

---

# APPLIED NONPARAMETRIC STATISTICAL METHODS

---

Third edition

---

P. Sprent

*Emeritus Professor of Statistics  
University of Dundee  
Scotland*

and

N.C. Smeeton

*Lecturer in Medical Statistics  
Guy's, King's, and St. Thomas' School of Medicine  
King's College, London, UK*

---

CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

## Library of Congress Cataloging-in-Publication Data

Sprent, Peter.

Applied nonparametric statistical methods.--3rd ed. I P Sprent, N.C. Smeeton.

p. cm. -- (Chapman & Hall/CRC texts in statistical science series)

Includes bibliographical references and index.

ISBN 1-58488-145-3 (alk. paper)

1. Nonparametric statistics. I. Smeeton, N.C. II. Title. III. Texts in statistical science.

QA278.8 S74 20W

519.5'4--dc21

00-055485

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Apart from any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the UK Copyright Designs and Patents Act, 1988, this publication may not be reproduced, stored or transmitted, in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without the prior permission in writing of the publishers, or in the case of reprographic reproduction only in accordance with the terms of the licenses issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the license issued by the appropriate Reproduction Rights Organization outside the UK.

All rights reserved. Authorization to photocopy items for internal or personal use, or the personal or internal use of specific clients, may be granted by CRC Press LLC, provided that \$.50 per page photocopied is paid directly to Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923 USA. The fee code for users of the Transactional Reporting Service is ISBN 1-58489-1453/0150.004.50. The fee is subject to change without notice. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

The consent of CRC Press LLC does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press LLC for such copying.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

Visit the CRC Press Web site at [www.crcpress.com](http://www.crcpress.com)

---

© 2001 by Chapman & Hall/CRC

No claim to original U.S. Government works

International Standard Book Number 1-58488-145-3

Library of Congress Card Number 00-05.5485

Printed in the United States of America 3 4 5 6 7 8 9 0

Printed on acid-free paper

# Contents

## Preface

### 1 Introducing nonparametric methods

- 1.1 Basic statistics
- 1.2 Samples and populations
- 1.3 Hypothesis tests
- 1.4 Estimation
- 1.5 Ethical issues
- 1.6 Computers and nonparametric methods
- 1.7 Further reading
- Exercises

### 2 Centrality inference for single samples

- 2.1 Using measurement data
- 2.2 Inferences about medians based on ranks
- 2.3 The sign test
- 2.4 Transformation of ranks
- 2.5 Asymptotic results
- 2.6 Robustness
- 2.7 Fields of application
- 2.8 Summary
- Exercises

### 3 Other single-sample inference

- 3.1 Inferences for dichotomous data
- 3.2 Tests related to the sign test
- 3.3 Matching samples to distributions
- 3.4 Angular data
- 3.5 A runs test for randomness
- 3.6 Fields of application
- 3.7 Summary
- Exercises

### 4 Methods for paired samples

- 4.1 Comparisons in pairs
- 4.2 A less obvious use of the sign test
- 4.3 Power and sample size
- 4.4 Fields of application
- 4.5 Summary
- Exercises

**5 Methods for two independent samples**

- 5.1 Centrality tests and estimates
  - 5.2 Rank based tests
  - 5.3 The median test
  - 5.4 Normal scores
  - 5.5 Tests for survival data
  - 5.6 Asymptotic approximations
  - 5.7 Power and sample size
  - 5.8 Tests for equality of variance
  - 5.9 Tests for a common distribution
  - 5.10 Fields of application
  - 5.11 Summary
- Exercises

**6 Three or more samples**

- 6.1 Comparisons with parametric methods
  - 6.2 Centrality tests for independent samples
  - 6.3 Centrality tests for related samples
  - 6.4 More detailed treatment comparisons
  - 6.5 Tests for heterogeneity of variance
  - 6.6 Some miscellaneous considerations
  - 6.7 Fields of application
  - 6.8 Summary
- Exercises

**7 Correlation and concordance**

- 7.1 Correlation and bivariate data
  - 7.2 Ranked data for several variables
  - 7.3 Agreement
  - 7.4 Fields of application
  - 7.5 Summary
- Exercises

**8 Regression**

- 8.1 Bivariate linear regression
  - 8.2 Multiple regression
  - 8.3 Nonparametric regression models
  - 8.4 Other multivariate data problems
  - 8.5 Fields of application
  - 8.6 Summary
- Exercises

**9 Categorical data**

- 9.1 Categories and counts
- 9.2 Nominal attribute categories

- 9.3 Ordered categorical data
- 9.4 Goodness-of-fit tests for discrete data
- 9.5 Extension of McNemar's test
- 9.6 Fields of application
- 9.7 Summary
- Exercises

## **10 Association in categorical data**

- 10.1 The analysis of association
- 10.2 Some models for contingency tables
- 10.3 Combining and partitioning of tables
- 10.4 Power
- 10.5 Fields of application
- 10.6 Summary
- Exercises

## **11 Robust estimation**

- 11.1 When assumptions break down
- 11.2 Outliers and influence
- 11.3 The bootstrap
- 11.4 M-estimators and other robust estimators
- 11.5 Fields of application
- 11.6 Summary
- Exercises

## **Appendix**

### **References**

### **Solutions to odd-numbered exercises**

# Preface

The second edition of this book was written by the first-named author to provide a then (1993) up-to-date introduction to nonparametric and distribution-free methods. It took a midway course between a bare description of techniques and a detailed exposition of the theory. Individual methods and links between them were illustrated mainly by examples, Mathematics was kept to the minimum needed for a clear understanding of scope and limitations. The book was designed to meet the needs both of statistics students making first contact with these methods and of research workers, managers, research and development staff, consultants and others working in various fields who had an understanding of basic statistics and who, although they had little previous knowledge of nonparametric methods, now found or thought they might find them useful in their work.

A positive response from readers and reviewers has encouraged us to retain the basic format while taking the opportunity to introduce new topics as well as changing the emphasis to reflect both developments in computing and new attitudes towards data analysis.

Nonparametric methods are basically analytic tools, but data collection, analyses and their interpretation are interrelated. This is why we have expanded the coverage of topics such as ethical considerations and calculation of power and of sample sizes needed to achieve stated aims. These make their main impact at the planning stage, but also influence the analytic and inferential phases.

There has been widespread criticism in recent years by many statisticians of inappropriate and even improper use of significance tests and the related concept of P-values. However, these tools have a positive role when properly used and understood. To encourage better use the section on hypothesis testing in Chapter I has been rewritten, and throughout the book there is more emphasis on how these concepts should be used and warnings about potential misuse.

The layout of Chapters I to 10 follows the broad pattern of the corresponding chapters in the second edition but there are many changes in order and other aspects of presentation including new and more detailed examples. One or two topics have been dropped or are treated in less detail, and new material has been inserted where appropriate. As well as comments on ethical considerations and discussions on power and sample size, there are new sections on the

analysis of angular data, the use of capture-recapture methods, the measurement of agreement between observers and several lesser additions. Examples have been chosen from a wider range of disciplines. For a few more advanced topics such as regression smoothing techniques and M-estimation we have not given details of specific methods but only a broad overview of each topic to enable readers to judge whether it may be relevant to their particular needs. In such cases references are given to sources that contain the detail needed for implementation.

Chapter 11 has been rewritten to give an elementary introduction to influence functions, the nonparametric bootstrap and robust estimation generally, again with references to source material for those who want to make full use of these ideas. Material that appeared in Chapter 12 of the second edition has been updated and incorporated at relevant points in the text.

We have not included tables for basic nonparametric procedures, mainly because more satisfactory information is provided by modern statistical software, making many standard tables insufficient or superfluous for serious users of the methods. Those who need such tables because they have no access to specialized software are well catered for by standard collections of statistical tables. We give references to these throughout the book and also when relevant to some specialized tables. We have retained the section outlining solutions to odd-numbered exercises.

We are grateful to many readers of the earlier editions who made constructive comments about the content and treatment, or sometimes about the lack of treatment, of particular topics. This input triggered many of the changes made in this edition. Our special thanks go to Jim McGanrick for helpful discussions on physiological measurements and to Professor Richard Hughes for advice on the Guillain-Barré syndrome. We happily renew the thanks recorded in the second edition to Timothy P. Davis and Chris Theobald who supplied us with data sets used initially in that edition for examples that we have retained.

P. Sprent  
N. C. Smeeton

July 2000

# 1

## Introducing nonparametric methods

### 1.1 BASIC STATISTICS

One need know only a little statistics to make sensible use of simple nonparametric methods. While we assume that many readers will be familiar with the basic statistical notions met in introductory or service courses of some 20 hours instruction, nevertheless those without formal statistical training should be able to use this book in parallel with one of the many introductory general statistical texts available in libraries or from most academic bookshops, or one recommended by a statistician colleague or friend. The choice may depend upon how familiar one is with mathematical terms and notation. For example, *Essential Statistics* (Rees, 1995) adopts a straightforward approach that should suffice for most purposes, but some may prefer a more advanced treatment, or an introductory text that emphasizes applications in particular areas such as medicine, biology, agriculture, the social sciences and so on. Readers with considerable experience in general statistics but who are new to nonparametric methods will be familiar with some background material we give, but we urge them at least to skim through this to see whether we depart from conventional treatments. For example, our approach to testing and estimation in Sections 1.3 and 1.4 differs in certain aspects from that in some general statistics courses.

In this chapter we survey some concepts relevant to nonparametric methods. It is the methods and **not** the data that are nonparametric.

#### 1.1.1 Parametric and nonparametric methods

Statistics students meet families of probability distributions early in their courses. One of the best known is the **normal** or **Gaussian** family, where individual members are specified by assigning constant values to two quantities called **parameters**. These are usually denoted by  $\mu$  and  $\sigma^2$  and represent the mean and variance.

The notation  $N(\mu, \sigma^2)$  denotes a normal distribution with these parameters. Normal distributions are often associated with continuous data such as measurements.

Given a set of independent observations (a concept explained more fully in [Section 1.2](#)) from a normal distribution, we often want to infer something about the unknown parameters. The sample mean provides a point (i.e. single value) estimate of the parameter  $\mu$  (sometimes called, in statistical jargon, the **population** mean). Here the well-known *t*-test is used to measure the strength of the evidence provided by a sample to support an a priori hypothesized value  $\mu_0$  for the population mean. More usefully, we may determine what is called a **confidence interval** for the ‘true’ population mean. This is an interval for which we have, in a sense we describe in [Section 1.4.1](#), reasonable confidence that it contains the true but unknown mean  $\mu$ . These are examples of **parametric inferences**.

The normal distribution is strictly relevant only to some types of continuous scale data such as measurements, but it often works quite well if the measurement scale is coarse (e.g. for examination marks recorded to the nearest integer). More importantly, it is useful for approximations in a wide range of circumstances when applied to other types of data.

The **binomial distribution** is relevant to some types of counts. The family also has two parameters  $n, p$ , where  $n$  is the total number of observations and  $p$  is the probability that a particular one from two possible events occurs at any observation. Subject to certain conditions, the number of occurrences,  $r$ , where  $0 \leq r \leq n$ , of that event in the  $n$  observations has a binomial distribution, which we refer to as a  $B(n, p)$  distribution.

With a binomial distribution if we observe  $r$  occurrences of an event in a set of  $n$  observations, then  $\hat{p} = r/n$  is a point estimate of  $p$ , the probability of success at each independent observation. We may want to assess how strongly sample evidence supports an a priori hypothesized value  $p_0$ , say, for  $p$  or obtain a confidence interval for the value of  $p$  for the population.

The binomial distribution is often relevant to counts in dichotomous outcome situations. For example, the number of male children in a family of size  $n$  is often assumed to have a binomial distribution with  $p = 1/2$ , but we see in [Section 1.1.3](#) that this is only approximate. The number of ‘sixes’ recorded in 10 casts of a fair die has a  $B(10, 1/6)$  distribution. The outcome of interest is sometimes called a ‘favourable’ event, but this is hardly appropriate if, for example, the event is a positive diagnosis of illness.

Other well-known families include the uniform (or rectangular), multinomial, Poisson, exponential, double exponential, gamma, beta and Weibull distributions. This list is not exhaustive and you may not be, and need not be, familiar with all of them.

It may be reasonable on theoretical grounds or on the basis of past experience to assume that observations come from a particular family of distributions. Also experience, backed by theory, suggests that for many measurements inferences based on the assumption that observations form a random sample from some normal distribution may not be misleading even if the normality assumption is incorrect. A theorem called the **central limit theorem** justifies these and other uses of the normal distribution, particularly in what are called **asymptotic approximations**. We often refer to these in this book.

**Parametric inference** is sometimes inappropriate or even impossible. To assume that samples come from any specified family of distributions may be unreasonable. For example, we may not have examination marks on, say, a percentage scale for each candidate but know only the numbers of candidates in banded and ordered grades designated Grade A, Grade B, Grade C, etc. Given these numbers for two different schools, we may want to know if they indicate a difference in performance between schools that might be due to unequal standards of teaching or the ability of one school to attract more able pupils. The method of inference is then usually **nonparametric**. Even when we have precise measurements it may be irrational to assume a normal distribution because normality implies certain properties of symmetry and spread. We may be able to see that a sample is not from a normal distribution simply by looking at a few characteristics like the sample mean, median, standard deviation and range of values. For example, if all observations are either zero or positive and the standard deviation is appreciably greater than the mean then, unless the sample is very small, it is unreasonable to assume it comes from a normal distribution (see [Exercise 1.9](#)). There are well-known distributions that are flatter than the normal (e.g. the continuous uniform, or rectangular, distribution) or skew (e.g. the exponential and gamma distributions). In practice we are often able to say little more than that our sample appears to come from a distribution that is skew, or very peaked, or very flat, etc. Here nonparametric inference may again be appropriate. In this latter situation some writers prefer the term **distribution-free** to nonparametric.

The terms ‘distribution-free’ and ‘nonparametric’ are sometimes regarded as synonymous. Indeed, in defining a distribution-free

method Marriott (1990) states that ‘Distribution-free inference or distribution-free tests are sometimes known as nonparametric but this usage is confusing and should be avoided’. We believe that view ignores subtle distinctions, though these usually need not worry us unduly. Indeed, the nature of these distinctions depends to some extent on how one defines terms like *parameter* and *distribution*. Statisticians agree that the constants like  $\mu$ ,  $\sigma^2$  that define family members in the normal distribution are parameters; many also argue that the mean, say, of any distribution that has a mean (precisely to what family the distribution belongs need not be specified) is a parameter even if it does not appear as a constant in the distribution function. This concept accords with the broader mathematical definition of a parameter as an unknown quantity that may take one of a set of possible values. Those who take this view argue that if we make inferences about the mean of a distribution without making any assumption about its nature or its belonging to a particular family it is still an inference about a parameter and so the name nonparametric is unjustified. However, it would be appropriate in these circumstances to say the inference is distribution-free when no assumption is made about the nature of the distribution.

In practice this is usually an over-simplification; in nearly all inferences we make some distributional assumptions. For example, we may assume a sample is from a symmetric distribution. This does not restrict us to a particular family such as the normal but it does mean we exclude all asymmetric distributions.

That there is room for argument about precisely what we mean by the terms nonparametric or distribution-free is reinforced by the growth in compromise terms for methods that do not fall clearly into areas covered by these names. Two examples are the descriptors **semiparametric** and **asymptotically distribution-free**.

Paradoxically, many tests that are generally regarded as nonparametric or distribution-free involve parameters and distributions (often the normal or binomial distributions). This is because the tags ‘nonparametric’ or ‘distribution-free’ relate **not** to the distribution of the test statistics, but to the fact that the methods can be applied to samples from populations having distributions which need only be specified in broad terms, e.g. as being continuous, symmetric, identical, differing only in median or mean, etc. They need not belong to a specified family. There is a grey area between what is clearly distribution-free and what is parametric inference. Some of the association tests described in [Chapters 9](#) and [10](#) fall in this area.

### 1.1.2 When we might use nonparametric methods

Even if a parametric test does not depend critically on an assumption that samples come from a distribution in a particular family, when in doubt we may prefer a nonparametric test needing weaker assumptions. More importantly, nonparametric methods are often the only ones available for data that simply specify order (ranks) or counts of numbers of events or of individuals in various categories.

Nonparametric methods are not assumption-free. In most statistical problems what we can deduce, by either parametric or nonparametric methods, depends upon what assumptions can validly be made. An example illustrates this.

#### *Example 1.1*

A contract for making metal rods specifies, among other requirements, that not more than 2.5 per cent should have a cross-sectional diameter exceeding 30 mm. Some quality control methods that are used to see if this condition is met are based on counting the proportion of defectives (i.e. rods exceeding 30 mm diameter) in samples. Such tests can be extended to assess whether it is reasonable to assume that two machines each produce the same proportion of defectives. Even if we conclude that they do, it does not follow that the distribution of diameters is the same for each.

For example, suppose that diameters of items produced by the first machine have a normal distribution with mean 27 mm and standard deviation 1.53 mm; normal distribution theory then tells us that 2.5 per cent of all items produced will have a diameter exceeding 30 mm. This is because, for any normal distribution, 2.5 per cent of all items have a diameter at least 1.96 standard deviations above the mean; in this case, 2.5 per cent exceed  $27 + 1.96 \times 1.53 \approx 30$ . If the second machine produces items with diameters that are uniformly distributed between 20.25 and 30.25 mm (i.e. with mean diameter 25.25 mm) it is easy to see that once again 2.5 per cent would have diameters exceeding 30 mm (since any interval of 0.25 mm between 20.25 and 30.25 mm contains a proportion  $1/40$ , i.e. 2.5 per cent, of all production). This uniform distribution situation is unlikely to be met in practice but this example shows that we may have the same proportion of defectives in two populations, yet each has a different mean and their distributions do not even belong to the same family. We consider a more realistic situation involving different means in [Exercise 1.1](#).

Clearly then the numbers of defectives alone cannot tell us whether the mean diameters of items produced by each machine are the same. However, if we also assume that the distributions of diameters for items from each machine differ, if at all, only in their means, then if we know the proportion of defectives in samples of, say, 200 from each machine, we could test whether the means can reasonably be supposed to be identical. The test would not be efficient. It would be better to measure the diameter of each item in smaller samples, and dependent upon the distributional assumptions one makes, use an appropriate parametric or nonparametric test. We say more about hypothesis tests in [Section 1.3](#).

Means and medians are widely used to indicate where distributions are centred. Both are often referred to broadly as ‘averages’. Because many people think of an average as the arithmetic mean we shall often be more formal and call them **measures of centrality**. Another name is **measures of location** but we use this only in a special context relating to what are called *location* and *scale* parameters. Tests and estimation procedures for measurement data are often about averages or centrality measures, e.g.:

- Is it reasonable to suppose that a sample come from a population with a pre-specified mean or median?
- Do two samples come from populations whose means differ by at least 10?
- Given a sample, what is an appropriate estimate of the population mean or median? How good is that estimate?

Increasingly, for example, in industrial quality control, interest is turning also to **spread** or **dispersion**, often measured by variance or standard deviation. Buyers of a new car or computer want not only a good average performance but also consistency. They expect their product to perform as well as those of other purchasers of that model. Success of a product often depends upon personal recommendations, so mixed endorsements – some glowing, others warning of niggling faults – are not good publicity. There are parametric and nonparametric methods for assessing spread or variability.

Some nonparametric techniques require little information. We may test if it is reasonable to assume that weights of items have a prespecified median 2 mg, say, if all we know is how many items in a sample of  $n$  weigh more than 2 mg. If it were difficult, expensive, or impossible to get exact weights, but easy to determine numbers above (or below) 2 mg this nonparametric approach may be cost effective. Simple nonparametric methods are also useful if data are in some way incomplete, like those in Example 1.2.

### *Example 1.2*

In medical studies the progress of patients is often monitored for a limited time after treatment; often anything from a few months to 5 or 6 years. Dinse (1982) gives data for survival times in weeks for 10 patients with symptomatic lymphocytic non-Hodgkin’s lymphoma. The precise survival time is not known for one patient who was alive after 362 weeks. The observation for that patient is said to be *censored*. Survival times in weeks were

49, 58, 75, 110, 112, 132, 151, 276, 281, 362\*

The asterisk denotes a censored observation.

Is it reasonable to suppose that these data are consistent with a median survival time of 200 weeks? Censored observations cause problems in many parametric tests but in [Section 1.3.2](#) we use a simple nonparametric test to show there is no strong evidence against the hypothesis that the median is 200. For that interpretation to be meaningful and useful we have to assume the data are a random sample from some population of patients with the disease.

To confirm that the median might well be 200 is not in itself very helpful. It would be more useful if we could say that the data imply that it is reasonable to assert that the median survival time is between 80 and 275 weeks or something of that sort. This is what confidence intervals ([Section 1.4.1](#)) are about. Dinse was interested, among other things, in whether the median survival times differed between symptomatic and asymptomatic cases; he used this sample and another for 28 asymptomatic cases to compare the survival time distributions in more detail. In this other sample 12 of the 28 observations were censored at values of 300 or more. We show in [Example 5.9](#) that, on the basis of his data, we would conclude that there is strong evidence that the medians for symptomatic and for asymptomatic cases are different. These data were also considered by Kimber (1990).

### 1.1.3 Historical note

The first chapter of the Book of Daniel records that on the orders of Nebuchadnezzar certain favoured children of Israel were to be specially fed on the king's meat and wine for 3 years. Reluctant to defile himself with such luxuries, Daniel pleaded that he and three of his brethren be fed instead on pulse for 10 days. After that time the four were declared 'fairer and fatter in flesh than all of the children which did eat the portion of the king's meat'. This evidence was taken on commonsense grounds to prove the superiority of a diet of pulse. In [Example 1.4](#) and throughout the book we illustrate how we test evidence like this more formally to justify the commonsense conclusion arrived at in the Old Testament. Although the biblical analysis is informal it contains the germ of a nonparametric, as opposed to a parametric, test.

John Arbuthnot (1710) observed that in each year from 1629 to 1710 the number of males christened in London exceeded the number of females. He regarded this as strong evidence against the probability of a male birth being  $\frac{1}{2}$ . The situation is somewhat akin to observing 82 heads in 82 consecutive tosses of a coin.

Francis Galton (1892) developed a measure – which he termed a ‘centesimal scale’ – to assess agreement between patterns (categorical data) on corresponding fingertips on left and right hands.

Karl Pearson (1900) proposed the well-known, and sometimes misused, chi-squared goodness-of-fit test applicable to any discrete distribution, and C. Spearman (1904) defined a rank correlation coefficient (see [Section 7.1.3](#)) that bears his name.

Systematic study of nonparametric inference dates from the 1930s when attempts were made to show that even if an assumption of normality stretched credulity, then at least in some cases making it would not greatly alter conclusions. This stimulated work by R.A. Fisher, E.J.G. Pitman and B.L. Welch on **randomization or permutation** tests which were then too time consuming for general use, a problem now overcome with appropriate statistical software.

About the same time it was realised that observations consisting simply of preferences or ranks could be used to make some inferences without too much computational effort. A few years later F. Wilcoxon and others showed that even if we have precise measurements, we sometimes lose little useful information by ranking them in increasing order of magnitude and basing analyses on these ranks. Indeed, when assumptions of normality are not justified, analyses based on ranks or on some transformation of them may be the most efficient available.

Nonparametric methods then became practical tools either when data were by nature ordinal (ranks or preferences) or as reasonably efficient methods that reduced computation even when measurements were available providing those measurements could be replaced by ranks. While hypothesis testing was usually easy, unfortunately the more important interval estimation ([Section 1.4](#)) was not, a difficulty that has now been largely overcome

In parallel with the above advances, techniques relevant to counts were developed. Counts often represent the numbers of items in categories which may be either **ordered**, e.g. examination grades, or **nominal** (i.e. unordered), e.g. in psychiatry characteristics like depression, anxiety, psychosis, etc.

Many advanced and flexible nonparametric methods are tedious only because they require repeated performance of simple but monotonous calculations, something computers do well.

The dramatic post-war development of feasible, but until recently often tedious to carry out, nonparametric procedures is described by Noether (1984). Kendall and Gibbons (1990) give more than 350 references for the one topic of rank correlation.

Computers have revolutionized our approach to data analysis and statistical inference. Hope, often ill-founded, that data would fit a restricted mathematical model with few parameters and emphasis on simplifying concepts such as linearity has been replaced by the use of robust methods ([Chapter 11](#)) and by exploratory data analyses to investigate different potential models. These are areas where nonparametric methods sometimes have a central role. Generalized linear models described by McCullagh and Nelder (1989) at the theoretical level and by Dobson (1990) at the practical level often blend parametric and nonparametric approaches.

Nonparametric methods are in no sense preferred methods of analysis for all situations. A strength is their applicability where there is insufficient theory or data to justify, or to test compatibility with, specific distributional models. At a more sophisticated level they are also useful, for example, in finding or estimating trends in large data sets that are otherwise difficult to detect due to the presence of disturbances usually referred to as ‘noise’.

Recent important practical developments have been in computer software ([Section 1.6](#)) to carry out permutation and other nonparametric tests. Results for these may be compared with those given by asymptotic theory, which, in the past, was often used where its validity was dubious.

Modern computing power has also led to increasing use of the bootstrap stemming from work by Efron and others since the early 1980s. Bootstrapping is a computer-intensive technique with both parametric and nonparametric versions. We give an introduction to this and related topics in [Section 11.3](#).

