the variances between the two groups is not more than 2, the t -test will generally remain valid at least when the sample sizes are similar. It is good practice, however, to carry out a test for equality of variance prior to conducting a t -test, and this is routinely done by software such as SPSS.

Since it may be difficult to decide just how far one can depart from the test assumptions, it is probably safer not to rely on robustness without seeking advice. If it is not safe to do so, we should consider either transforming the data or using a non-parametric test.

11.2.2 Transforming the Data

The most common departure from parametric assumptions that we come across is skewed data. We know that survival-type data are generally positively (right) skewed, and so are many other measurements. Skewed data from human populations are almost always positively skewed, that is, with a long right tail, as in the example of survival data shown in Figure 11.2.1, taken from Chapter 8.

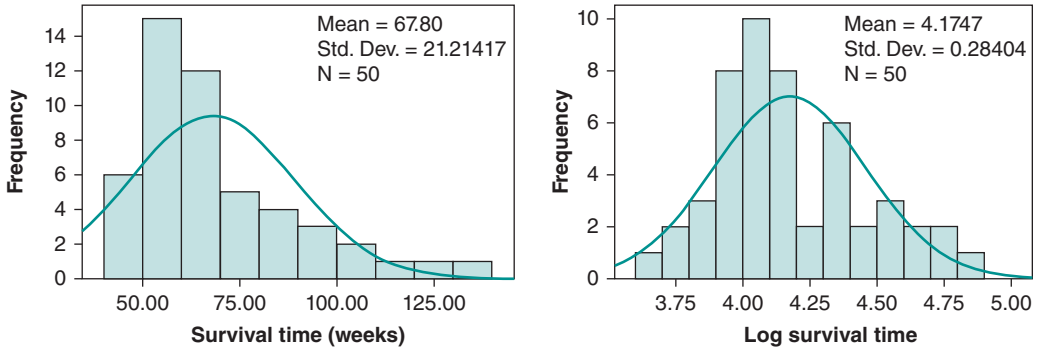


Figure 11.2.1 Effect of log transformation of skewed survival data.

We can sometimes overcome the problems of data that are skewed or not normally distributed by simply changing the scale of measurement; that is, by **transforming** the data. This means carrying out some calculation on the data. Examples of transformations include taking the logarithm of each data value (the most commonly used) and squaring each data value. If we can transform the data to a new set of values that meet the assumptions of a parametric test (e.g. that they have a normal distribution), we can carry out the appropriate hypothesis test on the transformed data.

The Log Transformation

The most widely used transformation for reducing positive skewness is the **log** (short for logarithmic) transformation, having the effect of stretching out the smaller values and squeezing together the larger values, and resulting in a more normally distributed set of data. This is known as the **log-normal probability distribution**. Figure 11.2.1 illustrates how log transformation results in a distribution that approximates the normal distribution far more closely than the original skewed distribution.

We usually use the **natural logarithm** for this transformation. This is sometimes written **ln** instead of **log**, and the relevant key on most calculators is labelled **ln**. Note that it does not matter whether we use natural logs or logs to some other base (for example, the next most commonly used is base 10): The effect on the data is the same. However, we do need to know which base has been used when it comes to interpreting the results. We now look at another example of skewed data and see how parametric assumptions can be applied.

Table 11.2.1 shows the vitamin D levels in the blood of 26 healthy men (Hickish, 1989). Plotting these data in a histogram shows them to be positively skewed, and taking logs

Table 11.2.1 Vitamin D levels in 26 healthy men.

14	25	30	42	54
17	26	31	43	54
20	26	31	46	63
21	26	32	48	67
22	27	35	52	83
24				

Source: Hickish 1989.

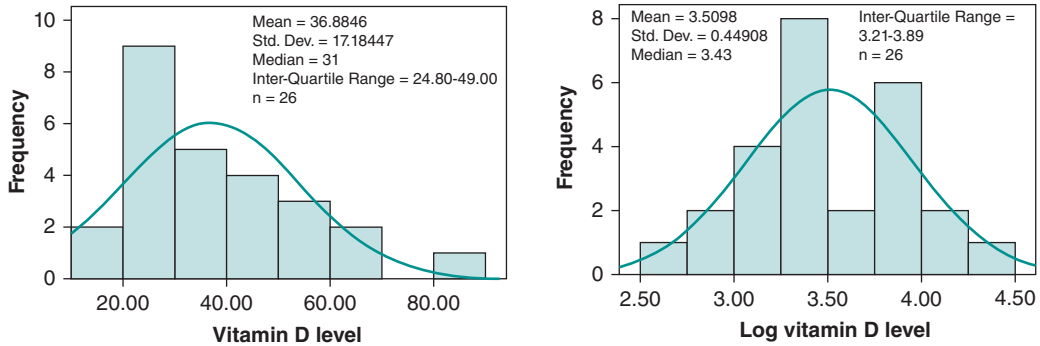


Figure 11.2.2 Skewed data on vitamin D levels (left) and natural log transformation (right).

(natural) results in a distribution that approximates quite closely to the normal distribution (Figure 11.2.2).

The mean of the untransformed data is 36.88 (SD = 17.18) and the median is 31, consistent with quite marked positive skew. In view of this degree of skew, we would carry out analyses on the *transformed data*. For example, we can state that the mean of the $\log(n)$ vitamin D data is 3.51, and the 95 per cent CI for the population mean is 3.33–3.69. At present, these are still $\log(n)$ values of the vitamin D levels.

Interpreting the Results

It is usually more meaningful to give results in the original scale. To do this, we have to do the opposite of taking logs, which is called *exponentiating*, or taking the *antilog*. We used this technique (exponentiation) to derive odds ratios from the regression coefficients in logistic regression, and hazard ratios in Cox regression. If x is an original data value, and $u = \log x$ is the transformed value, then $x = e^u$. This can also be written $x = \exp(u)$. When we exponentiate the mean of the logged data, we do not get back to the arithmetic mean of the original data. Instead, we obtain the *geometric mean*.

The geometric mean is similar to the median of the original data and is always less than the arithmetic mean. It is not greatly influenced by very large values in a skewed distribution, so it is a better representation of the average than the mean. Transforming our results for the mean and 95% confidence limits for the log-transformed data back to the original scale, we have

$$e^{3.5098} = 33.44; e^{3.3284} = 27.89; e^{3.6912} = 40.09$$

So we can say that the estimate of the geometric mean is 33.44 with a 95 per cent CI (27.89–40.09). The geometric mean of 33.44 compares with 36.88 and 31.0 for the original mean and median, respectively, so it lies between the mean and median but is slightly closer to the median. As an alternative, we could have summarised the untransformed vitamin D data by using the median (31) and interquartile range (IQR) (24.8–49.0). As we shall see when we look at non-parametric hypothesis tests in Section 11.3, the main advantage of using the mean and standard deviation derived from the transformed data is that we can still use *parametric* statistical methods, which not only produce CIs but are also generally more powerful.

Other Transformations of Data

There are a number of other transformations that we can apply to positively skewed data, such as the square root (\sqrt{x}) or reciprocal ($1/x$) transformations, but these are less commonly used.

The square root transformation is less dramatic than taking logs, but the reciprocal transformation is stronger and so can be useful for very skewed distributions.

Negatively skewed data can be made more normal with a power transformation, such as a square (x^2) or cube (x^3) transformation. There are many possible transformations; together, these are sometimes referred to as the ***ladder of powers***:

$$\dots, x^{-2}, x^{-1}, x^{-1/2}, \log x, x^{1/2}, x^1, x^2, \dots$$

The transformation x^1 leaves the value of x as it is. Provided $x > 1$, powers below 1 reduce the high values in a data set relative to the low values, and powers above 1 have the opposite effect of stretching out high values relative to low ones. The farther up or down the ladder from x^1 , the greater is the effect. Figure 11.2.3 illustrates the most common transformations for reducing skewness (x is the original data value).

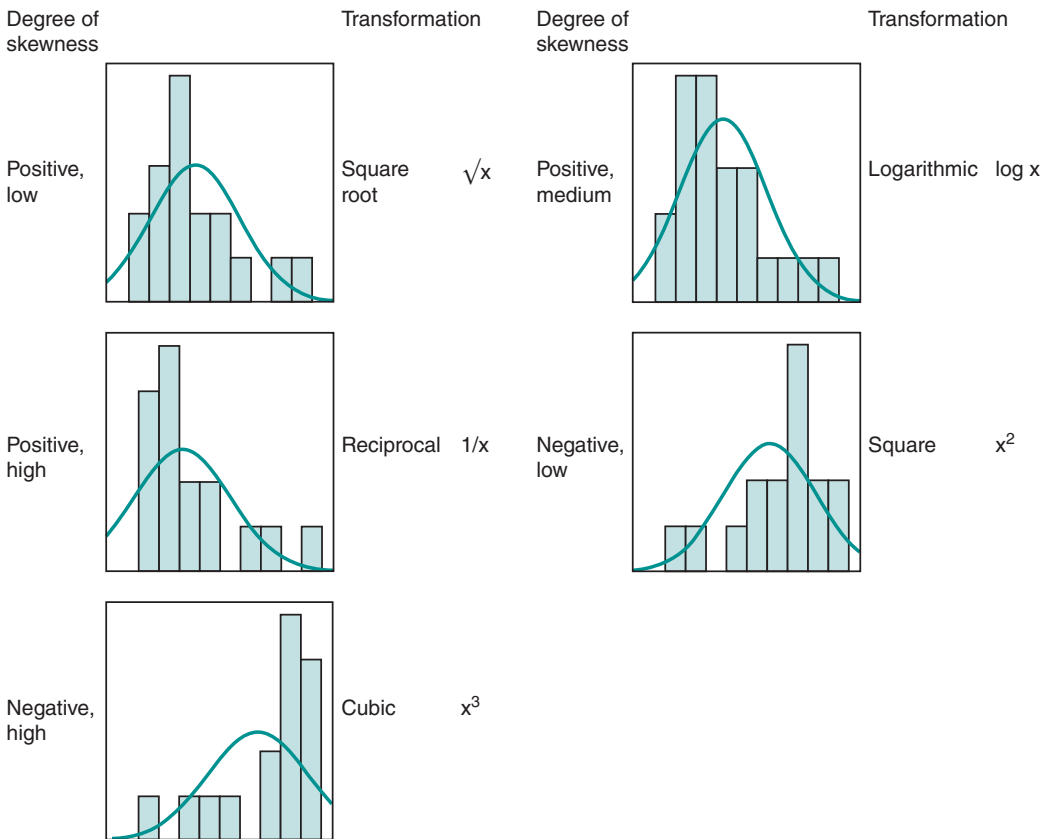


Figure 11.2.3 Common transformations for reducing the degree of skew in distributions.

Issues in Interpretation of Transformed Data

The main problem in applying a transformation is how to interpret the results afterwards. The advantage of the log transformation is that the geometric mean and CI are directly related to the original data in an interpretable way. Unfortunately, no other transformation allows easy interpretation of the back-transformation in this way, and CIs in the transformed units are very difficult to interpret.

It is usually interpretability and the importance of a summary estimate and CI that determine whether the data should be transformed or not. It is quite common, for example, to see clinical laboratory measures that have a skewed distribution, such as serum bilirubin or lipoprotein, summarised by a geometric mean and CI (that is, a log transformation has been carried out). On the other hand, a clinical measure, such as time since diagnosis, might not be so easy to interpret after transformation, and so a non-parametric method could be adopted instead.

So, the decision about whether or not to transform the data will depend on a number of factors. Of course, there is also the option of using both transformation and non-parametric methods for testing and presenting the results.

Summary: Transformation of Data

- Transformation is used to normalise data in order to apply parametric hypothesis tests.
- Data may be positively or negatively skewed: Positively skewed data are much more common.
- There are a range of transformations that can be used, lying on a ladder that defines the direction in which each alters the shape of the distribution and the strength of that change.
- Log transformation, which may be to any base (natural, base 10, etc.), is most commonly used for positively skewed data. It is the transformation that allows the most meaningful interpretation of means and CIs.
- Analysis is carried out on the transformed data, and (for log transformations) results are then exponentiated to obtain values in the original scale that are more meaningful to interpret.
- The exponentiated mean derived from log transformation is known as the geometric mean; it is always smaller than the mean of the original data, and it is usually closer to the median.
- The decision about whether to use transformation or non-parametric methods for skewed data depends on a number of factors, including how interpretable geometric means will be, and the importance of being able to provide 95 per cent CIs.
- There is no reason why a combination of methods should not be used.

11.2.3 Principles of Non-Parametric Hypothesis Testing

We have seen that parametric methods for hypothesis testing require that assumptions pertaining to a probability distribution should be met. Where this is not the case, or it is not easy to check (as with a small data set), or it is decided that transformation would not be appropriate, **non-parametric** methods can be used. Such methods do not require the assumption of the normal distribution or any other probability distributions and are intended to overcome these problems.

Non-parametric hypothesis tests are sometimes called **distribution-free** tests, because the distribution of the outcome variable can take any shape and may be very skewed or non-normal. However, even though they do not have the strict assumptions of parametric tests with the appropriate probability distribution, this does not mean that the tests are assumption free. When we compare two independent groups of continuous observations, for example, the distribution of the values within each group is assumed to be similar in shape and have similar variance, but the location of the distributions (e.g. medians) differs. Other assumptions will be outlined as we consider commonly used non-parametric hypothesis tests in more detail in Section 11.3.

Non-parametric methods can be used for continuous and ordered categorical data, such as scores. The most critical feature of these tests is that the **rank** ordering of the data is used, rather

than the actual values themselves. The rank of a value shows its position in an ordered list of the data. For example, the ages 31, 57, 46, 65, and 49 have the following ranks (Table 11.2.2):

Table 11.2.2 Rank ordering.

Age	Rank
31	1
46	2
49	3
57	4
65	5

If two or more values are the same (tied), their rank is the average. In the following example (Table 11.2.3) where two people are aged 49, ranks 2 and 3 are averaged, $(2+3)/2 = 2.5$.

Table 11.2.3 Rank ordering.

Age	Rank
31	1
49	2.5
49	2.5
57	4
65	5

Using rank ignores the information about the actual numerical values. This is why these methods are also particularly suitable for data that are scores rather than measurements, whether the scores have many or few values.

Non-parametric tests are very useful when assumptions of a normal distribution are not met, but they are less powerful than the corresponding parametric tests where assumptions for the latter are met. This means that if we use a non-parametric test, we are less likely to reject the null hypothesis when it is false. So, if the assumptions are valid, it is better to use a parametric test.

The choice of statistical method always depends upon several factors, such as the normality assumption, the importance of obtaining an estimate and CI, the ease of calculation and interpretation, and so on. You will, therefore, often encounter use of non-parametric methods in the medical and health literature. In the next section we look at non-parametric hypothesis tests alongside their parametric equivalents and discuss how and when each should be used.

11.3 Hypothesis Testing: Summary of Common Parametric and Non-Parametric Methods

11.3.1 Introduction

In this section we cover all the commonly used hypothesis tests. We have already introduced some of these in previous chapters, notably the chi-squared test and t -test, but others (including

non-parametric tests) are new. For tests that have been previously described, we revisit the circumstances in which they are used and refer to the relevant chapter and section of the book. Hypothesis tests that have not been covered elsewhere in the book are described in detail with the use of worked examples. Section 11.4 provides a guide in the form of a reference table to help you choose an appropriate hypothesis test for any given comparison of data.

To start with, we review the key principles of hypothesis testing we have covered in previous chapters.

11.3.2 Review of Hypothesis Tests

We have seen that we use different tests for different data types and sample sizes (for example, categorical or continuous data, paired data, and small or large samples), and to answer different types of questions, such as (i) is there a difference between means? and (ii) are these two variables related? When we want to test an hypothesis, we must use a test that is appropriate to the data and to the question we want to answer. The following exercise will help you to recall some of the features of the two most commonly used hypothesis tests: the chi-squared test (Chapter 5, Section 5.5.3) and the t -test (Chapter 5, Section 5.5.4).



Self-Assessment Exercise 11.3.1

1. What assumptions should be met for the chi-squared test to be valid?
2. In respect of the two-sample t -test, (a) what assumptions do we make? and (b) how might you check that these assumptions are reasonable for your data?

Answers in Section 11.6

It is important always to check that the data meet the requirements of the proposed test – for example, by viewing a histogram to see whether we can reasonably assume that the data are from a normal distribution. If this is not so, the test may not be valid, and the results could be misleading.

11.3.3 Fundamentals of Hypothesis Testing

Let us also recall why we carry out hypothesis tests in the first place, as well as some general rules that all hypothesis tests follow. These points are described in detail in Chapter 5 (cohort studies) with the introduction of the chi-squared test and t -test for looking at the results from the BRHS study.