## 1.2 SAMPLES AND POPULATIONS

When making statistical inferences we often assume that observations are a random sample from some population. Specification of that population may or may not be precise. If we select 20 books at random (i.e. so that each book in the population has the same probability of inclusion in the sample) from 100 000 volumes in a library, and record the number of pages in each book in the sample, then inferences made from the sample about the mean or median number of pages per book apply strictly only to the population of books in that library. If the library covers a wide range of fiction, non-fiction and reference works it is reasonable to assume that any inferences apply, at least approximately, to a wider population of

books, perhaps to all books published in the United Kingdom or the United States or wherever the library is situated. If the books in the library are all English language books, inferences may not apply to books in Chinese or in Russian.

We seldom deal with strictly random samples from a so clearly specified population. More commonly, data form samples that may be expected to have the essential properties of a random sample from a vaguely-specified population. For example, if a new diet is tested on pigs and we measure weight gains for 20 pigs at one agricultural experimental station we might assume that these are something like a random sample from all pigs of that or similar breeds raised *under such conditions*. This qualification is important. Inferences might apply widely only if the experimental station adopted common farming practices and if responses were fairly uniform for many breeds of pig. This may not be so if the experimental station chose an unusual breed, adopted different husbandry practices from those used on most pig farms, or if the 20 pigs used in the experiment were treated more favourably than is usual in other respects, e.g. if they were kept in specially heated units during the experiment.

The abstract notion of random sampling from an infinite population (implicit in most inference based upon normal distribution theory) often works well in practice, but is never completely true! At the other extreme there are situations where the sample is essentially the whole population. For example, at the early stages of testing a new drug for treating a rare disease there may be just, say, nine patients available for test and only four doses of the drug. One might choose at random the four patients from nine to receive the new drug, the remaining five are untreated or may be treated with a drug already in use. Because of the random selection, if the drug has no effect, or is no better than one currently in use, it is unlikely that a later examination would show that the four patients receiving the new drug had responded better than any of the others. This is possible, but it has a low probability, which we can calculate on the assumption that the new drug is ineffective or is no better than the old one. Clearly, if the drug is beneficial or better than that currently in use the probability of better responses among those treated with it is increased (see [Example 1.4](#), [Section 1.3.3](#)).

Because of the many ways data may be obtained we must consider carefully the validity and scope of inferences. For example, the ten patients for whom survival times were measured in [Example 1.2](#) came from a study conducted by the Eastern Co-operative

Oncology Group in the USA and represented all patients afflicted with symptomatic lymphocytic non-Hodgkin's lymphoma available for that study. In making inferences about the median or other characteristics of the survival time distribution, it is reasonable to assume these inferences are valid for all patients receiving similar treatment and alike in other relevant characteristics, e.g. with a similar age distribution. The patients in this study were all male, so clearly it would be unwise to infer, without further evidence, that the survival times for females would have the same distribution.

Fortunately the same nonparametric procedures are often valid whether samples are from an infinite population, a finite population or when the sample is the entire relevant population. What is different is how far the inferences can be generalized. The implications of generalizing inferences is described for several specific tests by Lehmann (1975, Chapters 1–4) and is discussed in more detail than in this book in Sprent (1998, Chapter 4).

In [Section 1.1.1](#) we referred to a set of  $n$  independent observations from a normal distribution. Independence implies that knowing one value tells us nothing about other values. For any distribution a sequence of  $n$  independent observations all having the same distribution is a **random sample** from that distribution. Independence is an important, and by no means trivial, requirement for the validity of many statistical inferences.

## 1.3 HYPOTHESIS TESTS

### 1.3.1 Basic concepts

Estimation ([Section 1.4](#)) is often a key aim of statistical analysis. It is easy to explain in terms of testing a range of hypotheses, so we need to understand testing even though it is a technique that tends to be both overused and misused.

We assume familiarity with simple parametric hypothesis tests such as the  $t$ -test and chi-squared test, but we review here some fundamentals and changes in interpretation made possible by modern computer software. Until such software became widely available hypothesis testing was often based on tables.

The familiar  $t$ -test is often used to test hypotheses about the unknown mean  $\mu$  of a normal distribution. We specify a **null hypothesis**,  $H_0$ , that  $\mu$  takes a specific value  $\mu_0$  and an alternative hypothesis,  $H_1$ , is that it takes some other value. Formally this is stated as:

$$\text{Test } H_0: \mu = \mu_0 \text{ against } H_1: \mu \neq \mu_0 \quad (1.1)$$

The test is based on a **statistic**,  $t$ , that is a function of the sample values calculated by a formula given in most statistics textbooks. The classic procedure using tables was to compare the magnitude, without regard to sign, of the calculated  $t$ , often written  $|t|$ , with a value  $t_\alpha$  given in tables, the latter chosen so that when  $H_0$  was true

$$\Pr(|t| \geq t_\alpha) = \alpha. \quad (1.2)$$

In practice  $\alpha$  nearly always took one of the values 0.05, 0.01 or 0.001. These probabilities are often expressed as equivalent percentages, i.e. 5, 1 or 0.1 and are widely known as **significance levels**. Use of these particular levels was dictated at least in part by available tables. In this traditional approach if one obtained a value of  $|t| \geq t_\alpha$  the result was said to be **significant** at probability level  $\alpha$  or at the corresponding  $100\alpha$  per cent level, and these levels were often referred to respectively as ‘significant’, ‘highly significant’ and ‘very highly significant’. If significance at a particular level was attained one spoke of rejecting the hypothesis  $H_0$  at that level. If significance was not attained the result was described as **not significant** and  $H_0$  was said to be **accepted**. This is unfortunate terminology giving – especially to nonstatisticians – the misleading impression that nonsignificance implies that  $H_0$  is true, while significance implies it is false.

The rationale behind the test (1.1) is that if  $H_0$  is true then values of  $t$  near zero are more likely than large values of  $t$ , either positive or negative. Large values of  $|t|$  are more likely to occur under  $H_1$  than under  $H_0$ . It follows from (1.2) that if we perform a large number of such tests on different independent random samples when  $H_0$  is true we shall in the long run incorrectly reject  $H_0$  in a proportion  $\alpha$  of these; e.g. if  $\alpha = 0.05$  we would reject the null hypothesis when it were true in the long run in 1 in 20 tests, i.e. in 5 per cent of all tests.

The traditional approach is still common, especially in certain areas of law, medicine and commerce, or to conform with misguided policy requirements of some scientific and professional journals.

Modern statistical software lets us do something more sensible though by itself still far from satisfactory. The output from any good modern software for a  $t$ -test relevant to (1.1) gives the **exact** probability of obtaining, when  $H_0$  is true, a value of  $|t|$  equal to or greater than that observed. In statistical jargon this probability is usually called a  $P$ -value and it is a measure of the strength of evidence against  $H_0$  provided by the data – the smaller  $P$  is, the stronger is that evidence.

Using such evidence one formal possibility when it is strong is to decide  $H_0$  is implausible. When we decide what  $P$ -values are sufficiently small for  $H_0$  to be implausible we may speak formally of rejecting  $H_0$  at the **exact**  $100P$  per cent significance level. This avoids the difficulty that rigid application of the 5 per cent significance level leads to the unsatisfactory situation of  $H_0$  being rejected for a  $P$ -value of 0.049 but accepted for the slightly larger  $P$ -value of 0.051. It does not, however, overcome a more serious objection to this approach, namely, that with small experiments one may never observe small  $P$ -values even when  $H_0$  is **not** true. Suppose a coin is tossed 5 times to test the hypothesis  $H_0$ : *the coin is fair* (equally likely to fall heads or tails) against  $H_1$ : *the coin is biased* (either more likely to fall heads or more likely to fall tails). Clearly in 5 tosses the strongest evidence against  $H_0$  is associated with the outcomes 5 heads or 5 tails. Under  $H_0$  the probability ( $P$ ) of getting one of these outcomes is given by the sum of the probabilities of  $r = 0$  or  $r = 5$  ‘heads’ for a binomial  $B(5, \frac{1}{2})$  distribution, so  $P = 2 \times (\frac{1}{2})^5 = 0.0625$ . This is the smallest attainable  $P$ -value when  $n = 5$  and  $p = \frac{1}{2}$  and so we never reject  $H_0$  at a conventional  $P = 0.05$  level whatever the outcome of the 5 tosses – even if the coin is a double-header! That the experiment is too small is the only useful information given by the  $P$ -value in this example. The situation is very different if we increase the experiment to 20 tosses and get 20 heads or 20 tails. This weakness of hypothesis testing together with the perpetration of myths such as equating accepting a hypothesis to proof it is true has led to justified criticism of what is sometimes called the  $P$ -value culture. Krantz (1999) and Nelder (1999) both highlight dangers arising from inappropriate use of and misunderstandings about what a  $P$ -value implies. We draw attention also to an ethical danger near the end of [Section 1.5](#).

In the real world policy decisions are often based on the findings of statistical analyses; if the evidence against  $H_0$  is strong a formal rejection of  $H_0$  on the ground that it is implausible may be appropriate. To appreciate the implications of either a formal rejection of  $H_0$  or of a decision not to reject it at a given significance level we need some additional concepts. If we decide to reject  $H_0$  whenever our  $P$ -value is less than some fixed value  $P_0$ , say, then, if in all cases where we do so  $H_0$  is true, we would in the long run reject it in a proportion  $P_0$  of those cases. Rejection of  $H_0$  when it is true is an **error of the first kind**, or Type I error. What a  $P$ -value therefore tells us is the probability that we are making an error of the first kind by rejecting  $H_0$ . In a  $t$ -test if we make a decision to reject

$H_0$  whenever we observe a  $P \leq P_0$  we do so when  $|t| \geq k$  where  $k$  is such that when  $H_0$  holds  $\Pr(|t| \geq k) = P_0$  and these values of  $t$  define a **critical** or **rejection** region of **size**  $P_0$ . Using a critical region of size  $P_0$  implies we continue to regard  $H_0$  as plausible if  $|t| < k$ . If we follow this rule we shall sometimes (or as we saw above for the 5 coin tosses, in extreme cases, always) continue to regard  $H_0$  as plausible when, in fact,  $H_1$  is true. Continuing to accept  $H_0$  when  $H_1$  is true is an **error of the second kind**, or Type II error. Let  $\beta$  denote the probability of a Type II error. The probability of a Type II error depends in part on the true value of  $\mu$  if indeed it is not equal to  $\mu_0$ . Intuition correctly suggests that the more  $\mu$  differs from  $\mu_0$ , the more likely we are to get large values of  $|t|$ , i.e. values in the critical region, so that  $\beta$  decreases as  $|\mu - \mu_0|$  increases. If we decrease  $P_0$  (say from 0.03 to 0.008) our critical region becomes smaller, so that for a given  $\mu$  we increase  $\beta$  because the set of values of  $t$  for which we accept  $H_0$  is larger. The other factor affecting  $\beta$  is the sample size,  $n$ . If we increase  $n$  we decrease  $\beta$  for a given  $\mu$  and  $P_0$ . Thus  $\beta$  depends on the true value of  $\mu$  (over which we have no control) and the value of  $n$  and of  $P_0$  determining the size of the critical region (for both of which we often have some choice).

Despite obvious limitations  $P$ -values used properly and constructively have a basic role in statistical inference. In [Section 1.4](#) we shall see that a null hypothesis that specifies one single value of a parameter is usually just one of many possible hypotheses that are not contradicted by the sample evidence. Donahue (1999) and Sackrowitz and Samuel-Cahn (1999) discuss various distributional properties of the  $P$ -value that relate indirectly to uses we discuss here and in [Section 1.4](#).

Fixing the probability of an error of the first kind, whether we denote it by the conventional symbol  $\alpha$  or the alternative  $P_0$  does not determine  $\beta$ . We want  $\beta$  to be small because, in the  $t$ -test for example, we want the calculated  $t$ -value to be in the critical region when  $H_0$  is not true. Clearly  $1 - \beta$  is the probability of getting a  $t$ -value in the critical region when  $H_0$  is not true. The quantity  $1 - \beta$  is called the **power** of the test; we want this to be large. For samples from a normal distribution and all choices of  $n$ ,  $P$  and for any  $\mu$ , the  $t$ -test is more powerful than any other test of the form (1.1).

The historical choice of significance levels 5, 1 and 0.1 per cent as the basis for tables was made on the pragmatic grounds that one does not want to make too many errors of the first kind. It clearly would be silly to choose a significance level of 50 per cent, for then we would be equally likely to accept or to reject  $H_0$  when it were

true. Even with conventional significance levels or other small  $P$ -values we may often make errors of the second kind in a  $t$ -test (or any other test) if it has low power for one or more of these reasons:

- $\mu$  is close to the value specified in  $H_0$ , or
- the sample size is small, or
- we specify a very small  $P$ -value for significance, or
- assumptions required for the test to be valid are violated.

How we proceed when a test gives a nonsignificant result depends on our aims and objectives. In later chapters we consider for some tests the often non-trivial problem of determining how big an experiment is needed to ensure reasonable power to achieve given objectives.

Using small  $P$ -values in place of traditional 5, 1 and 0.1 per cent significance levels gives more freedom in weighing evidence for or against a null hypothesis. Remembering that  $P = 0.05$  corresponds to the traditional 5 per cent significance level long used as a reasonable watershed, one should not feel there is strong evidence against a null hypothesis if  $P$  is much greater than 0.05, but values of  $P$  not greatly exceeding 0.05 often point at least to a case for further studies – often to a need for larger experiments. In this book we shall usually discuss the evidence for or against hypotheses in terms of observed  $P$ -values but in some situations where it is appropriate to consider a hypothetical fixed  $P$ -value we use for this the notation  $\alpha$  with an implication that we regard any  $P$ -value less than that  $\alpha$  as sufficient evidence to prefer  $H_1$  to  $H_0$ . This complies with certain long-established conventions.

The test in (1.1) is called a **two-tail** test because the critical region consists both of large positive and large negative values of the statistic  $t$ . More specifically, large positive values of  $t$  usually imply  $\mu > \mu_0$  and large negative values of  $t$  imply  $\mu < \mu_0$ .

Specification of  $H_0$  and  $H_1$  is determined by the logic of a problem. Two other common choices are

$$(i) \quad \text{Test } H_0: \mu = \mu_0 \text{ against } H_1: \mu > \mu_0 \quad (1.3)$$

$$(ii) \quad \text{Test } H_0: \mu \leq \mu_0 \text{ against } H_1: \mu > \mu_0 \quad (1.4)$$

both leading to a **one-tail** (here *right* or *upper-tail*) test, since in each case when the  $t$ -test is relevant large *positive* values of  $t$  favour  $H_1$ , whereas a small positive value or *any* negative value indicates that  $H_0$  is more likely to hold. The modifications to a one-tail test if the inequalities in (1.3) or (1.4) are reversed are obvious. The critical region then becomes the left, or lower, tail.

For example, if the amount of a specified impurity in 1000g ingots of zinc produced by a standard process is normally distributed with a mean of 1.75g and it is hoped that a steam treatment will remove some of this impurity we might steam-treat a sample of 15 ingots and determine the amount of impurity left in each ingot. If the steam is free from the impurity the treatment cannot increase the level and either it is ineffective or it reduces the impurity. It is therefore appropriate to test

$$H_0: \mu = 1.75 \text{ against } H_1: \mu < 1.75.$$

If ingots have an unknown mean impurity level, but a batch is acceptable only if  $\mu \leq 1.75$ , the appropriate test is

$$H_0: \mu \leq 1.75 \text{ against } H_1: \mu > 1.75.$$

For the  $t$ -test some computer packages give a  $P$ -value appropriate for a one-tail test, e.g.  $\Pr(t \geq t_P) = P$ . Because the distribution of  $t$  is symmetric one doubles this probability to obtain  $P$  for a two-tail test. The doubling of one-tail probabilities to give the corresponding two-tail test probability or significance level applies in other parametric tests such as the  $F$ -test for equality of variance based on samples from two normal populations, but in these cases the two subregions are not symmetric about the mean. However, in many applications where relevant statistics have a chi-squared or  $F$ -distribution a one-tail (upper-tail) test is appropriate.

A common misconception is that a low  $P$ -value indicates a departure from the null hypothesis that is of practical importance. We show why this is not necessarily true in [Section 1.4.2](#). Knowing exact tail probabilities is, however, useful when comparing the performance of different tests.

### 1.3.2 Some nonparametric test statistics

Some statistics used for nonparametric tests have, at least approximately, familiar continuous distributions such as the normal,  $t$ ,  $F$  or chi-squared distribution. However, we shall use many statistics that have discontinuous distributions and this raises a further problem that has a practical impact mainly with small samples.

*Example 1.2 (continued)*

In this example we want to test the hypothesis that the median  $\theta$  of survival times for the population is 200 against the alternative of some other value,

$$\text{i.e. to test } H_0: \theta = 200 \text{ against } H_1: \theta \neq 200 \quad (1.5)$$

A simple test needs only a count of the number of sample values exceeding 200 (recording each as a ‘plus’). By the definition of a random sample and that of a population median, if we have a random sample from **any** distribution with median 200 each sample value is equally likely to be above or below 200. This means that under  $H_0$  the number of plus signs has a binomial  $B(10, \frac{1}{2})$  distribution. The probability of observing  $r$  plus signs in 10 observations when  $p = \frac{1}{2}$  is given by the binomial formula

$$p_r = \Pr(X = r) = \binom{10}{r} (\frac{1}{2})^{10}$$

where the binomial coefficient  $\binom{10}{r} = 10!/[r!(10-r)!]$  and  $r!$  is the product of all integers between 1 and  $r$  (called *factorial r*) and  $0! = 1$ .

The values of these probabilities,  $p_r$ , for each value of  $r$  between 0 and 10, correct to 4 decimal places, are

$R$	0	1	2	3	4	5	6	7	8	9	10
$P$	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010

In the data 3 observations exceed 200 so there are 3 plus signs and from the table above we see that the probability of 3 or less plus signs in a sample of 10 is  $0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719$  when  $H_0$  is true. There is clearly no strong evidence against  $H_0$ , since the probability of getting 3 or less plus signs or 7 or more plus signs (the two relevant ‘tail’ probabilities for our observed statistic, the number of plus signs) is  $2 \times 0.1719 = 0.3438$ , implying that departures from the expected number of plus signs, 5, as large or larger than that observed will occur in slightly more than one-third of all samples when  $H_0$  is true. This simple test, called the **sign test**, is discussed more fully in [Section 2.3](#).

In  $t$ -tests all values of  $P$  between 0 and 1 are possible, but for the sign test only certain discrete  $P$ -values occur. In this example for a two-tail test the three smallest are  $P = 2 \times (0.0010) = 0.0020$  (rounded to 4 decimal places) corresponding to 0 or 10 plus; and  $P = 2 \times (0.0010 + 0.0098) = 0.0216$  corresponding to 1 or 9 plus; then  $P = 2 \times (0.0010 + 0.0098 + 0.0439) = 0.1094$  corresponding to 2 or 8 plus; next comes the observed  $P = 0.3438$ . For a one-tail test these  $P$ -values are all halved. Our statistic – the number of plus signs – has a discrete distribution. This means there is no direct way of obtaining a critical region of exact size 0.05 for a two-tail test; we must choose between regions of size 0.0216 or 0.1094.

A device called a randomization procedure has been proposed with the property that in the long run an error of the first kind occurs in repeated testing with a probability at some prechosen nominal level, e.g. at 5 per cent, rather than at an exact level in each case. In practice our prime interest is what happens in our one test, so it is better, when we know them, to use exact levels, rather than worry about nominal arbitrary levels. There is, however, when there are discontinuities, a case for forming a tail probability by allocating

only one half of the probability that the statistic equals the observed value to the ‘tail’ when determining the size of the ‘critical’ region. This approach has many advocates. We do not use it in this book, but if it is used this should be done consistently.

Most modern computer programs calculate exact  $P$ -values for many parametric and for some nonparametric tests, but pragmatism means we may deplore, but cannot completely ignore, the established concept of nominal levels like 5 or 1 per cent entirely, for if we have no program that gives an exact  $P$ -value we must use tables where information for many tests (e.g. the Wilcoxon–Mann–Whitney test described in [Section 5.2](#)) is sometimes only given for such nominal levels, thus forcing users to quote these nominal levels rather than the exact  $P$ -value.

Care is needed in interpreting  $P$ -values especially in one-tail tests, because most computer programs for nonparametric tests quote the probability that a value greater than or equal to the test statistic will be attained if this probability is less than  $1/2$ , otherwise they give the probability that a value less than or equal to the test statistic is obtained. This is the probability of errors of the first kind in a one-tail test if we decide to reject at a significance level equal to that probability. In practice the evidence against  $H_0$  is only rated strong if this ‘tail’ probability is sufficiently small **and is in the appropriate tail!** In general, we recommend doubling a one-tail probability to obtain the actual significance level for a two-tail test, but see [Example 1.3](#) and the remarks following it. If the test statistic has a symmetric distribution doubling is equivalent to considering equal deviations from the mean value of the statistic. If the statistic does not have a symmetric distribution, taking tails equidistant from the mean is not equivalent to doubling a one-tail probability.

[Example 1.3](#) shows another difficulty that sometimes arises due to discontinuities in  $P$ -values; namely, that if we only regard the evidence against  $H_0$  as strong enough to reject it if  $P \leq 0.05$  (or at any rate a value not very much greater than this), we may never get that evidence because no outcome provides it, a problem we have already alluded to with small experiments.

### *Example 1.3*

In a dental practice, experience has shown that the proportion of adult patients requiring treatment following a routine inspection is  $3/4$ , so the number of individuals requiring treatment,  $S$ , in a sample of 10 independent patients has a binomial  $B(10, 3/4)$  distribution. Here the probabilities for the various values  $r$  of the statistic  $S$ , where  $r$  takes integral values between 0 and 10 are given by

$$p_r = \Pr(S = r) = \binom{10}{r} \left(\frac{3}{4}\right)^r \left(\frac{1}{4}\right)^{10-r}$$

The relevant probabilities are

$r$	0	1	2	3	4	5	6	7	8	9	10
$p_r$	0.0000	0.0000	0.0004	0.0031	0.0162	0.0584	0.1460	0.2503	0.2816	0.1877	0.0563

Thus, if we had data for another practice and wanted for that practice to test  $H_0: p = \frac{3}{4}$  against  $H_1: p > \frac{3}{4}$ , the smallest  $P$ -value for testing is in the upper tail and is associated with  $r = 10$ . i.e.  $P = 0.0563$ , so if we only regard  $P \leq 0.05$  as sufficiently strong evidence to discredit  $H_0$  such values are never obtained. There is no problem here for a one-tail test of  $H_0: p = \frac{3}{4}$  against  $H_1: p < \frac{3}{4}$  since in the appropriate lower tail  $P = \Pr(S \leq 4) = 0.0162 + 0.0031 + 0.0004 = 0.0197$ . This example also shows a logical difficulty associated with the rule that the appropriate level for a two-tail test is twice that for a one-tail test, for if we get  $S = 4$  the two-tail test level based on this rule is  $2 \times 0.0197 = 0.0394$ . This presents a dilemma, for there is no observable upper tail area corresponding to that in the lower tail. This means that if a two-tail test is appropriate we shall in fact only be likely to detect departures from the null hypothesis if they are in one direction; there may well be a departure in the other direction, but if so we are highly unlikely to detect it at the conventional level  $P \leq 0.05$ , and even if we did it would be for the wrong reason (*see Exercise 1.10*). This is not surprising when, as shown above, the appropriate one-tail test must fail to detect it, for generally a one-tail test at a given significance level is more powerful for detecting departures in the appropriate direction than is a two-tail test at the same level. The implication is that we need a larger sample to detect departures of the form  $H_1: p > \frac{3}{4}$  in this example. Again the fairly large  $P$ -value associated with the result most likely under this  $H_1$  only tells us our sample is too small.

The stipulation that the patients be independent is important. If the sample included three members of the same family it is quite likely that if one of them were more (or less) likely to require treatment than the norm, this may also be the case for other members of that family.

Statisticians do not all agree that one should double a one-tail probability to get the appropriate two-tail significance level [*see* for example, Yates (1984) and the discussion thereon]. An alternative to doubling the one-tail probability is that once the exact size of a one-tail region has been determined, we should, for a two-tail test, add the probabilities associated with an opposite tail situated equidistant from the mean value of the test statistic to that associated with our observed statistic value. In the symmetric case we have already pointed out that this is equivalent to doubling the probability, but it seems inappropriate with a non-symmetric distribution. In [Example 1.3](#) the region  $r \leq 4$  is appropriate for a lower-tail test. The mean of the test statistic (the binomial mean  $np$ ) is 7.5. Since  $7.5 - 4 = 3.5$ , the corresponding deviation above the mean is  $7.5 + 3.5 = 11$ .

Because  $\Pr(r \geq 11) = 0$ , the two-tail test based on equidistance from the mean would have the same exact significance level as the one-tail test.

### 1.3.3 Permutation tests

#### *Example 1.4*

In Section 1.2 we envisaged a situation where four from nine patients are selected at random to receive a new drug. After three weeks all nine patients are examined by a skilled consultant who, on the basis of various tests and clinical observations ‘ranks’ the patients’ conditions in order from least severe (rank 1) to most severe (rank 9). If there is no beneficial effect of the new drug, what is the probability that the patients who received the new drug are ranked 1, 2, 3, 4?

Selecting four patients ‘at random’ means that any four are equally likely to be given the new drug. If there really is no effect one would expect some of those chosen to end up with low ranks, some with moderate or high ranks, in the post-treatment assessment. From a group of nine patients there are 126 ways of selecting a set of four. This may be verified by using the well-known mathematical result that the number of ways of selecting  $r$  objects from  $n$  is  $n!/[r!(n-r)!]$ . We give all 126 selections in Table 1.1. Ignore for the moment the numbers in brackets after each selection.