Inference

We observed a number of interesting results in the BRHS; for example, that smoking was associated with an increased risk of IHD and that blood pressure differed between men with and without IHD. The fundamental question, however, was whether these results, observed in this sample of men, reflected the real situation in the population of middle-aged men in Britain from whom the sample was drawn. We noted that estimates from any possible sample vary from sample to sample (sampling error) and that hypothesis testing allows us to assess objectively whether the results observed in our sample are evidence of a real difference in the population or whether they are simply due to chance. Hypothesis testing, therefore, is an essential tool in

the process of *inference*, relating findings from our sample to the population from which the sample was drawn.

The Null Hypothesis

Hypothesis testing essentially examines the probability that an estimate (e.g. the difference in mean systolic blood pressure) seen in the sample has occurred by chance and therefore does not reflect a true difference in the population. This premise means the starting point of hypothesis testing is to state the null hypothesis (H_0) (that there is no real association, difference in means, etc., in the population) and then the alternative hypothesis (H_1) (that there is a real association, etc.). If the null hypothesis is found to be unreasonable – that is, the data are not in agreement with such an assumption – we reject the null hypothesis in favour of the alternative hypothesis.

The Test Result: The p -Value

We use an hypothesis test to determine the probability that the observed estimates from our sample occurred by chance, under the assumption that there is no true difference or association (H_0). We do this by using either probability distributions (parametric methods) or non-parametric methods. The usual convention is that if the probability of obtaining the observed difference (in this case), or in a more extreme one, is less than 0.05 (5 per cent, or 1 in 20) under the assumption of the null hypothesis (H_0), then it is sufficiently unlikely that we can reject the null hypothesis.

The acceptable level of probability for rejecting the null hypothesis is known as the significance level or p -value (alpha, α) and can be set at different levels depending on how certain we wish to be. However, while we might conclude that it is unlikely that estimates from our sample have arisen by chance under the assumption of H_0 , we cannot say this establishes beyond all doubt that our sample estimates show there is a real difference in the population, but only that it is very likely. For example, with a significance level of 0.05, there remains a 1 in 20 chance that the observed difference could have arisen by chance; that is, it is a false positive, or type I error.

11.3.4 Summary: Stages of Hypothesis Testing

Another important aspect we covered was the steps that should be taken in carrying out an hypothesis test:

1. Summarise data from the sample (difference in means, or in proportions with contingency table, etc.).
2. State the null hypothesis (H_0) and the alternative hypothesis (H_1) in relation to the sample estimate.
3. Carry out an hypothesis test appropriate to the type of data, having first checked that the relevant assumptions are met.
4. Obtain the test statistic, which will most commonly be done by computer; this provides the result for step 5.
5. Assess the probability (p -value) of obtaining the observed test result (or one more extreme) by probability distribution-based (parametric) or non-parametric methods.
6. If the p -value is small (e.g. less than 0.05), reject the null hypothesis.
7. Where parametric methods have been used, state the estimate of difference or association, with the 95 per cent (or other level, as appropriate) CI.

We are now ready to look at specific tests in more detail, and we start with the type of data, comparison, and test with which you are probably most familiar, a comparison of means from two groups, using the independent samples t -test.

11.3.5 Comparing Two Independent Groups

We often want to compare two groups for a particular attribute measured on a continuous scale; for example, is the mean blood pressure of male smokers higher than that of non-smokers? These are independent samples because the groups are mutually exclusive; individuals cannot be in both groups. If the data conform to a particular probability distribution, we can use parametric hypothesis tests (the z -test or t -test) as long as the assumptions of the tests are met. If the data do not meet these assumptions, we need to use a non-parametric, or distribution-free, method (e.g. the Mann–Whitney U test). Table 11.3.1 describes the main characteristics of hypothesis tests for comparing continuous data for two independent groups.

Table 11.3.1 Hypothesis tests used for comparing continuous data for two independent groups.

Test	Parametric hypothesis tests data/assumptions	Test	Non-parametric hypothesis data/assumptions
z-test	<ul style="list-style-type: none"> • Continuous data (sample size >29) • Data have approximately normal distribution • Probability distribution (standard normal) 	Mann–Whitney U test	<ul style="list-style-type: none"> • Continuous or ordered categorical data • No distribution assumption • Ranked data
Reference: described in Chapter 5, Section 5.5.4		Reference: described in Section 11.3.5	
t-test	<ul style="list-style-type: none"> • Continuous data (any sample size) • Groups have similar standard deviations • Data has approximately normal distributions • Probability distribution (t-distribution) 	Wilcoxon signed rank test	<ul style="list-style-type: none"> • Continuous or ordered categorical data • No distribution assumption • Ranked data
Reference: described in Chapter 5, Section 5.5.4		Reference: not described in this text. Produces results identical to those of the Mann–Whitney U test	

The z -test and t -test were described in Chapter 5, Section 5.5.4. We now look in detail at the non-parametric hypothesis test for continuous data (or ordered categorical data) for two independent groups: the Mann–Whitney U test. An alternative (Wilcoxon signed rank test) provides identical results to the Mann–Whitney U test and is therefore not considered further.

The Mann–Whitney U Test

As with all non-parametric hypothesis tests, the Mann–Whitney U test uses ranking of the data, rather than comparing the actual distributions. The following example describes how the test statistic is calculated and the result is interpreted.

Calculation of the Mann–Whitney U Statistic

Suppose that we have two groups of individuals, with n_1 in the first group and n_2 in the second group. To carry out the test,

1. Rank all the measurements in ascending order
2. Sum the ranks of group 1 to give R_1
3. Calculate $U_1 = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - R_1$, and $U_2 = n_1 n_2 - U_1$
4. Set U to be the smaller of U_1 and U_2

Tables of critical values of the Mann–Whitney U statistic are available to determine the probability of obtaining the observed data under (H_0) . These are tabulated for n_1 and n_2 less than 20. For larger values of n , the U statistic approximates the standard normal distribution. Computer software provides the test result and p -value.

Example

A study was set up to measure synthesis of alkaline phosphatase (an enzyme in the bloodstream that helps break down bodily proteins) in two groups of patients: healthy subjects ($n = 6$) and patients with coeliac disease ($n = 7$). The data for each group (including a histogram) are presented and displayed in Figure 11.3.1.

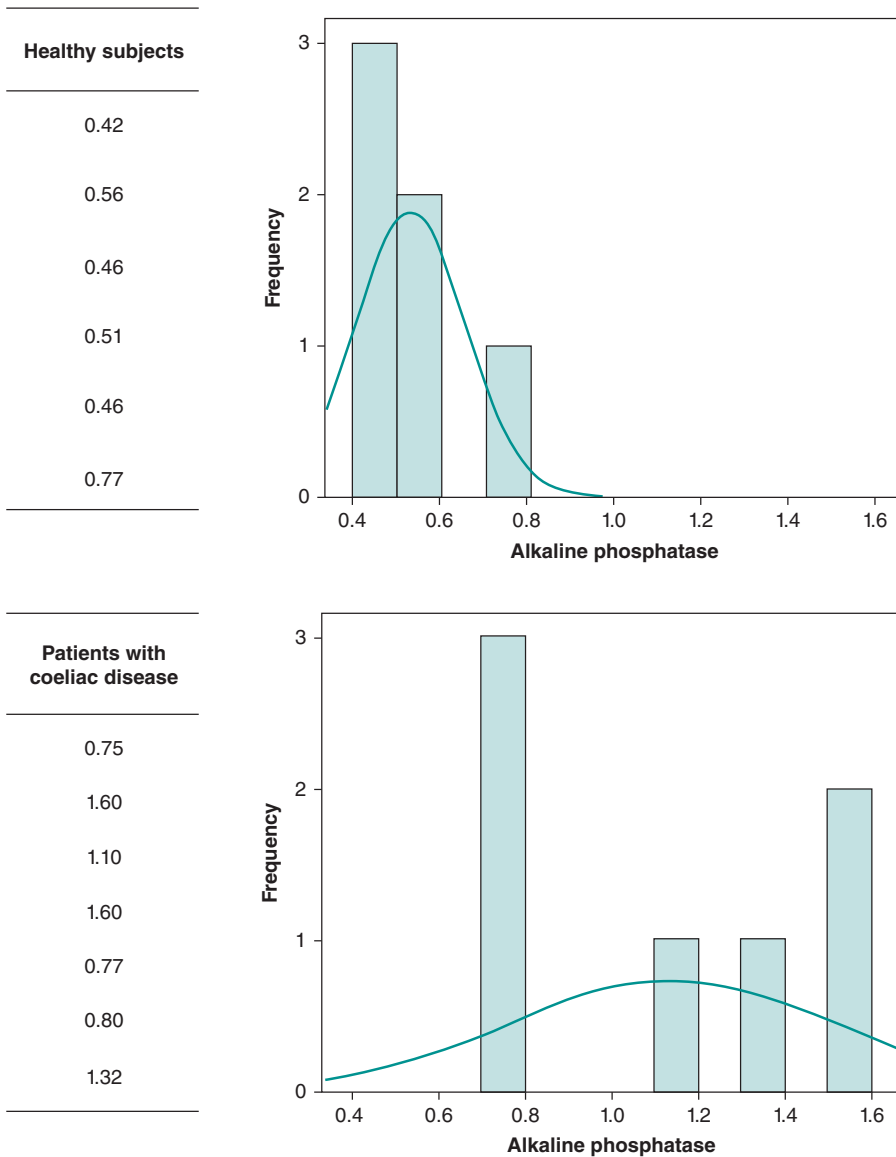


Figure 11.3.1 Alkaline phosphatase levels in (a) healthy subjects and (b) patients with coeliac disease.

We can see that the data for both groups are not normally distributed, and therefore we should not compare the means of the two groups by the two-sample t -test. Instead we could compare the medians of the two groups and use a non-parametric test (Table 11.3.2). The median and interquartile range (IQR) of alkaline phosphatase level for the two groups of patients are shown below.

Table 11.3.2 Comparison of medians.

Group	Median	IQR
Healthy subjects	0.485	0.450–0.613
Patients with coeliac disease	1.100	0.770–1.600

There does appear to be a substantial difference between the medians of the groups, and a non-parametric test can be used to test whether this difference is statistically significant. We now calculate the Mann–Whitney U statistic for ($n_1 = 6$ and $n_2 = 7$) to check for significance, as follows:

1. Assign ranks to all the data (from both groups) (Table 11.3.3).

Table 11.3.3 Assignment of ranks.

Healthy subjects	Rank	Patients with coeliac disease	Rank
0.42	1	0.75	6
0.56	5	1.60	12.5
0.46	2.5	1.10	10
0.51	4	1.60	12.5
0.46	2.5	0.77	7.5
0.77	7.5	0.80	9
		1.32	11
Total	$22.5 = R_1$		

2. Calculate the value of the U statistic, in this case,

$$U_1 = 6 \times 7 + \frac{1}{2} \times 6 \times (6 + 1) - 22.5 = 40.5$$

and

$$U_2 = 6 \times 7 - 40.5 = 1.5$$

Therefore, $U = 1.5$.

Note that either R_1 or R_2 can be used to find U_1 and U_2 (it does not matter which group is labelled group 1), and that the ranks of equal (tied) values are averaged.

3. Compare the value of the U statistic to tabulated values of U . Table 11.3.4 shows an extract of the table of critical values for the U statistic (the values relevant to our example are shaded).

For a significance level of 0.01 with sample sizes of $n_1 = 6$ and $n_2 = 7$, the critical value in the table is 3. The calculated value of U is 1.5, which is lower than the tabulated value. Note that

Table 11.3.4 Selected critical values of the Mann–Whitney U statistic.

n_2	p -value	n_1					
		3	4	5	6	7	8
3	0.05	–	0	0	1	1	2
	0.01	–	0	0	0	0	0
4	0.05	–	0	1	2	3	4
	0.01	–	–	0	0	0	1
5	0.05	0	1	2	3	5	6
	0.01	–	–	0	1	1	2
6	0.05	1	2	3	5	6	8
	0.01	–	0	1	2	3	4
7	0.05	1	3	5	6	8	10
	0.01	–	0	1	3	4	6
8	0.05	2	4	6	8	10	13
	0.01	–	1	2	4	6	7

with this test, a value lower than the critical value denotes that the corresponding probability is achieved. Thus, the null hypothesis (there is no difference between the medians) is rejected ($p < 0.01$), and we can accept the alternative hypothesis (that there is a difference between the medians). The box below summarises the main attributes of the Mann–Whitney U test.

Summary: The Mann–Whitney U Test

Key Features

- It is used to test the null hypothesis that the medians of two populations are equal.
- It is a non-parametric (distribution-free) hypothesis test.
- The data must be continuous or ranked (e.g. scores).
- The samples must be random.
- The samples must be independent (that is, a person cannot be in both samples).

Procedure

1. State the hypotheses

$$H_0 : med_1 = med_2$$

$$H_1 : med_1 \neq med_2$$

defining med_1 and med_2 as the population medians.

2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. Calculate the U statistic.
4. Compare the value of U with tabulated critical values of the U statistic, which is tabulated according to the sample size of the two groups (n_1 and n_2) to obtain the p -value, or review output from computer software.
5. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

11.3.6 Comparing Two Paired (or Matched) Groups

In the description of intervention studies in Chapter 7, we introduced the idea of *paired* continuous data, with the example of before-and-after assessments of a sample of people with asthma. In this situation, we were interested in differences between the two assessments of a knowledge score in the same people. When our data are paired, we require a method designed specifically to test the null hypothesis of no difference between the paired measurements. The reason for using a test specific to this purpose is that the pairing (or matching) results in much less variation between data obtained on the same individual before and after an intervention (or between matched individuals) compared to where data being analysed are from two independent groups. Also, the test is applied to the distribution of those differences.

The parametric hypothesis test is the *paired t-test*, and the non-parametric equivalent is the *Wilcoxon signed rank test*. Table 11.3.5 describes the main characteristics of hypothesis tests for comparing continuous data for two paired groups.

Table 11.3.5 Hypothesis tests comparing continuous data for two matched/paired groups.

Test	Parametric hypothesis tests data/assumptions	Test	Non-parametric hypothesis tests data/assumptions
Paired <i>t</i>-test	<ul style="list-style-type: none"> • Continuous paired data • Sample differences have normal distribution • Probability distribution (<i>t</i>-distribution) 	Wilcoxon signed rank test	<ul style="list-style-type: none"> • Continuous or other ranked (e.g. scores) paired data • No distribution assumption • Ranked data
	Reference: described in Section 11.3.6		Reference: described in Section 11.3.6

The paired *t*-test and the Wilcoxon signed rank test are now described with examples in the following sections.

The Paired *t*-Test

For this hypothesis test, we return to the before-and-after assessments in the sample of people with asthma. In this example, the two test scores for each patient are paired continuous data and are shown in Table 11.3.6.

We saw that this gives us a series of differences ($A - B$) for each person. The mean of these differences is 3.85, which is the difference between the mean of the after scores (A) and the mean of the before scores (B). We need to use a paired hypothesis test; in this example we have continuous data (the scores), so we use a *paired t-test* rather than the independent-samples *t*-test. The calculation of this test is now described, using the data in Table 11.3.6 as an example.

Calculation of the Paired *t*-Test

As this is a paired test, we do not require the assumption of independence, nor do we need the assumption that each set of observations is drawn from populations that are normally distributed. We do need to be able to assume, however, that the differences between the paired observations (right-hand column in Table 11.3.6) are normally distributed. The distribution of these differences is examined below.

In the example of subject 1 in Table 11.3.6, the difference between scores (After talks (A) minus Before talks (B)) is $55 - 51 = 4$. After calculating the differences between scores for each of the 13 subjects, we have one sample of differences d_i with mean \bar{d} . To meet the assumptions

Table 11.3.6 Test scores of asthma patients before and after attending a series of talks about asthma and its treatment.

Subject	Test scores		Difference (A – B)
	Before talks (B)	After talks (A)	
1	51	55	4
2	43	49	6
3	48	52	4
4	19	32	13
5	57	62	5
6	39	44	5
7	37	40	3
8	46	46	0
9	43	39	-4
10	43	52	9
11	53	50	-3
12	58	61	3
13	49	54	5
<i>Mean</i>	45.08	48.92	3.85

of the paired *t*-test, we should check the distribution of the differences, and this is shown in Figure 11.3.2. Allowing for the small numbers, this can be accepted as normal. Generally, the distribution of differences tends to be normal, even if the original samples are skewed.

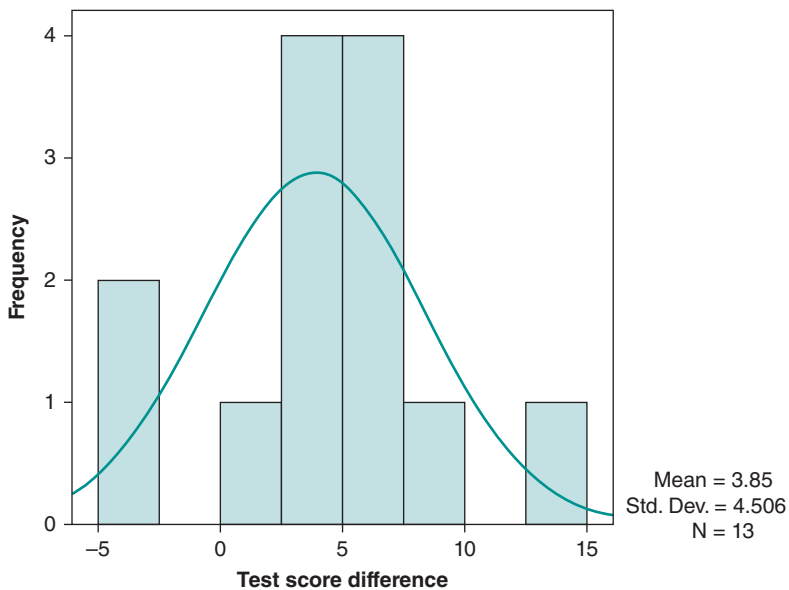


Figure 11.3.2 Distribution of differences in the asthma test scores.

The next step is to test whether the mean difference is zero (that is, whether there is, on average, no difference between the measurements on each person). The paired t -test statistic is given by

$$t = \frac{\bar{d}}{SE(\bar{d})}$$

where $SE(\bar{d}) = \frac{s}{\sqrt{n}}$, which is the standard error of \bar{d} . The term s is the sample standard deviation of the differences and n the sample size – that is, the number of differences. Under the null hypothesis, this value comes from the same Student's t -distribution we used previously, on $n - 1$ degrees of freedom: Note this is the number of pairs minus 1 (and of course is also the number of differences $- 1$). Therefore, we reject the null hypothesis if our test statistic (ignoring whether it is negative or positive) is larger than the tabulated value at the required level of significance. Returning to our asthma example, we want to test whether the difference between the test scores for each individual is, on average, zero; that is, the talks have no effect on the patients' knowledge of asthma. More precisely, we test the null hypothesis:

$$H_0 : \mu_d = 0$$

where μ_d is the mean difference in test scores for the population of asthma patients. From Table 11.3.6, the mean of the sample of differences is as follows (note that four decimal places have been used to minimise rounding errors):

$$\bar{d} = \frac{50}{13} = 3.8462$$

and the standard deviation is $s = 4.506405697$.

The standard error of the mean difference is therefore

$$SE(\bar{d}) = \frac{s}{\sqrt{n}} = \frac{4.5064}{\sqrt{13}} = 1.2497$$

and the value of the t -statistic is

$$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{3.8462}{1.2497} = 3.0777 = 3.08$$

We can then look up the value of t from tabulated critical values of the **t -distribution** (Table 11.3.7), the relevant parts of which are shaded.

Table 11.3.7 Critical values of the t -distribution.

Degrees of freedom	Two-tailed p -value			
	0.10	0.05	0.01	0.001
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140

From the tabulated t -distribution on 12 degrees of freedom, we obtain $0.001 < p < 0.01$. There is strong evidence that the test scores differ, on average, in the population of asthma patients: The score obtained after attending the talks is higher than the score obtained before (positive mean difference). A 95 per cent CI for the population mean difference in test scores is given by

$$\bar{d} \pm t_{n-1} SE(\bar{d})$$

where t_{n-1} is the two-sided 0.05 value of the t -distribution on $n - 1$ degrees of freedom. The following exercise will help consolidate what you have learned about the comparison we are making and the hypothesis test.



Self-Assessment Exercise 11.3.2

1. Calculate a 95 per cent CI for the population mean difference in scores (after – before) for the asthma example. You will need to use the critical values of the t -distribution in Table 11.3.7 to do this.
2. Do you think the two-sided test was appropriate?

Answers in Section 11.6

Summary: The Paired t -Test

Key features

- The data must be continuous.
- It is used to test the null hypothesis that the population mean of the differences between matched pairs is zero.
- For this test, the samples must be paired: pairs may be two measurements on the same individual on two occasions, or they may be measurements on two well-matched individuals from separate samples.
- The samples must be random.
- It is based on the t -probability distribution: it will depend on the degrees of freedom ($n - 1$), where n is the number of pairs.
- The sample differences must be from a population with a normal distribution.

Procedure

1. State the hypotheses

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

defining μ_d , the population mean difference.

2. Decide on the significance level. Call this α . (Typically $\alpha = 0.05$).
3. Calculate the t -statistic.
4. Compare the value of t with the t -distribution on $n - 1$ degrees of freedom and determine the p -value or review output from computer software.
5. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

The Wilcoxon Signed Rank Test

The non-parametric equivalent to the paired t -test is the *Wilcoxon signed rank test*. This test is used with paired data when the distribution of differences in a continuous variable is not normal, or the variable is ordered categorical (e.g. scores).

Calculation of the Wilcoxon Signed Rank Test

For n pairs of observations, the procedure is as follows:

1. Calculate the signed differences between the pairs of observations (that is, observe whether the differences are negative or positive).
2. Rank the differences, ignoring the signs.
3. Sum the ranks of the negative differences to give T^- and the positive differences to give T^+ .
4. Set T to be the smaller of T^- and T^+ .

Tables of critical values of T for the Wilcoxon signed rank test are available to ascertain the probability that the observed data would arise under the assumption of the null hypothesis (H_0). These are tabulated for $n < 50$. For larger values of n , the T statistic approximates the standard normal distribution. Computer software provides the test result and p -value.

Example

Eight pairs of identical twins were assessed for the effect of nursery school attendance on children's social perceptiveness scores. One twin was chosen at random to attend nursery school. After one term, the following results were obtained (Table 11.3.8):

Table 11.3.8 Wilcoxon signed rank test.

Pair	Score of 'nursery' twin	Score of 'home' twin	Difference	Rank ⁺	Rank ⁻
a	82	63	19	7	
b	69	42	27	8	
c	73	74	-1		1
d	42	37	5	3	
e	58	51	7	4	
f	56	43	13	5	
g	76	80	-4		2
h	82	65	17	6	

The median scores for nursery and home groups are 72.5 and 55, respectively, quite a substantial difference. The Wilcoxon signed rank test is used here to test for significance because the data are paired by virtue of these being identical twins (so although two separate groups of individuals are being compared, they are closely matched and can be treated as paired), and the distribution of the differences is not normal. In fact, the very small sample size makes it difficult to determine whether the distribution is normal, but it is assumed to be non-normal in this case.

Here $T^- = 3$ and $T^+ = 33$, and so $T = 3$. We can then look up this value in a table of critical values of T (Table 11.3.9), the relevant parts of which are shaded.

Table 11.3.9 Critical values of T (for Wilcoxon signed rank test distribution).

n	Two-tailed p -value			
	0.10	0.05	0.01	0.001
5	0	–	–	–
6	2	0	–	–
7	3	2	0	–
8	5	3	1	0
9	8	5	3	1

From the tables, we can see that this result is just outside statistical significance at $p = 0.05$. Therefore, there is not quite sufficient evidence to reject the null hypothesis that nursery school experience does not affect the social perceptiveness of children. This could be regarded as an overcautious interpretation, as the p -value is right on the cut-off for statistical significance.

It should be noted that zero differences are discounted, n is decreased accordingly, and (as with the Mann–Whitney U test), tied values are assigned the average of the ranks covered. In an alternative approach, rather than rank the differences, sometimes the number of positive differences is used to calculate a test statistic that is then compared with a binomial distribution (or normal distribution for large samples). This is called the *sign test* and is not considered further here.

Summary: The Wilcoxon Signed Rank Test

Key features

- It is used to test the null hypothesis that the population median of the differences between matched pairs is zero.
- It is a non-parametric (distribution-free) hypothesis test.
- The data must be continuous or ranked (e.g. scores).
- The samples must be random.
- The samples must be paired: Each subject appears in both samples or has a well-matched partner in the other sample.

Procedure

1. State the hypotheses

$$H_0 : med_d = 0$$

$$H_1 : med_d \neq 0$$

defining med_d as the population median difference.

2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. Calculate the T statistic.
4. Compare the value of T with tabulated critical values of the T statistic, tabulated according to the sample size of the number of pairs (n), or review output from computer software.
5. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

11.3.7 Testing for Association Between Two Groups

We described how to summarise the relationship between two continuous variables when we introduced the principles of correlation in Chapter 2, Section 2.4. We found that if two continuous variables have an approximately linear relationship when examined in a scatterplot, a correlation coefficient can be calculated to indicate the strength and direction of the relationship. The hypothesis test for assessing the probability of the null hypothesis being true is based on the value of the correlation coefficient (r). The method used for calculating r and its related p -value is the **Pearson (product moment) correlation**. This is the parametric hypothesis test for calculating a correlation coefficient. The non-parametric alternative is **Spearman's rank correlation** (Table 11.3.10).

Table 11.3.10 Hypothesis tests to study strength and direction of association between two groups.

Test	Parametric hypothesis tests data/assumptions	Test	Non-parametric hypothesis tests data/assumptions
Pearson (product moment) correlation	<ul style="list-style-type: none"> • Continuous data • Both groups have approximately normal distributions • Linear relationship • Probability distribution (t-distribution) 	Spearman's rank correlation	<ul style="list-style-type: none"> • Continuous or ordered categorical data • No distribution assumption • No assumption of linearity • Ranked data
Reference: described in Chapter 2, Section 2.4.5		Reference: described in Section 11.3.7	

Spearman's Rank Correlation

Like the Pearson correlation coefficient, **Spearman's rank correlation coefficient** (r_s) can take a value between -1 and $+1$, and the interpretation of the value of the coefficient is essentially the same (although note that an equivalent coefficient of determination cannot be derived by taking the square of Spearman's coefficient). We may prefer this measure of correlation if any of the following are true:

- The data are not normally distributed.
- One or both variables are ordinal (have ordered categories).
- We require a measure that is not dependent on linearity.

Calculation of Spearman's Rank Correlation Coefficient

For n pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$, this coefficient is most efficiently calculated as follows:

1. Rank each variable in ascending order.
2. Calculate the difference, d_i , between the two ranks for each individual i .
3. Calculate the sum of the squared differences; i.e. $D = \sum d_i^2$
4. Spearman's rank correlation coefficient is then $r_s = 1 - \frac{6D}{n(n^2 - 1)}$

Example

A sample of 10 students training as clinical psychologists were evaluated by a tutor at the end of the course according to suitability for their career (measure X) and their knowledge of psychology (measure Y). The scores on both measures were found to be highly positively

skewed, so Spearman's correlation was chosen to study the association. The first two steps, the ranks and their differences, are presented in Table 11.3.11:

Table 11.3.11 Spearman's correlation.

Student	A	B	C	D	E	F	G	H	I	J
Rank on X	4	10	3	1	9	2	6	7	8	5
Rank on Y	5	8	6	2	10	3	9	4	7	1
d_i	-1	2	-3	-1	-1	-1	-3	3	1	4
d_i^2	1	4	9	1	1	1	9	9	1	16

Hence, $D = 52$ and

$$r_s = 1 - \frac{6 \times 52}{10(100 - 1)} = 0.68.$$

We can then use a table of critical values of Spearman's correlation coefficient (r_s) to determine whether or not r_s differs significantly from zero (Table 11.3.12), the relevant parts of which are shaded. Computer software provides the test result and p -value.

Table 11.3.12 Critical values of r_s (for Spearman's rank correlation).

n	Two-tailed p -value		
	0.05	0.01	0.001
5	1.000		
6	0.886	1.000	
7	0.786	0.929	1.000
8	0.738	0.881	0.976
9	0.700	0.833	0.933
10	0.648	0.794	0.903

We can see that our value of r (0.68) lies between the critical values for 0.05 and 0.01 with $n = 10$, so it is statistically significant. This can be expressed as $0.01 < p < 0.05$, and we can reject the null hypothesis that the correlation between suitability for career (measure X) and knowledge of psychology (measure Y) is equal to zero.

In addition to Spearman's rank correlation, you may also come across Kendall's tau (τ), which is a similar non-parametric correlation coefficient; we will not discuss that test further here.

Summary: Spearman's Rank Correlation Hypothesis Test

Key Features

- It is used to test the null hypothesis that the population correlation coefficient is equal to zero.
- It is a non-parametric (distribution-free) hypothesis test.
- The data must be continuous, or ordered categorical (e.g. scores).

- The samples must be random.
- There is no assumption of linearity between the two variables, as it is the ranks that are being correlated. The data should be pictured on a scatterplot to allow meaningful interpretation of the Spearman correlation coefficient.

Procedure

1. State the hypotheses

$$H_0 : r_p = 0$$

$$H_1 : r_p \neq 0$$

defining r_p as the population correlation coefficient.

2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. Calculate the r_s correlation coefficient, and obtain the p -value using tabulated critical values, or review output from computer software.
4. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
5. State the conclusion and interpret the result.

11.3.8 Comparing More Than Two Groups

We have seen how to compare continuous data for two groups by using the independent-samples z -test or t -test for normally distributed data, or alternatively by using the non-parametric Mann–Whitney U test for data when the assumptions for the parametric tests are not met. We might also wish to compare continuous data between more than two groups. The parametric method for comparing more than two means is *analysis of variance (ANOVA)*, and its non-parametric equivalent for continuous or ordered categorical data is the *Kruskal–Wallis test* (Table 11.3.13).

Table 11.3.13 Hypothesis tests to compare continuous data for more than two groups.

Test	Parametric hypothesis test data/assumptions	Test	Non-parametric hypothesis test data/assumptions
Analysis of variance (ANOVA)	<ul style="list-style-type: none"> • Continuous data • Three or more independent groups with normal distributions • Groups should have similar standard deviations • Probability distribution (F-distribution) 	Kruskal-Wallis	<ul style="list-style-type: none"> • Continuous data • Three or more groups • No distribution assumption • Ranked data
	Reference: described in Section 11.3.8		Reference: described in Section 11.3.8

Analysis of Variance (ANOVA)

We will study analysis of variance with an example of data on systolic blood pressure (SBP) obtained from samples of four occupational groups in an industrial setting. The distributions for each group are almost normal, and the variances are similar (Figure 11.3.3), so we could use the t -test for comparing the mean SBP values for any two groups. If we wish to compare mean SBP values for all four groups, we need to use *one-way ANOVA*. This is simple enough to carry out with software such as SPSS.

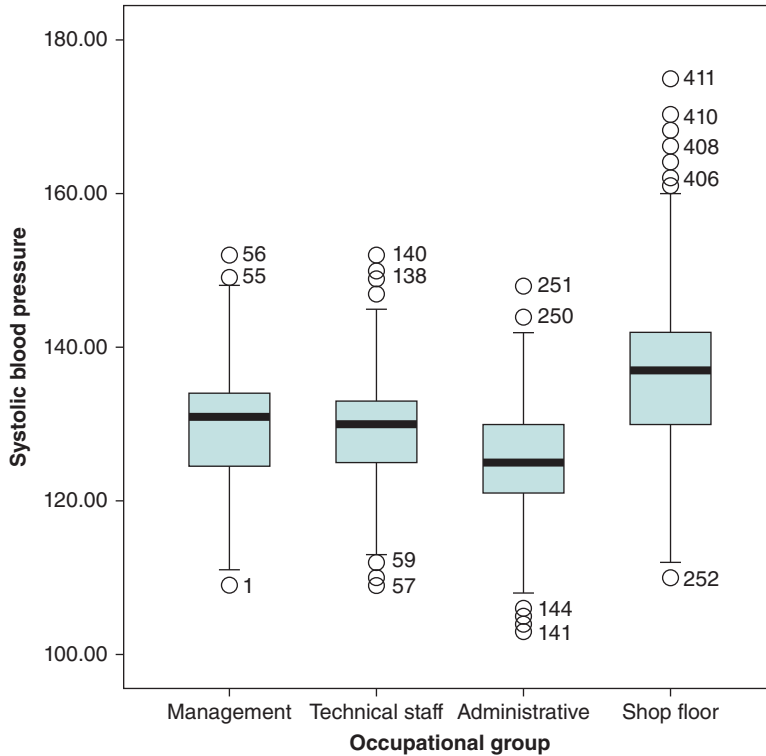


Figure 11.3.3 Box and whisker plot (see text) showing systolic blood pressure for four occupational groups.

ANOVA separates the total variability in the data into that which can be attributed to differences between the individuals from the different groups (the *between-group variation*) and that which can be attributed to the random variation between the individuals within each group (the *within-group variation*).

A box-and-whisker plot illustrates the distribution of continuous data. The box shows the median (thick black horizontal line) and the 25th and 75th centiles (bottom and top of the box, respectively) – hence the inter-quartile range (IQR). The whiskers extending above and below the box show the range of values within 1.5 IQR below the 25th centile and 1.5 IQR above the 75th centile. Values outside this range are denoted as outliers and are indicated by the subject identification number, so these can easily be checked.

These components of variation are measured by variances (hence ANOVA). Under the null hypothesis, the group means are the same. If, however, there are important differences between the groups, the between-group variance is larger relative to the within-group variance. The hypothesis test for one-way ANOVA is based on the ratio of these two variances and follows the *F-distribution*. In fact, it is the mean square of the variances that are compared, as explained in the example below.