If the new drug were ineffective the set of ranks associated with the four patients receiving it are equally likely to be any of the 126 quadruplets listed in

**Table 1.1** Possible selections of four individuals from nine labelled 1 to 9 with the sum of the ranks in brackets.

1,2,3,4 (10)	1,2,3,5 (11)	1,2,3,6 (12)	1,2,3,7 (13)	1,2,3,8 (14)	1,2,3,9 (15)	1,2,4,5 (12)
1,2,4,6 (13)	1,2,4,7 (14)	1,2,4,8 (15)	1,2,4,9 (16)	1,2,5,6 (14)	1,2,5,7 (15)	1,2,5,8 (16)
1,2,5,9 (17)	1,2,6,7 (16)	1,2,6,8 (17)	1,2,6,9 (18)	1,2,7,8 (18)	1,2,7,9 (19)	1,2,8,9 (20)
1,3,4,5 (13)	1,3,4,6 (14)	1,3,4,7 (15)	1,3,4,8 (16)	1,3,4,9 (17)	1,3,5,6 (15)	1,3,5,7 (16)
1,3,5,8 (17)	1,3,5,9 (18)	1,3,6,7 (17)	1,3,6,8 (18)	1,3,6,9 (19)	1,3,7,8 (19)	1,3,7,9 (20)
1,3,8,9 (21)	1,4,5,6 (16)	1,4,5,7 (17)	1,4,5,8 (18)	1,4,5,9 (19)	1,4,6,7 (18)	1,4,6,8 (19)
1,4,6,9 (20)	1,4,7,8 (20)	1,4,7,9 (21)	1,4,8,9 (22)	1,5,6,7 (19)	1,5,6,8 (20)	1,5,6,9 (21)
1,5,7,8 (21)	1,5,7,9 (22)	1,5,8,9 (23)	1,6,7,8 (22)	1,6,7,9 (23)	1,6,8,9 (24)	1,7,8,9 (25)
2,3,4,5 (14)	2,3,4,6 (15)	2,3,4,7 (16)	2,3,4,8 (17)	2,3,4,9 (18)	2,3,5,6 (16)	2,3,5,7 (17)
2,3,5,8 (18)	2,3,5,9 (19)	2,3,6,7 (18)	2,3,6,8 (19)	2,3,6,9 (20)	2,3,7,8 (20)	2,3,7,9 (21)
2,3,8,9 (22)	2,4,5,6 (17)	2,4,5,7 (18)	2,4,5,8 (19)	2,4,5,9 (20)	2,4,6,7 (19)	2,4,6,8 (20)
2,4,6,9 (21)	2,4,7,8 (21)	2,4,7,9 (22)	2,4,8,9 (23)	2,5,6,7 (20)	2,5,6,8 (21)	2,5,6,9 (22)
2,5,7,8 (22)	2,5,7,9 (23)	2,5,8,9 (24)	2,6,7,8 (23)	2,6,7,9 (24)	2,6,8,9 (25)	2,7,8,9 (26)
3,4,5,6 (18)	3,4,5,7 (19)	3,4,5,8 (20)	3,4,5,9 (21)	3,4,6,7 (20)	3,4,6,8 (21)	3,4,6,9 (22)
3,4,7,8 (22)	3,4,7,9 (23)	3,4,8,9 (24)	3,5,6,7 (21)	3,5,6,8 (22)	3,5,6,9 (23)	3,5,7,8 (23)
3,5,7,9 (24)	3,5,8,9 (25)	3,6,7,8 (24)	3,6,7,9 (25)	3,6,8,9 (26)	3,7,8,9 (27)	4,5,6,7 (22)
4,5,6,8 (23)	4,5,6,9 (24)	4,5,7,8 (24)	4,5,7,9 (25)	4,5,8,9 (26)	4,6,7,8 (25)	4,6,7,9 (26)
4,6,8,9 (27)	4,7,8,9 (28)	5,6,7,8 (26)	5,6,7,9 (27)	5,6,8,9 (28)	5,7,8,9 (29)	6,7,8,9 (30)

**Table 1.2** Sums of ranks of four items from nine.

Sum of ranks	10	11	12	13	14	15	16	17	18	19	20
Number of occurrences	1	1	2	3	5	6	8	9	11	11	12
Sum of ranks	21	22	23	24	25	26	27	28	29	30	
Number of occurrences	11	11	9	8	6	5	3	2	1	1	

**Table 1.1.** Thus, if there is no treatment effect there is only 1 chance in 126 that the four showing greatest improvement (ranked 1, 2, 3, 4 in order of condition after treatment) are the four patients allocated to the new drug. It is more plausible that such an outcome may reflect a beneficial effect of the drug.

In a hypothesis testing framework we have a group of 4 treated with the new drug and a group of 5 (the remainder) given nothing or a standard treatment in what is called a *two independent sample* experiment. We discuss such experiments in detail in Chapter 5. The most favourable evidence for the new drug would be that those receiving it are ranked 1, 2, 3, 4; the least favourable that they are ranked 6, 7, 8, 9. Each of these extremes has a probability of 1/126 of occurring when there is no real effect.

If we consider a test of

$$H_0: \text{new drug has no effect}$$

against the two-sided alternative

$$H_1: \text{new drug has an effect (beneficial or deleterious)}$$

the outcome 1, 2, 3, 4 and 6, 7, 8, 9 are extremes with a total associated probability  $P = 2/126 \approx 0.0159$  if  $H_0$  is true. In classic testing terms we might speak of rejecting  $H_0$  at an exact 1.59 per cent significance level if we observed either of these outcomes. This small  $P$ -value provides strong evidence that the new drug has an effect. What if the patients receiving the new drug were ranked 1, 2, 3, 5? Intuitively this evidence looks to favour the new drug, but how do we test this?

We seek a statistic, i.e. some function of the four ranks, that has a low value if all ranks are low, a high value if all ranks are high and an intermediate value if there is a mix of ranks for those receiving the new drug. An intuitively reasonable choice is the sum of the four ranks. If we sum the ranks for every quadruplet in **Table 1.1**, and count how many times each sum occurs we easily work out the probability of getting any particular sum, and hence the distribution of our test statistic when  $H_0$  is true.

In **Table 1.1** the number in brackets after each quadruplet is the sum of the ranks for that quadruplet, e.g. for 1, 2, 7, 9 the sum is  $1 + 2 + 7 + 9 = 19$ . The lowest sum is 10 for 1, 2, 3, 4 and the highest is 30 for 6, 7, 8, 9. **Table 1.2** gives the numbers of quadruplets having each given sum.

Because there are 126 different but equally likely sets of ranks the probability that the rank sum statistic, which we denote by  $S$ , takes a particular value is obtained by dividing the number of times that value occurs by 126. Thus, for

example,  $\Pr(S = 17) = 9/126 \approx 0.0714$ . To find what outcomes are consistent with a  $P$ -value not exceeding 0.05 we select a region in *each* tail (since  $H_1$  implies a two-tail test) with a total associated probability not exceeding 0.025. Clearly, if we select in the lower tail  $S = 10$  and 11, the associated probability is 2/126 and if we add  $S = 12$  the associated total probability, i.e.  $\Pr(S \leq 12) = 4/126 \approx 0.0317$ . This exceeds 0.025, so our lower-tail critical region should be  $S \leq 11$  with  $P = 2/126 \approx 0.0159$ . By symmetry, the upper-tail region is  $S \geq 29$  also with  $P = 0.0159$ . Thus, for a two-tail test the largest symmetric critical region with  $P \leq 0.05$  is  $S = 10, 11, 29, 30$  and the exact  $P = 4/126 \approx 0.0317$ .

Some statisticians suggest choosing a critical region with probability as close as possible to a target level such as  $P = 0.05$  rather than the more conservative choice of one no larger. In this example adding  $S = 12$  and the symmetric  $S = 28$  to our critical region gives a two-tail  $P = 8/126 \approx 0.0635$ . This is closer to 0.05 than the size (0.0317) of the region chosen above. We reaffirm that ideally it is best to quote the exact  $P$ -value obtained and point out again that the practical argument (though there are further theoretical ones) for quoting nominal sizes such as 0.05 is that many tables give only these, although a few, e.g. Hollander and Wolfe (1999, Table A6) give relevant exact  $P$ -values for many sample size combinations and different values of  $S$ . Computer programs giving exact  $P$ -values overcome any difficulty if the latter type of table is not readily available.

Unless there were firm evidence before the experiment started that an effect, if any, of the new drug could only be beneficial, a two-tail test is appropriate. We consider a one-tail test scenario in [Exercise 1.5](#).

In [Section 1.2](#) we suggested that in preliminary testing of drugs for treating a rare disease our population may be in a strict sense only the cases we have. However, if these patients are fairly typical of all who might have the disease, it is not unreasonable to assume that findings from our small experiment may hold for any patients with a similar condition *providing* other factors (nursing attention, supplementary treatments, consistency of diagnosis, etc.) are comparable. When our experiment involves what is effectively the whole population and the only data are ranks, a permutation test is the best test available. Random allocation of treatments is essential for the test to be valid; this may not always be possible in the light of some ethical considerations that we discuss in [Section 1.5](#).

Tests based on permutation of ranks or on permutation of certain functions of ranks (including as a special case the original measurements) are central to nonparametric methods and are called **permutation** or **randomization** tests. Not only do many of these tests have a commonsense intuitive appeal but they comply with well established theoretical criteria for sound inference. This theoretical basis is discussed in the literature and is summarized by Hettmansperger and McKean (1998) for many different procedures.

Small scale tests of a drug like that in [Example 1.4](#) are often called *pilot studies*. Efficacy of a drug in wider use may depend on factors like severity of disease, treatment being administered sufficiently early, the age and sex of patients, etc., all or none of

which may be reflected in a small group of ‘available patients’. An encouraging result with the small group may suggest further experiments are desirable; a not very small  $P$ -value associated with what looks to be an intuitively encouraging result may indicate that a larger experiment is needed to tell us anything useful.

### 1.3.4 Pitman efficiency

We pointed out in [Section 1.3.1](#) that the power of a test depends upon the sample size,  $n$ , the choice of the largest  $P$ -value to indicate significance in the classical sense (usually denoted in power studies by  $\alpha$ ), the magnitude of any departure from  $H_0$  and whether assumptions needed for validity hold. Most intuitively reasonable tests have good power to detect a true alternative that is far removed from the null hypothesis providing the data set is large enough. We sometimes want tests to have as much power as possible for detecting alternatives close to  $H_0$  even when these are of no practical importance because such tests are usually also good at detecting larger departures, a desirable state of affairs. If  $\alpha$  is the probability of a Type I error and  $\beta$  is the probability of a Type II error (the power is  $1 - \beta$ ), then the efficiency of a test  $T_2$  relative to a test  $T_1$  is the ratio  $n_1/n_2$  of the sample sizes needed to obtain the same power for the two tests with these values of  $\alpha, \beta$ . In practice, we usually fix  $\alpha$  at some  $P$ -value appropriate to the problem at hand; then  $\beta$  depends on the particular alternative *as well as* the sample sizes. Fresh calculations of relative efficiency are required for each particular value of the parameter or parameters of interest in  $H_1$  and for each choice of  $\alpha, \beta$ .

Pitman (1948), in a series of unpublished lecture notes, introduced the concept of **asymptotic relative efficiency** for comparing two tests. He considered sequences of tests  $T_1, T_2$  in which we fix  $\alpha$  but allow the alternative in  $H_1$  to vary in such a way that  $\beta$  remains constant as the sample size  $n_1$  increases. For each  $n_1$  we determine  $n_2$  such that  $T_2$  has the same  $\beta$  for the particular alternative considered. Increasing sample size usually increases the power for alternatives closer to  $H_0$ , so that for large samples Pitman studied the behaviour of the efficiency,  $n_1/n_2$ , for steadily improving tests for detecting small departures from  $H_0$ . He showed under very general conditions that in these sequences of tests  $n_1/n_2$  tended to a limit as  $n_1 \rightarrow \infty$ , and more importantly, that this limit which he called the asymptotic relative efficiency (ARE) was the same for all choices of  $\alpha, \beta$ . Bahadur (1967) proposed an alternative definition that is less widely

used, so for clarity and brevity we refer to Pitman's concept simply as the **Pitman efficiency**. The concept is useful because when comparing two tests the small sample relative efficiency is often close to, or even higher, than the Pitman efficiency

The Pitman efficiency of the sign test relative to the  $t$ -test when the latter is appropriate is  $2/\pi \approx 0.64$ . Lehmann (1975, p.173) shows that for samples of size 10 and a wide range of values of the median  $\theta$  relative to the value  $\theta_0$  specified in  $H_0$  with  $\alpha$  fixed, the relative efficiency exceeds 0.7, while for samples of 20 it is nearer to, but still above, 0.64. Here Pitman efficiency gives a pessimistic picture of the performance of the sign test at realistic sample sizes.

We have already mentioned that when it is relevant and valid the  $t$ -test for a mean is the most powerful test for any mean specified in  $H_0$  against any alternative. When the  $t$ -test is not appropriate, other tests may have higher efficiency. Indeed, if our sample comes from the double exponential distribution, which has much longer tails than the normal, the Pitman efficiency of the sign test relative to the  $t$ -test is 2. That is, a sign test using a sample of  $n$  is (at least for large samples) as efficient as a  $t$ -test applied to a sample of size  $2n$ .

## 1.4 ESTIMATION

### 1.4.1 Confidence intervals

The sample mean is widely used as a **point** estimate of a population mean. The sample mean varies between samples so we need a measure of the precision of this estimate. A **confidence interval** is one such measure. One way to specify a  $100(1 - \alpha)$  per cent confidence interval for a parameter  $\theta$  is to define it as the set of all values  $\theta$  for which, if any such  $\theta$  were specified in  $H_0$ , for the given data the test would lead to a  $P > \alpha$ . Thus, if a confidence interval includes the value of a parameter that is specified in  $H_0$  there is no strong evidence against  $H_0$ , whereas a value specified in  $H_0$  that lies well outside that confidence interval indicates strong evidence against  $H_0$ . If a  $P$ -value notation is preferred, setting  $P = P_0$  as the critical value for determining significance, one may use the notation  $P_0$  in place of  $\alpha$ .

#### *Example 1.5*

We saw in [Section 1.3.2](#), [Example 1.2](#) that for a sign test for a median using a sample of 10 and a two-tail test at the 2.16 per cent level we accept  $H_0$  if we get between 2 and 8 plus signs.

Consider the data in [Example 1.2](#), i.e.

49, 58, 75, 110, 112, 132, 151, 276, 281, 362\*

where the asterisk represents a censored observation. Clearly we have between 2 and 8 plus signs if the median specified in  $H_0$  has any value greater than 58 but less than 281, so the interval (58, 281) is a  $100(1 - 0.0216) = 97.84$  per cent confidence interval for  $\theta$ , the population median survival time. Since we would retain an  $H_0$  that specified any value for the median greater than 58 but less than 281 there is considerable doubt about the population median value. It is almost an understatement to say our estimate lacks precision.

Another common interpretation of a  $100(1 - \alpha)$  per cent confidence interval is in terms of the property that if we form such intervals for repeated samples, then in the long run  $100(1 - \alpha)$  per cent of these intervals would contain (or cover) the true but unknown  $\theta$ . Confidence intervals are useful because:

- They tell us something about the precision with which we estimate a parameter.
- They help us decide (a) whether a significant result is likely to be of practical importance or (b) whether we need more data before we decide if it is.

We elaborate on these points in [Section 1.4.2](#).

A useful way of looking at the relationship between hypothesis testing and estimation is to regard testing as answering the question:

- **Given** a hypothesis  $H_0: \theta = \theta_0$  about, say, a parameter  $\theta$ , what is the probability ( $P$ -value) of getting a sample as or less likely than that obtained if  $\theta_0$  is indeed the true value of  $\theta$ ?

whereas estimation using a confidence interval answers the question:

- **Given** a sample, what values of  $\theta$  are consistent with the sample data in the sense that they lie in the confidence interval?

## 1.4.2 Precision and significance and practical importance

### *Example 1.6*

Doctors treating hypertension are often interested in the decrease in systolic blood pressure after administering a drug. When testing an expensive new drug they might want to know whether it reduces systolic blood pressure by at least 20 mm Hg. Such a minimum difference could be of **practical importance**. Two clinical trials (I and II) were carried out to test the efficacy of a new drug (A) for reducing blood pressure. A third trial (III) was carried out with a second new drug (B). Trial I involved only a small number of patients, but trials II and III

involved larger numbers. The 95 per cent confidence intervals for mean blood pressure reduction (mm Hg) after treatment at each trial were:

Drug A	Trial I	(3, 35)
Drug A	Trial II	(9, 12)
Drug B	Trial III	(21, 25)

In each trial a hypothesis  $H_0$ : *drug does not reduce blood pressure* would be rejected at a 5 per cent significance level (because the confidence intervals do not include zero), implying strong evidence against  $H_0$ . Trial I is imprecise; we would accept in a significance test at the 5 per cent level any mean reduction between 3 and 35 units. The former is not of practical importance to a doctor; the latter is. This small trial only answers questions about the ‘significant’ mean reduction with **low precision**. The larger Trial II, using the same drug, indicates an average reduction between 9 and 12 units, a result of statistical significance but not of practical importance in this context. Compared to Trial I, it has **high precision**. Other relevant factors being unchanged, increasing the size of a trial increases the precision, this being reflected in shorter confidence intervals. Trial III using drug B also has high precision and tells us the mean reduction is likely to be between 21 and 25 units, a difference of practical importance. Drug B appears to be superior to Drug A.

For a given test, increasing sample size increases the probability that small departures from  $H_0$  may provide strong evidence against  $H_0$ . The art of designing experiments is to take enough observations to ensure a good chance of detecting with reasonable precision departures from  $H_0$  of practical importance, but to avoid wasting resources by taking so many observations that trivial departures from  $H_0$  provide strong evidence against it. An introduction to sample size calculation is given by Kraemer and Thiemann (1987) and it is also discussed with examples by Hollander and Wolfe (1999). Practical design of experiments is best done with guidance from a trained statistician although many statistical software packages include programs giving recommendations in specific circumstances. In later chapters we show for some tests how to find sample sizes needed to meet specified aims.

Our discussion of hypothesis testing and estimation has used the frequentalist approach to inference. The Bayesian school adopt a different philosophy, introducing subjective probabilities to reflect prior beliefs about parameters. Some statisticians are firm adherents of one or other of these schools, but a widely accepted view is that each has strengths and weaknesses and that one or the other may be preferred in certain contexts. However, for the procedures we describe sensible use of either approach will usually lead to similar conclusions despite the different logical foundations, so for consistency we use the frequentalist approach throughout.

## 1.5 ETHICAL ISSUES

Ethical considerations are important both in general working practices (Gillon, 1986) and in the planning and conduct of investigations (Hutton, 1995). The main principles are respect for autonomy, nonmaleficence, beneficence and justice. Many research proposals need to be assessed by ethical committees before being approved. This applies particularly in the field of medicine but increasing attention is being given to ethical issues in environmentally sensitive fields like biotechnology and also in the social sciences where questions of legal rights or civil liberties may arise. The role of statistics and statisticians in what are known as research ethics committees has recently been discussed by Williamson et al. (2000). The related issue of the development of guidelines for the design, execution and reporting of clinical trials is described by Day and Talbot (2000).

It would be unacceptable to study some issues by allocating individuals to one of the possible groups at random as we did in [Example 1.4](#). For instance, in a study of the effects of smoking on health, one could not instruct individuals to smoke or to abstain from smoking as this disregards the autonomy of the study participants. In such a situation, the individual's choice of whether to smoke cigarettes or not must be respected, and an alternative type of study which makes this possible must be chosen.

Many ethical issues are less clear and in practice until recently some have often been ignored. At the early stages of development, a new treatment for an illness may be associated with harmful side-effects and may clash with the principle of nonmaleficence. When the benefits of a new treatment are fairly clear, many doctors reason from the principle of beneficence that it is wrong to deny patients care by giving them the previously standard treatment. If early results from a trial indicate that one treatment is clearly superior, it is unethical to allow the trial to continue.

All new studies should be based on information gleaned from a comprehensive search of findings from related work by using, for instance, MEDLINE, a continually updated source of information on articles from medical and biological journals. It is unethical to conduct research that ignores previous work that may be relevant because it is then likely that any or all of time, money and scarce resources will not be used to best effect. The results of literature searches nevertheless need to be interpreted with caution. Studies with interesting findings are more likely to appear in print, leading to publication bias (*see* Easterbrook et al., 1991).

When there is little relevant prior knowledge it may be prudent to conduct an initial pilot study to highlight potential problems that might arise. Results from a pilot study can also be useful in choosing an appropriate number of participants for the main study. A sufficiently large number should be involved even at the pilot stage to have a reasonable chance of finding the expected difference between the two groups if it really exists. The intended method of statistical analysis also influences the sample size requirement. Small studies often fail to yield useful findings and are thus a poor use of resources. On the other hand, resources can be wasted by recruiting more participants than needed. In medical research, in either situation more patients than necessary are at risk of receiving an inferior treatment. Careful planning should consider the composition of the sample with respect to age, sex, ethnic group, etc. as this will enable problems under investigation to be answered more effectively.

In medical investigations each potential participant should receive a written information sheet that outlines the main points about the study. Ideally, complete information about the possible efficacy and side-effects of the treatments involved in the study should be given to the patient. In practice, not all patients will understand or even wish to receive details beyond those given in the information sheet, particularly in a sophisticated trial. In this situation, the patient should be given a choice about what information is supplied. Once a trial has been completed, patients who feel that they have received an effective treatment for their health problem may wish to continue with it. Financial constraints and/or the concerns of the patient's general practitioner may prevent long-term use of the treatment; this should be discussed in advance as part of the patient information.

The autonomy of the patient should be respected and the patient should only make a decision on whether or not to enter the trial following careful consideration of the information provided. This is particularly important with tests for inherited diseases that only become evident in later life. A positive finding may distress the patient, have serious implications for any children and prejudice life assurance proposals. Patients should give informed written consent to the investigator(s) prior to being entered into a trial and they should be allowed to withdraw from the trial at any time.

It has been suggested, for instance, that in a comparison of two treatments patients who do not wish to be entered into a randomized controlled trial should be allowed instead to choose which of the treatments they wish to receive. Doing this would produce three

groups: the patients who request Treatment A, the patients who request Treatment B and those who are randomized. This must be taken into account in any statistical analysis.

Data collected in studies should be kept confidential. In the United Kingdom, for example, computer records should adhere to the principles laid down in the Data Protection Act. Data used for statistical purposes should not contain patients' names or addresses.

Suppose that in the situation described in [Example 1.4](#), there were high hopes that the new drug might greatly relieve suffering in severe cases but only enough doses were available to treat four patients. The principles of beneficence and justice may suggest that the four patients to receive the drug should be those with the most severe symptoms. In a situation like this, the drug may reduce suffering, but such patients may still, after treatment, be ranked 6, 7, 8, 9 because, although their condition may have improved, their symptoms may still be more severe than those of patients not given the drug. The consequent statistical difficulties might be overcome by basing ranks not on relative condition after treatment but on 'degree of improvement' shown by each patient.

At the other extreme, an experimenter might allocate the new drug to the patients with the least severe symptoms. From a research point of view this is misleading as even if it were ineffective or no better than an existing treatment, these patients may still be ranked 1, 2, 3, 4 after treatment. However, if it is likely that only patients in the early stages of the disease will benefit it will be more appropriate from an ethical viewpoint to give the new drug to these patients.

Even when patients are allocated to treatments at random and we find strong evidence to suggest we should abandon a hypothesis of 'no treatment effect', the statistically significant outcome may be of no practical importance, or there may be ethical reasons for ignoring it. A doctor would be unlikely to feel justified in prescribing the new treatment if it merely prolonged by three days the life expectation of terminally-ill patients suffering considerable distress, but may from the principle of beneficence feel bound to prescribe it if it substantially improved survival prospects and quality of life.

An example of unethical statistical behaviour is that where a statistician performs a number of competing tests – parametric or nonparametric – each producing a different  $P$ -value, but only publishes the  $P$ -value that is most favourable to the conclusion he or she wants to establish, regardless of whether it is obtained by using an appropriate test. This is unethical suppression of evidence.

Ethical considerations are relevant no matter whether parametric

or nonparametric methods of data analysis are used and they may influence not only how an experiment is carried out (the experimental design) but also what inferences are possible and how these should be made.

## 1.6 COMPUTERS AND NONPARAMETRIC METHODS

Software that rapidly computes exact  $P$ -values for permutation tests for small to medium-sized samples and that provides accurate estimates of tail probabilities by simulation for larger samples has revolutionized application of nonparametric methods. The software and hardware situation is changing rapidly. At the time of writing StatXact 4.0 distributed by Cytel Software Corporation, Cambridge, MA gives exact  $P$ -values or Monte Carlo estimates of these for a range of tests. Large sample or **asymptotic** results are also given and there are facilities for computing confidence intervals and also the power of some of the tests for assigned sample sizes and specified alternative hypotheses. Some of the exact tests in StatXact are now available as options in the general statistical packages SAS and SPSS.

Testimate, distributed by IDV Daten-analyse und Versuchsanplanung, Munich, Germany has considerable overlap with StatXact but some methods are included in one but not both these packages.