As with other parametric (distribution-based) tests, there are assumptions that should be met, namely: The data should be (approximately) normally distributed in the population, and the variance for each group is (approximately) equivalent. We saw from Figure 11.3.3 that this was the case. The mathematics involved in carrying out one-way ANOVA is complex and is not considered further here. Standard software packages, including SPSS, provide output in the form of descriptive information and an ANOVA table displaying the value of the *F-ratio* and the associated *p*-value, and this is shown for the current example in Table 11.3.14. This

Table 11.3.14 Results from one-way ANOVA comparing mean SBP for four occupational groups.

Staff group	N	Mean	SD	SE	95% CI		Min	Max
					Lower bound	Upper bound		
Management	56	129.9	9.5	1.27	127.4	132.5	109.0	152.0
Technical staff	84	129.4	8.4	0.92	127.6	131.3	109.0	152.0
Administrative	111	125.2	8.3	0.78	123.6	126.8	103.0	148.0
Shop floor	160	137.0	11.8	0.93	135.2	138.9	110.0	175.0
Total	411	131.3	11.1	0.55	130.2	132.4	103.0	175.0

Table 11.3.15 ANOVA table: Systolic blood pressure.

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	9796.420	3	3265.473	32.849	.000
Within groups	40458.825	407	99.407		
Total	50255.246	410			

table shows mean values, data on the distributions, and 95 per cent CIs for each mean. The second (ANOVA) table, Table 11.3.15, shows the results for the significance test of the difference between groups ($p < 0.0005$).

A clearer understanding of this test can be gained from looking at the second part of the ANOVA output (Table 11.3.15) in more detail. This shows the sums of squares (variance), degrees of freedom, mean squares (sums of squares divided by the df), the *F*-ratio (the ratio of mean squares), and the *p*-value.

The degrees of freedom for between groups is calculated as the number of groups (*k*) minus 1, so in this case, $4 - 1 = 3$ (the numerator for the *F*-ratio calculation). The degrees of freedom for within groups is calculated as the total number of subjects (*n*) minus the number of groups being compared (*k*); that is, $411 - 4 = 407$ (the denominator for the *F*-ratio calculation). The *F*-ratio ($3265.473/99.407$) is compared to tables of the *F*-distribution with ($k - 1 = 3$) and ($n - k = 407$) degrees of freedom, as shown in Table 11.3.16, the relevant sections of which are shaded.

Table 11.3.16 Critical values of the *F*-distribution.

df of denominator	p-value	Degrees of freedom (df) of the numerator				
		1	2	3	4	5
10	0.05	4.96	4.10	3.71	3.48	3.33
50	0.05	4.03	3.18	2.79	2.56	2.40
100	0.05	3.94	3.09	2.70	2.46	2.31
407	0.05	3.86	3.02	2.63	2.39	2.24
	0.01	6.70	4.66	3.83	3.37	3.06
	0.001	10.99	7.03	5.53	4.71	4.19
1000	0.05	3.85	3.00	2.61	2.38	2.22

We can see that our value of the F -ratio is considerably greater than the value for $p < 0.001$ for the relevant degrees of freedom in the table and is sufficiently large to indicate that the H_0 of no difference between groups is very unlikely. This result is consistent with the SPSS output, which shows $p < 0.0005$. We can therefore reject the H_0 , although we need to look to the data to see which mean varies from which other mean. Exercise 11.3.3 takes you through this.



Self-Assessment Exercise 11.3.3

1. From Figure 11.3.3, which groups do you think might have significantly different mean SBP?
2. Using the 95 per cent CIs in Table 11.3.14, summarise the differences between groups that appear to be significant.

Answers in Section 11.6

Summary: One-Way Analysis of Variance (ANOVA)

Key Features

- Is used to test the null hypothesis that the means of three or more populations are equal.
- Is based on the F probability distribution (will depend on the number of degrees of freedom).
- The data must be continuous.
- The samples must be random.
- The samples must be independent (that is, a person cannot be in more than one group).
- The samples must be from populations with normal distributions.
- The populations must have similar standard deviations.

Procedure

1. State the hypotheses:
 H_0 all group means in the population are equal
 H_1 at least one group mean in the population differs from the others.
2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. Calculate the F -ratio.
4. Compare the value of the F -ratio with the **F -distribution** for $k - 1$ and $n - 1$ degrees of freedom and determine the p -value ($k =$ no. of groups; $n =$ no. of observations), or review output from computer software.
5. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

The Kruskal–Wallis Test

Staying with our example of the four occupational groups from the previous section, one of the other variables available from this occupational data set is alcohol consumption, in units per week. Figure 11.3.4a shows the differences between the groups and Figure 11.3.4b shows the distribution for one of these groups (administrative workers). You will see that the distribution is positively skewed due to a large number of non-drinkers, in addition to a small number of very heavy drinkers. We could try log-transformation to normalise the distribution, but this is unlikely to help greatly due to the shape. A reciprocal transformation may be more effective,

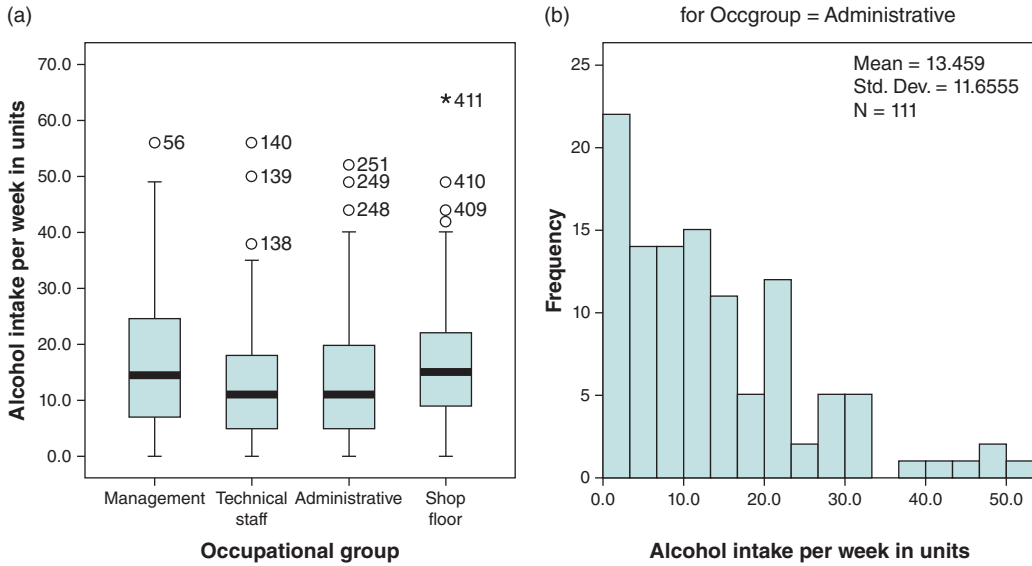


Figure 11.3.4 (a) Box-and-whisker plot showing alcohol intake (units per week) by four occupational groups and (b) distribution of alcohol intake (units per week) for administrative workers.

but will be difficult to interpret. In this situation it is more appropriate to use a non-parametric test to compare the average alcohol consumption among the four groups. The non-parametric equivalent to one-way ANOVA is the Kruskal–Wallis test.

Like other non-parametric hypothesis tests, the Kruskal–Wallis test uses ranking to compare groups and is an extension of the Wilcoxon signed rank test. Under the null hypothesis of no differences in the distributions between the groups, the sums of the ranks in each of the k groups should be comparable. After all the values in each of the groups have been ranked and the ranks summed, a test statistic can be calculated and compared to the chi-squared probability distribution for k (number of groups) – 1 degrees of freedom.

For the alcohol-consumption data, the chi-squared statistic is 15.84 and there are three degrees of freedom. The p -value for obtaining a chi-squared statistic this large by chance (under the H_0) is $p = 0.001$. Hence, we can reject the null hypothesis of no difference in alcohol consumption between groups.

This result does not tell us which group differs from which other group, and we have to go back to the data to assess this (Figure 11.3.4a and Table 11.3.17). Mann–Whitney U tests could be carried out to compare pairs of groups that appear to differ the most.

Table 11.3.17 Distributions for alcohol consumption by occupational group.

Group	Median	IQR	Mean	SD	95% CI	Min	Max
Management	14.5	17.8	17.7	13.6	14.1, 21.4	0	56.0
Technical	11.0	13.0	13.0	11.2	10.6, 15.4	0	56.0
Administrative	11.0	15.0	13.5	11.6	11.3, 15.6	0	52.0
Shop floor	15.0	13.0	16.9	10.8	15.2, 18.6	0	64.0

Summary: Kruskal–Wallis Test**Key Features**

- Is used to test the null hypothesis that each group has the same distribution of values in the population.
- Is a non-parametric (distribution-free) hypothesis test.
- The data must be continuous or ordered categorical.
- The samples must be random.
- The samples must be independent (that is, a person cannot be in more than one sample).
- The sample size should be sufficiently large for the analysis.

Procedure

1. State the hypotheses:
 H_0 : each group has the same distribution of values in the population
 H_1 : each group does not have the same distribution of values in the population.
2. Decide on the significance level. Call this α (typically $\alpha = 0.05$).
3. Calculate the test statistic.
4. Compare the test statistic with the chi-squared probability distribution for $k - 1$ degrees of freedom and determine the p -value ($k = \text{no. of groups}$) or review output from computer software.
5. If $p < \alpha$, reject the null hypothesis. Otherwise do not reject H_0 .
6. State the conclusion and interpret the result.

11.3.9 Association Between Categorical Variables

We have already come across the hypothesis tests for comparing whether there is a difference in proportions (or per cent) between independent groups (*chi-squared test*) or matched groups (*McNemar's test*) in a 2×2 contingency table. We used the chi-squared test to investigate whether there was a significant relationship between exercise and cancer in the chapter on cohort studies (Chapter 5, Section 5.5.3), and we used McNemar's test to investigate the null hypothesis that the population odds ratio was 1 for a matched case–control study (Chapter 6, Section 6.3.3). In this section we see how a variation of the chi-squared test can be used to look for evidence of a *dose–response* relationship between an ordered categorical variable and an outcome of interest (*the chi-squared test for trend*). We also introduce another hypothesis test that should be used if the assumptions of the chi-squared test are not met (*Fisher's exact test*). These tests are summarised in Table 11.3.18.

Chi-Squared Test for Trend

For the chi-squared test for trend, we will use data from the occupational back pain data set (see Preface). Table 11.3.19 shows the proportion of women working in the manual occupational settings with back pain, stratified by age in three groups. We can see that there appears to be a greater proportion of women with back pain in the older ages.

Using the chi-squared test detailed in Chapter 5, Section 5.5.3, we could test the null hypothesis that there is no relationship between self-reported back pain and age against the alternative hypothesis that there is a relationship. The chi-squared statistic is calculated as 9.88 for two degrees of freedom ($p < 0.01$); therefore, the null hypothesis can be rejected.

Table 11.3.18 Hypothesis tests to compare categorical data for two or more independent groups.

Test	Data/assumptions
Chi-squared test	<ul style="list-style-type: none"> • Categorical data • Independent groups • >5 expected in 80% of cells • Probability distribution (chi-squared distribution)
Reference: described in Chapter 5, Section 5.5.3	
McNemar’s test	<ul style="list-style-type: none"> • Categorical data • Matched/paired groups • Probability distribution (chi-squared distribution)
Reference: described in Chapter 6, Section 6.3.3	
Chi-squared test for trend	<ul style="list-style-type: none"> • Ordered categorical data • Independent groups • >5 expected in 80% of cells • Probability distribution (chi-squared distribution)
Fisher’s exact test	<ul style="list-style-type: none"> • Categorical data • Independent groups • Typically used when < 5 expected in >20% of cells • Probability distribution (<i>F</i>-distribution)
Reference: described in Section 11.3.6	

The chi-squared statistic would be the same no matter what the order of the rows and columns, and the test ignores the natural ordering of the columns. It can be seen that the percentage of women with back pain increases progressively with age. It is of interest to know whether there is evidence that the prevalence of back pain increases with increasing age; that is, whether there is evidence of a *trend*. You will recall that demonstrating a *dose–response relationship* is one of the viewpoints that Hill described as supporting causal inference (see Section 5.7.1 of Chapter 5). To test the null (H_0) hypothesis of no trend, we use the *chi-squared test for trend*, which has one degree of freedom. Calculating this by hand is somewhat laborious, and we will just look at the result of calculating this using computer software. This test yields a chi-squared statistic of 8.54 ($df = 1$), which is significant at $p < 0.01$. Therefore, we can reject the null hypothesis that there is no linear association between age and back pain in this sample of women.

Table 11.3.19 Proportion of female manual workers with low back pain by age group.

Group	Age group						Total
	18–29 years		30–39 years		40–49 years		
	No	%	No	%	No	%	
Low back pain	30	22.1	49	24.6	64	36.6	143
No low back pain	106	77.9	150	75.4	111	63.4	367
Total	136	100	199	100	175	100	510

Fisher's Exact Test

We saw that for 2×2 tables, the chi-squared test is not valid if any of the expected frequencies are less than 5. In this situation, we can use **Fisher's exact test**. The test works by evaluating the probability associated with all possible 2×2 tables that have the same row and column totals (sometimes called marginal totals) as the observed data, under the assumption that the null hypothesis is true.

Calculating Fisher's Exact Test

We start with the 2×2 table, for which the row totals r_1 and r_2 and column totals c_1 and c_2 are fixed at those in our data (Table 11.3.20).

Table 11.3.20 Fisher's exact test.

	Yes	No	Total
Group A	a	b	r_1
Group B	c	d	r_2
	c_1	c_2	N

The probability of obtaining the cell frequencies a , b , c , and d under the null hypothesis is given by:

$$\frac{r_1!r_2!c_1!c_2!}{N!a!b!c!d!}$$

where $x!$ is called ' x factorial' and means that we multiply together all integers from x down to 1; e.g. $4! = 4 \times 3 \times 2 \times 1 = 24$. Note that if one of the cells is zero (0), then $0! = 1$. We use this formula to calculate the probability of observing each of the different tables into which the N individuals can be arranged, keeping the row and column totals the same. From this information, we can calculate the **exact probability** (p -value) of obtaining the observed set of frequencies, or a set that is more extreme, under the assumption of the null hypothesis.

Example

Data from a study of teenagers' eating behaviour ($n = 16$) were collected. Table 11.3.21 shows the number of males and females who were or were not dieting at the time of the study. We might hypothesise that the proportion of dieting teenagers is higher among women than men. Therefore the null hypothesis is that there is no difference between the proportion of male and female teenagers who are dieting. It can be seen that at least one of the expected values would be less than 5, so we will need to use the Fisher's exact test.

Table 11.3.21 Numbers of male and female teenagers who are currently dieting.

	Males	Females	Total
Dieting	1	5	6
Not dieting	8	2	10
Total	9	7	16

Table 11.3.22 All possible 2 × 2 tables (sub-tables) with fixed marginal totals based on data in Table 11.3.21, with exact probabilities.

(i)	<table border="1"><tr><td>0</td><td>6</td></tr><tr><td>9</td><td>1</td></tr></table>	0	6	9	1	$p = 0.00087$	(v)	<table border="1"><tr><td>4</td><td>2</td></tr><tr><td>5</td><td>5</td></tr></table>	4	2	5	5	$p = 0.33042$
0	6												
9	1												
4	2												
5	5												
(ii)	<table border="1"><tr><td>1</td><td>5</td></tr><tr><td>8</td><td>2</td></tr></table>	1	5	8	2	$p = 0.02360$	(vi)	<table border="1"><tr><td>5</td><td>1</td></tr><tr><td>4</td><td>6</td></tr></table>	5	1	4	6	$p = 0.11014$
1	5												
8	2												
5	1												
4	6												
(iii)	<table border="1"><tr><td>2</td><td>4</td></tr><tr><td>7</td><td>3</td></tr></table>	2	4	7	3	$p = 0.15734$	(vii)	<table border="1"><tr><td>6</td><td>0</td></tr><tr><td>3</td><td>7</td></tr></table>	6	0	3	7	$p = 0.01049$
2	4												
7	3												
6	0												
3	7												
(iv)	<table border="1"><tr><td>3</td><td>3</td></tr><tr><td>6</td><td>4</td></tr></table>	3	3	6	4	$p = 0.36713$							
3	3												
6	4												

To carry out the test, we start by rearranging the 2 × 2 table so that the smallest value is in the top left-hand cell if this is not the case. The number of possible sets of frequencies that add up to the observed row and column totals (including the observed data) is shown in Table 11.3.22, together with the probabilities of observing such data if the null hypothesis were true, calculated by the above formula.

The observed data are shown in Table 11.3.22, sub-table (ii), and the probability of observing such data if the null hypothesis were true is calculated as

$$\frac{6!10!9!7!}{16!1!5!8!2!} = \frac{6 \times 9 \times 7 \times 6 \times 5 \times 4 \times 3}{16 \times 15 \times 14 \times 13 \times 12 \times 11} = 0.02360$$

The other probabilities are calculated in a similar way. Notice how the factorials simplify by cancelling out sequences that appear on the top and bottom of the formula. If we do not simplify, the factorial multiplication might exceed the storage capacity of a calculator. The exact test is available on standard computer software such as SPSS, but if done by hand, we can check our calculations by adding up the probabilities for each table, which should sum to 1.

The probability of obtaining a difference between the two groups as large as, or larger than, the observed difference is found by adding up the probabilities for the tables that correspond to the observed data and those that would give a more extreme difference between the two groups, which in this case is just Table 11.3.22, sub-tables (i) and (ii):

$$p = (0.00087 + 0.02360) \times 2 = 0.049.$$

The probability is doubled (×2) because we are using a *two-sided* test. This is the most common approach, and it allows for the fact that the difference between groups could be in either direction; that is, more or fewer females are dieting than males. A *one-sided* test is less common and is used when the difference could only be in one direction. In our example, therefore, we can conclude that there is some evidence ($p < 0.05$, just!) to suggest that the proportion of dieting teenagers is higher in females than in males.

An alternative way of deriving the p -value for a two-sided test, which is sometimes used, is to add together the probabilities of all the tables that have probabilities less than or equal to that for the observed data. In our example this would give

$$p = 0.00087 + 0.02360 + 0.01049 = 0.035.$$

This method always gives a p -value that is less than or equal to the first method, so it is less conservative (more likely to result in rejecting the null hypothesis).

11.4 Choosing an Appropriate Hypothesis Test

11.4.1 Introduction

This section is designed to help in the process of selecting the appropriate hypothesis test (or tests – often there is more than one suitable method) for the commonly encountered types of data and analytic procedures. Table 11.4.1 serves as a guide for making these choices: We will shortly work through one example to illustrate how to use the table, and this is followed by some exercises for you to try.

We have emphasised on several occasions the importance of checking, summarising, and displaying data before carrying out any hypothesis test, and the methods for doing this have been covered in previous chapters. Once we are satisfied that the data are clean, and we are familiar with how they are distributed and so on, it is time to select the right test. The following exercise provides some revision on the correct steps, before we look in detail at Table 11.4.1.



Self-Assessment Exercise 11.4.1

List all of the issues that need to be considered when choosing an appropriate hypothesis test.

Answers in Section 11.6

11.4.2 Using a Guide Table for Selecting a Hypothesis Test

Table 11.4.1 presents all of the basic hypothesis tests discussed in this chapter, together with the main features of each, and the situations for which they are most appropriate. The following example will illustrate how the table is organised.

Let's say we have two groups of primary school-age children living in deprived areas of two cities (A and B) of a middle-income country where lead-based paint is still commonly used. Both groups have been sampled using random methods so that they represent the two cities. In one of the cities (B), a campaign has been running for the last few years to educate families about the danger of lead and to reduce the availability of paints with high lead content. The children's blood lead levels have been measured, and the distributions are found to be highly skewed, though more so in city A than city B, but they are more or less normalised by log transformation. The variances for the log values in the two samples are not dissimilar, but that for city B is about 75 per cent of that in city A.

We wish to compare the blood lead levels obtained from the two cities to see whether the campaign has started to have any impact, given that a similar study carried out 5 years earlier found no difference between the cities. We have calculated means and SDs, and medians and

Table 11.4.1 Table for selecting hypothesis tests according to main attributes: data, comparison groups, data type, distribution, and purpose of test. In selecting an individual test, be sure that the assumptions for that test are met (these are not stated in full in this table).

Data comparison groups	CONTINUOUS (and ordered categorical)				CATEGORICAL			
	Two groups		Three or more groups		Two groups		Three or more groups	
	Independent	Matched or paired	Independent	Independent	Independent	Matched or paired	No order	Ordered
Distribution	Parametric Non-parametric	Parametric Non-parametric	Parametric Non-parametric	Parametric Non-parametric	Chi-square/ binomial	Chi-square/ binomial	Chi-square/ binomial	Ordered
	Normal Skewed	Normal Skewed	Normal Skewed	Normal Skewed				
<i>If distributional assumptions not met (e.g. skewed), we can consider normalisation with transformation and then use the appropriate parametric hypothesis test. Otherwise use a non-parametric hypothesis test.</i>								
Hypothesis test for comparison	z-test ($n \geq 30$) Mann-Whitney U Test(\$)	Paired t-test Wilcoxon signed rank test(\$)	Analysis of variance (ANOVA)	Kruskal-Wallis test(\$)	Chi-square test* ($n \geq 5$ expected in 80% of cells) Fisher's exact ($n < 5$ expected in $\geq 20\%$ of cells)	McNemar's test	Chi-square test* ($n \geq 5$ expected in 80% of cells) Fisher's Exact ($n < 5$ expected in $\geq 20\%$ of cells)	Chi-square test for trend
Hypothesis test for association (~)	Pearson product moment correlation	Spearman's rank correlation	Pearson correlation matrix (#)	Spearman's correlation matrix (#)	Chi-square test* ($n \geq 5$ expected in 80% of cells) Fisher's exact ($n < 5$ expected in $\geq 20\%$ of cells)	McNemar's test	Chi-square test* ($n \geq 5$ expected in 80% of cells) Fisher's exact ($n < 5$ expected in $\geq 20\%$ of cells)	Chi-square test for trend

~ Although correlation coefficients indicate the strength of an association for continuous (or ordered categorical) data, hypothesis tests for associations with categorical data do not. * Note: With small numbers, a Yates continuity correction is required.

\$ Non-parametric tests can also be used if assumptions of parametric tests are not met (e.g. equality of variance for independent t-test).

Although correlation matrices have not been covered in this book, they present the results of multiple simple correlations (depending on the number of groups being looked at).

IQRs, and it does appear that the levels in city B are lower than those in city A. Which test should we use to determine whether this difference is statistically significant?

So let's look at Table 11.4.1 and work through the steps. Blood lead levels are continuous data, and the children are in two independent groups; this takes us to the first two columns on the left of the table. We are making a comparison of means, so we look in the row labelled 'hypothesis test for comparison', which shows t -test, z -test, and Mann–Whitney U test. Although the standard deviations are not too dissimilar for use of the independent sample t -test, we know the distributions are markedly skewed, and in fact more or less log normal, which raises doubts about using one of the parametric tests (t -test or z -test). So now we have two options, either to transform the data and use an independent sample t -test, or to use the non-parametric Mann–Whitney U test.

The final choice depends on how we wish to present the data, and at this point you might like to refer back to Section 11.2.2 on transformation. If we want to retain information about a mean difference and 95 per cent CI, we may opt for log transformation and the t -test, though this will provide a geometric mean and 95 per cent CI. If we want to retain the untransformed values, we can present medians and IQRs and use the Mann–Whitney U test. Alternatively, we can use both approaches.

Try using Table 11.4.1 in answering the questions in the next exercise. Detailed information about the nature of the data and whether assumptions are met is not given, so you will need to think about the options you have for different tests.



Self-Assessment Exercise 11.4.2

Which statistical test would you use to analyse the following:

- a. To compare the number of cigarette smokers among cancer cases and age- and sex-matched healthy controls?
- b. To look at the relationship between cigarette smoking and presence or absence of respiratory symptoms in a group of people with asthma?
- c. To look at the relationship between cigarette smoking (yes/no) and sex (male/female) in a small pilot study of 40 undergraduate students in their first year?
- d. To examine the change in respiratory symptom prevalence from winter to summer in a group of people with asthma?
- e. To compare the serum thyroxine levels during pregnancy of the mothers of two groups of babies? The first group of seven mothers had babies who died and the second group of 13 mothers had babies who survived.
- f. To compare the number of cigarette smokers among a group of cancer cases and the number in a random sample of the general population?
- g. To compare mean peak flow rate values (litres/min) in a group of women as measured by two different instruments.
- h. To compare knowledge scores about dangers of alcohol among groups of Year 10 children in four schools?
- i. To investigate the association between lead levels in a sample of children from one of the two cities in the example discussed above, and the same children's scores from a 15-point reading aptitude test?

Answers in Section 11.6

11.4.3 The Problem of Multiple Significance Testing

When we carry out multiple hypothesis tests on a number of different comparisons, there is an increased probability of finding a significant difference just by chance. Using a significance (alpha) level of 0.05, each test has a 5 per cent chance of a false-positive result when there is no real difference (type I error). So, if in a study we carry out, say, 22 comparisons with hypothesis tests, the probability of at least one false positive is greater than 5 per cent.

To deal with this problem there are a number of statistical methods aimed at controlling the overall type I error rate at no more than 5 per cent (or some other specified level). The disadvantage of all of these methods is that they are conservative and are likely to err on the side of safety (non-significance).

For example, we will consider a commonly used method of correction for multiple significance testing, the Bonferroni method. This is the most conservative method of correction and is carried out as follows: If we perform k comparisons, we should multiply the p value from each test carried out on the data by k ; that is, we calculate $p(\text{corrected}) = kp$. So, if one of 10 hypothesis test results is $p = 0.032$, we multiply by 10 to obtain an adjusted p -value of 0.32. In this example, a p -value will have to be $<0.05/10$ or <0.005 to be accepted as significant.

There are other, less-conservative, methods of correction for multiple significance testing (for example, Duncan's multiple range test), and you will find that different statistical packages use different multiple comparison procedures. It is advisable to consult a statistician before using one of these statistical correction methods.

We should also distinguish between a priori comparisons (that is, decided upon in advance of data collection and analysis) and those that are post-hoc (that is, decided upon after the data are available). For example, specifying five important comparisons in the protocol of a trial presents a very different scenario for multiple significance testing than would choosing five comparisons once the results are available. Adjustment for multiple testing may not be required for the former, but it should certainly be considered for the latter.

11.5 Bayesian Methods

11.5.1 Introduction: A Different Approach to Inference

In all the material we have covered so far, there has been a common theme of inference from sample to population, using hypothesis tests to determine the probability that we could obtain the observed data (e.g. a difference in means) under an assumption of the null hypothesis (H_0 ; no effect). If that probability is low (less than 5 per cent), we conventionally reject the null hypothesis and conclude that there is a real difference. This approach is sometimes referred to as *frequentist*, since it is based on frequency distributions of statistics (e.g. the mean) seen with all possible samples from a given population (see Chapter 4).

Bayesian methods, named after Thomas Bayes, whose work on these ideas was first published (posthumously) in 1763, takes an alternative approach. In essence, the existing view about the probability of an event or effect (known as the *prior probability*) is first stated, and then it is modified by the current assessment (for example, a diagnostic test, or a trial of a drug) to produce a new judgement about the probability of that effect or event (known as the *posterior probability*). The process is described by a formula, which is discussed in the following section, together with an exercise to work through.

11.5.2 Bayes' Theorem and Formula

A reasonably accessible way to understand the application of Bayes' theorem can be appreciated through the example of a screening test to establish the probability of a person having a disease. In this context, the theorem is described by the following formula:

$$\text{Posterior odds of a disease} = \frac{\text{prior odds} \times \text{likelihood}}{\text{ratio of a positive test result.}}$$

The likelihood ratio was introduced and calculated in Chapter 10 (validation of screening tests), and formulae for this and for the odds required to make the calculations are shown below:

$$[1] \text{ Prior odds} = \frac{\text{Prior probability}}{1 - \text{prior probability}}$$

$$[2] \text{ Likelihood ratio (for positive test)} = \frac{\text{Sensitivity}}{1 - \text{specificity}}$$

$$[3] \text{ Posterior probability} = \frac{\text{Posterior odds}}{1 + \text{posterior odds}}$$

We will now see how to apply Bayes' theorem through the following exercise.



Self-Assessment Exercise 11.5.1

A woman aged 36 attends the GP for contraceptive advice. She smokes 20 cigarettes per day, has had more than five sexual partners in the last few years, and has not used barrier contraception. She also mentions that she may have noticed some vaginal bleeding in addition to her monthly menstruation. The GP sees from the records that the woman is due for a cervical smear test and notes that she has several risk factors for cervical cancer (smoking, multiple partners, and lack of barrier contraception) and possibly an important symptom (bleeding). The GP establishes a moderately high prior probability, which we will state as 40 per cent (0.40).

1. Calculate the prior odds.
2. The GP takes a smear test. Assuming the sensitivity of the test is 80 per cent and the specificity is 70 per cent, calculate the likelihood ratio for a positive test.
3. The smear test is positive. Calculate the posterior odds, and hence the posterior probability that the woman has cervical cancer.
4. Can you obtain the same result (approximately) by using Fagan's nomogram (Figure 11.5.1 below)? This chart is used to obtain an estimate of the posterior probability by running a line from the prior probability, through the likelihood ratio, to find the posterior (post-test) probability.
5. Interpret the result in the light of what we have discussed about Bayesian methods.
6. If the woman had a history indicating a very low risk (she attends for a routine smear, she is a non-smoker, she is unmarried and without a sexual partner for more than 5 years, and she has no bleeding), the GP may assume a much lower prior probability; let's say 0.05 (5 per cent). This implies the GP thinks cervical cancer is unlikely, although it is still possible. This woman also has a positive smear test. Use the nomogram (Figure 11.5.1) to obtain the posterior probability that she has cervical cancer, and interpret the result.

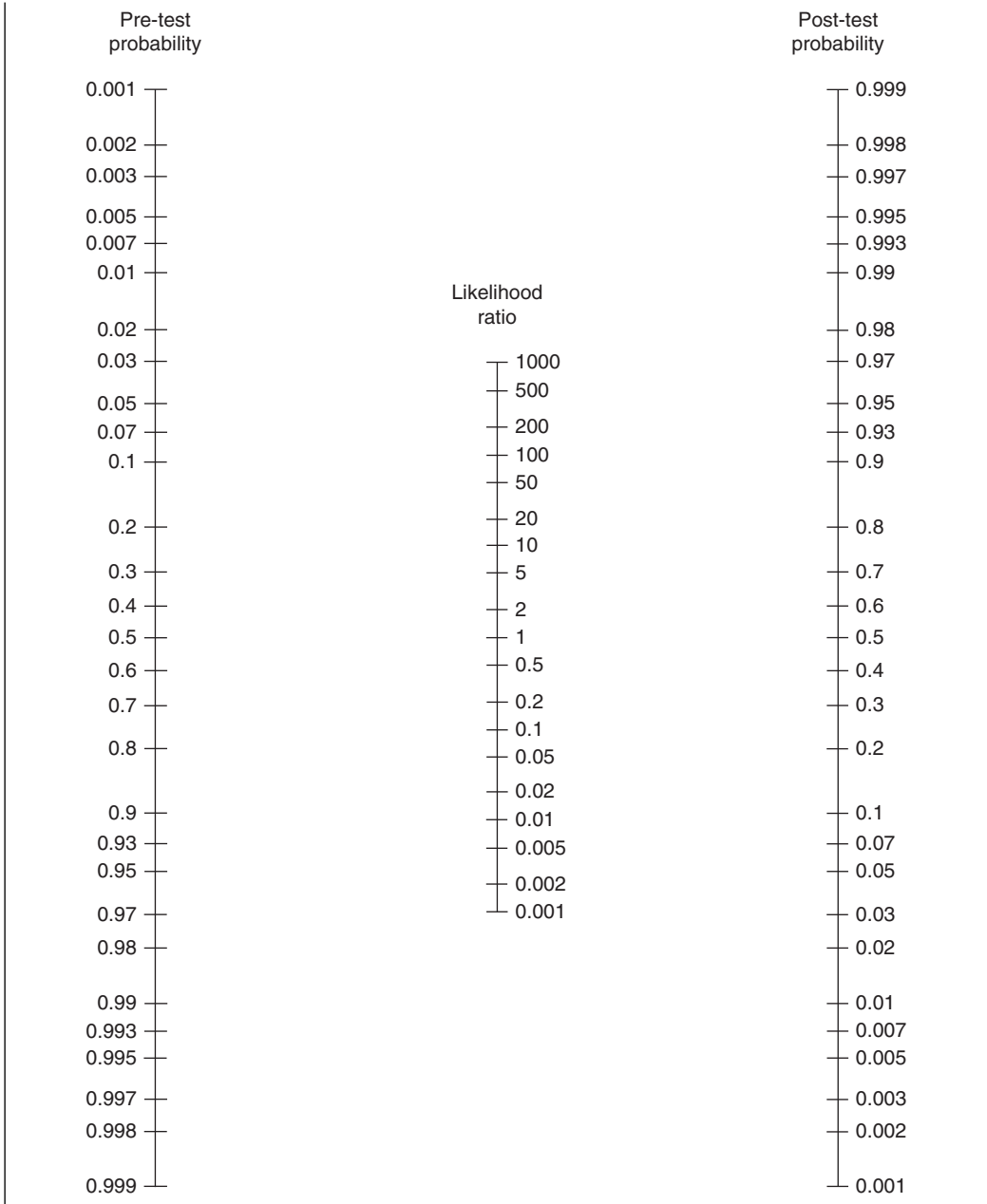


Figure 11.5.1 Fagan's nomogram.