There are specialized programs dealing with particular aspects of the broad fields of nonparametric and semiparametric inference. These include LogXact and EGRET which are especially relevant to logistic regression, a topic not covered in this book. Popular general statistical packages such as Stata, Minitab and the standard versions of SPSS, SAS, etc., include some nonparametric procedures. In a few of these exact tests are given but many rely heavily on asymptotic results sometimes with insufficient warning about when, particularly with small or unbalanced sample sizes, these may be misleading.

S-PLUS is particularly useful for the bootstrap described in Chapter 11. Monte Carlo approximations to exact  $P$ -values or for bootstrap estimation can often be obtained from standard packages by creating macros that make use of inbuilt facilities for generating many random samples with or without replacement.

Developments in statistical computer software are rapid and much of what we say about this may be out of date by the time you read it. Readers should check advertisements for statistical software in

relevant journals and look for reviews of software in publications such as *The American Statistician* to trace new products.

The efficiency of StatXact programs stems from the use of algorithms based on the work of Mehta and his co-authors in a series of papers including Mehta and Patel (1983; 1986), Mehta, Patel and Tsiantis (1984), Mehta, Patel and Gray (1985), Mehta, Patel and Senchaudhuri (1988, 1998). Similar and other efficient algorithms are used in Testimate, but understanding the algorithms is not needed to use these packages.

Users should test all programs using examples from this book and other sources to ensure that they can interpret the output. In some cases the output will inevitably be different, being either more or less extensive than that given in the source of the examples. For instance, output may give nominal (usually 5 or 1 per cent) significance levels rather than exact  $P$ -values.

Appropriate software is virtually essential for implementation of all but the simplest methods. This book is mainly about long-established methods, but only modern computing facilities allow us to use them in the way we describe

## 1.7 FURTHER READING

Hollander and Wolfe (1999), Conover (1999) and Gibbons and Chakraborti (1992) give more background for many of the procedures described here. Each book covers a slightly different range of topics, but all are recommended for those who want to make a more detailed study of nonparametrics. Daniel (1990) is a general book on applied nonparametric methods. A moderately advanced mathematical treatment of the theory behind basic nonparametric methods is given by Hettmansperger and McKean (1998). Randles and Wolfe (1979) and Maritz (1995) are other recommended books covering the theory at a more advanced mathematical level than that used here. Two classics are the pioneering Bradley (1968) and Lehmann (1975). The latter repays careful reading for those who want to pursue the logic of the subject in more depth without too much mathematical detail. Books on applications in the social sciences include Marascuilo and McSweeney (1977), Leach (1979) and more recently Siegel and Castellan (1988), an update of a book written by Siegel some 30 years earlier. Noether (1991) uses a nonparametric approach to introduce basic general statistical concepts. Although dealing

basically with rank correlation methods, Kendall and Gibbons (1990) give an insight into the relationship between many nonparametric methods. Agresti (1984; 1990; 1996) and Everitt (1992) give detailed accounts of various models, parametric and nonparametric, used in categorical data analysis. A sophisticated treatment of randomization tests with emphasis on biological applications is given by Manly (1997). Good (1994) and Edgington (1995) cover randomization and permutation tests. Sprent (1998) extends the treatment of some of the topics in this book and shows how they interconnect with other aspects of statistical analysis. Books dealing with the bootstrap are Efron and Tibshirani (1993), Davison and Hinkley (1997) and Chernick (1999). Many nonparametric procedures use ranks, and the theory behind rank tests is given by Hájek, Sidák and Sen (1999). There are a number of advanced texts and reports of conference proceedings that are mainly for specialists.

## EXERCISES

- 1.1 As in [Example 1.1](#), suppose that one machine produces rods with diameters normally distributed with mean 27 mm and standard deviation 1.53 mm, so that 2.5 per cent of the rods have diameter 30 mm or more. A second machine is known to produce rods with diameters normally distributed with mean 25 mm and 2.5 per cent of rods it produces have diameter 30 mm or more. What is the standard deviation of rods produced by the second machine?
- 1.2 A library has on its shelves 114 books on statistics. I take a random sample of 12 and want to test the hypothesis that the median number of pages,  $\theta$ , in all 114 books is 225. In the sample of 12, I note that 3 have less than 225 pages. Does this justify retention of the hypothesis that  $\theta = 225$ ? What should I take as the appropriate alternative hypothesis? What is the largest critical region for a test with  $P \leq 0.05$  and what is the corresponding exact  $P$ -level?
- 1.3 The numbers of pages in the sample of 12 books in Exercise 1.2 were:

126 142 156 228 245 246 370 419 433 454 478 503

- Find a confidence interval at a level not less than 95 per cent for the median  $\theta$ .
- 1.4 Use the sum of ranks given in brackets after each group in [Table 1.1](#) to verify the correctness of the entries in [Table 1.2](#).
  - 1.5 Suppose that the new drug under test in [Example 1.4](#) has all the ingredients of a standard drug at present in use and an additional ingredient which has proved to be of use for a related disease, so that it is reasonable to assume that the new drug will do at least as well as the standard one, but may do better. Formulate the hypotheses leading to an appropriate one-tail test. If

- the post-treatment ranking of the patients receiving the new drug is 1, 2, 4, 6 assess the strength of the evidence against the relevant  $H_0$ .
- 1.6 An archaeologist numbers some articles 1 to 11 in the order he discovers them. He selects at random a sample of 3 of them. What is the probability that the sum of the numbers on the items he selects is less than or equal to 8? (You do not need to list **all** combinations of 3 items from 11 to answer this question.) If the archaeologist believed that items belonging to the more recent of two civilizations were more likely to be found earlier in his dig and of his 11 items 3 are identified as belonging to that more recent civilization (but the remaining 8 come from an earlier civilization) does a rank sum of 8 for the 3 matching the more recent civilization provide reasonable support for his theory?
- 1.7 In [Section 1.4.1](#) we associated a confidence interval with a two-tail test. As well as such two-sided confidence intervals, one may define a one-sided confidence interval composed of all parameter values that would not be rejected in a one-tail test. Follow through such an argument to obtain a confidence interval at level not less than 95 per cent based on the sign test criteria for the 12 book sample values given in [Exercise 1.3](#) relevant to a test of  $H_0: \theta = \theta_0$  against a one-sided alternative  $H_1: \theta > \theta_0$ .
- 1.8 We wish to compare a new treatment with a standard treatment and only 6 patients are available. We allocate 3 to each treatment at random and after an appropriate interval rank the patients in order of their condition. What is the situation (i) for testing  $H_0: \text{treatments do not differ}$  against  $H_1: \text{the new treatment is better}$  and (ii) for testing the same  $H_0$  against  $H_1: \text{the two treatments differ in effect}$ ?
- 1.9 In [Section 1.1.1](#) we stated that if all observations in a not-too-small sample are zero or positive and the standard deviation is appreciably greater than the mean, then the sample is almost certainly not one from a normal distribution. Explain why this is so. (Hint: consider the probability that a normally distributed variable takes a value that differs from the mean by more than one standard deviation.)
- 1.10 In [Example 1.3](#) we remarked that a situation could arise where we might reject a null hypothesis for the wrong reason. Explain how this is possible in that example.
- 1.11 State the appropriate null hypothesis for the example from the book of Daniel about diet (p.7). How could you use ranks to calculate the probability that those receiving the diet of pulses were ranked 1, 2, 3, 4? Obtain this probability assuming that there were 20 young men involved altogether.
- 1.12 In [Section 1.3.1](#) we pointed out that 5 tosses of a coin would never provide evidence against the hypothesis that a coin was fair (equally likely to fall heads or tails) at a conventional 5 per cent significance level. What is the least number of tosses needed to provide such evidence using a two-tail test, and what is the exact  $P$ -value? If in fact a coin that is tossed this least number of times is such that  $\Pr(\text{heads}) = \frac{2}{3}$  what is the probability of an error of the second kind? What is the power of the test?

# 2

## Centrality inference for single samples

### 2.1 USING MEASUREMENT DATA

In this book we illustrate the logic of many inference methods using examples where, with a few exceptions, we discuss various aspects under the headings:

*The problem*  
*Formulation and assumptions*  
*Procedure*  
*Conclusion*  
*Comments*  
*Computational aspects*

The summary and exercises at the ends of chapters are preceded by indicative but not exhaustive lists of fields of application.

This chapter covers inferences about centrality measures for a single sample. These provide simple examples for both parametric and nonparametric methods, although in most real problems we compare two or more samples. The methods then used are often easily described as extensions of or developments from those for single samples.

#### 2.1.1 Some general considerations

Choosing the right inference procedure depends both on the type of data and the information we want. In this chapter the initial data are usually measurements but some of the methods are directly applicable to other kinds of data such as ranks, or various types of counts.

In [Example 1.2](#) we wanted to make inferences about average survival time, but there was the complication that the precise survival time was not known for one patient. The sign test let us use the partial information we had for that patient but the test used only a small part of the total information available; namely, whether or not each observation was above or below 200. We could have

carried out the test if we had been given only this partial information. Even when data sets are measurements occasionally we do as well or better with nonparametric methods that do not use all the available information. We develop several nonparametric tests in this chapter and compare their performance using simple data sets.

In [Sections 2.1 — 2.4](#) we show how to compute and interpret some exact  $P$ -values and confidence intervals. In practice even the best currently available software gives exact  $P$ -values only for small to medium-sized samples. We present in [Section 2.5](#) asymptotic results that often provide good approximations for larger samples. If programs for exact tests are not available a sensible compromise is to use Monte Carlo sampling to get good estimates of exact  $P$ -values. StatXact provides this facility, but if this package is not available one may still create the relevant samples for Monte Carlo approximation using any software that allows one to generate rapidly many random samples.

### 2.1.2 A raw-data randomization test for centrality

Pitman (1937a) developed ideas put forward by Fisher (1935) for a randomization test for hypotheses about the mean or median of a symmetric distribution when we have a random sample of observations from a population that has that distribution. For a symmetric distribution if the mean and median both exist they coincide at the point of symmetry. A few symmetric distributions have no mean, but this complication is seldom relevant in real data problems, so for the symmetric case we tend to use the terms population mean or median almost as though they are interchangeable. For asymmetric (skew) distributions the mean and median are no longer equal. Pitman's prime aim was to show that his test, which only assumed symmetry and did not require the distribution to belong to a specific family such as the normal family would, at least for samples that were not very small, lead to similar inferences to those based on a  $t$ -test which requires a normality assumption for strict validity. In effect he verified an earlier assertion by Fisher that using the  $t$ -test or other normal theory tests gave a good approximation to an *exact* randomization test even when the normality assumption was no longer valid. This justified continued use of the then easier-to-compute  $t$ -test based inference. Without modern computers the calculations for the Pitman test were prohibitive except for trivially small samples. The test is seldom used in practice (see Comment 5 on [Example 2.3](#) for some reasons why) but we describe it in some detail partly because of its historical

interest but also because it highlights several principles that are taken into account in more useful nonparametric methods.

We call the test a **Pitman test**, but the names **raw data test** and **Fisher–Pitman test** are also used. Modern computer packages such as StatXact or Testimate make computation easy. However, illustrating the test with a simple numerical example helps us understand its main features and comparing it with other tests highlights some weaknesses.

The basic idea behind the test is that if a random sample is from a symmetric distribution with unknown mean or median  $\theta$  then symmetry implies it is equally likely that any sample value will differ from  $\theta$  by some positive amount  $d$ , say, or by the same negative amount,  $-d$ , for all values of  $d$ . The test is valid even if the observations  $x_1, x_2, \dots, x_n$  are not each from the same distribution providing they are independent of one another and all have symmetric distributions with the same value for  $\theta$ . This situation might arise, for example, if  $n$  observatories each independently measure the diameter of the same star. Because each observatory uses different equipment, some of it more sophisticated than other equipment, it is likely that the precision with which observations are made will differ between observatories. If one assumes that these measurements are all ‘centred’ symmetrically about the true diameter,  $\theta$ , but differ only in the (usually unknown) spread about  $\theta$  and makes no further assumptions, the Pitman test is still valid.

Suppose  $n$  observations

$$x_1, x_2, \dots, x_n$$

satisfy the conditions above and we want to test the hypothesis that the median  $\theta$  has a fixed value  $\theta_0$ , i.e.  $H_0: \theta = \theta_0$  against the two-sided alternative  $H_1: \theta \neq \theta_0$ ; the test proceeds on the basis that under  $H_0$  the sign of each of the  $n$  differences

$$d_i = x_i - \theta_0, i = 1, 2, \dots, n$$

is equally likely to be positive or negative. Clearly, when  $H_0$  does not hold there is more likely to be a preponderance *either* of positive *or* of negative signs associated with the  $d_i$ . The situation has similarities to the sign test scenario in [Example 1.2](#), but there we considered only the numbers of positive or negative signs and *not* the magnitudes of the associated  $d_i$  and this required no assumption of symmetry for its validity.

### Example 2.1

*The problem.* The heart rate (beats per minute) when standing was recorded for seven members of a tutorial group. We assume that in the population of such students the distribution of heart rates when standing is symmetric. Given observations of 73, 82, 87, 68, 106, 60 and 97 test the hypothesis  $H_0: \theta = 70$  against the alternative  $H_1: \theta > 70$  using the Pitman test.

*Formulation and assumptions.* Symmetry implies that under  $H_0$  sample values should have a near-symmetric scatter about  $\theta = 70$ , so the magnitudes of the sums of positive and of negative deviations from 70 (the  $d_i$ ) should not differ greatly. The one-sided alternative  $H_1$  implies that when it holds there are likely to be more positive deviations from 70 than there would be under  $H_0$  and that these positive deviations will tend to be larger than the negative deviations.

*Procedure.* For the given data deviations from 70 are  $73 - 70 = 3$ ,  $82 - 70 = 12$  and similarly those remaining are 17, -2, 36, -10, 27. Since, under  $H_0$  the sum of the positive and negative deviations should be nearly equal whereas under  $H_1$  we expect the sum of the positive deviations to be appreciably larger than the sum of the negative deviations, it is clear that the sum of the positive deviations (i.e. differences from 70) is an intuitively reasonable statistic to help assess the strength of the evidence against  $H_0$  in support of the specified one-sided alternative  $H_1$ . If the alternative had been the two-sided  $H_1: \theta \neq 70$  small as well as large sums of positive deviations would also have supported  $H_1$ . For a randomization test based on  $S_+$ , the sum of the positive deviations, we form the appropriate randomization distribution in a manner similar to that in [Example 1.4](#) for the statistic  $S$  used there. The key to obtaining the randomization distribution of  $S_+$  when  $H_0$  holds lies in the fact that under  $H_0$  the sign of each deviation  $d_i$  is equally likely, on account of the symmetry assumption, to be positive or negative. Thus any combination of plus and minus signs attached to a set of deviations having the observed magnitudes is equally likely. Clearly in the current example with 7 observations there are  $2^7 = 128$  possible allocations of sign (2 to each of the 7 differences) each allocation being equally likely under  $H_0$ . Recording and ordering all of them without a computer is tedious. We need not do so if we only want to obtain the relevant  $P$ -value when that is small because we then only need to know how many of the equally likely allocations of signs give an  $S_+$  greater than or equal to that for our sample, i.e.  $S_+ = 3 + 12 + 17 + 36 + 27 = 95$ . Obviously the greatest possible value of  $S_+$  occurs when all the deviations have positive signs and then  $S_+ = 3 + 12 + 17 + 2 + 36 + 10 + 27 = 107$ . Clearly the next highest sum (105) is attained when only the smallest deviation, 2, has a negative sign. In order of decreasing magnitude it is easily verified that other sums greater than the observed  $S_+ = 95$  occur only when negative signs are attached to 3 only (sum 104), 2 and 3 (sum 102) and 10 only (sum 97). If a negative sign is attached to 12 only or to both 10 and 2 (as above) the sum is 95. Thus there are seven sums greater than or equal to the observed  $S_+ = 95$ . Since all are equally likely the relevant  $P$ -value for assessing the strength of evidence against  $H_0$  is  $P = 7/128 \approx 0.055$ .

*Conclusion.* The evidence against  $H_0$  when  $P = 0.055$  is sufficiently strong to perhaps warrant a similar study with a larger group of students.

*Comments.* 1. These data show that, despite our warning in [Section 1.3.1](#) that small experiments may never produce  $P$ -values indicating strong evidence against  $H_0$ , an interesting effect may sometimes be suggested even with very small samples; this is encouraging given the current emphasis on small group work in teaching, where data concerning individuals in the group are often used for illustrative purposes. A teaching advantage of using such data from small groups is that in working through the arithmetic one is not bogged down as would be the case with larger data sets. However, this example draws attention to a major limitation of a conventional hypothesis test. While it tells us that there is some doubt about the wisdom of retaining  $H_0$  it tells us nothing about what other hypotheses might be plausible. We need a confidence interval to give this information. Unfortunately, while modern computer programs enable us to perform a Pitman hypothesis test with relative ease it is harder to obtain a confidence interval for the true  $\theta$ , a matter discussed in more detail in [Example 2.3](#).

2. Modifications for a two-tail test when that is appropriate are straightforward and were mentioned under *Procedure*.

3. We chose  $S_+$  as our test statistic but we could equally well have chosen  $S_-$ , the sum of the negative deviations. This is because the sum of the **magnitudes** of the deviations remains constant for all possible allocation of signs (in this example always having the value 107). Thus the statistic has a symmetric distribution because for any of the 128 sign permutations we always have  $S_+ + S_- = 107$  or  $S_- = 107 - S_+$ . Computer software sometimes calculates the smaller of  $S_+$ ,  $S_-$  but each leads to the same relevant  $P$ -value. Another statistic that is often used is  $S_d = |S_+ - S_-|$ . For this example  $S_d = |2S_+ - 107|$  so there is a one-to-one ordered relationship between  $S_d$  and  $S_+$ . One may even use the one sample  $t$ -test statistic instead of  $S_+$  or  $S_-$  because it is not difficult to show ([Exercise 2.2](#)) that again there is a one-to-one ordered relationship between the values of  $t$  and  $S_+$ . Remember though that the distribution of this statistic  $t$  is not that given by the usual  $t$ -tables that assume normality, but is identical with that for  $S_+$ .

4. The approach has a strong intellectual appeal because it uses all the information in the data but it lacks robustness against departures from symmetry. This is explained in [Section 2.6](#).

5. Because the possible values of  $S_+$  depend upon values of  $d_i$  which in turn depend upon the data values it follows that the  $S_+$  corresponding to a given  $P$ -value varies between data sets even for samples of the same fixed size. The test is called a **conditional test** because it is conditional upon the data values. A practical implication of this is that the  $P$ -value associated with  $H_0$  must be computed afresh for each data set.

6. If we had followed the classic convention requiring  $P \leq 0.05$  before asserting *significance* we would not have rejected  $H_0$  in a formal significance test at the 5 per cent level. We see no merit in following such an arbitrary rule in this example, preferring to argue that the case for retaining  $H_0$  is not strongly supported even though 0.055 slightly exceeds 0.050. Remember also our comment in [Section 1.3.1](#) that a not very small  $P$ -value may sometimes only indicate that our experiment is too small to yield valuable information.

*Computational aspects.* 1. Using StatXact the  $P$ -value obtained above and the complete distribution of  $S_+$  can be calculated for this example in less than one

second. That program also gives an asymptotic approximation that is strictly valid only for large samples. Such approximations are discussed in [Section 2.5](#). For these data the asymptotic  $P$ -value for the one-tail test is  $P = 0.051$ , which, despite the small sample size, is close to the exact  $P$ -value, but such close agreement is not always obtained for small samples. In passing we note that the normal theory  $t$ -test for this example also gives  $P = 0.051$ . We comment further on the relation between the Pitman test and the  $t$ -test in [Example 2.2](#).

2. To estimate  $P$  by Monte Carlo sampling one would randomly allocate, each with probability  $p = 0.5$ , a plus or a minus sign to every  $d_i$  for, say, 10 000 samples and record the number,  $r$ , of these samples giving values of  $S_+$  greater than or equal to 95. Then  $P^* = r/10\,000$  is an estimate of the exact one-tail  $P$ .

Students of literary style often want to compare characteristics of the work of a particular author with those of other authors. The distribution of the number of words per sentence (sentence length) is one facet that is often studied. The mean or the median sentence length and the spread in sentence length, for example, varies between authors, especially in scientific or technical writing where average sentence length may even differ between several works by the same author, reflecting differing complexity of subject matter. When writing about some subjects there may even be varying patterns of sentence length in different positions in paragraphs or in different chapters. For example, the first sentence in any paragraph may tend to be longer than subsequent ones. This might happen in technical writing if each paragraph introduces a new basic idea and the first sentence may have to be long to expound that idea, while later sentences are subsidiary in nature to point out exceptions, limitations, fields of usage or other relevant matters. Both hypothesis testing and estimation play a part in studies of this kind. For example, suppose that an expert on writing skills asserts that for clarity in technical writing the average length of sentences should not exceed 30 words. A sample of sentences could be taken from the work of some author. Looking at the median sentence length,  $\theta$ , the strength of evidence against the hypothesis  $H_0: \theta \leq 30$  would be of interest; if the evidence is strong, the alternative  $H_1: \theta > 30$  seems more acceptable. If no prior assertion about ideal sentence length had been made one might be more interested in the median or mean sentence length of the work of a particular author and this leads naturally to an estimation type problem where finding a confidence interval is appropriate.

[Examples 2.2](#) and [2.3](#) bring out some difficulties that arise if the Pitman test is used when a symmetry assumption may not be reasonable. The consequences are discussed under the *Comments* headings.

### *Example 2.2*

*The problem.* The number of words in the first sentences in 12 randomly selected paragraphs from Fisher (1948) were:

12 18 24 26 37 40 42 47 49 49 78 108

These are arranged in ascending order but the analysis is not affected by this re-ordering of the original data. Test the hypothesis  $H_0: \theta \leq 30$  against  $H_1: \theta > 30$  where  $\theta$  is the population mean or median using the Pitman test.

*Formulation and assumptions.* We assume that sentence lengths are symmetrically distributed (but see Comment 1 below).

*Procedure.* We apply the Pitman randomization test to the differences formed by subtracting 30 from each datum, giving

-18 -12 -6 -4 7 10 12 17 19 19 48 78

Clearly the sum of the negative differences is smaller than the sum of the positive differences so we may use as our test statistic  $S_- = 18 + 12 + 6 + 4 = 40$ . If  $\theta = 30$  each of the 12 differences is equally likely to be positive or negative and there are now  $2^{12} = 4096$  possible allocations of signs to the differences and it is a formidable task to find how many of these lead to an  $S_- \leq 40$ . Software such as that in StatXact comes to our rescue and using that program for a one-tail test gives  $P = 0.0405$ .

*Conclusion.* The  $P$ -value indicates fairly strong evidence that the median exceeds 30.

*Comments.* 1. We remarked that Pitman (1937a) showed that results for his test are usually close to those given by a  $t$ -test. A one-tail  $t$ -test in this example gives an exact  $P = 0.0462$ , not markedly different from that given by the Pitman test. However it is doubtful whether the distribution of sentence length is symmetric, and certainly doubt about whether it is normally distributed. There are good grounds for these doubts. Firstly, for a symmetric distribution the means and medians coincide and thus in a sample one expects them to be close. For these data the median is 41 and the mean 44.17 so there is a recognizable difference between them. Inspecting the sample values shows that this difference is largely due to the influence on the mean of the value 108 which is 67 more than the median, whereas the minimum observed length is only 29 less than the median. Secondly, when one considers the nature of these data population skewness is not unreasonable. The shortest possible sentence must contain at least one word, but there is no fixed upper limit for the number of words in a sentence. In practice most writers sometimes use a sentence of 100 or more words even if their average sentence length is between 25 and 40, a common range for the median for many writers. A formal test for normality developed in [Section 3.3.3](#) provides strong evidence against these data being a sample from a normal distribution.

2. A breakdown of the symmetry assumption affects the performance of both the Pitman test and the  $t$ -test in an undesirable way, and although we do not prove it, the general effect of asymmetry is to reduce the power of both tests.

3. *Number of words* is only a crude measure of sentence length, especially when comparing works by different authors because some habitually use longer

words than others. Alternative measures are the total number of letters in a sentence or, if all sentences are set in exactly the same type-face, the length in centimetres between the start of the first word and the end of the last word. This last method has the disadvantage that for typographic reasons letters and words are usually not evenly spaced throughout a sentence. It is likely that any of the three methods for measuring sentence length mentioned here would lead to broadly similar inferences providing one could agree on the appropriate equivalent null hypothesis value of  $\theta$  for each measure – that in itself might call for some sample-based experiments! For example, is a 30-word sentence broadly equivalent to a 180-letter sentence?

*Computational aspects.* StatXact gives exact one- or two-tail permutation test  $P$ -values for small to moderate sample sizes and also an asymptotic result which is reliable only for larger samples. The program also provides a Monte Carlo approximation for  $P$  if required. A few general statistical programs also provide an asymptotic approximation to a Pitman test  $P$ -value.

Confidence intervals for  $\theta$  based on the Pitman test are seldom used for reasons that will soon become apparent so we omit the theory for computing them. It is given by Sprent (1998, Section 5.2) and by Maritz (1995, Section 2.4). We know of no readily available software package that gives confidence intervals directly in this case, but a trial-and-error method based on a hypothesis testing program may be used.

### *Example 2.3*

*The problem.* Given the data in [Example 2.2](#) determine a confidence interval with at least 95 per cent coverage for the mean or median  $\theta$  based on the Pitman procedure. Remember that we are, with little justification, assuming symmetry.