Answers in Section 11.6

11.5.3 Application and Relevance

Bayesian methods are used in many different disciplines, including engineering, image processing, expert systems, decision analysis, gene sequencing, financial predictions, and increasingly

in complex epidemiological models. The method is thought to be intuitive and conforms to practice: a common example (as we have seen with Exercise 11.5.1) being clinical practice in which a physician has some judgement (prior probability) about a patient having a particular disease, which is modified by current assessment (history, clinical examination and diagnostic tests) to help the doctor decide on the probability of the diagnosis (the posterior probability).

To explore how Bayesian methods can be used in practice we will look an example of risk prediction for suicide among hospital psychiatric inpatients, as described in the resource paper by Powell *et al.* (2000). The abstract for this paper is reproduced below; although this does not make specific reference to Bayesian methods (those are described in the Methods and Results), you will see that adjusted likelihood ratios are reported:

Abstract

Background: Psychiatric hospital inpatients are known to be at high risk of suicide, yet there is little reliable knowledge of risk factors or their predictive power.

Aims: To identify risk factors for suicide in psychiatric hospital in-patients and to evaluate their predictive power in detecting people at risk of suicide.

Method: Using a case–control design, 112 people who committed suicide while in-patients in psychiatric hospitals were compared with 112 randomly selected controls. Univariate analysis and multivariate analyses were used to estimate odds ratios and adjusted likelihood ratios.

Results: The rate of suicide in psychiatric in-patients was 13.7 (95% CI 11.7–16.1) per 10,000 admissions. There were five predictive factors with likelihood ratios >2, following adjustment: planned suicide attempt, 4.1; actual suicide attempt, 4.9; recent bereavement, 4.0; presence of delusions, 2.3; chronic mental illness, 2.2; and family history of suicide, 4.6. On this basis, only two of the patients who committed suicide had a predicted risk of suicide above 5%.

Conclusions: Although several factors were identified that were strongly associated with suicide, their clinical utility is limited by low sensitivity and specificity, combined with the rarity of suicide, even in this high-risk group.

The application of a Bayesian approach to addressing the aim of this study, namely to identify risk factors for suicide and evaluate their predictive power, is described in the Methods. We will not look at this in detail, but if you are interested, you may wish to review this section of the paper. The following brief extract introduces the approach taken by the authors in applying Bayes' theorem to multiple risk factors:

Bayes' theorem states that the post-test odds of a condition, given the presence of a risk factor, can be found by multiplication of pre-test odds by the appropriate likelihood ratio (LR) for that risk factor. Direct application of the LRs from more than one risk factor could be achieved by using extensions of Bayes' theorem such as:

$$\text{Post-test odds} = \text{LR}_1 \times \text{LR}_2 \times \dots \text{LR}_n \times \text{pre-test odds}$$

They go on to discuss the fact that, unless the likelihood ratios (LRs) are adjusted for confounding, this approach would result in over-prediction. Adjustment of the LRs was therefore carried out using logistic regression. It is recognized that this method does not account for any possible interaction (effect modification) between predictors.

Table 11.5.1 (Table 3 from the paper) shows the adjusted likelihood ratios and the post-test probabilities based on the final model estimates and obtained using Bayes' theorem. The pre-test probability used in this study was the overall incidence rate of suicide among psychiatric in-patients, reported in the abstract as 13.7 per 10,000 admissions (equivalent to a probability of 0.00137, or 0.14%). The post-test probabilities for the most important predictors were in the range 0.2% to 0.7%, as shown in Table 11.5.1.

Table 11.5.1 Likelihood ratios and post-test probabilities of the factors predictive of inpatient suicide.

Risk factor	Crude likelihood ratio (95% CI)	Adjusted likelihood ratio (ALR) ¹	Final model likelihood ratio (FMLR)	Post-test probability (using FMLR)
Suicidal thoughts				
None (baseline)	0.3 (0.2–0.5)	0.4	0.3	
Some ideas, no plan	1.6 (0.8–3.2)	1.5	1.6	0.002
Plan, not acted upon	3.7 (0.6–24.5)	3.1	4.1	0.006
Act of self-harm either leading to, or during, admission	4.5 (2.6–8.1)	3.7	4.9	0.007
Recent bereavement	2.9 (1.0–8.2)	5.3	3.9	0.006
Depressed mood	1.4 (1.2–1.8)	1.0		
Delusions	2.3 (1.4–4.0)	2.3	2.2	0.003
Hopelessness	3.0 (1.9–5.0)	1.8		
Worthlessness or guilt	2.2 (1.4–3.5)	1.3		
Previous self-harm (not including acts leading to or during index admission)	2.1 (1.4–3.2)	1.6		
Chronic mental illness (>5 years)	1.9 (1.1–3.5)	1.9	2.2	0.003
One or more previous admission	1.3 (1.0–1.5)	1.2		
First-degree relative committed suicide	3.8 (1.2–12.2)	3.3	4.6	

The authors concluded that none of these factors were very good predictors from the point of view of those assessing and caring for psychiatric in-patients. Even though a combination of the five factors with a LR of >2 resulted in a risk of suicide of 5% (as reported in the Abstract) – considerably higher than the pre-test probability of 0.14% – only two of the suicide cases had all five of these predictors.

Bayesian methods have also been applied quite extensively with trials and meta-analyses, and – in an example we referred to in Chapter 10 – for modelling incidence rates for the Global Burden of Disease (GBD) Study. In all these cases, Bayesian methods are used to incorporate existing evidence and knowledge, so that the result of the analysis provides an estimate reflecting the totality of knowledge on a topic.

A number of advantages of Bayesian methods are described. For example, decisions about the prior probability (which, remember, must be stated) can make explicit assumptions and raise

issues that often are not openly acknowledged in traditional approaches. Sources for the prior probability can include a wide range of information, including previous studies in the literature, meta-analyses, and expert opinion. The prior probability can be deliberately set as sceptical if caution seems warranted, or alternatives can be used in sensitivity analysis to see how much influence variation across a range of prior probabilities might have.

Another advantage is that the conclusion – that is, the posterior probability – tells us how the latest piece of evidence should change what is currently believed and/or done. This has been argued to be of greater relevance to policy and practice than the output of traditional methods.

It is not possible in a brief introduction to Bayesian methods to do justice to the debates around the use and interpretation of this method, and interested readers should look to more detailed publications on the topic (for example, Broemling, 2014).

11.6 Answers to Self-Assessment Exercises

Section 11.1

Exercise 11.1.1

- True. Events are independent, so the probabilities are multiplied.
 - False. The probability for both must be less than that for each one.
 - False. The events are independent. This means that the probability of having a cold will still be 0.10.
- No. This is a continuous variable and therefore will follow a continuous probability distribution.
 - It is the number of adult male smokers that will follow the binomial distribution, not the proportion. Proportions follow the chi-squared distribution.
 - No. As it is rare (count data), this variable should follow the Poisson distribution.
 - Yes. This variable will follow the binomial distribution, as there are only two possible events (back pain or no back pain).
- True.
 - False. Continuous random variables can also be positively or negatively skewed. We consider this in more detail in section 11.2.
 - True.
 - False. The Poisson distribution becomes more normal as its mean increases.
 - True.

Section 11.3

Exercise 11.3.1

- Data should be from independent groups, the comparison must be of categorical data and in the form of counts (numbers), and at least 80 per cent of cells must have an expected value of 5 or more.
- Data should be from independent groups and continuous. Data are from population(s) with (approximately) symmetric normal distribution(s). It is the appropriate test for small samples ($n < 30$), but it can be (and usually is) used for larger sample sizes. The standard deviations of the two groups being compared should be similar.
 - It should be clear from the description of the data whether they are continuous and the samples are independent. Histograms could be used to assess whether the data appear to be from normal distributions, although this can be difficult to assess with very small

samples. The sample variances should be checked to see whether they are similar: If the larger sample variance is more than twice the smaller variance, the samples may be from populations with unequal standard deviations, and the t -test may therefore not be appropriate. This can be tested by the F -test (and F -distribution), a check made routinely by statistical analysis software such as SPSS.

Exercise 11.3.2

1. The 95 per cent CI is calculated with the critical value of the t -distribution for the relevant degrees of freedom, and the level of certainty we require (5 per cent). The critical value for 12 df at 0.05 probability level is 2.179, so the 95 per cent CI is given by

$$\bar{d} \pm 2.179SE = 3.8462 \pm (2.179 \times 1.2497) = (1.1, 6.6)$$

This should be stated to one more decimal place than the data.

2. It could be argued in this example that attending the talks cannot decrease patients' knowledge, so if there is any change in test score it can only be an increase; that is, a one-sided test is appropriate. But is it possible that patients could become more confused by the talks, and do worse in the second test? As we have noted, one-sided tests are rarely appropriate in practice. If in doubt, use a two-sided test!

Exercise 11.3.3

1. Administrative and shop-floor staff are most likely to differ from each other, but each may also differ from management and technical staff. Management and technical staff have very similar means and distributions.
2. The 95 per cent CIs for administrative and shop-floor staff clearly do not overlap. The upper limit for administrative staff is also below the lower limits for both management and technical staff. The lower limit for shop-floor staff is also above the upper limits for both management and technical staff.

Section 11.4

Exercise 11.4.1

The following need to be considered:

- the number of groups for analysis
- whether the groups are independent of each other, such as boys and girls (the general rule is that an individual cannot be in more than one group), or whether each set of observations are paired (that is, made on the same individual) or closely matched (when individuals in one group are individually matched with a member of the other group)
- the type of data, whether continuous, categorical or ordinal
- the distribution of the data
- the size of the sample
- the objective of the analysis.

Exercise 11.4.2

- a. Matched/paired categorical data; therefore McNemar's test.
- b. Independent categorical data; therefore chi-squared test (symptoms may be present/absent or on an ordinal scale, in which case adjust for trend).
- c. Chi-squared test, but we may get <5 expected in one cell; hence, we should consider Fisher's exact test.

- d. Matched/paired categorical data; therefore McNemar's test.
- e. Independent continuous data; therefore two-sample t -test (or could consider using Mann–Whitney U test if the data did not meet assumptions of t -test).
- f. Independent categorical data; therefore chi-squared test.
- g. Paired continuous data; therefore paired t -test (or could consider using Wilcoxon signed rank test if the data did not meet the assumptions of the paired t -test).
- h. We could use ANOVA or the Kruskal–Wallis test, depending on distributions and variances and whether it is appropriate to treat the scores as continuous.
- i. We know that the distribution of blood lead is markedly skewed, though we could transform this and use log lead levels with Pearson correlation so long as the reading test score distributions meet assumptions, but Spearman's rank correlation may be more suitable.

Section 11.5

Exercise 11.5.1

1. Calculation of prior odds $\frac{0.4}{1 - (0.4)} = 0.66$
2. Calculation of likelihood ratio (for positive test) $\frac{0.8}{1 - (0.7)} = 2.67$
3. Hence, the posterior odds = $0.66 \times 2.67 = 1.76$, and posterior probability = $\frac{1.76}{1 + 1.76} = 0.64$
4. Fagan's nomogram produces a consistent result (it is difficult to read this precisely).
5. This example shows how the probability of the woman having cancer (posterior probability) is determined by not only the result of the smear test but also the GP's view that, given the risk factor history and symptoms of bleeding, there is a moderately high chance of cancer.
6. In the case of the second woman, the history led the GP to a very low prior suspicion of cancer (prior probability = 5 per cent); using the nomogram, the posterior probability is only (approximately) 0.15 (15 per cent) despite the positive smear test. In other words, there is a high chance that the smear result is a false positive (though, of course, the smear result would need to be followed up carefully in the usual way).

Bibliography

- Andermann, A., Blancquaert, I., Beauchamp, S., Déry V. (2008). Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years. *Bull WHO* **86**(4): 317–319.
- Biddle, L., Brock, A., Brookes, S.T., Gunnell, D., *et al.* (2008) Suicide rates in young men in England and Wales in the 21st century: time trend study. *BMJ* **336**: 539. doi: <http://dx.doi.org/10.1136/bmj.39475.603935.25>
- Bingham, S., Riboli, E. (2004). Diet and cancer – the European prospective investigation into cancer and nutrition. *Nat Rev Cancer* **4**, 206–215. doi: 10.1038/nrc1298. Last accessed March 2016.
- Bradburn, N.M., Sudman, S., Wansink, B. (2004). *Asking questions: the definitive guide to questionnaire design – for market research, political polls, and social and health questionnaires* (rev edn). Jossey–Bass: San Francisco.
- Bronowski J. (1956) *The common sense of science*. Pelican: London.
- Brown, W.J., Hockey, R., Dobson, A. (2007). Rose revisited: a “middle road” prevention strategy to reduce noncommunicable chronic disease risk. *Bull WHO* **85**(11): 886–887. doi: 10.2471/BLT.07.041566
- Bruce, N., Dherani, M., Liu, R., Hosgood III, H.D., Sapkota, A., *et al.* (2015). Does household use of biomass fuel cause lung cancer? A systematic review and evaluation of the evidence for the GBD 2010 study. *Thorax* **70**: 433–441.
- Burton, H., Sagoo, G.S., Pharoah, P., Zimmern, R.L. (2012). Time to revisit Geoffrey Rose: strategies for prevention in the genomic era? *Ital J Public Hlth* **9**(4): e8665–1 to 9. doi: 10.2427/8665
- Craig, P., Dieppe, P., Macintyre, S., Mitchie, S., Nazareth, I., Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* **337**: a1655. doi: 10.1136/bmj.a1655
- Cresswell, J., Plano Clark, V.L. (2011). *Designing and conducting mixed methods research*. Sage: London.
- Deeks, J.J., Higgins, P.T. (2010). Statistical algorithms in Review Manager 5. Statistical Methods Group of The Cochrane Collaboration 2010. 1–11.
- Doll, R., Peto, R., Boreham, J., Sutherland, I. (2005). Mortality from cancer in relation to smoking: 50 years observations on British doctors. *Br J Cancer* **92**: 426–429.
- Erens, B., Phelps, A., Clifton, S., Mercer, C.H., Tanton, C., *et al.* (2013). Methodology of the Third British National Survey of Sexual Attitudes and Lifestyles (Natsal-3). (2013) *Sex Transm Infect* **90**: 84–89. Available at <http://sti.bmj.com/content/90/2/84.full?sid=1bb4a477-2fc6-40b2-a218-f2bcb4ea681d>. Last accessed 29 June 2017.
- Feigl, H. (1969). The origin and spirit of logical positivism. In P. Achinstein and S.F. Baker (eds.). *The legacy of logical positivism* (pp. 3–24). Johns Hopkins University Press: Baltimore.
- Guba, E.G., Lincoln, Y.S. (1994). Competing paradigms in qualitative research. In N.K. Denzin and Y.S. Lincoln (eds.). *Handbook of qualitative research* (pp. 105–117). London: Sage.

- Halfpenny, P. (1982). *Positivism and sociology. Explaining social life*. Routledge: New York.
- Hickish, T., Colston, K.W., Bland, J.M., Maxwell, J.D. (1989). Vitamin D deficiency and muscle strength in male alcoholics. *Clinical Science* **77**: 171–176.
- Hill, A.B. (1965). The environment and disease: association or causation? *Proc R Soc Med* **58**: 295–300.
- Howick, J., Glasziou, P., Aronson, J.K. (2009). The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? *J R Soc Med* **102**: 186–194. doi: 10.1258/jrsm.2009.090020
- Kuhn, T. (1970). *The structure of scientific revolutions* (2nd edn). University of Chicago Press: Chicago.
- Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., *et al.* (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**(9859): 2224–2260. doi: 10.1016/S0140-6736(12)61766-8
- Miettinen, S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* **25**: 339–355.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.; The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**(7): e1000097. doi: 10.1371/journal.pmed.1000097
- Murray, C.J.L., Richards, M.A., Newton, J.N., Fenton, K.A., Anderson, H.R., *et al.* (2013). UK health performance: findings of the Global Burden of Disease Study 2010. *Lancet* **381**(9871): 997–1020.
- Murray, C.J.L., Vos, T., Lozano R., Naghavi, M., Flaxman, A.D., *et al.* (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**: 2197–2223. doi: 10.1016/S0140-6736(12)61689-4
- Phillips, D. (1990) Postpositivistic science myths and realities. In E. Guba (ed.). *The Paradigm Dialog* (pp 31–45). (Guba E., ed). Sage: London.
- Popper K. (1959) *The logic of scientific discovery*. Hutchinson: London.
- Public Health England. (2010). Notifiable diseases as causative organisms, how to report. Retrieved from <https://www.gov.uk/guidance/notifiable-diseases-and-causative-organisms-how-to-report>. Last accessed 29 June 2017.
- Public Health England. (2014). Statutory Notifications of Infectious Diseases (NOIDS). Available at <https://www.gov.uk/government/collections/notifications-of-infectious-diseases-noids>. Last accessed 29 June 2017.
- Salomon, J., Vos, T., Hogan, D.R., Gagnon, M., Naghavi, M., *et al.* (2012). Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *Lancet* **380**: 2129–2143. doi: 10.1016/S0140-6736(12)61680-8
- Simonian, R., Laird, N. (1986). Meta-analysis in clinical trials. *Control Clin Trials* **7**: 177–188.
- Steenland, K., Armstrong, B. (2006). An overview of methods for calculating the burden of disease due to specific risk factors. *Epidemiology* **17**(5):1–8. doi 10.1097/01.ede.00002229155.05644.43
- Thomson Reuters. IDS Employment Law Brief. Private sector distribution of gross weekly earnings for full time employees. Available at: <https://ids.thomsonreuters.com/taxonomy/term/16/www.dti.gov.uk?page=7>. Accessed 29 June 2017.
- UK Ministry of Justice. (2014). Coroners Statistics 2014 England and Wales. Ministry of Justice Statistics Bulletin. Ministry of Justice: London. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427720/coroners-statistics-2014.pdf. Last accessed 29 June 2017.