*Formulation and assumptions.* Assuming symmetry we may obtain the limits that define the interval by determining values  $\theta_1$  and  $\theta_2$  of  $\theta$  where  $\theta_1 < \theta_2$  such that we would retain the hypothesis that this were the true parameter value if  $\theta = \theta_1$  but not for any  $\theta < \theta_1$  or if  $\theta = \theta_2$  but not for any  $\theta > \theta_2$ , in each case with a relevant one-tail probability as close as possible to, but not exceeding 0.025. The interval is then based on a two-tail test with  $P$  not exceeding 0.05. We consider the modification for a one-sided interval in Comment 2.

*Procedure.* A trial and error method is used to establish the required limits starting with suitable initial values. Because the Pitman test usually gives results not very different from a  $t$ -test applied (however inappropriately) to the same data it is reasonable to start with 95 per cent confidence limits based on the  $t$ -test procedure. These are 27.2 and 61.1 and are obtainable from most statistical software packages. Our trial and error procedure begins by using a Pitman test program to obtain one-tail  $P$ -values for testing the hypotheses  $\theta = 27.2$  and  $\theta = 61.1$ . The data editor in most relevant statistical programs will generate the relevant deviations directly by subtraction of the appropriate constants 27.2 and 61.1 from the data. The relevant  $P$ -values found are 0.016 in the lower tail and 0.026 in the upper tail. To adjust each of these probabilities to be as close as

possible to 0.025 without exceeding that value clearly we should try a higher value for  $\theta$  for the lower limit (to increase  $P$ ) and also a higher value for the upper limit (to decrease  $P$ ). We first chose 28 and 61.5 giving respectively  $P = 0.021$  and  $P = 0.025$ . Our next choice was 28.4 and 61.6 giving  $P = 0.024$  and  $P = 0.024$ . We then took 28.5 as a revised lower value giving  $P = 0.025$ . Both lower and upper limits are close to the target  $P = 0.025$  but for further refinement we took 28.6 giving  $P = 0.026$ . It is now clear that in two-tail tests at the 5 per cent level we would favour the hypotheses  $\theta = 28.6$  or perhaps  $\theta = 61.5$ , but we would not retain the hypotheses if  $\theta = 28.5$  or  $\theta = 61.6$ . As one further piece of fine tuning, really unnecessary in practice, we found that the  $P$ -values associated with  $\theta = 28.59$  and  $\theta = 61.51$  were respectively  $P = 0.025$  and  $P = 0.024$ .

*Conclusion.* A Pitman test based interval  $(28.6, 61.5)$  has confidence level  $[1 - (0.025 + 0.024)]100 = 95.1$  per cent. This follows because if we take values of  $\theta$  just outside these limits, i.e. 28.59 or 61.51 we would reject these hypothetical values while we would accept the values 28.6 and 61.5 in conventional tests if we used an exact 4.9 per cent significance level.

*Comments.* 1. Although discontinuities in possible  $P$ -values stop us forming an exact 95 per cent interval our coverage of 95.1 per cent is close to 95 per cent and the interval  $(28.6, 61.5)$  is not very different from the  $t$ -test 95 per cent interval  $(27.2, 61.1)$ . The Pitman interval is slightly shorter and shifted a little to the right, but these differences are of no practical importance. Another effect of discontinuities is that the exact  $P$ -values may not be equal in each tail.

2. Both the  $t$ -test and the Pitman test intervals indicate that with a two-tail test there is no strong evidence against the hypothesis  $\theta \leq 30$ . Indeed the rule for doubling the one-tail probability for getting an appropriate two-tail probability leads to the same conclusion if we had modified our  $H_1$  appropriately in [Example 2.2](#). Confidence intervals are usually applied in a two-tail context but we can form intervals relevant to a one-tail test. Corresponding to the test in [Example 2.2](#), for such a 95 per cent interval we find a lower cut-off point that gives a critical region of size  $P = 0.05$  in the lower-tail. The upper limit becomes infinity and a trial and error process similar to that used for the two-tail situation gives a one tail  $P = 0.050$  corresponding to  $\theta = 30.74$  and  $P = 0.051$  corresponding to  $\theta = 30.75$ . Thus a 95 per cent one-sided confidence interval is  $(30.75, \infty)$ . In other words, hypothetical values of  $\theta \geq 30.75$  appear more plausible than values of  $\theta < 30.75$ , consistent with our finding in [Example 2.2](#) that a one-tail test gave fairly strong evidence against  $H_0: \theta = 30$ . The upper limit  $\infty$  is in practice somewhat meaningless, being a mathematically symbolic way of saying that the true median is more likely to be some (unspecified) value above 30.75 rather than one at or below that value. One only need glance at the sample values to see that the median is extremely unlikely to be 200!

3. In this example it is sensible and suffices to give limits only to one decimal place. Also, when quoting  $P$ -values, one or two significant figures suffice, although some programs may produce values like  $P = 0.025631$  in some examples. It is sensible to report  $P$ -values below 0.001 simply as  $P < 0.001$ .

4. We show in [Example 2.11](#) that using a method not assuming a symmetric distribution gives a shorter confidence interval for the median. Indeed, in that example we obtain a 96.2 per cent interval  $(24, 49)$  for these data. Although the lower limit is only slightly below that for the  $t$ -test or Pitman test intervals the

upper limit is markedly lower. We discuss why this is so in [Example 2.11](#) remarking here only that breakdown of the symmetry assumption may have a considerable effect on confidence intervals obtained when assuming symmetry.

5. Inferences based on the  $t$ -test and the Pitman test tend to be similar and both are unsatisfactory when there is asymmetry. The Pitman test procedure for computing confidence intervals is cumbersome.

## 2.2 INFERENCES ABOUT MEDIAN BASED ON RANKS

An alternative to the Pitman test is to apply a similar test not to the signed deviations from the hypothesized median but to the signed ranks of the absolute deviations. This leads to the signed-rank **test** proposed by Wilcoxon (1945). The population distribution is again assumed symmetric and ideally, continuous. In theory this means there is zero probability of two sample values coinciding. This may not be the case in practice; equal or **tied** values may be present in a data set. A modified form of the test allows for ties.

### 2.2.1 The Wilcoxon signed-rank test

Given a sample of  $n$  independent measurements we first determine, as we did in [Section 2.1.1](#), the magnitude of departures from the hypothetical mean or median  $\theta_0$  specified in  $H_0$ . We arrange these **absolute deviations** in order of **magnitude** and assign ranks to these in ascending order (1 for the smallest,  $n$  for the largest). We next attach a negative sign to each rank that corresponds to a negative deviation (i.e. to data values below  $\theta_0$ ). We expect a near equal scatter of positive and negative ranks if  $\theta_0$  is the true mean or median, implying that the sum of all positive ranks and the sum of all negative ranks should not differ greatly. A high sum of the positive (negative) ranks relative to that of the negative (positive) ranks implies  $\theta_0$  is unlikely to be the population mean. We now apply a test that is mechanically equivalent to the Pitman test where signed ranks of deviations replace the signed deviations themselves.

*Example 2.4*

*The problem.* For a group of 12 female students, the changes in heart rate (beats per minute) when standing up from lying down are:

$$-2 \ 4 \ 8 \ 25 \ -5 \ 16 \ 3 \ 1 \ 12 \ 17 \ 20 \ 9$$

If we assume population symmetry, we may use the Wilcoxon signed-rank test to test  $H_0: \theta = 15$  against the alternative  $H_1: \theta \neq 15$ .

*Formulation and assumptions.* We arrange the deviations from 15 in ascending order of magnitude and rank these, associating with each rank the sign of the corresponding deviation. We calculate the lesser of the sum of positive and negative signed ranks and, if available, use appropriate computer software to determine  $P$ .

*Procedure.* Subtracting 15 from each sample value gives the deviations  $-17, -11, -7, 10, -20, 1, -12, -14, -3, 2, 5, -6$ . We rearrange these in increasing order of magnitude while retaining signs, i.e.  $1, 2, -3, 5, -6, -7, 10, -11, -12, -14, -17, -20$ , whence the signed ranks are  $1, 2, -3, 4, -5, -6, 7, -8, -9, -10, -11, -12$ . The sum of the positive ranks,  $S_+ = 1 + 2 + 4 + 7 = 14$ , is less than the sum,  $S_-$  of the negative ranks. StatXact, Testimate and some general statistical programs give exact  $P$ -values in programs designed specifically for the Wilcoxon test where data may be input either in its original form or as signed ranks and the program then effectively carries out a Pitman test using signed ranks in place of the exact signed deviations. For this example these programs indicate that the two-tail  $P$ -value corresponding to  $S_+ = 14$  is  $P = 0.052$ .

*Conclusion.* The evidence against  $H_0$  is not very strong, but the  $P$ -value is sufficiently small for a further study involving a larger group to be considered.

*Comments.* 1. Doctors often assume that heart rate measurements are symmetrically distributed. In this example, the median, 8.5, is close to the mean, 9.0, indicating that symmetry is possible. There are several tests for symmetry when we are given only the sample values. One proposed by Randles et al. (1980) is described by Hollander and Wolfe (1999, Section 3.9) and also by Siegel and Castellan (1988, Section 4.4) but caution is needed in assessing significance if  $n < 20$  in cases where the asymptotic result is on the borderline of significance. However, for these data the test statistic has a value well below any that would suggest asymmetry. There is, however, some evidence for a departure from normality because there is no particular cluster of values around the mean of 9.0; instead several values are close to zero and there is another group of values around 17. Given the likely degree of symmetry in the distribution the Wilcoxon test should not be seriously misleading.

2. The two clumps of values could reflect two distinct groups among female students: those who are heavily involved in sports such as hockey or tennis and those who rarely participate. The females who are physically fit may not experience a large increase in heart rate on standing compared with those who follow a more sedentary life style.

3. One must be cautious about generalizing the above results to the community at large. For a variety of reasons female students are probably not typical of young adults in general.

*Computational aspects.* 1. Not all statistical packages give an exact  $P$ -value for the Wilcoxon signed-rank test, some only having a program that calculates  $S_+$  or  $S_-$  and then sometimes gives only a  $P$ -value based on asymptotic theory (see Section 2.5). If one has no program to give exact  $P$ -values and is not prepared to seek a Monte Carlo approximation for small to medium samples it may be better to refer to published tables even though many give only values required for significance at a ‘nominal’ 5 or 1 per cent level in one- or two-tail tests. Due to discontinuities in the distribution of  $P$ -values the exact levels are not precisely 5

or 1 per cent. For most published tables the exact values are the closest possible values that do not exceed the nominal significance levels although a few tables give a value as close as possible to the nominal level even if that closest value exceeds the nominal level. Most major published statistical tables [e.g. Neave (1981) or Lindley and Scott (1995)] and some books on nonparametric methods include tables for the Wilcoxon signed-rank test and usually describe which of the above options they use for defining the nominal level. A more satisfactory table in Hollander and Wolfe (1999, Table A4) gives most of the exact  $P$ -values likely to be of interest for sample sizes  $3 \leq n \leq 60$ , but these tables are not needed if programs like that in StatXact or Testimate are available.

2. The asymptotic  $P$ -value given by StatXact for these data for a two-tail test is  $P = 0.050$  but this does not include a continuity correction we describe in [Section 2.5](#). That correction increases the value to  $P = 0.0545$ . For these data the two-tail  $t$ -test value is  $P = 0.046$ . These values do not differ greatly from the exact value but give a timely warning on the dangers of adhering to a rigid ‘significance’ level such as  $\alpha = 0.05$  as a decision-making basis for accepting or rejecting  $H_0$ .

## 2.2.2 Theory of the Wilcoxon signed-rank test

We pointed out in [Section 2.2.1](#) that the Wilcoxon signed-rank test is a special case of the Pitman test with signed ranks of the differences replacing the signed differences themselves. Using ranks gives theoretical simplifications and also has implications for what happens if an assumption of symmetry is not justified. Availability of computer software to give exact  $P$ -values means that mastery of the theory and its implications is not so important as it once was, but a broad understanding helps to remove misconceptions and is needed to show how to get confidence intervals.

We describe the exact permutation test for a sample size  $n = 7$ , assuming no deviations are tied in magnitude. As already indicated, when  $H_0$  is true the symmetry assumption implies the sums of ranks for positive and negative ranked deviations should be nearly equal, whereas one sum should be much larger than the other if  $H_0$  is not true. In the extreme case when all sample values are above the mean or median hypothesized in  $H_0$  all ranks will be positive so  $S_- = 0$  and  $S_+ = 1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$ . Clearly, for any allocation of signs to ranks  $S_+ = 28 - S_-$  when  $n = 7$ . More generally, for  $n$  observations  $S_+ = \frac{1}{2}n(n + 1) - S_-$  since the sum of the ranks 1, 2, 3, . . . ,  $n$  is  $\frac{1}{2}n(n + 1)$ . This implies a symmetry between  $S_+$  and  $S_-$  in the sense that

$$\Pr(S_- = s) = \Pr[S_+ = \frac{1}{2}n(n + 1) - s].$$

As in the Pitman test we may use any of the statistics  $S_+$ ,  $S_-$  or  $S_W = |S_+ - S_-|$ , since there is an ordered linear relationship between them.

When  $n = 7$ , signs may be allocated randomly to ranks in  $2^7 = 128$  different and equally likely ways. The distribution of the test statistic  $S_+$  now depends only on the sample size, 7, since all data sets then give the rank magnitudes 1, 2, 3, 4, 5, 6, 7. To get the distribution of  $S_+$  we write down the  $S_+$  for all possible sign allocations and it is left as an exercise for the reader to establish that the relevant sums and associated probabilities are those in [Table 2.1](#). Exact  $P$ -values for one- or two-tail tests may be obtained using that table and are applicable to a test for any data without tied values when  $n = 7$ . For example, in a one-tail test where a small lower-tail value of  $S_+$  is relevant the exact  $P$ -value corresponding to  $S_+ = 3$  is  $P = \Pr(S_+ \leq 3) = (1 + 1 + 1 + 2)/128 = 5/128 \approx 0.039$ . This probability is doubled for a two-tail test.

Manual calculation of probabilities associated with all values of  $S_+$  or  $S_-$  for larger  $n$  is time consuming and error prone. StatXact obtains these values rapidly for sample sizes of practical importance.

The number of different values of the rank sums increases as the sample size,  $n$ , increases and the differences between the discrete probabilities associated with each become smaller, especially in the tails. Also, the distributions of  $S_+$ ,  $S_-$  approach that of a normal distribution as  $n$  increases. When  $n = 12$  there are  $2^{12} = 4096$  possible associations of signs with ranks and the possible signed-rank sums range from 0 to 78. Symmetry means we need only know the probabilities for each sum between 0 and 39. [Table 2.2](#) is adapted from StatXact output and gives sufficient information to calculate the complete permutation test distribution of  $S_+$  or  $S_-$ . From [Table 2.2](#) we easily find that  $\Pr(S_- \leq 13) = 87/4096 \approx 0.021$  while  $\Pr(S_- \leq 14) = 107/4096 \approx 0.026$ . For two-tail tests we double these probabilities giving 0.042 and 0.052. Thus values of the smaller signed-rank sum not exceeding 13 indicate significance at level 4.2 per cent or less, i.e. below a conventional 5 per cent level. Again from [Table 2.2](#) we see that  $\Pr(S_- \leq 9) = 33/4096 \approx 0.008$ , while  $\Pr(S_- \leq 10) = 43/4096 \approx 0.0105$ ; it follows that for a one-tail test at a level not exceeding 1 per cent that  $S_- \leq 9$  forms an appropriate critical region with associated  $P = 0.008$ .

[Figure 2.1](#), based on [Table 2.2](#), shows the probability mass function (also known as the frequency function) of  $S_+$  or  $S_-$  for  $n = 12$  in the form of a bar chart. The shape resembles that of the familiar probability density function for the normal distribution. In [Exercise](#)

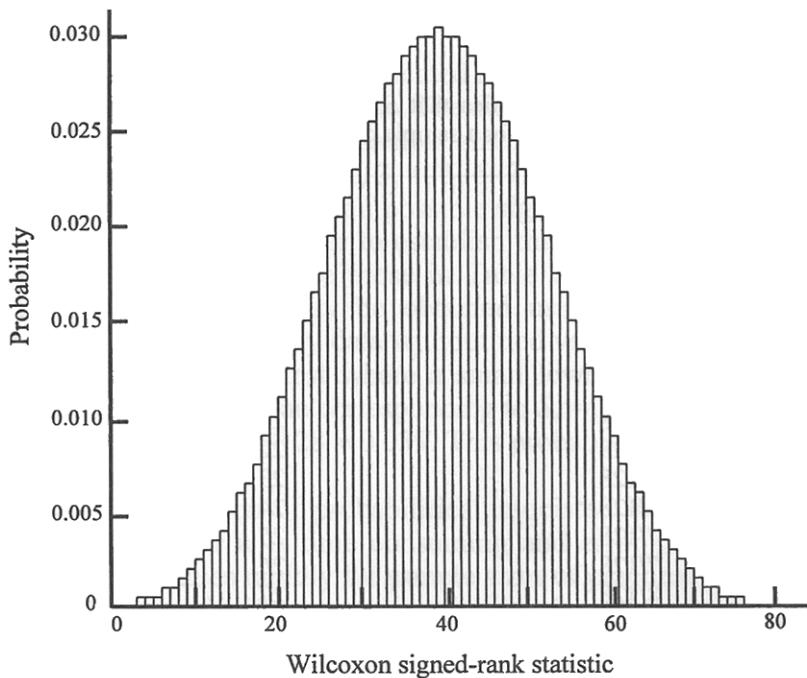
[2.4](#) we ask the reader to draw and comment upon a bar chart based on Table 2.1 for the case  $n = 7$ .

**Table 2.1** Exact probabilities that  $S_+$ ,  $S_-$  equal a given value  $k$  in a Wilcoxon signed-rank test for a sample of size  $n = 7$  when  $H_0$  is true.

$k$	<i>Probability</i>	$k$	<i>Probability</i>
0	1/128	15	8/128
1	1/128	16	8/128
2	1/128	17	7/128
3	2/128	18	7/128
4	2/128	19	6/128
5	3/128	20	5/128
6	4/128	21	5/128
7	5/128	22	4/128
8	5/128	23	3/128
9	6/128	24	2/128
10	7/128	25	2/128
11	7/128	26	1/128
12	8/128	27	1/128
13	8/128	28	1/128
14	8/128		

**Table 2.2** Exact probabilities that  $S_+$ ,  $S_-$  equal a given value  $k$  in a Wilcoxon signed-rank test for a sample of size  $n = 12$  when  $H_0$  is true. Probabilities for  $k > 39$  follow from the symmetry property  $\Pr(S_- = k) = \Pr(S_- = 78 - k)$ .

$k$	<i>Probability</i>	$k$	<i>Probability</i>	$k$	<i>Probability</i>
0	1/4096	13	17/4096	26	78/4096
1	1/4096	14	20/4096	27	84/4096
2	1/4096	15	24/4096	28	89/4096
3	2/4096	16	27/4096	29	94/4096
4	2/4096	17	31/4096	30	100/4096
5	3/4096	18	36/4096	31	104/4096
6	4/4096	19	40/4096	32	108/4096
7	5/4096	20	45/4096	33	113/4096
8	6/4096	21	51/4096	34	115/4096
9	8/4096	22	56/4096	35	118/4096
10	10/4096	23	61/4096	36	121/4096
11	12/4096	24	67/4096	37	122/4096
12	15/4096	25	72/4096	38	123/4096
				39	124/4096



**Figure 2.1** A bar chart illustrating the distribution of the Wilcoxon signed-rank statistic when  $n = 12$ .

### 2.2.3 The Wilcoxon test with ties

In theory, when we sample from a continuous distribution both the probabilities of tied observations or of a sample value exactly equaling the population mean or median are zero. Real observations never have a distribution that is strictly continuous in the mathematical sense either because of their nature or as a result of rounding or limited measuring precision. We measure lengths to the nearest centimetre or millimetre; weights to the nearest kilogram, gram or milligram; the number of pages in a book in complete pages, although chapter layout often results in part-pages of text; numbers of words in a sentence must be integers. These realities may produce data with rank ties or zero departures from a hypothesized mean or median. If they do, the exact distribution of  $S$ , the lower of the positive or negative rank sums, requires fresh computation for different numbers of ties and ties in different positions in the rank order. Before the advent of software to compute exact  $P$ -values for permutation distributions with ties, a common

procedure was, after adjusting the scoring method based on ranks to allow for ties, to use the asymptotic normal approximation we give in [Section 2.5](#). This usually works well for a few ties if  $n$  is reasonably large, but may have bizarre consequences for small  $n$ . Another approach for smaller samples was to use tables for the no-tie case assuming that the relevant value of  $S_+$  would be little affected by ties. This need not be so.

When two or more ranks are equal in magnitude, but not necessarily of the same sign, we replace them by their mid-rank, an idea easily explained by an example. If 7 observations are 1, 1, 5, 5, 8, 8, 8 and we want to use a Wilcoxon signed-rank test with  $H_0 : \theta = 3$  the relevant signed differences from 3 are  $-2, -2, 2, 2, 5, 5, 5$ . The magnitude of the smallest difference is 2 and four differences have that magnitude. The mid-rank rule assigns to these ties the mean of the four smallest ranks 1, 2, 3, 4; i.e. we allocate to these observations the signed ranks  $-2.5, -2.5, 2.5, 2.5$ . The remaining three differences are all 5 and we give each the mean of the remaining ranks 5, 6 and 7; i.e. each is ranked 6. The exact permutation test is based on the signed mid-ranks  $-2.5, -2.5, 2.5, 2.5, 6, 6, 6$ . An appropriate test statistic is  $S_- = 5$ , the sum of the negative mid-ranks.

If we test using the permutation distribution for the ‘no-tie’ case given in [Table 2.1](#) we find  $\Pr(S_- \leq 5) = 10/128 = 0.078$ . However, with ties, the distribution of  $S_-$  under  $H_0$  is **not** that in [Table 2.1](#). The exact distribution with ties depends both on how many ties there are and where they lie in the rank sequence. We can work out exact

**Table 2.3** Exact probabilities that  $S_+, S_-$  equal a given value  $k$  in a Wilcoxon signed-rank test for a sample of size  $n = 7$  when  $H_0$  is true and mid-ranks are 2.5 (four times) and 6 (three times).

$k$	Probability	$k$	Probability
0	1/128	14.5	12/128
2.5	4/128	16	3/128
5	6/128	17	18/128
6	3/128	18	1/128
7.5	4/128	19.5	12/128
8.5	12/128	20.5	4/128
10	1/128	22	3/128
11	18/128	23	6/128
12	3/128	25.5	4/128
13.5	12/128	28	1/128

significance levels with ties for small to medium sized samples. For any tie pattern StatXact or Testimate give exact  $P$ -values for one- and two-tail tests and the former will also compute the complete distribution of  $S_+$  or  $S_-$  for samples that are not too large. For the tie pattern above for a sample of 7, StatXact gives the exact distribution in [Table 2.3](#).

Comparing [Tables 2.1](#) and [2.3](#) we see that in each case the distribution of the test statistic is symmetric (that is why we may base our test on either  $S_+$  or  $S_-$ ). However, the distribution with no ties is unimodal (i.e. it has one maximum) with values of the statistic confined to integral values increasing in unit steps, but that in [Table 2.3](#) is heavily multimodal with  $S$  taking unevenly spaced and not necessarily integer values. Discontinuities are also more marked.

From [Table 2.3](#) we easily find that  $\Pr(S_- \leq 5) = 11/128$ , compared to the no-tie probability  $10/128$  found from [Table 2.1](#). Despite many ties in this example and the differences in the distribution of the test statistic in the two cases, it may appear that ties have little effect on our conclusion. However, this is not always the case.

### *Example 2.5*

Ties often result from rounding. The data  $1, 1, 5, 5, 8, 8, 8$  considered above could arise from rounding to the nearest integer in either of the sets:

$$\begin{aligned} & 0.9, 1.1, 5.2, 5.3, 7.9, 8.0, 8.1 \\ & 0.9, 1.0, 4.8, 4.9, 7.9, 8.0, 8.1 \end{aligned}$$

For a Wilcoxon test with  $H_0: S_- = 3$  the signed differences from 3 are:

$$\begin{aligned} & -2.1, -1.9, 2.2, 2.3, 4.9, 5.0, 5.1 \\ & -2.1, -2.0, 1.8, 1.9, 4.9, 5.0, 5.1 \end{aligned}$$

with signed ranks:

$$\begin{aligned} & -2, -1, 3, 4, 5, 6, 7 \\ & -4, -3, 1, 2, 5, 6, 7 \end{aligned}$$

There are now no tied ranks in either set and for the first set  $S_- = 3$ , while for the second set  $S_- = 7$ . [Table 2.1](#) gives  $\Pr(S_- \leq 3) = 5/128 \approx 0.039$  and  $\Pr(S_- \leq 7) = 19/128 \approx 0.148$ . Thus a one-tail test at a level not exceeding 5 per cent would indicate an implausible hypothesis in the first case but not in the second. Rounding to the nearest integer thus greatly affects the apparent strength of evidence against the null hypothesis.

This example is a simple illustration of difficulties of interpretation with ties due to rounding, a process that makes small perturbations or changes to most observations. Many statistical analyses are sensitive to such changes. The impact of ties due to rounding in this example is more extreme than one is likely to meet in practice, resulting in only two different magnitudes for ranks. For larger samples and not too many ties, the effect on  $P$ -values is less marked.

### Example 2.6

*The problem.* Consider the data in [Example 2.2](#) for the numbers of words in sentences in Fisher (1948), i.e.

12 18 24 26 37 40 42 47 49 49 78 108

Use the Wilcoxon signed-rank test to test for the population median  $\theta$  the hypothesis  $H_0: \theta \leq 30$  against the alternative  $H_1: \theta > 30$ .

*Formulation and assumptions.* To justify the Wilcoxon test, we assume the population is symmetric. The deviations of observations from 30 are ranked, with due regard to sign using mid-ranks for ties. The lesser of  $S_+$  and  $S_-$  is then found.

*Procedure.* The signed deviations from 30 are  $-18, -12, -6, -4, 7, 10, 12, 17, 19, 19, 48, 78$  and the corresponding signed ranks using mid-ranks for the tied magnitudes at 12 and 19 are

-8 -5.5 -2 -1 3 4 5.5 7 9.5 9.5 11 12

giving  $S_- = 16.5$  (or  $S_+ = 61.5$ ). A one-tail test is appropriate. StatXact gives the exact probability of observing  $S_- \leq 16.5$  with this pattern of tied deviations when  $n = 12$  as  $P = 0.0405$ .

*Conclusion.* The evidence implies  $H_0$  is implausible at a 4.05 per cent significance level, so there is fairly strong evidence against  $H_0$ .

*Comments.* 1. Although the  $P$ -value here is identical to that for the Pitman test in [Example 2.2](#), in general this is not the case. Large differences between results for the two tests, although possible, are unusual (see [Example 2.13](#)).

2. Before we had software that gave the exact distribution of the test statistic when there are ties, the advice to resort to asymptotic (large sample) results of a type we describe in [Section 2.5](#) had little justification for small samples.

3. The data values 78, 108 throw doubt on the validity of the symmetry assumption. Had these observations been replaced by, say, 51, 53 our Wilcoxon rank-sum test would give the same result but there then would have then been no intuitive reason to doubt the symmetry assumption. However the Pitman test would be affected by such a change and in [Exercise 2.5](#) we ask you to confirm that with those alterations for that test  $P = 0.0505$ .

*Computational aspects.* With packages like StatXact it is a useful exercise to generate exact distributions for a number of tied situations for small sample sizes to develop a feeling for the effect of ties on standard results. An extreme case arises with the observations 4, 4, 8, 8, 8, 8, 8 for the Wilcoxon test when we set  $H_0: \theta = 6$ . Then all ranks of deviations are tied apart from signs and it is easy to see ([Exercise 2.6](#)) that the Wilcoxon test statistic distribution is now equivalent to that for the sign test. In this sense a sign test is a special case of the Wilcoxon test where all deviation magnitudes are tied at the same value.

If one or more sample values equal the mean hypothesized under  $H_0$ , the associated ‘deviations’ are zero. Statisticians are divided on how best to proceed in this case. Some argue that we should omit such observations from our computations because they are

uninformative about any alternative hypothesis. The test is then applied to a reduced sample. An alternative we recommend is to rank all deviations temporarily giving any zero the rank 1 because it is the smallest difference (or the appropriate tied rank if there is more than one zero). After all signed ranks are allocated we then change the rank(s) associated with zero difference to 0, leaving the other ranks as allocated. To obtain exact tests in the latter situation a suitable computer program is needed for the exact distribution of the test statistic, else we may resort to asymptotic results in a way we indicate in [Section 2.5](#). Deleting zero differences is the simplest procedure if no suitable program for exact probabilities of  $S$  for the modified rank statistic is available, but the second procedure has the theoretical advantage of providing a more powerful test.

#### 2.2.4 Confidence intervals based on signed ranks

For many nonparametric methods the more important aim of calculating confidence intervals is not so simple as hypothesis testing, but the procedures, once mastered, are often straightforward and not difficult to program by writing a macro if no package program is available. Although at the time of writing only a few packages deal in a completely satisfactory way with confidence intervals based on Wilcoxon signed-rank test theory, the position is improving steadily.

The trial-and-error approach suggested in [Section 2.1.2](#) for the Pitman test is avoided by using an alternative way of calculating the statistic  $S_-$  (or  $S_+$ ). We outline the theory, but a reader prepared to take this on trust may move directly to [Example 2.7](#) where this approach is applied to both hypothesis testing and determining confidence intervals.

For convenience we assume a sample of  $n$  observations with no ties has been arranged in **ascending** order  $x_1, x_2, \dots, x_n$ . The null hypothesis for the mean or median is specified as  $H_0: \theta = \theta_0$ . In practice  $\theta_0$  will usually lie between  $x_1$  and  $x_n$ , but this is not a theoretical requirement. A deviation  $x_i - \theta_0$  will have a negative signed rank for any  $x_i$  less than  $\theta_0$ . Since the  $x_i$  are in ascending order, among all **negative** ranks, if there are any, that of greatest magnitude will be associated with  $x_1$ , that of next greatest magnitude with  $x_2$ , and so on. Similarly, all  $x_j$  greater than  $\theta_0$  have positive signed ranks and of these  $x_n$  will have the highest rank,  $x_{n-1}$  the next highest and so on.

Consider now the paired averages of  $x_1$  with each of  $x_1, x_2, \dots, x_n$ . Suppose  $x_1$  has associated signed rank  $-p$ . A moment's reflection shows that each of the averages  $\frac{1}{2}(x_1 + x_q)$  will be less than  $\theta_0$  providing the deviation associated with  $x_q$  has either a negative rank or a positive rank less than  $p$ . Since we are assuming there are no rank ties no observation has positive rank  $p$  (since the rank  $p$  is already associated with  $x_1$ ). If  $x_q$  is the smallest observation associated with a positive rank greater than  $p$ , this implies  $|x_q - \theta_0| > |x_1 - \theta_0|$  and it follows that the average of  $x_1$  and  $x_q$  is greater than  $\theta_0$ . Then, because the data are in ascending order, so is the average of  $x_1$  with any  $x_r > x_q$ . Thus the number of averages involving  $x_1$  that are less than  $\theta_0$  is equal to the negative rank associated with  $x_1$ . Similarly, if we form all the paired averages of any  $x_i$  less than  $\theta_0$  with each of  $x_i, x_{i+1}, \dots, x_n$  the number of these averages less than  $\theta_0$  will equal the (negative) rank associated with  $x_i$ . Clearly, when  $x_i$  has a positive signed rank none of the averages with itself or greater sample values will be less than  $\theta_0$ . The complete set of averages  $\frac{1}{2}(x_i + x_j)$  for  $i = 1, 2, \dots, n$  and all  $j \geq i$  are often called **Walsh averages**, having been proposed by Walsh (1949a, b). Clearly the number of Walsh averages less than  $\theta_0$  equals  $S_-$  and the number greater than  $\theta_0$  equals  $S_+$ . We may use Walsh averages to calculate the test statistic  $S$  for a given  $H_0$  and to obtain confidence limits for a population mean or median.

### *Example 2.7*

*The problem.* Given the changes in heart rate for the female students in [Example 2.4](#), viz.

-5 -2 1 3 4 8 9 12 16 17 20 25

obtain an estimate of the population mean with confidence intervals at levels at least 95 and 99 per cent, assuming a symmetric distribution.

*Formulation and assumptions.* We use Walsh averages and the principle enunciated in [Section 1.4.1](#) that the  $100(1 - \alpha)$  per cent confidence interval contains those values of the parameter that would be accepted in a significance test at the  $100\alpha$  per cent significance level.

*Procedure.* [Table 2.4](#) shows a convenient way to set out Walsh averages. Both the top row and the first column (in italics) of that table give the sample values in ascending order. The triangular matrix forming the body of the table gives the Walsh averages. Comment 2 below gives tips for calculating these if a program to do so is not available.

An intuitively reasonable point estimator of the population median is the median of the Walsh averages since it follows from the arguments given before this example that for this estimate the sums of both the negative and the positive rank deviations are equal, both being equal to the number of Walsh averages

**Table 2.4** Walsh averages for changes in heart rate from lying to standing.

	-5	-2	1	3	4	8	9	12	16	17	20	25
-5	-5.0	-3.5	-2.0	-1.0	-0.5	1.5	2.0	3.5	5.5	6.0	7.5	10.0
-2		-2.0	-0.5	0.5	1.0	3.0	3.5	5.0	7.0	7.5	9.0	11.5
1			1.0	2.0	2.5	4.5	5.0	6.5	8.5	9.0	10.5	13.0
3				3.0	3.5	5.5	6.0	7.5	9.5	10.0	11.5	14.0
4					4.0	6.0	6.5	8.0	10.0	10.5	12.0	14.5
8						8.0	8.5	10.0	12.0	12.5	14.0	16.5
9							9.0	10.5	12.5	13.0	14.5	17.0
12								12.0	14.0	14.5	16.0	18.5
16									16.0	16.5	18.0	20.5
17										17.0	18.5	21.0
20											20.0	22.5
25												25.0

above (or below) this median. The median of the Walsh averages may be determined by inspecting Table 2.4. We see that entries in any row increase from left to right and those in any column increase as we move down, so that the smallest entries are at the top left of the table and the largest are at the bottom right. For a sample size  $n$  there are  $\frac{1}{2}n(n + 1)$  Walsh averages, i.e. 78 in this example. Thus the median of these lies between the 39th and 40th ordered Walsh average and inspection of the table establishes that both equal 9.0; so this is the appropriate estimate of the population median (or mean). To establish a 95 per cent confidence interval we select as end points values of  $\theta$  that will just be acceptable if  $P = 0.05$ . From [Example 2.4](#),  $P$  is only slightly more than 0.05 if the sum of the lesser of positive or negative ranks is 14. Thus, if we choose a value less than the 14th smallest Walsh average we would reject  $H_0$ . Similarly if we choose a value greater than the 14th largest Walsh average we also reject  $H_0$  since in either case our test statistic  $S$  would be such that  $S \leq 13$ . Thus for  $n = 12$  a confidence interval at not less than a 95 per cent confidence level is the interval (14th smallest Walsh average, 14th largest Walsh average). From Table 2.4, by counting from the most extreme value we see that this is the interval (2.5, 16.0). We noted in discussing [Table 2.2](#) that using  $S \leq 13$  as a criterion for significance corresponded to an exact significance level of 4.25 per cent (two tails). It is intuitively reasonable therefore to regard the interval (2.5, 16.0) as an actual 95.75 per cent confidence interval. Owing to discontinuities in the distribution of the test statistic and the fact that the end points of the confidence interval often coincide with observed values we must regard values outside this interval as ones indicating rejection at a 4.25 per cent significance level. We must also, if our exact level is critical, test the end values to see if they would be accepted or rejected at this significance level. Exact tests using StatXact show that we would find evidence against the hypothesized means of 2.5 and 16.0. Thus in this case all points in the open interval  $2.5 < \theta < 16.0$  would be acceptable, but the closed interval (2.5, 16.0) is a conservative 95 per cent interval, having a confidence level exceeding 95 per cent. In practice it is usual to cite a confidence interval

simply as the interval (2.5, 16.0) without being specific as to whether it is open or closed, bearing in mind that discontinuities inherent in the Wilcoxon test statistic distribution may imply (and it will vary from case to case) that the end points may or may not be included as **acceptable** values at a nominal significance level such as 5 per cent. One should not get excited about subtleties due to discontinuity at the end point of intervals. If relevant programs are available it may be worthwhile checking the actual confidence level associated with any quoted interval. It is left as an exercise to establish that a nominal 99 per cent confidence interval (i.e. with at least 99 per cent coverage) is the closed interval (0.5, 18).

*Conclusion.* An appropriate point estimator of the population median is 9.0 and actual 95.75 and 99.08 per cent confidence intervals are respectively (2.5, 16.0) and (0.5, 18), where it is understood that if values of  $\theta$  outside these intervals are specified in  $H_0$  there will be evidence against these at levels implied by the relevant  $P$ -values.

*Comments.* 1. A reasonable alternative point estimator to the median of the Walsh averages is the mean of all sample values. It is unbiased in the sense that if we sampled repeatedly and took the mean of the sample values as our estimate in all cases, the mean of these estimates would converge to the population mean. However, the median of the Walsh averages is in a certain technical sense, which we do not discuss here, generally closer to the population median; this estimator is called the **Hodges–Lehmann estimator**, having been proposed by Hodges and Lehmann (1963). A more formal justification for its use when basing the interval on signed ranks is given by Hettmansperger and McKean (1998, Section 1.3).

2. If confidence limits only are required we need write down only a few Walsh averages in the top left and bottom right of the triangular matrix in [Table 2.4](#). If a computer program is available for forming this matrix there is no problem in computing all the averages. If the matrix has to be formed manually once we have the Walsh averages in the first row, the required averages in the second row may be obtained by adding the same constant, namely half the difference  $x_2 - x_1$ , to the average immediately above it. To verify this denote by  $x_{ij}$  the average of  $x_i$  and  $x_j$ . This is the entry in row  $i$  and column  $j$  of the matrix of Walsh averages. Now for any  $k > 2$

$$x_{2k} = \frac{1}{2} (x_2 + x_k) = \frac{1}{2} (x_2 - x_1 + x_1 + x_k) = \frac{1}{2} (x_2 - x_1) + x_{1k}.$$

The idea generalizes for all subsequent rows, each entry in row  $r + 1$ , say, being obtained by adding the same constant  $\frac{1}{2} (x_{r+1} - x_r)$  to the entry immediately above it.

3. In [Example 2.4](#) we found when testing  $H_0: \theta = 15$  that  $S_+ = 14$ . In [Table 2.4](#) there are 14 Walsh averages exceeding 15.0.

*Computational aspects.* 1. Minitab includes a program for computing Walsh averages. It gives asymptotic confidence intervals (see [Section 2.5](#)). These may not be satisfactory for small  $n$ , especially when there are many ties. If one has the matrix of Walsh averages generated by some available program one may easily establish nominal 95 or 99 per cent confidence limits using tables of critical values of the statistic  $S$ .

2. StatXact includes a program giving the Hodges-Lehmann estimator and confidence intervals at any pre-assigned nominal level and confirms the values obtained above.

With ties in signed ranks the method for confidence intervals in this section needs modification. If computer software for confidence intervals in these circumstances is not available one might use the Walsh average method described above, then use an available computer program to study appropriate end point adjustments to those for the untied case along the lines indicated in [Example 2.3](#) for the Pitman interval. Such adjustments are usually small unless there is heavy tying.

### *Example 2.8*

*The problem.* Consider again the numbers of words in sentences used by R. A. Fisher with the sample given in [Example 2.6](#), i.e.

12 18 24 26 37 40 42 47 49 49 78 108

Obtain a confidence interval at a level not less than 95 per cent for the median sentence length based on the Wilcoxon signed-rank test.

*Formulation and assumptions.* We form the Walsh averages. We already know from [Example 2.7](#) that for samples of 12 in the no-tie case an interval with nominal 95 per cent coverage is that between the 14th smallest and 14th largest Walsh average. With a small number of ties one hopes these will be good approximations and this may be tested and adjustments made as indicated under *Procedure* if an exact hypothesis testing program is available.

*Procedure.* Walsh averages may be calculated using either some appropriate computer program or manually in the way used to obtain [Table 2.4](#). This is left as an exercise ([Exercise 2.8](#)). The 14th smallest average turns out to be 27.5 and the 14th largest to be 63 suggesting an interval (27.5, 63) slightly longer than that obtained in [Example 2.3](#) based on the Pitman test. Fine tuning of end points using a trial-and-error method with StatXact leads to an interval of (27.5, 62.5) as an exact 95.3 per cent interval (lower-tail  $P = 0.022$ , upper-tail  $P = 0.025$ ).

*Conclusion.* The 95.3 per cent confidence interval is (27.5, 62.5).

*Comments.* This interval differs little from those associated with the *t*-test or the Pitman test approach and is again strongly influenced by the outlying values of 78 and more especially 108. This may seem surprising after we intimated in the comments on [Example 2.6](#) that the hypothesis test result would not have been altered if these two observations were replaced by values appreciably more in line with the rest of the data. The reason these values influence the upper confidence limit is evident from a full table of Walsh averages ([Exercise 2.8](#)), for it will be apparent from this that the 14 highest averages all include at least one of the values 78, 108.

*Computational aspects.* Obtaining Walsh averages is reasonably straightforward even when no program is available. Ties create a situation where fine

tuning of the limits is best achieved by a trial-and-error procedure using an exact test program such as those in StatXact or Testimate, commencing with limits that ignore tying, i.e. that are appropriate to a no-tie case. However, if an exact program for a confidence interval such as that in StatXact is available fine tuning is not needed.

## 2.3 THE SIGN TEST

### 2.3.1 The sign test and the effect of sample size

We look more closely at assumptions and possible complications and extend concepts developed in [Example 1.2](#). For the sign test – like most tests – the larger the sample the higher the power and the shorter the confidence interval for a given confidence level.

A symmetry assumption is not needed and providing all sample  $x_i$  values are independent they need not all have the same distributions providing that they are from distributions with identical medians. Remember that for skew distributions the median does not equal the mean.

#### *Example 2.9*

*The problem.* The percentage water content of agricultural land often varies only slightly over an area such as a large field apart from a few local patches of exceptionally high or low percentages associated with especially poor or good drainage. The potential range of percentage water content is exemplified in a set of readings for nearly 400 soil quadrats on 22 June 1992 given by Gumpertz, Graham and Ristiano (1997). We use two small subsets (with values slightly rounded) from this large data set to illustrate some points about single sample nonparametric analysis. The use we make of these data has no direct relevance to the paper from which our samples were taken. The percentage water contents for our sampled quadrats, arranged in ascending order of magnitude, are:

Sample I	5.5	6.0	6.5	7.6	7.6	7.7	8.0	8.2	9.1	15.1
Sample II	5.6	6.1	6.3	6.3	6.5	6.6	7.0	7.5	7.9	8.0
	8.0	8.1	8.1	8.2	8.4	8.5	8.7	9.4	14.3	26.0

We use these data to test in each case  $H_0: \theta = 9$  against the alternative  $H_1: \theta \neq 9$  for the population median  $\theta$  using the sign test.

*Formulation and assumptions.* If the population median is 9, then any observation in a random sample is equally likely to be either above or below 9. Proceeding as in [Example 1.2](#) we associate a plus sign with *values greater than* 9, and if the hypothesis  $H_0: \theta = 9$  holds, the number of plus signs has a binomial distribution with  $p = \frac{1}{2}$ ,  $n = 10$ , i.e. a  $B(10, \frac{1}{2})$  distribution for the first sample and a  $B(20, \frac{1}{2})$  distribution for the second sample. For the alternative  $H_1: \theta \neq 9$ , a two-tail test is appropriate.

*Procedure.* In the first sample there are 2 values, 9.1 and 15.1, greater than 9, implying 2 plus signs. We gave a table of  $B(10, \frac{1}{2})$  probabilities in [Example 1.2](#) (p. 17) from which we easily calculate the probability of 2 or less or 8 or more plus signs is  $P = 0.109$ , a value confirmed by any computer program for the sign test. In the second sample there are 3 values, 9.4, 14.3 and 26.0 greater than 9. The probability of 3 or less plus signs for a  $B(20, \frac{1}{2})$  distribution is easily calculated although there is seldom any need to do this with the ready availability of computer software or tables for a sign test, and any relevant program should give  $P = 0.003$  for the relevant two-tail test.

*Conclusion.* With the sample of 10, since  $P = 0.109$  there is no convincing evidence against  $H_0$  since this clearly exceeds the conventional 5 per cent level. With a sample of 20 the evidence against  $H_0$  becomes very strong since  $P = 0.003$  corresponding to a 0.3 per cent significance level.

*Comments.* 1. We used a two-tail test since there was no a priori reason to confine alternatives to only greater or only less than the value specified in  $H_0$ .

2. The results are consistent with our remark in [Section 1.3](#) that increasing sample size generally increases the power of a test.

*Computational aspects.* With adequate tables of binomial probabilities when  $p = \frac{1}{2}$  for various  $n$ , computer programs might almost be regarded as optional extras for the sign test but StatXact and Testimate and nearly all major general statistics packages include programs for this or equivalent procedures.

If one or more sample values coincide with the value of  $\theta$  specified in  $H_0$  we cannot logically assign either a plus or minus to these observations. They may be ignored and the sample size reduced by 1 for each such value, e.g. if in the sample of 10 in [Example 2.9](#) we specify  $H_0: \theta = 8.0$  we treat our problem as one with 9 observations giving 3 plus signs. Doing this rejects evidence which strongly supports the null hypothesis, but which is uninformative about the direction of possible alternatives. Another proposed approach is to toss a coin and allocate a plus to a value equal to  $\theta$  if the coin falls heads and a minus if it falls tails: yet another is to assign a plus or minus to such a value in a way that makes rejection of  $H_0$  less likely. The coin-toss approach usually makes only a small difference and has little to commend it since it only introduces additional randomness, being in a sense more ‘error’; the last approach is ultraconservative. Lehmann (1975, p. 144) discusses pros and cons of these choices in more detail.

### 2.3.2 Confidence intervals

*Example 2.10*

*The problem.* Obtain approximate 95 per cent confidence intervals for the median water content using the sign test with the samples in [Example 2.9](#).

Determine the exact confidence level in each case.

*Formulation and assumptions.* We seek all values for the population median  $\theta$  that would not be rejected in a two-tail test at a 5 per cent significance level.

*Procedure.* We use the argument developed in [Section 1.4.1](#). Consider first the sample of 10. From the table given in [Example 1.2](#) (or any published tables or using a computer program that calculates probabilities for any given  $n, p$ ) we easily find that a critical region consisting of 0, 1, 9 or 10 plus signs has size  $2 \times (0.0010 + 0.0098) = 0.0216$  and that this is the largest exact size less than 0.05. Thus we reject  $H_0$  if it leads to 1 or fewer or 9 or more plus signs. We retain  $H_0$  if we get between 2 and 8 plus signs. It is clear from the sample values

5.5 6.0 6.5 7.6 7.6 7.7 8.0 8.2 9.1 15.1

in [Example 2.9](#) that we have between 2 and 8 plus signs if  $\theta$  lies in the open interval from 6 to 9.1, i.e.  $\theta > 6$  and  $\theta < 9.1$ . This implies an exact confidence level of  $100(1 - 0.0216) = 97.84$  per cent for the interval (6, 9.1). For any  $\theta$  not in this interval we would reject the hypothesis that it is the population median at the  $100 - 97.84 = 2.16$  per cent significance level.

We leave it as an exercise for the reader, using appropriate tables or a relevant computer program, to show that for the sample of 20 in [Example 2.9](#) the interval (6.6, 8.4) is an actual 95.86 per cent confidence interval.

*Conclusion.* For the sample of 10 an exact 97.84 per cent confidence interval for the median is (6, 9.1) and for the sample of 20 an exact 95.86 per cent interval is (6.6, 8.4).

*Comments.* 1. The sample median is an appropriate point estimator of the population median when we make no assumption of symmetry. For our samples these are respectively 7.65 and 8.0. The rationale for this choice is that it gives equal numbers of plus and minus signs, the strongest supporting evidence for  $H_0$ . Hettmansperger and McKean (1998, Section 1.3) justify this choice on even stronger theoretical grounds.

2. Increasing the sample size shortens the confidence interval.

3. The samples contain values 14.3, 15.1 and 26.0 that all appreciably exceed the median, suggesting a skew distribution of water content with a long upper tail. This is in line with the common experience that after rain, for example, one often finds a few small areas in a field excessively wet due to poor local drainage.

4. Skewness makes normal theory ( $t$ -distribution) tests or confidence intervals inappropriate. Remember that the mean and median do not coincide for a skew distribution. In passing we note that normal theory  $t$ -test based estimation for the sample of 10 in this example gives a mean of 8.13 with a 95 per cent confidence interval (6.2, 10.0) and for the sample of 20 a mean of 8.77 and a 95 per cent confidence interval (6.7, 10.9). Both intervals are slightly longer than those obtained using the sign test and lead to the somewhat bizarre result that the confidence interval for the larger sample is longer than that for the smaller one. Further they apply to the means rather than the medians and thus are shifted to the right, being symmetric about the sample mean; a moment's reflection will show this shift is a consequence of the presence of the high right-tail values 14.3, 15.1 and 26.0, i.e. those that suggest skewness.

*Computational aspects.* Given suitable tables there is little need for a computer program to calculate sign-test confidence intervals. Many general statistical packages will generate the relevant binomial probabilities if these are not available in tables. In Comment 4 above we considered  $t$ -distribution based 95 per cent confidence intervals. For direct comparison with the nonparametric intervals, many statistical packages provide parametric (normal theory,  $t$ -distribution) confidence intervals at any specified level such as 95.86 or 97.84 per cent if one wants to make closer comparisons at exact levels.

In Comment 4 above we remarked that  $t$ -distribution confidence intervals are not appropriate because the data are skew. We suggested when discussing in [Examples 2.3](#) and [2.8](#) the Fisher sentence length data that there is doubt there about the  $t$ -test based intervals. However, we found in those examples that the Pitman,  $t$ -test and Wilcoxon test based intervals were not very different. It is interesting to consider the sign-test based interval for those data.

### *Example 2.11*

Following arguments similar to those in [Example 2.10](#) and applying them to the sentence length data set given in [Example 2.2](#), i.e.

12 18 24 26 37 40 42 47 49 49 78 108

it is easily verified (using appropriate binomial tables when  $n = 12$ ,  $p = \frac{1}{2}$  or a suitable computer program) that the interval  $(24, 49)$  is an exact 96.14 per cent confidence interval and this is appreciably shorter than the interval  $(27.5, 62.5)$  obtained in [Example 2.8](#). The sign test does not assume a symmetric distribution and the consequence when that assumption is violated is sometimes (but not always) that it has higher Pitman efficiency resulting in greater power in the hypothesis testing context which often translates to shorter confidence intervals. However, where a symmetry assumption is valid the sign-test approach usually leads to wider confidence intervals than either the Pitman or Wilcoxon test methods. An exception is certain symmetric distributions with very long tails, e.g. the double exponential, where the sign test has higher Pitman efficiency than the signed-rank test and may give a shorter confidence interval.