- UK Ministry of Justice. Guide to Coroners Services. Ministry of Justice. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/363879/guide-to-coroner-service.pdf. Last accessed 29 June 2017.
- UK Office for National Statistics. Mid-year population estimates available. Retrieved from <http://www.ons.gov.uk>. Last accessed 8 December 2015.
- UK Office for National Statistics. (2013). Statistical Bulletin: Childhood, Infant and Perinatal Mortality in England and Wales: 2013. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/childhoodinfantandperinatalmortalityinenglandandwales/2015-03-10>. Last accessed 081215. Last accessed 29 June 2017.
- US Department of Health and Human Services. (2004). The health consequences of smoking: a report of the U.S. surgeon general, 2004. Smoking data: <http://www.lung.org/lung-disease/lung-cancer/resources/facts-figures/lung-cancer-fact-sheet.html>. Last accessed 29 June 2017.
- Wadsworth, J., Field, J., Johnson, A.M., Bradshaw, S., Wellings, K. Methodology of the National Survey of Sexual Attitudes and Lifestyles. (1993) *J R Stat Soc A* **156**: 407–421.
- Wareham, N.J., Jakes, R.W., Rennie, K.L., Mitchell, J., Hennings, S., Day, N.E. Validity and repeatability of the EPIC-Norfolk Physical Activity Questionnaire (2002) *Int J Epidemiol* **31**(1): 168–174.
- World Health Organisation. (2005). Reported information on mortality statistics. Retrieved from http://www.who.int/healthinfo/statistics/survey_2005/en/. Last accessed 29 June 2018.
- World Health Organisation. (2015). World Health Statistics 2015. Part II: Global Health Indicators. Retrieved from http://www.who.int/gho/publications/world_health_statistics/EN_WHS2015_Part2.pdf?ua=1. Last accessed 29 June 2017.
- World Health Organisation. (2016). WHO Mortality Database. Retrieved from http://www.who.int/entity/healthinfo/mortality_data/en/index.html. Last accessed 29 June 2017.

Index

a

abridged life table 380
 accuracy, test validity 454
 Adult Social Care Outcomes Framework 77
 AF. *See* attributable fraction
 African Index Medicus (AIM) 394
 age distribution, analysis of 102–104
 aggregate-level health service data 77–78
 aggregate measures, and risk of ecological fallacy 83
 aim, study 126
 air pollution
 available data on 39–41
 and health 39
 in London during the 1950s 80–81
 all-cause mortality 101
 Allied and Alternative Medicine (AMED) 393
 alternative hypothesis 201
 analysis of variance (ANOVA) 233, 323, 484, 508–511
 one-way ANOVA 508–509
 ANOVA. *See* analysis of variance
 antilog 490
 AR. *See* attributable risk
 asthma, case study on 38–39
 attributable fraction (AF) 432
 for continuous and multiple category exposures 434
 for dichotomous exposure 432–434
 exposed 433
 population 433–434
 attributable risk (AR) 430–432
 autocorrelation 332

b

bar chart 173, 174
 baseline hazard function 372
 Bayesian methods 520–525
 advantages of 524–525
 application and relevance 522–524
 Bayes' theorem and formula 521–522
 Bayes' theorem 521–522. *See also* Bayesian methods
 Bayes, Thomas 520
 benchmark tools 78
 beta coefficients 231, 283
 bias 128, 170. *See also specific types*
 observer variation and 167
 bimodal distribution 44
 binomial distribution 484–485
 binomial regression 240, 241
 biological abstracts (BIOSIS) 393
 birth cohort 463
 births, registration of 72
 blinding 308
 blood pressure, measurement of 191
 bridging tables 72
 British Library Conference Collections 394
 British Regional Heart Study (BRHS) 186. *See also* cohort studies

c

CAB Direct 394
 Cancer Outcomes and Services Dataset (COSD) 75
 cancer, registration of 35
 carry-over 318
 Carstairs Index 80
 case ascertainment 192, 193

- case-control studies 251–252
 - applications of design 258
 - approach to analysis 255–257
 - case definition and selection 259
 - confounding and logistic regression 276–289
 - logistic regression 278–287
 - stratification 277–278
 - control group selection 260–261
 - exposure assessment 262–263
 - bias in 263–264
 - hypothesis tests for 271–272
 - matching 261–262
 - multiple controls 262
 - objectives 253
 - odds ratio (OR) 265
 - in matched analysis 269–271
 - in unmatched analysis 265–268
 - sample size calculation 273–276
 - structure of 254–255, 256
 - time perspective 257–258
 - unmatched and matched analysis 265–273
- case definition 192
- case finding 192
- case studies 84
- categorical data 172
 - displaying and summarising 173–174
- categorical variables codings 282
- causal pathway 222
- causal risk difference 431
- cells 203
- censoring 357–358
 - indicator 358, 362
 - left 358
 - right 358
- census data 71
- Central Limit Theorem 144
- central register 193
- cervical cancer screening programme 450
- China Knowledge Resource Integrated Database (CNKI) 394
- chi-squared distribution 205, 484
- chi-squared statistic 205
- chi-squared test 201–209, 271–272, 313, 513
- chi-squared test for trend 513–514
- CI. *See* confidence interval
- Clean Air Act 82
- Clinical Commissioning Group (CCG) 78
- Clinical Practice Research Datalink (CPRD) 35, 75
- closed questions 158
- cluster randomised trial 328–330
- cluster random sampling 132–133
- clusters 328
- Cochrane, Archie 419
- Cochrane Central Register of Controlled Trials (CENTRAL) 393, 423
- Cochrane Collaboration 386, 419, 422–423
 - Cochrane library 422–423
 - collaborative review groups 422
 - logo of 422
 - review software 423
- Cochrane Database of Systematic Reviews 422–423
- Cochrane Information Management System (IMS) 423
- Cochrane Methodology Register (CMR) 423
- Cochrane website 393
- Cochran's Q 406
- coefficient of determination 67–68
- cohort effect 463
 - suicide trends in UK, analysis of 464–468
- cohort life table 377
- cohort studies 185
 - confounding 220–224
 - follow-up 193–198
 - Hill viewpoints on association 218–220
 - measurement 190–193
 - multiple linear regression 235–242
 - objectives of 186–187
 - obtaining the sample 188–190
 - presentation and results analysis 198–214
 - chi-squared test 201–209
 - relative risk 199–200
 - t*-test 209–214
 - z*-test 209–214
 - sample size 214–217
 - simple linear regression 224–234
 - structure of 188
- communicable diseases
 - food poisoning, time trends in 36, 37
 - notifiable in UK 36–38
 - seasonal and age patterns in 36–38
- Community Information Data Set (CIDS) 74
- community randomised trial 330–331
- complex interventions 319

- computer-assisted personal-interview (CAPI) 128
- computer-assisted self-interview (CASI) 128, 157
- concealment 307
- concordant pairs 269
- conditional logistic regression 269, 287
- confidence interval (CI) 107, 357
 - 95 per cent CI for population mean 147–148
 - for regression coefficients 239
 - for sample mean 146–149
- confounding 220–224
 - in intervention study 306
 - methods for dealing with 224
 - in physical activity and cancer study 222–223
- confounding variables 309
- Consolidated Standards of Reporting Trials (CONSORT) 343
- CONSORT Statement 343–344
- constant 231
- contamination 328
- contingency tables 202–203
 - degrees of freedom of 206
- continuous data 42, 56, 172, 176
 - hypothesis testing for 209–214
- continuous measurements 172
- continuous outcome 279
- continuous probability distributions 482–484
 - chi-squared distribution 484
 - F*-distribution 484
 - normal distribution 483
 - t*-distribution 483–484
- continuous variables 60
- control group selection 260
 - bias arising from 260–261
- controlled before-and-after study 332
- convenience sampling 133
- coroner 28
- correlation coefficient 64–67, 224
 - calculation of 65–67
- Cox proportionate hazards regression 355
- Cox regression 240, 371–377
 - application 375–376
 - hazard function 371–372
 - interpretation of model 373–374
 - model 372
 - prediction 374–375
 - proportional hazards assumption 372
 - checking 372–373
- critical value 406
- crossover trial 315, 317–318
- crude death rates 102, 111
- cumulative incidence 355
- cumulative incidence rate 13
- Cumulative Index to Nursing and Allied Health Literature (CINAHL) 394
- cumulative meta-analysis 387
- current life tables 355, 377–381
 - abridged life table 380
 - disability-free life expectancy 379–380
 - and life expectancy at birth 377–379
 - life expectancy at other ages 379
- d**
- DALYs. *See* disability-adjusted life years
- data 7, 172. *See also* frequency distribution
 - displaying 41–42
 - and summarising 173–176
 - morbidity 73
 - summarised, example of 56
 - types of 172–173
- Database of Abstracts of Reviews of Effectiveness (DARE) 423
- datum 7
- death 26–27
 - certificate 72
 - data (*see* mortality data)
 - notification to coroner 28, 72
 - process of recording of, in England and Wales 29, 72
- death rates, age-specific 104
- deduction 4
- degrees of freedom 53, 54, 206, 231, 272
- demographic data 69, 71–73
- demography 71
- denominator 11
- dependent variable 225
- deprivation indices 79
 - Carstairs Index 80
 - Indices of Multiple Deprivation 79–80
- Derwent Drug File 394
- description 7
- descriptive epidemiology 33–39
 - definition of 33
 - ecological studies 82–83
 - London Smogs of the 1950s 80–82

diagnostic fashion 38
 direct standardisation 104, 110–113
 deaths from stroke, change in 111–112
 European standard population, use of 112–113
 and indirect method, comparison of 113
 disability-adjusted life years (DALYs) 436, 439
 disability weights 436
 discordant pairs 269
 discrete data 172
 displaying and summarising 174–176
 discrete measurements 172
 discrete probability distributions 482, 484–487
 binomial distribution 484–485
 Poisson distribution 485–487
 disease burden 434–435
 attributable to specific risk factors 438, 440
 disability-adjusted life years (DALYs) 436, 439
 years lived with disability (YLD) 435–436, 437
 years of life lost (YLL) 435, 437
 disease frequency, measures of 12
 disease registers 78
 diseases, notifiable 36
 disease surveillance 75–76
 DisMod-MR model 436
 distribution 43. *See also* frequency distribution
 distribution-free tests. *See* non-parametric hypothesis tests
 dose–response relationship 415, 514
 dummy variables 279–280

e

ecological fallacy 83
 ecological study 82–83
 effect modification 319
 efficacy of treatment 305. *See also* intervention studies
 EMBASE (Excerpta Medica dataBASE) 393
 empirical observations 4
 empiricism 2
 end event 356
 environmental pollution data 39–41
 epidemiological study designs 84–86

epidemiology 2
 definition of 6
 functions of 6
 scientific reasoning and 2
 European standard population 112–113
 Eurostat 80
 exact probability 515
 expected frequency, calculation of 203
 explanatory variables 225, 281, 371
 exponentiation 490
 exposed attributable fraction 433
 exposure assessment 262–263
 bias in 263–264
 external consistency 163
 external migration 71
 external validity of trial 309

f

factorial design 319, 322–323
 factors 203
F-distribution 484, 509
 finite population correction factor 142
 first quartile 52–53
 Fisher's exact test 515–517
 fixed-effect meta-analysis 407
 follow-up, in cohort studies 193–198
 food poisoning, time trends in 36, 37
 forest plot 406, 408–409
F-ratio 231, 237, 282, 509
 frequency distribution 42, 173, 479, 480
 calculation of 42–43
 description of 44–57
 location 47–51
 shape 44–47
 spread 51–55
 frequency of disease, measures of 12
 frequentist approach 520
 funnelling 159
 funnel plot 403–404
 asymmetrical 403, 404
 publication bias and 404–405
 linear regression method 404
 rank correlation method 404

g

genomic research 448–449
 geometric mean 490
 Global Burden of Disease (GBD)–2010 study 435–437

Global Health 394
 goodness of fit of linear regression model
 229–231
 gradient, straight line 225
 green prescription intervention 329
 grey literature 394
 electronic databases 394
 guide table, use of 517–519

h

hazard 371
 hazard ratio 371, 372
 Health and Social Care Act 2012, 73
 Health and Social Care Information Centre
 (HSCIC) 73, 74, 76, 79
 Healthcare Quality Improvement Partnership
 (HQIP) 76
 health event data 69, 73–78. *See also* health
 service data
 health inequalities, in Merseyside 101–104
 health information
 routine collection of 26–32
 sources 69 (*see also* routine data sources)
 Health Protection (Notification) Regulations
 2010 35
 health service data
 aggregate-level 77–78
 patient-level 73–77
 routinely available 73
 Health Survey for England (HSE) 79
 heterogeneity 405–407
 Hill viewpoints 218–220
 histogram 41–43. *See also* frequency
 distribution; relative frequency
 distribution
 homogeneity 406
 Hospital Episode Statistics (HES) 74
 HSCIC. *See* Health and Social Care
 Information Centre
 hypothesis 4, 201
 hypothesis testing 201, 228, 313, 357, 488,
 493–494
 association between categorical variables
 513–517
 chi-squared test 513–514
 Fisher's exact test 515–517
 association between two groups 506–508
 Spearman's rank correlation 506–508
 choosing appropriate test 517–519

comparing more than two groups 508–513
 analysis of variance 508–511
 Kruskal–Wallis test 511–513
 comparing two independent groups 496
 Mann–Whitney *U* test 496–499
 comparing two matched/paired groups
 500–505
 paired *t*-test 500–503
 Wilcoxon signed rank test 504–505
 data transformation 488–489
 interpretation 491–492
 log transformation 489–490
 reciprocal transformation 491
 square root transformation 491
 fundamentals of 494–495
 inference 494–495
 null hypothesis 495
p-value 495
 multiple testing 520
 non-parametric 492–493
 robustness of test 488
 stages of 495
 hypothetico-deductive method 4

i

ICC. *See* intra-class correlation coefficient
 illness. *See* morbidity
 IMR. *See* infant mortality rate
 incidence density 14–15, 199, 355
 incidence rate 12–15
 incidence density 14–15
 person-time 14–15
 and prevalence rate, relationship between
 15–16
 Income Deprivation Affecting Children Index
 (IDACI) 80
 Income Deprivation Affecting Older People
 Index (IDAOPI) 80
 independent groups 315
 independent variable 225
 index population 101
 Indices of Multiple Deprivation 79–80
 indirectly standardised mortality rates 110
 indirect standardisation 104, 105–110. *See
 also* standardised mortality ratio (SMR)
 induction 4
 inductive approach 5
 infant mortality rate (IMR) 31–32
 infectious disease, notification of 35–38

- inference 7, 494–495
 - influenza, seasonal trends in 37
 - informed consent, intervention studies 308–309
 - instrument error 191
 - intention-to-treat analysis 312–313, 324
 - interaction 319
 - intercept, straight line 225
 - internal consistency 163, 164
 - internal migration 71
 - International Classification of Disease (ICD) 31, 72
 - International Committee of Medical Journal Editors (ICMJE) 392
 - International Passenger Survey 71
 - interpretative approach 5
 - and positivist approach 5
 - interquartile range (IQR) 52–53
 - calculation of 52–53
 - median and 55
 - interrupted time-series 332
 - interval measurements 172
 - intervals 42
 - intervention studies 297–344
 - analysis 309–310
 - adjustment for variables 315
 - compliance with treatment allocation 311–312
 - conclusions, drawing 315
 - crossover trial 317–318
 - effect of intervention 313–314
 - by intention-to-treat 312–313
 - loss to follow-up 311
 - paired comparisons 315–316
 - per-protocol analysis 313
 - review of variables at baseline 310–311
 - cluster design for analysis of, use of 334–337
 - complex interventions, testing 318–334
 - analysis and interpretation 323–326
 - cluster randomised trial 328–330
 - community randomised trial 330–331
 - factorial design 322–323
 - methodological opportunities and constraints 327
 - natural experiment 333
 - non-randomised intervention designs 332
 - randomised trial of individuals 319–322
 - ethical issues for 308–309
 - management 342–343
 - purpose of 299–300
 - registration 342
 - reporting 343–344
 - sample size calculation 337–341
 - categorical outcomes 337–339
 - cluster study designs 340–341
 - continuous outcome 339–340
 - one-sided and two-sided tests 339
 - structure of 300–301
 - study design
 - blinding 308
 - inclusion and exclusion criteria 303–304
 - informed consent 308–309
 - intervention and control 304–306
 - outcome assessment 307
 - placebo effect 305–306
 - randomisation 306–307
 - terminology 298
 - interviewer bias 158, 263
 - interviews 157–158
 - intra-class correlation coefficient (ICC) 334–335, 340
 - IQR. *See* interquartile range
 - I² statistic 406
- k**
- Kaplan–Meier survival curve 355, 359–362
 - interpretation of 365–370
 - Kruskal–Wallis test 511–513
- l**
- ladder of powers 491
 - Latin American and Caribbean Health Sciences Information System (LILACS) 394
 - lead-time bias 461
 - least squares regression 226
 - length-based sampling bias 460–461
 - life expectancy 377. *See also* current life tables
 - life tables 355, 377–381. *See also* current life tables
 - likelihood ratios (LRs) 454–455, 523–524
 - linear association 279