## 2.4 TRANSFORMATION OF RANKS

### 2.4.1 Normal scores

In essence transformation of a sample of  $n$  continuous ordered data to ranks replaces the sample values by something like values we would expect when sampling from a uniform or rectangular distribution over  $(0, n)$  – more precisely the rank  $r$  corresponding to the  $r/(n + 1)$ th quantile of that distribution. Might further

transformation of these ranks increase the Pitman efficiency or have other advantages?

In many parametric analyses a normality assumption is a key feature and when it can validly be made it often leads to mathematical and computational simplicity. In [Section 2.1](#) we indicated that exact permutation tests were only practical for small to medium sample sizes even with the best software. We show in [Section 2.5](#) how we sometimes face this difficulty using asymptotic results which themselves carry an assumption of normality. It is not unreasonable to hope that if we start with data that resemble a sample from a normal distribution that such asymptotic or ‘large sample’ results will approach exact results more rapidly in the sense that they may be reasonable for not very large samples.

Such hopes inspired a study of transformations to give data more like those from a normal distribution as a basis for exact tests while retaining some of the desirable properties such as the test being unconditional on the actual data as was the case for the Wilcoxon signed-rank test (but not for the Pitman test). A possible transformation is suggested by our remark that ranks correspond to quantiles of a uniform distribution. Why not transform ranks to values (often called scores) corresponding to quantile values for a normal distribution? More specifically, why not choose the standard normal distribution, for then it is easy to make the transformation using tables of the standard normal **cumulative distribution function** (which we now abbreviate to cdf). For example, if we have the very nonnormal sample 2, 3, 7, 21, 132 we first replace these values by the ranks 1, 2, 3, 4, 5 and transform these to normal scores which are the  $1/6$ th,  $2/6$ th,  $3/6$ th,  $4/6$ th and  $5/6$ th quantiles of the standard normal distribution. These may be read from standard normal cumulative distribution function (cdf) tables, but many software programs compute these as one of a set of possible data transformations. The normal score corresponding to the  $1/6 = 0.1667$  quantile is the  $x$  value such that the standard normal cdf, commonly denoted by  $\Phi(x)$ , is  $\Phi(x) = 0.1667$ . From tables (e.g. Neave, 1981, p.18) or from software we find  $x = -0.97$ . For the  $2/6 = 0.3333$  quantile we find  $x = -0.43$  and for the  $3/6$ th (mean or median)  $x = 0$ . By symmetry the remaining quantiles are  $x = 0.43$  and  $x = 0.97$ . These quantile scores are often called **van der Waerden scores** having been proposed by van der Waerden (1952; 1953).

Alternatives to the above scores are discussed by Conover (1999, Section 5.10) and others. These include **expected normal scores** where the  $i$ th ordered sample value in ascending order, sometimes

called the  $i$ th order statistic, is replaced by the expectation of the  $i$ th order statistic for the standard normal distribution. The  $i$ th order statistic is often written  $x_{(i)}$ . Fisher and Yates (1957, Table XX) give expected normal scores corresponding to ranks for  $n \leq 50$ . In practice using van der Waerden or expected normal scores usually lead to similar conclusions, but both need adapting for use in the one sample situation. We consider here only van der Waerden scores.

Direct application of van der Waerden scores in an analogue to the Wilcoxon signed-rank test is not possible because the rationale of the test demands that we allocate signs to **magnitudes** of ranks but van der Waerden scores are of equal magnitudes but opposite signs. One practical way around this difficulty is first to add a constant  $k$  to all van der Waerden scores so that they are nonnegative. The choice  $k = 3$  achieves this for sample sizes less than 700. Adding 3 in effect gives normal scores for a distribution of mean 3 and standard deviation 1. The next step is to attach a negative sign to the score corresponding to a negative deviation. Programs for raw data permutation tests analogous to the Pitman test for the original data may be used with these scores now treated as the raw data.

### *Example 2.12*

*The problem.* Use the normal scores test procedure outlined above to test the median hypothesis  $H_0 : \theta = 30$  against  $H_1 : \theta > 30$  for the sentence length data given in [Example 2.2](#), *viz*:

12 18 24 26 37 40 42 47 49 49 78 108

*Formulation and assumptions.* We replace the signed ranks in [Example 2.6](#) by van der Waerden scores, then add three to each and perform a randomization test on these regarding them as raw data.

*Procedure.* Statistical software may be used to form the scores and add the constant  $k = 3$ . In StatXact (or most software programs) this is a one-step process but for completeness we give the relevant scores here. The unmodified van der Waerden scores in order of increasing absolute ranks of differences used in the Wilcoxon test in [Example 2.6](#) are:

−1.43 −1.02 −0.74 −0.50 −0.19 −0.19 0.10 0.29 0.62 0.62 1.02 1.43

Adding 3 to each of the above scores gives a set of all positive scores and we then assign negative signs to those that correspond to negative deviations, giving the modified scores

−1.57 −1.98 2.26 2.50 −2.81 2.81 3.10 −3.29 3.62 3.62 4.02 4.43

The raw data permutation test (effectively a Pitman test) applied to these scores as data gives a one-tail  $P = 0.062$ , not in close agreement with the values in [Examples 2.2](#) and [2.6](#).

*Conclusion.* There is only fairly weak evidence against  $H_0$  and in favour of  $H_1$ .

*Comments.* 1. The result is somewhat out of line with common experience that results using these scores usually differ little from those using the Wilcoxon test. In [Section 2.5](#) we consider the situation regarding asymptotic tests for each.

2. Discontinuities in possible  $P$ -values tend generally to be smaller for this test than for the Wilcoxon test because the test statistic used here takes more possible values.

3. The distribution is symmetric in the sense that we may use either the sum of the positive scores or the sum of the negative scores as our test statistic. Ties have been dealt with in this example by assigning quantiles corresponding to the mid-ranks. An alternative sometimes recommended is to assign instead the mean of the quantiles corresponding to the ranks that are tied, e.g. if there is a tie for the  $r$ th and the  $(r + 1)$ th ordered data values we take the mean of the quantiles corresponding to each. Generally the difference between these approaches is small.

4. So far as we know there is no completely satisfactory way of getting a confidence interval based on these scores. Not only do the same difficulties arise as in the Pitman test situation for determining end points but the interval pertains to the transformed differences and there appears to us no satisfactory way of back-transforming to the original data. Problems of this sort arise in many data transformation situations both in parametric and nonparametric inference.

5. Normal score tests in this and other situations met in later chapters have Pitman efficiency 1 when the normal theory  $t$ -test is appropriate and in any other case the Pitman efficiency is at least 1. Any euphoria generated by this property must be tempered by a realization that there are other tests with Pitman efficiency greater than 1 relative to normal theory tests when certain other appropriate distributional assumptions are made and also that an asymptotic result does not guarantee equivalent small sample efficiency, although Pitman efficiency is usually a good guide. We remarked in Comment 3 on [Example 2.6](#) that these data suggest skewness.

So far as we are aware the procedure in [Example 2.12](#) was first formally described in the second edition of this book but it is so straightforward we would be surprised if it or something similar has not been used earlier. An alternative method for obtaining positive scores to which appropriate signs may be allocated is to use only some of the positive quantiles (i.e. those lying above the median) of the standard normal distribution as scores. Appropriate signs are then attached to these. Recommended quantiles corresponding to the ranks 1, 2, 3, . . . ,  $n$  are the  $(n + 2)/(2n + 2)$ th,  $(n + 3)/(2n + 2)$ th,  $(n + 4)/(2n + 2)$ th, . . . ,  $(2n + 1)/(2n + 2)$ th quantiles. The attached signs are those for Wilcoxon signed ranks. One then proceeds using these scores as though they were the original data in a Pitman-type test. The fact that these quantiles are not symmetrically distributed about their median detracts from the intuitive appeal of the

transformation, but again our experience with a few examples is that it generally gives results similar to a Wilcoxon signed-rank test. Maritz (1995, Example 2.15) gives an example using these scores or closely related scores based on expected values and proceeds to discuss hypothesis testing and point estimation using these and even more general scores.

#### 2.4.2 Other scores

Scores other than normal scores are sometimes appropriate. We use some of these in more sophisticated problems. They are needed when data that are essentially nonnormal occur or when there are complications due to censoring of data. An example is given in [Section 5.5.1](#). In analyzing survival data what are called Savage or exponential scores are often appropriate. We introduce these in [Section 5.5.2](#) and they are also discussed by Lehmann (1975, p. 103) and by Sprent (1998, Section 4.9).

### 2.5 ASYMPTOTIC RESULTS

Widely available software for exact permutation tests for small to medium-sized samples removes the need to use asymptotic results for such samples where one knows these may not be reliable unless samples are moderately large (how large depending partly upon the procedure being used).

All centrality test and estimation procedures given so far in this chapter – Pitman, Wilcoxon, sign test, rank-transformations – have been based on statistics that are the sum of certain scores, e.g. the sum of the positive deviations from a hypothesized mean or median  $\theta_0$  in the Pitman test, the sum of the positive signed ranks in the Wilcoxon test, the number of positive signs in the sign test (i.e. the sum of positive scores if each deviation is scored as +1 or -1). Here the deviations, the signed ranks, the  $\pm 1$ s and likewise the signed modified van der Waerden scores are each relevant scores for a particular procedure.

In general for a sample of  $n$  observations our procedures have been based on the sum,  $S$ , of positive (or negative) scores where the score allocated to observation  $i$  may in general be denoted by  $s_i$ ,  $i = 1, 2, \dots, n$ . Under the null hypothesis a key feature is that each score  $s_i$  is equally likely to have a positive or negative sign. It follows ([Exercise 2.16](#)) that any such statistic  $S$  has mean value  $\frac{1}{2}\sum|s_i|$  and

variance  $\frac{1}{4}\sum s_i^2$  where the summation is over all  $n$  scores. From the central limit theorem it follows that for large  $n$  the distribution of

$$Z = \frac{S - \frac{1}{2}\sum |s_i|}{\frac{1}{2}\sqrt{\sum s_i^2}} \quad (2.1)$$

is approximately standard normal.

Formula (2.1) simplifies for a number of special cases ([Exercise 2.16](#)). For example, with no rank ties it is easy to verify for the Wilcoxon signed-rank test  $E(S_+)$  or  $E(S_-)$  reduces to  $\frac{1}{4}n(n + 1)$  and the  $\text{var}(S_+)$  or  $\text{var}(S_-)$  to  $n(n + 1)(2n + 1)/24$ . The formula in this case can be improved by a continuity correction that allows for the fact that we are approximating to a continuous distribution with a discrete distribution where  $S$  can take only integer values. If  $S$  is the smaller of the sums of the positive or negative ranks the appropriate correction is to replace  $S$  by  $S + \frac{1}{2}$ , while if  $S$  is the larger sum it is replaced by  $S - \frac{1}{2}$ . When there are tied ranks it is easiest to revert to the general formula (2.1) using the tied ranks as scores. If the statistic  $S_W = |S_+ - S_-|$  is used it is easily verified that now (2.1) reduces to  $Z = S_W/\sqrt{[\sum(s_i^2)]}$  and if there are no ties the denominator reduces to  $\sqrt[n]{[n(n+1)(2n+1)/6]}$ . For the Wilcoxon test the asymptotic approximation is usually satisfactory for  $n \geq 20$ , and sometimes for smaller values of  $n$  also.

For the sign test where  $S$  is the number of observations greater than (or less than) the hypothesized median it is easily verified that  $E(S) = \frac{1}{2}n$  and  $\text{var}(S) = \frac{1}{4}n$  whence (2.1) reduces to

$$Z = \frac{S - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}. \quad (2.2)$$

Again the approximation is improved by a continuity correction replacing  $S$  by  $S + \frac{1}{2}$  if  $S$  is less than its mean and by  $S - \frac{1}{2}$  otherwise.

Programs giving asymptotic results for many standard nonparametric tests are included in most major statistical software packages and as options in specialist programs like StatXact and Testimate. They work well for large samples but the approximations may or may not be good for medium or small samples. In [Examples 2.13](#) and [2.14](#) we consider some uses of asymptotic results in testing and estimation, comparing them where possible with exact results. Manual computation is possible but tedious if software is not available to carry out asymptotic tests automatically.

### Example 2.13

*The problem.* For the data sets on water content of soil used in [Example 2.9](#), viz.

5.5 6.0 6.5 7.6 7.6 7.7 8.0 8.2 9.1 15.1

and

5.6 6.1 6.3 6.3 6.5 6.6 7.0 7.5 7.9 8.0 8.0 8.1 8.1 8.2  
8.4 8.5 8.7 9.4 14.3 26.0

explore the use of asymptotic results for the Pitman, Wilcoxon and sign tests for testing the hypothesis that the median content is  $\theta = 9$  against a two-tail alternative. Compare results with those for exact tests where possible.

*Formulation and assumptions.* Any assumptions of symmetry implicit for relevant exact tests also apply to asymptotic tests. Generally increasing sample size will bring the asymptotic result closer to the exact result.

*Procedure.* Although manual computation of  $Z$  is feasible using (2.1) or (2.2) as appropriate, we assume relevant statistical software is available. For testing  $H_0: \theta = 9$  against  $H_1: \theta \neq 9$  StatXact gives the following exact and asymptotic  $P$ -values:

Test	Sample of 10		Sample of 20	
	Exact	Asymptotic	Exact	Asymptotic
Pitman	0.352	0.305	0.916	0.817
Wilcoxon	0.102	0.092	0.015	0.017
Sign	0.109	0.057	0.0026	0.0017

StatXact does not use the continuity correction for the sign test. That has quite a large effect in these examples. For instance for the sample of 20 its use increases the two-tail asymptotic  $P$  to 0.0037, but this is not closer to the exact value, whereas surprisingly for the smaller sample it does better, increasing  $P$  to 0.114 which is closer to the exact value. This indicates the need for caution about using asymptotic values even for moderate  $n$ . It is particularly in the tails (i.e. for low values of  $P$ ) that asymptotic values sometimes show the greatest relative departure from exact probabilities. When there are ties a continuity correction is not usually used for the Wilcoxon test as it has little theoretical justification, although in practice it sometimes tends to bring results closer to those for the exact permutation test. The  $t$ -test  $P = 0.33$  and 0.82 are close to the values for the Pitman test.

*Conclusion.* The data in both samples are, as already pointed out, slightly skewed to the right, especially in the sample of 20. Both the Pitman and Wilcoxon tests indicate lower efficiency relative to the sign test in the larger sample, but the effect of skewness is less for the Wilcoxon test than it is for the Pitman test because the transformation to ranks does not give the more extreme values of 14.3, 15.1 and 26.0 the influence they have in the Pitman test. Indeed they would have had the same rankings had they had the values 12.6, 12.7 and 12.8.

*Comment.* In view of the conclusion that our rankings would not have been altered if the three largest observations had been replaced by 12.6, 12.7 and 12.8 when there would be only a small indication of skewness, it is interesting to see how the Pitman test would perform if these values are used in it. Intuition suggests that the  $P$  values should then be nearer those given by the Wilcoxon test (which, of course, are unaltered by the change). Indeed with this change the exact  $P$ -value for the Pitman test reduces to  $P = 0.1230$  for the smaller sample and to  $P = 0.0355$  for the larger, both closer to the results for the Wilcoxon test. This should cause no surprise because there is now little indication of skewness and so a key validity requirement for both tests is nearer to being satisfied.

*Computational aspects.* The wide availability of software to carry out asymptotic tests makes them appealing. Three notes of caution are:

- (i) Large sample sizes are needed to ensure consistently good approximations, especially in the tails, so beware of asymptotic tests when  $n < 20$ .
- (ii) A breakdown of an assumption such as a requirement for symmetry may make asymptotic results of little value (as is also true for exact tests).
- (iii) Practice varies regarding use of continuity corrections; it is usually easy to check whether or not these are used either from documentation included with software or from the given output which often includes the test statistic value together with its mean and standard deviation.

Asymptotic confidence intervals are obtainable using the form of (2.1) relevant to the Wilcoxon test without ties or for the sign test using (2.2). For the Wilcoxon test the asymptotic result enables us to obtain an approximate value of  $S$  for significance at a suitable level and Walsh differences may then be used to get the approximate confidence interval. This works quite well if there are no ties and it gives a starting point for trial-and-error refinement when there are many ties. For the sign test we can estimate the number of positive or negative signs that determine the end points of our interval. An example shows how this works.

#### *Example 2.14*

*The problem.* For the larger data set in Example 2.13 use the asymptotic formulae for the Wilcoxon test and that for the sign test to establish approximate 95 per cent confidence intervals for the population median.

*Formulation and assumptions.* The key to establishing end points of 95 per cent confidence intervals is that if the statistic used for computing  $P$ -values for some hypothetical value of the median  $\theta$  is such that  $|Z| = 1.96$  then that  $\theta$  is the end point of the approximate confidence interval.

*Procedure.* Consider first the Wilcoxon test. For the sample of 20 observations in Example 2.13 either direct computation or use of statistical software for the asymptotic test establishes that the statistic  $S$  has mean 105 and standard deviation 26.78. Thus, if for some hypothesized mean a calculated  $S$  is such that

$$(S - 105)/26.78 = \pm 1.96$$

it follows that such  $S$  are appropriate values for calculating the end points of the confidence interval using Walsh averages. Solutions are  $S = 52.82$  and  $157.18$ . For practical purposes we round these to 53 and 157. In fact we only need calculate one of these in view of the symmetry in counts of positive and negative ranks; the smaller value 53 indicates that we should reject the 53 largest and 53 smallest Walsh averages. If no program is available to compute these averages it is only a matter of tedious arithmetic computing the more extreme averages to establish that the approximate interval is  $(7.15, 8.45)$ .

For the sign test it follows immediately from (2.2) that when  $n = 20$  the mean value of the test statistic (the number of positive signs or of negative signs) is 10 and the standard deviation is  $\frac{1}{2}\sqrt{20} = 2.236$  whence the lower limit is the value  $S$  such that  $(S - 10)/2.236 = -1.96$ , i.e.  $S = 5.62$ . Rounding to the nearest integer suggests that we should reject  $H_0$  if there are 6 or fewer positive or negative signs leading for these data to the approximate interval  $(7, 8.2)$ .

*Conclusion.* The asymptotic results give approximate 95 per cent confidence intervals based on the Wilcoxon statistic as  $(7.15, 8.45)$  and based on the sign test statistic as  $(7, 8.2)$ .

*Comments.* 1. The 95 per cent confidence interval based on the  $t$ -test statistic is  $(6.7, 10.9)$ ; this is considerably longer than those obtained in this example and reflects the inappropriateness of the  $t$ -test procedure when data are clearly skew. Despite the fact that we found in Examples 2.3 and 2.8 for a different data set that the Wilcoxon based intervals may be similar to the  $t$ -test based intervals in the presence of skewness, its performance here is better. In practice how well the Wilcoxon method performs in estimation problems depends upon the number of outlying observations in the sample as these influence only the more extreme Walsh averages. Only a few outliers when  $n$  is large are less influential than the same number of outliers for smaller  $n$ .

2. Given programs for exact tests such as that in StatXact it is easy to determine the exact coverage of the intervals and also to adjust the limits by a trial and error method to give better coverage. In this example the Wilcoxon interval  $(7.15, 8.45)$  is associated with a lower-tail  $P = 0.0235$  and an upper-tail  $P = 0.030$ . Although one is below and the other above 0.025 the interval is reasonable for most practical purposes. For the sign-test interval with these end points some observations coincide with the limits so it is useful to perform an exact test for  $P$ -values taking values just above and below these end points. Using StatXact we find the interval  $(6.95, 8.25)$  has only an 88.46 per cent coverage but the interval  $(7.05, 8.15)$  has an even lower 73.7 per cent coverage. Again there is an indication here that one needs to be cautious about asymptotic results. Use of a continuity correction in determining the asymptotic limits may help. In Example 2.10 we established the 95 per cent interval based on the sign test was  $(6.6, 8.4)$ .

*Computational aspects.* Although programs giving exact  $P$ -values or good Monte Carlo approximations thereto are needed to refine asymptotic limits, limits based on asymptotic theory are usually reasonable for moderately large samples and can be computed using any program that deals adequately with the relevant asymptotic nonparametric tests. Some programs (e.g. Minitab) quickly

generate tables of Walsh averages, so if the Wilcoxon intervals are required once  $S$  corresponding to  $Z = \pm 1.96$  has been computed it is easy to get the Walsh averages corresponding to end points, although Minitab will itself give asymptotic limits for the Wilcoxon test and exact limits for the sign test.

## 2.6 ROBUSTNESS

In [Example 2.14](#) we found an approximate 95 per cent confidence interval for the population mean using the Wilcoxon signed-rank procedure was (7.15, 8.45) while that based on the sign test was displaced slightly to the left and was slightly shorter being (7, 8.2). However, the confidence levels were not quite identical. The interval (6.7, 10.9) based on the normal theory  $t$ -test is appreciably longer, reflecting the inappropriateness of the  $t$ -test approach when applied to obviously skew data. On the other hand one commonly meets situations where the sign-test interval is longer than that given by the Wilcoxon test which in turn is longer than that based on the  $t$ -test. In such examples although there might be some evidence of departure from normality, if the departure is not severe and there is not much evidence of skewness these results are in general agreement with the finding based on Pitman efficiency that when a sample comes from a normal distribution the efficiency of the  $t$ -test is greater than that of any other test. The Pitman efficiency of the Wilcoxon test can be shown to be  $3/\pi$  (approximately 95.5 per cent) while that of the sign test is  $2/\pi$  (approximately 63.7 per cent) when a sample is from a normal distribution. Two kinds of departure from normality that are common are one where the sample may come from some distribution such as the exponential or gamma distribution which are known to be asymmetric, or one where the sample comes from a mixture of distributions. Mixtures may arise in several ways. For example our data might be results of chemical analyses carried out in two different laboratories, one of these giving less precise results than the other and perhaps also incorporating a consistent error (called bias) due to an equipment fault. Mixtures may also occur in accurate data. We explained why this was so for the data in [Example 2.10](#) (*Comment 3*). Another way skew or long-tail distributions may occur is by the intrusion of rogue observations often called outliers. If we know why they are rogues there may be a good case for rejection, but often we do not know if suspect observations are really rogues (e.g. incorrectly recorded readings, observations on units that are not members of the population of interest, etc.) or if we are sampling from a population where a small

proportion of outlying values is the norm. While many parametric tests like the  $t$ -test are sensitive to certain departures from assumptions, the Wilcoxon test may be less so because it makes fewer assumptions, and the sign-test even less sensitive because it in turn makes even fewer assumptions. These latter tests exhibit a property called **robustness** which we discuss more fully in [Chapter 11](#). Test and estimation procedures are robust if they are little influenced by fairly blatant departures from assumptions. In [Example 2.15](#) we indicate that the Wilcoxon test may be more robust than the  $t$ -test if there is a marked departure from symmetry. We also see that in such circumstances the sign test, not unexpectedly, may do even better since it makes no symmetry assumption.

### *Example 2.15*

Suppose that the data in [Example 2.4](#), i.e.

−5 −2 1 3 4 8 9 12 16 17 20 25

are amended by omitting the observation 25 and replacing it (i) by 35 and (ii) by 65. These changes induce obvious upper-tail skewness. [Table 2.5](#) gives the nominal 95 per cent and 99 per cent confidence intervals based on the normal theory  $t$ , the Wilcoxon signed-rank statistic and the sign test in each case.

For the moderately skewed set (i) we have a mixed picture, the  $t$ -test does best at the 95 per cent level and the Wilcoxon test at the 99 per cent level but the other methods do nearly as well in each case. For the more markedly skewed set (ii), the  $t$  limits are highly unsatisfactory, and those based on Wilcoxon are moderately satisfactory at the 95 per cent level but not satisfactory at the 99 per cent level. However, the binomial-based sign-test limits are the same as those for set (i) and are not influenced by the extreme observation at 65. For set (ii) they are the same width as the Wilcoxon limits at the 95 per cent level although the exact coverage (96.14 per cent) is slightly greater than the exact Wilcoxon coverage (95.5 per cent) but this is more a reflection of the discontinuities in possible available levels than anything else. At the 99 per cent level the sign test limits are the clear winner. This is not surprising since no symmetry assumption is needed.

Are such extreme values as that in data sets (i) and (ii) realistic from a medical point of view? A change in heart rate of 35 beats per minute might just be possible but the value of 65 in set (ii) is a data error. It is good statistical practice to check data carefully to eliminate errors, but one often has to analyze data where there is no indication that an extreme value (often referred to as an outlier) is not a correct observation. This is when one may do better with the simple sign test than with a test requiring more stringent assumptions. On the other hand it is generally inappropriate to use the sign test when there is little or no evidence of skewness.

**Table 2.5** 95 and 99 per cent confidence intervals for data sets (i) and (ii) in [Example 2.15](#) based on normal theory ( $t$ ), Wilcoxon signed ranks ( $W$ ) and the binomial sign test ( $B$ ). For each set the shortest interval at each confidence level is indicated by an asterisk.