- linear regression 240
 - multiple 235–242
 - simple 224–234
- line chart 174, 175
- literature search 393–395
- Liverpool Quality Assessment Tools (LQATs) 398, 416
- location of distribution, measures of 47–51
 - mean 49
 - median 47–49
 - mode 47
- log hazard ratio 372
- logistic regression 240, 278–280
 - conditional 287
 - multivariable 281–287
- log-normal probability distribution 489
- log odds 279, 283
- log-rank test 362–364
- log transformation 489–490
- LRs. *See* likelihood ratios
- lung cancer death rates, variations in 33–34

- m**
- Mann–Whitney *U* test 496–499
- matched analysis 262, 265
- matched pairs, in case–control study 269, 270
- matching, case-control studies 261–262
- Maternal, Infant and Perinatal programme 76
- MBRRACE-UK 76
- McNemar’s test 272, 316
- mean 49
 - measures of spread for 55
- mean number of events 486
- mean sums of squares 231
- measurement 157
 - accuracy of 157
 - checking success of 161, 163–164
 - cohort study 190–193
 - development of questionnaire 161, 162
 - interviews 157–158
 - quality, assessment of 165–169
 - questionnaire design 158–161
 - self-completed questionnaires 157–158
 - sources of error 169–171
- measurement error 191
 - types of 191
- median 47–49
 - and mean, comparison of 49–50
 - measures of spread for 55
- median survival time 361
- Medicine and Healthcare Products Regulatory Agency 75
- Medline 393
- meningitis, age and seasonal trends in 36–37
- Mental Health Services Data Set (MHSDS) 74
- meta-analysis 385–386, 399, 402–414
 - assessment of publication bias 403–405
 - calculating pooled estimate 407–408
 - fixed-effect model 407
 - random-effects model 407–408
 - forest plot 408–409
 - methods for 402–403
 - of observational studies 415–418
 - sensitivity analysis 409–410
 - statistical software for 410–411
 - statistical test for heterogeneity 405–407
 - value of 411–414
- methodological quality, assessment of 396–398
- mixed-methods approaches 5
- modal category 173
- modal group 47
- mode 44, 47
- model sum of squares 230
- morbidity 34–35
 - cancer registration 35
 - in general practice 38–39
 - infectious disease notification 35–38
 - recording of 35
 - sources of information on 35
- mortality
 - datasets 76–77
 - follow-up for 193
- mortality data 26, 77
 - collection and recording of 26–29
 - infant mortality rate 31–32
 - suicide rates 29–31
- mortality statistics in UK and Austria, preparation of 34
- multi-level modelling 335
- multiple linear regression 235–242
- multistage sampling 132, 133
- multivariable analysis 224
- multivariable logistic regression 281–287

n

National Air Quality Strategy 40
 National Cancer Intelligence Network (NCIN) 75
 National Confidential Enquiry into Patient Outcome and Death (CEPOD) 76–77
 National Confidential Inquiry into Suicide and Homicide (NCISH) 76
 National Drug Treatment and Monitoring System (NDTMS) 75
 National Health Service (NHS) 74
 Central Register 71
 cervical screening programme 450
 Outcomes Framework 77
 National Institute for Health Research (NIHR) 75
 National Morbidity Studies 35, 38
 national population census 71
 National Society for Smoke Abatement 82
 National Survey of Sexual Attitudes and Lifestyles (Natsal) 123–126
 natural experiment 86, 298, 333
 natural logarithm 489
 negatively skewed data 45, 491
 negative predictive value 454
 Newcastle–Ottawa Scale 398, 416
 NHS Economic Evaluation Database (EED) 423
 non-parametric hypothesis tests 492–493
 non-parametric methods 487
 non-randomised intervention study designs 332
 non-responders 153
 non-response bias 153, 154
 normal distribution 144, 145–146, 483
 Notification of Infectious Disease System (NOIDS) 75
 null hypothesis 201–202, 271, 314, 495
 number needed to treat (NNT) 326
 numerator 11

o

objectives, study 126
 observational studies. *See also specific studies*
 bias in 415
 confounding in 415
 designs 251, 298
 meta-analysis of 415–418
 systematic review of 414–418

observations 41
 observer bias 192, 263
 observer error 191
 observer variation 165, 167
 odds ratio (OR) 256, 265
 in matched analysis 269–271
 pooled, adjusted OR 278
 in unmatched analysis 265–268
 Office of National Statistics (ONS) Longitudinal Study 79
 Omnibus Tests of Model Coefficients (OTMC) table 282, 285
 one-sided test 516
 OpenEpi statistical software 274
 open questions 158, 159
 OR. *See* odds ratio
 ordered categorical data 172
 order effects 318
 outcomes frameworks 77
 outcome variable 225
 outlier 46
 overall response rate 153
 oversampling 131
Oxford Textbook of Medicine, 387

p

PAF. *See* population attributable fraction
 paired data 315
 paired tests 316
 paired *t*-test 316, 484, 500–503
 paradigm shifts 3
 parameters 482
 parametric methods 487
 Pascal Biomed 394
 patient information systems 35
 patient-level health service data 73–77
 peaks, distribution 44
 Pearson correlation 506
 Pearson correlation coefficient 64
 pencil and paper interviews (PAPI) 157
 period effect 463–464
 suicide trends in UK, analysis of 464–468
 period of time 11
 period prevalence 12, 13
 per-protocol analysis 313
 person-time 14–15
 PICO (population, intervention, comparator, and outcome) 389

- PICO model 124
- pie chart 173
- piloting 159
- placebo effect 305–306
- point prevalence 12, 13
- Poisson distribution 485–487
 - examples of 485–486
 - features of 486–487
- Poisson regression 240, 241
- pooled standard deviation 213
- Popper, Karl 4, 8
- population 11, 127
 - estimates 71–72
 - projections 72
 - at risk 13
 - size 71
 - structure 71
- population attributable fraction (PAF) 433–434
- population-based health information 69, 78–79
- positively skewed 45
- positive predictive value 454
- positivism 3
- posterior probability 520, 525
- post-positivistic approach 3
- post-stratification 278
- pragmatic trials 305
- prediction 374–375
- predictive value 454
- Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) 419. *See also* PRISMA statement
 - checklist for reporting of systematic reviews and meta-analyses 420–421
- pretesting 159
- prevalence rate 12
 - incidence rate, illness duration, and 15–16
- prevention
 - approaches to 16–17
 - primary 18
 - secondary 18
 - strategies of
 - distribution of risk in populations 440–443
 - genomic research, implications of 448–449
 - high-risk approach 443–444
 - population approach 444–445
 - Rose hypotheses 447–448
 - safety and population strategy 445
 - tertiary 18
- Primary Care Mortality Database 76
- prior probability 520
- PRISMA statement 389, 395, 419
- probability 129–130, 202, 478–479
 - main features of 479
- probability density 481
- probability distributions 479–481
 - continuous 482–484
 - chi-squared distribution 484
 - F*-distribution 484
 - normal distribution 483
 - t*-distribution 483–484
 - discrete 482, 484–487
 - binomial distribution 484–485
 - Poisson distribution 485–487
 - statistical methods, implications for 487
 - types of 481–487
- product–moment correlation. *See* Pearson correlation coefficient
- prognostic index 374
- proportional hazards 372
- proportional hazards regression 371, 372. *See also* Cox regression
- prospective data collection 257, 258
- PsychInfo 394
- publication bias 391–392
- Public Health England (PHE) 75, 76
 - Data and Analysis tools 77
- Public Health Outcomes Framework 77
- p*-value 206, 211–212, 495
- q**
- Q statistic 406
- qualitative methods 5
- Quality and Outcomes Framework (QOF) 78
- quantitative information 7
- quantitative methods 5
- quasi-experimental designs 332
- questionnaire
 - checking success of 161, 163–164
 - closed questions 158
 - design 158–161
 - development of 161, 162
 - general layout 159

questionnaire (*Continued*)
 length of 159
 open questions 158, 159
 ordering of questions 159
 phrasing of questions 160
 skip questions 161
 quotas 134–135
 quota sampling 134–135

r

random-effects meta-analysis 407–408
 random error 107, 169
 randomisation 297, 306–307
 method of 306–307
 purpose 306
 randomised blind (placebo) control trial 301
 randomised control trial (RCT) 298
 random sampling 129. *See also* sampling
 methods
 range 51–52
 rank ordering 492–493
 rate
 definition of 11
 incidence 12–15
 need of 11
 numerator and denominator 11
 prevalence 12
 raw data 41
 recall bias 263
 receiver-operator characteristic (ROC)
 curve 456–460
 reciprocal transformation 491
 records, bias from 263–264
 reference category 280, 283
 regression 222, 225
 regression coefficient, adjusted 238
 regression line
 interpreting 227–228
 using 228
 regression model 357
 relative frequency 173
 relative frequency distribution 57–59
 relative risk (RR) 199–200, 265, 371, 430–431
 reliability 165
 repeatability 165
 reproducibility 165
 research 10
 political context of 126
 study designs 84–86

research ideas 8
 development of 10
 research paradigms 5
 research question
 formulation of 8–10, 124
 PICO model 124
 residual confounding 218, 415
 residuals 204
 resistant measure 50
 responders 153
 response bias 138, 154
 response rate 153
 determining 153–154
 maximising 154–156
 retrospective data collection 257–258
 Review Manager (RevMan) 410–411, 423
 risk
 attributable (*see* attributable risk (AR))
 relative (*see* relative risk (RR))
 risk difference 431
 risk ratio. *See* relative risk (RR)
 Rose, Geoffrey 447
 routine data sources 69, 70
 for countries other than UK 80
 demographic data 71–73
 deprivation indices 79–80
 health event data 73–78
 population-based health information 78–79
 strengths and weaknesses of 70
 RR. *See* relative risk

s

sample 7, 127
 inadequate 127–128
 representative 154
 sample mean
 confidence interval for 146–149
 sampling distributions of 139–145
 standard error of 140–145
 sample size 148
 for case–control studies 273–276
 cohort study 190, 214–217
 estimating
 population mean 149–150
 population proportion 151–153
 intervention studies 337–341
 sampling 127–129
 losses in 137–138
 in Natsal-3 study 135–136

- sampling distribution 139–145, 201
- sampling error 128, 132, 139
- sampling fraction 142
- sampling frame 137–138
- sampling methods 129
 - cluster random sampling 132–133
 - convenience sampling 133
 - multistage random sampling 133
 - people difficult to contact 133–134
 - quota sampling 134–135
 - simple random sampling 130–131
 - snowball sampling 134
 - stratified sampling 131–132
 - systematic sampling 133
- scatterplot 60–62, 67, 224
- Science Citation Index (SCI) 394
- scientific reasoning
 - and epidemiology 2
 - forms of 3
 - deduction 4
 - induction 4
- scientific research
 - approaches to 2–8
 - epidemiology 6
 - history and nature of 2–5
 - learning, approach to 8
 - statistics 7
- screening programme 450
 - effectiveness of, issues in 460–462
 - lead-time bias 461
 - length-based sampling bias 460–461
 - selection bias 460
 - evaluation of, criteria for 451–452
 - purpose of 451
 - validity of 452–460
 - accuracy 454
 - likelihood ratio 454–455
 - predictive value 454
 - receiver-operator characteristic (ROC) curve 456–460
 - sensitivity 453
 - specificity 453–454
- SE. *See* standard error
- Secondary Uses Service (SUS) 73
- Second Generation Surveillance System (SGSS) 75–76
- secular trend 332
- selection bias 128, 460
- self-completed questionnaires 157–158
- sensitivity analysis 72, 397, 398
 - meta-analysis 409–410
- shape of distribution 44–47
- significance level of test 206
- sign test 505
- simple linear regression 224–234
 - best fit for straight line, finding 226–227
 - describing associations 224–226
 - goodness of fit of model 229–231
 - hypothesis testing 228–229
 - interpreting regression line 227–228
 - SPSS output, interpreting 231–234
 - ANOVA table 233
 - coefficients table 233–234
 - model summary 232–233
 - variables entered/removed 231–232
 - using regression line 228
- simple random sampling 130–131
- skewed data 488
- skewed distribution 45, 50
- skip questions 161
- smoking
 - and heart disease 222
 - and suicide, association between 415–416
- SMR. *See* standardised mortality ratio
- snowball sampling 134
- Spearman correlation coefficient 64
- Spearman's rank correlation hypothesis test 506–508
- spending and outcomes tool (SPOT tool) 78
- spread of distribution 51–55
- square root transformation 491
- standard deviation 54
 - mean and 55
- standard error (SE) 108, 140–145
 - of difference between means 211
 - and 95 per cent CI for population proportion 150–151
- standardisation 101, 104, 224
 - age distribution analysis and 102–113
 - direct 104, 110–113
 - for factors other than age 114
 - indirect 104, 105–110
 - purpose of 101
- standardised mortality ratio (SMR) 104
 - calculation of 105–110
 - comparison of 110
 - increasing precision of 108
 - indirectly standardised rate 110

- standardised mortality ratio (*Continued*)
 interpretation of 107
 mortality in Liverpool 105
 mortality in Sefton 108–110
 95 per cent confidence interval for 107–108
 standard error of 108
- standard normal distribution 483
- statistical heterogeneity 405–407
- statistics 7
 use of 7
- step function 362
- straight line 225
- stratification 277–278
 post-stratification 278
 during study design 277–278
- stratified analysis 277
- stratified sampling 131–132
- Strengths and Difficulties Score (SDQ) 334
- The Structure of Scientific Revolutions* (Thomas Kuhn) 3
- suicide, mortality rates for
 among men 29–31
 among young women 31
- suicide, trends in, study on 464
 age-specific suicide rates 464
 cohort effects, analysis of 466–467
 period effects, analysis of 465–466
 three-year moving averages 465
- Summary Hospital-Level Mortality Indicator (SHMI) 76
- summary measure, usefulness of 104
- sum of squared residuals 230
- sum of squares 53
- surveillance systems for diseases 75–76
- surveys 84, 123
 sources of error in 170
- survival analysis 355, 356–370
 censoring 357–358
 Kaplan–Meier survival curve 359–362
 interpretation of 365–370
 log-rank test 362–364
 need of 356–357
- survival time 356
- symmetric distribution 44–45
- systematic error 169
- systematic reviews 385, 387–402. *See also* meta-analysis
- data extraction 399
- descriptive presentation of results 399–401
- flow of information through different phases of 396
- identifying relevant studies 391–396
 literature search 393–395
 publication bias 391–392
 specifying search terms 392
- importance of conducting of 387–388
- inclusion and exclusion criteria 390–391
- intervention studies *versus* observational studies 417
- methodological quality, assessment of 396–398
- of observational studies 414–418
- reporting and publishing 418–419
- research question and study objectives 389–390
- structured protocol, development of 388–389
- systematic sampling 133
- System for Information on Grey Literature in Europe (SIGLE) 394
- t**
- t*-distribution 483–484
- terminal event 356
- test of significance 207
- theoretical positions 5
- third quartile 52–53
- total sum of squares 230
- t*-test 313
- two-sample *t*-test 212–214
- two-sample *z*-test 210–212
- two-sided test 516–517
- type I error 215
- type II error 215
- U**
- UK Cochrane Centre 419
- UK Medical Research Council (MRC) 319
- UK Office of National Statistics (ONS) 72
- uncontrolled before-and-after study 332
- underreporting 36
- unimodal distribution 44
- univariate analysis 237

univariate model 283
unmatched analysis 262, 265

V

validity 165, 166
 of screening test 452–460
variables 60
 continuous 60
 relationship between 62
 linear relationship 62, 63
 negative relationship 63
 non-linear relationship 62
 positive relationship 63
 scatterplot 60–62
variance 53, 54

W

Wilcoxon signed rank test 496, 504–505
Wilson–Jungner criteria, screening
 programmes 451, 461–462
World Health Organisation Statistical
 Information System (WHOSIS) 80

Y

years lived with disability (YLD) 435–436, 437
 disability weights 436
 DisMod-MR model 436
years of life lost (YLL) 435, 437

Z

z-test 209–214