		<i>95 per cent</i>	<i>99 per cent</i>
<i>Data set (i)</i>	$t$	(2.79, 16.88)*	(−0.12, 19.79)
	$W$	(2.5, 17.0)	(0.5, 20.0)*
	$B$	(1.0, 17.0)	(−2.0, 20.0)
<i>Data set (ii)</i>	$t$	(0.70, 23.97)	(−4.10, 28.77)
	$W$	(2.5, 18.5)*	(0.5, 34.5)
	$B$	(1.0, 17.0)*	(−2.0, 20.0)*

## 2.7 FIELDS OF APPLICATION

### Insurance

The median of all motor policy claims paid by a company in 1999 is £870. Early in 2000 the management thinks claims are higher. To test this, and to estimate the likely rise in mean or median, a random sample of 25 claims is taken. The distribution of claims will almost certainly be skew, so a sign test would be appropriate.

### Medicine

The median systolic blood pressure of a group of boys prior to physical training is known. If the blood pressure is taken for a sample after exercise the sign test could be used to test for a shift in median. Would you consider a one- or a two-tail test appropriate? If it appears reasonable to assume a symmetric distribution a Wilcoxon test or even a normal theory  $t$ -test is more appropriate. Even if an assumption of symmetry in systolic blood pressures before exercise is reasonable, this may not be so after exercise. For instance, the increase after exercise might be relatively higher for those above the median blood pressure at rest and this could give rise to skewness. A physiologist asking questions about changes should know if this is likely.

### Engineering

Median noise level under the flight path to an airport might be known for aircraft with a certain engine type (the actual level will

vary from plane to plane and from day to day depending on weather factors, the precise height each plane flies over the measuring point, etc.). If the engine design is modified a sample of measurements under similar conditions to that for the old engine may indicate a noise reduction. A one-tail test would be appropriate if it were clear the modification could not increase noise. We are unlikely to be able to use a true random sample here, but if taken over a wide range of weather conditions, the first, say, 40 approaches using the new engine may broadly reflect characteristics of a random sample.

## **Biology**

Heartbeat rates for female monkeys of one species in locality A may have a symmetric distribution with known mean. Given heartbeat rates for a sample of similar females from locality B, the Wilcoxon test could be used to detect a shift or to obtain confidence limits for the true mean.

## **Education**

A widely used test of numerical skills for 12-year-old boys gives a median mark of 83. A new method of teaching such skills is used for a class of 27 such pupils. The asymptotic approximation to the Wilcoxon test could be used to test for a shift in median if symmetry could be assumed. If not, a sign test would be preferred.

## **Management**

Records give the mean and median number of days absent from work for all employees in a large factory for 1998. The number of days absent for a random sample of 20 is noted in 1999. Do you think such data are likely to be symmetric? In the light of the answer to this question one may select the appropriate test for indications of changes in the absentee pattern.

## **Geography and environment**

Estimates of the amount of cloud cover at the site of a proposed airport are taken at a fixed time each day over a period. The site might be rated unsuitable if the median cover were too high. A confidence interval for the true median would be useful.

## **Local differences**

A psychologist is told that the ‘national average IQ of drug abusers aged between 16 and 18 is 103’. He assesses the IQ of a sample of

abusers in that age group from an area where drug abuse is rife. He might use a Wilcoxon test to assess whether it is reasonable to assume the mean is 103 for that area. It is not unusual in a number of contexts to find that *national averages* applicable to a large area do not apply locally. In the UK, for instance, the average price per litre of petrol in London is very different from that in the Scottish Highlands and both (especially the latter) differ from the nationwide average. The average price in euros of a litre of milk in Italy, France and Germany may each differ from the average for the whole European community.

## Industry

The median time people stay in jobs in a large motor assembly plant in Germany is known to be 5.2 years. Employment times for all 55 employees who have left a UK plant in the last 3 months are available. Although this may not be a random sample it may be reasonable to use the data to test whether the median time for UK workers is also 5.2 years. If the distribution appears skew a sign test would be appropriate.

## Combining information from different sources

Tests are often carried out in more than one laboratory to determine, for example, the mean or median level of some impurity in a product. If it is reasonable to consider that the population means are equal for all the material sent for analysis and the laboratory means are centred about this value even if the distributions of the test results differ in other respects an appropriate nonparametric test of the combined data may be used depending upon whether assumptions of symmetry are or are not appropriate.

## Astronomy

Astronomers frequently estimate quantities such as the mass of a particular star or the diameter of a planet in several observatories each using non-identical equipment. The resulting measurements are often far from normally distributed, forming either a longer tailed or skewed distribution of readings. Providing reasonable assumptions can be made about the nature of measurement errors nonparametric methods of analysis may often be appropriate. Some comments on this situation were made in [Section 2.1.2](#).

## 2.8 SUMMARY

**Pitman test.** A suitable test statistic ([Section 2.1.2](#)) is the sum  $S_+$  of the positive data deviations from a hypothesized mean or median  $\theta$  assuming the sample is from a symmetric distribution. Appropriate computer software is required for significance tests and even with such software confidence limits can only be obtained by a cumbersome trial-and-error approach. The procedure lacks robustness and often gives similar results to a  $t$ -test even when the latter is not appropriate. It is conditional upon each particular data set and is not widely used in practice.

**Wilcoxon signed-rank test.** The test statistic  $S_+$ , the sum of the positive signed-rank differences from the mean or median specified in  $H_0$  ([Section 2.2.1](#)), is widely used. Tables, or more satisfactorily, appropriate computer software may be used for significance tests. Ties require special treatment ([Section 2.2.3](#)) and suitable computer software is then needed to determine exact permutation distributions.

**Walsh averages** ([Section 2.2.4](#)) provide an alternative hypothesis testing procedure and they are particularly useful for estimation and calculating confidence intervals. An assumption of population symmetry is needed for validity of test and estimation procedures. Asymptotic normal approximations ([Section 2.5](#)) work well for sample sizes  $n \geq 20$  if there are no (or very few) ties; modification of the denominator in the test statistic is required for numerous ties even for relatively large  $n$ .

**Sign test.** The test statistic is the number of observations above or below a median specified in  $H_0$  (see e.g. [Example 2.9](#)). Significance is determined using binomial probability tables for  $p = \frac{1}{2}$  and various  $n$  or else from a suitable computer program. The method is easily used to obtain confidence intervals ([Section 2.3.2](#)). Normal theory approximations ([Section 2.5](#)) may be used for sample sizes greater than 20. Modifications are required to deal with values equal to the median specified in  $H_0$  ([Section 2.3.1](#)). No assumption of symmetry is required.

**Normal scores.** These scores ([Section 2.4](#)) aim to make data more like samples from a normal distribution. Test results are usually not very different from those given by the Wilcoxon test. There appear to be no simple satisfactory methods for obtaining confidence intervals using these transformations.

## EXERCISES

- 2.1 For the data in [Example 2.1](#) carry out a test of  $H_0: \theta = 80$  against  $H_1: \theta \neq 80$  using the Pitman test.
- 2.2 In Comment 3 on [Example 2.1](#) we asserted that the usual  $t$ -statistic could be used in place of  $S_+$  as the test statistic for the Pitman test because there was a one-to-one correspondence between the ordering of the two statistics. Establish that this is so. [Hint: show that the denominator of the  $t$ -statistic is invariant under all permutations of the signs of the  $d_i$ .]
- 2.3 In Comment 3 on [Example 2.4](#) we suggested that for a variety of reasons one should be cautious about extending inferences about heartbeat rates for female students to the population at large. What might some of these reasons be?
- 2.4 Using the data in [Table 2.1](#) for the distribution of the Wilcoxon  $S$  when  $n = 7$ , construct a bar chart like that in [Figure 2.1](#) showing the probability function for  $S$ . Discuss the similarity, or lack of similarity, to a normal distribution probability density function.
- 2.5 Verify the  $P$ -value quoted in Comment 3 in [Example 2.6](#) for the data modified in the way suggested in that comment.
- 2.6 Establish that the permutation distribution of the Wilcoxon signed-rank statistic for testing the hypothesis  $H_0: \theta = 6$ , given the observations 4, 4, 8, 8, 8, 8 has a distribution equivalent to that for the sign test of the same hypothesis. Would this equivalence hold if the null hypothesis was changed to  $H_0: \theta = 7$ ?
- 2.7 Establish nominal 95 per cent confidence intervals for the median based on the Wilcoxon signed-rank test for the following data sets. If an appropriate computer program is available use it to comment on the discontinuities at the end points of your estimated intervals based on Walsh averages.

*Set I*    1, 1, 1, 1, 1, 3, 3, 5, 5, 7, 7  
*Set II*   1, 2, 2, 4, 4, 4, 5, 5, 5, 7

- 2.8 Form a table of Walsh averages for the Fisher sentence length data given in [Example 2.8](#) and use it to confirm the confidence interval obtained in that example and also to obtain an approximate 99 per cent confidence interval.
- 2.9 The numbers of pages in the sample of 12 books referred to in [Exercise 1.2](#) were

126 142 156 228 245 246 370 419 433 454 478 503

Use the Wilcoxon signed-rank test to test the hypothesis that the mean number of pages in the statistics books in the library from which the sample was taken is 400. Obtain a 95 per cent confidence interval for the mean number of pages based on the Wilcoxon test and compare it with the interval obtained on an assumption of normality.

- 2.10 Apply the sign test to the data in [Example 2.4](#) for the hypotheses considered there.
- 2.11 For the sample of 20 in [Example 2.9](#) if  $\theta$  is the population median test the hypothesis  $H_0: \theta = 9$  against the alternative  $H_1: \theta \neq 9$  using the sign test by

computing any relevant binomial probabilities directly from the binomial probability formula. Also perform the test for  $H_0: \theta = 7.5$  against the alternative  $H_1: \theta > 7.5$ .

- 2.12 Before treatment with a new drug 11 people with sleep problems have a median sleeping time of 2 hours per night. A drug is administered and it is known that if it has an effect it will increase sleeping time, but some doctors doubt if it will have any effect. Are their doubts justified if the hours per night slept by these individuals after taking the drug are:

$$3.1 \ 1.8 \ 2.7 \ 2.4 \ 2.9 \ 0.2 \ 3.7 \ 5.1 \ 8.3 \ 2.1 \ 2.4$$

- 2.13 Assuming symmetry, carry out a test relevant to the situation and data in Exercise 2.12 using a normal scores procedure.
- 2.14 Kimura and Chikuni (1987) give data for lengths of Greenland turbot of various ages sampled from commercial catches in the Bering Sea as aged and measured by the Northwest and Alaska Fisheries Center. For 12-year-old turbot the numbers of each length were:

---

<i>Length (cm)</i>	64	65	66	67	68	69	70	71	72	73	75	77	78	83
<i>No. of fish</i>	1	2	1	1	4	3	4	5	3	3	1	6	1	1

---

Would you agree with someone who asserted that, on this evidence, the median length of 12-year-old Greenland turbot was almost certainly between 69 and 72 cm?

- 2.15 Use (i) the Wilcoxon signed-rank test and (ii) modified van der Waerden scores to test the hypothesis that the median length of 12-year-old turbots is 73.5 using the data in Exercise 2.14.
- 2.16 Establish the values given in [Section 2.5](#) for the mean and variance of the statistic  $S$  relevant to certain one-sample location tests, considering in particular the simplifications for the no-tie Wilcoxon test and for the sign test.
- 2.17 Determine a  $P = 0.01$  two-tail test critical region for the Wilcoxon statistic  $S_+$  when  $n = 12$  based on the asymptotic result using equation [\(2.1\)](#).
- 2.18 The first application listed in [Section 2.7](#) involved insurance claims. The 1999 median was £870. A random sample of 14 claims from a large batch received in the first quarter of 2000 were for the following amounts (in £):

$$475 \ 483 \ 627 \ 881 \ 892 \ 924 \ 1077 \ 1224 \ 1783 \ 1942 \ 2013 \ 2719 \ 4650 \ 6915$$

What test do you consider appropriate for a shift in median relative to the 1999 median? Would a one-tail test be appropriate? Obtain a 95 per cent confidence interval for the median based upon these data.

- 2.19 The weight losses in kilograms for 16 overweight women who have been on a diet for 2 months are as follows:

$$4 \ 6 \ 3 \ 1 \ 2 \ 5 \ 4 \ 0 \ 3 \ 6 \ 3 \ 1 \ 7 \ 2 \ 5 \ 6$$

The firm sponsoring the diet advertises ‘Lose 5 kg in 2 months’. In a consumer affairs radio programme they claim this is an ‘average’ weight loss. You may be unclear about what is meant by ‘average’, but assuming

the sample is effectively random do the data support a median weight loss of 5 kg in the population of dieters? Test this without an assumption of symmetry. What would be a more appropriate test with an assumption of symmetry? Carry out this latter test.

- 2.20 A pathologist counts the numbers of diseased plants in randomly selected areas each 1 metre square on a large field. For 40 such areas the numbers of diseased plants are:

---

21	18	42	29	81	12	94	117	88	210
44	39	11	83	42	94	2	11	33	91
141	48	12	50	61	35	111	73	5	44
6	11	35	91	147	83	91	48	22	17

---

Use an appropriate nonparametric test to find whether it is reasonable to assume the median number of diseased plants per square metre might be 50 (i) without assuming population symmetry, (ii) assuming population symmetry. For these data do you consider the latter assumption reasonable?

- 2.21 A traffic warden notes the time cars have been illegally parked after their metered time has expired. For 16 offending cars he records the time in minutes as:

10 42 29 11 63 145 11 8 23 17 5 20 15 36 32 15

Obtain an appropriate 95 per cent confidence interval for the median overstay time of offenders prior to detection. What assumptions were you making to justify using the method you did? To what population do you think the confidence interval you obtained might apply?

- 2.22 Knapp (1982) gives the percentage of births on each day of the year averaged over 28 years for Monroe County, New York State. Ignoring leap years (which make little difference), the median percentage of births per day is 0.2746. Not surprisingly, this is close to the expected percentage on the assumption that births are equally likely to be on any day, that is,  $100/365 = 0.274$ . We give below the average percentage for each day in the month of September. If births are equally likely to be on any day of the year this should resemble a random sample from a population with median 0.2746. Do these data confirm this?

---

0.277 0.277 0.295 0.286 0.271 0.265 0.274 0.274 0.278 0.290 0.295  
0.276 0.273 0.289 0.308 0.301 0.302 0.293 0.296 0.288 0.305 0.283  
0.309 0.299 0.287 0.309 0.294 0.288 0.298 0.289

---

# 3

## Other single-sample inferences

### 3.1 INFERENCES FOR DICHOTOMOUS DATA

#### 3.1.1 Binomially distributed data

The binomial distribution is relevant to certain counts associated with only two possible outcomes —often referred to as dichotomous data. These counts may be the raw or **primary** data as in [Example 1.3](#), where the *count* was the number of patients in a sample of 10 requiring treatment. In that example we used binomial distribution properties to assess the strength of evidence against a hypothesis that the sample came from a  $B(10, \frac{3}{4})$  distribution, i.e. to test the hypothesis  $H_0: p = \frac{3}{4}$  against  $H_1: p > \frac{3}{4}$ . In [Example 1.2](#) we applied the sign test to hypotheses about the median of measured survival times (the primary data), basing the test not on these primary data but on the derived or **secondary** data of numbers of observations above or below a hypothesized median. For the sign test the parameter  $p = \frac{1}{2}$  is relevant to  $H_0$ , and although the test was based on that parameter, we were interested not in the parameter itself but in the median of an underlying continuous distribution of survival times. In this section we give other examples of inferences about a binomial parameter  $p$ .

It is convenient to refer to each of the  $n$  observations as the outcome of a trial. The key assumptions needed for a count to have a binomial distribution are :

- the  $n$  trials are mutually independent;
- only one of two possible outcomes  $A, B$  is observed at each trial;
- at each trial there is a fixed probability  $p$  associated with the outcome  $A$  and a probability  $q = 1 - p$  associated with the outcome  $B$ .

The outcomes, or events  $A, B$  are often labelled *success* and *failure*, but for reasons explained in [Section 1.1.1](#) this may be inappropriate, so we designate the outcomes as *event A* and *event B*.

The requirement that  $p$  be constant is often violated in a strict sense. For example, when UK children are skin-pricked for sensitization to grass, dust mites and cats around 7 per cent show a skin reaction to all three allergens. This does not imply that there is a probability  $p = 0.07$  that a child given the skin tests will react to all three tests positively. For instance, for children with asthma the probability of three positive reactions is around 0.14, whereas for children without asthma it is about 0.05. However, if a random sample is taken from a large group of UK children which is representative of the population in terms of asthma, it is reasonable to expect approximately 7 per cent of those in the random sample to have reactions to all three allergens. For a similar sample of children from a country with a desert climate coupled with few domestic cats there is medical evidence to suggest that the proportion of children reacting to the grass and cat allergens may be lower. An appropriate statistical test could be based on the binomial distribution corresponding to a sample of  $n$  with the null hypothesis  $H_0: p = 0.07$  against the alternative  $H_1: p < 0.07$ . However, in a comment on [Example 3.5](#) we point out another complication that may invalidate the test.

Tests based on the binomial distribution also arise in quality control problems where we are interested in the proportion of items in large batches having a specific attribute —often a defect. A widely used procedure is to take a random sample of  $n$  items from a batch of  $N$  items and for a buyer to accept the whole batch if the number of defective items,  $r$ , in that sample does not exceed some small fixed number  $s$ . Otherwise the batch is rejected. The consequences of such a procedure can be studied using binomial distribution theory. The results are only approximate because the usual sampling method of selecting items without replacing each before we select later items (called *sampling without replacement*) does not lead to a binomial distribution for  $r$ . However, if  $N$  is very much larger than  $n$ , as it often is in practice, the binomial approximation is satisfactory. The scheme we outline below is usually embellished in practice, but we do not consider here the many possible elaborations.

### *Example 3.1*

*The problem.* A contract for supplying rechargeable batteries specifies that not more than 10 per cent should need recharging before 100 hours of use. To test compliance with this requirement a potential purchaser takes a random sample of 20 from a large consignment and finds that 3 need recharging after

being used for less than 100 hours. Assess the strength of evidence that the batch is below specification.

*Formulation and assumptions.* The test is distribution-free in the sense that we make no assumption about the distribution of times before a recharge is needed. However, we assume that there is a fixed probability  $p$ , say, that any one battery will need recharging in less than 100 hours and that the failure time for each battery is independent of that of any other battery. The null hypothesis is  $H_0: p \leq 0.10$  and the alternative is  $H_1: p > 0.10$ .

*Procedure.* Tables for the binomial distribution or standard statistical software indicate that when  $p = 0.10$  the probability that a sample of 20 contains 3 or more batteries that fail to meet specifications is  $P = 0.323$ .

*Conclusion.* Since the probability that 3 or more batteries will need recharging in less than 100 hours when  $p = 0.10$  is nearly one-third there is little evidence for preferring the alternative that the failure rate in the large consignment exceeds 10 per cent.

*Comments.* 1. This example does little more than highlight serious limitations of hypothesis testing. Three failures in 20 items represents a 15 per cent failure rate in the sample, so the data would clearly support a null hypothesis that specified  $p = 0.15$  and also many higher values. Confidence intervals for  $p$  will be informative. We show in [Example 3.2](#) that a two-sided confidence interval with at least 95 per cent coverage for the true  $p$  is the interval (0.032, 0.379). This implies that we would only reject a hypothesized value of  $p$  falling outside that interval when we use the conventional  $P \leq 0.05$  as sufficient evidence against that hypothesis, i.e. the sample evidence here suggests only that the batch percentage failing to live up to specification is likely to be between about 3.2 and 37.9 per cent.

2. Having obtained a confidence interval, two further questions of interest are (i) what is the power of a test for  $H_0: p = 0.1$  against various alternatives and (ii) how may we use larger samples to make a test more precise (in the sense of shortening the confidence interval or increasing the power for specific alternatives)? We address these questions in [Examples 3.3](#) and [3.4](#).

*Computational aspects.* In [Examples 3.1–3.4](#) we discuss computational aspects under *Procedure* as they form an integral part of that.

### 3.1.2 Confidence intervals for binomial proportions

For a binomial distribution, if an event  $A$  occurs in  $r$  out of  $n$  trials the sample estimate of the binomial probability,  $p$ , is  $\hat{p} = r/n$ . We use the notation  $X = r$  where  $r = 0, 1, 2, \dots, n$  to indicate that the event  $A$  occurs in  $r$  among  $n$  trials. To find a two-sided confidence interval for  $p$  with at least 95 per cent coverage we seek for its upper limit a value  $p_u$  such that if the binomial parameters are  $n, p_u$ , then the value of  $\Pr(X \leq r)$  is as close as possible to, but does not exceed, 0.025. Well-known binomial distribution theory indicates that  $p_u$  must satisfy the equation

$$\sum_i \binom{n}{i} (p_u)^i (1 - p_u)^{n-i} = 0.025 \quad (3.1)$$

where the summation is over  $i$  from 0 to  $r$ . A similar approach provides an equation for a lower limit,  $p_l$ . These equations may be modified to give any required level by replacing 0.025 by another appropriate value, e.g. 0.005 for a 99 per cent interval.

The equation (3.1) generally has no simple analytic solution for  $p_u$ . The classic way of overcoming this difficulty was recourse to tables or charts obtained by numerical methods where these limits were presented for several confidence levels (usually 90, 95 and 99 per cent) for a range of values of  $n$  and  $r$ . Table A4 in Conover (1999) is one such table. Many computer packages including Minitab and StatXact compute exact confidence limits at any level chosen by the user.

For reasonably large  $n$  if neither exact tables nor relevant computer programs are available an asymptotic approximation may be used. If we observe  $X = r$  occurrences of event  $A$  in  $n$  trials for any  $p$  such that  $0 < p < 1$ , then

$$Z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

has for large  $n$  approximately a standard normal distribution. The approximation is good for  $n > 20$  if  $p$  is close to 0.50, but appreciably larger values of  $n$  are needed if  $p$  is close to 0 or 1. To obtain an approximate 95 per cent interval for  $p$  we follow the usual normal distribution confidence interval procedure and find the upper and lower limits as the value of  $p$  that satisfy the equations

$$Z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1 - \hat{p})/n}} = \pm 1.96 \quad (3.2)$$

To obtain limits for other confidence levels 1.96 is replaced by the appropriate value, e.g. 1.64 if a 90 per cent interval is required.

If the event does not occur at all ( $r = 0$ ) then the estimate of the binomial probability is zero. An upper limit for  $p$  can be readily obtained (using a tail probability of 0.05 as the lower tail is not meaningful) and equation (3.1) reduces to

$$(1 - p_u)^n = 0.05$$

Correspondingly, if  $r = n$  a lower limit can be obtained by solving

$$(p_l)^n = 0.05.$$

Sackett et al. (1991) tabulate  $p_l$  and  $p_u$  when  $r = 0$  and  $r = n$  for a selection of sample sizes between 1 and 300.

### *Example 3.2*

*The problem.* For the data in [Example 3.1](#) confirm that an exact 95 per cent confidence interval for  $p$  is (0.032, 0.379). Determine also 90 and 99 per cent intervals. Obtain an asymptotic approximation to the 95 per cent interval.

*Formulation and assumptions.* Assuming that the number of batteries in a sample of 20 that fail to last 100 hours before needing a recharge has a binomial distribution, the exact 95 per cent confidence limits are given by equation [\(3.1\)](#) and the corresponding equation for the lower limit. For the 90 per cent and 99 per cent limits the 0.025 on the right-hand side is replaced by 0.05 and 0.005 respectively. Because the equations have no analytical solution the required limits are obtained in practice from tables or by using appropriate software. The relevant asymptotic limits are given by [\(3.2\)](#).

*Procedure.* From Conover (1999) Table A4 one finds the following exact intervals for this problem where  $n = 20$  and  $r = 3$ .

90 per cent	(0.042, 0.344)
95 per cent	(0.032, 0.379)
99 per cent	(0.018, 0.450)

Both Minitab and StatXact give the same values and these should also be obtained from any other software correctly claiming to give exact intervals.

For the asymptotic limits we use [\(3.2\)](#) with  $n = 20$  and  $\hat{p} = 3/20 = 0.15$  and the solutions give the 95 per cent interval (-0.0065, 0.306). Most computer software will give these limits as an alternative to exact limits, but not all include the timely warning in Minitab that these may be unreliable for small samples.

*Conclusion.* Evidence from this sample of 20 indicates only that it seems likely that the overall proportion in a large batch lies somewhere between about 3 – 4 per cent and about 35 – 40 per cent.

*Comments.* 1. We have already indicated that the asymptotic approximation requires  $n$  to be considerably greater than 20 for small  $p$ . The asymptotic 95 per cent lower limit, -0.0065 is a statistical nonsense, being outside the permissible range  $0 \leq p \leq 1$ . Tables of exact values in Conover cover values of  $n \leq 30$  but with small values of  $p$  the asymptotic result may still be unreliable even when  $n > 30$ , meaning that unless one has access to a table for larger values of  $n$  one needs statistical software for an exact interval. For the case  $n = 30$  and  $r = 2$  an exact 95 per cent interval for  $p$  is (0.008, 0.221) while the normal approximation gives the nonsense interval (-0.023, 0.156). However, if  $n = 30$  and  $r = 13$  the exact 95 per cent interval is (0.255, 0.626) and the asymptotic approximation is a reasonable (0.256, 0.611).

2. In this example we considered a two-sided confidence interval whereas in [Example 3.1](#) we were interested in a one-tail hypothesis test. The lower limit of the one-sided 95 per cent confidence interval for the test considered there corresponds to the lower limit for the two-sided 90 per cent interval, i.e. the 95 per cent one-sided interval is in this example (0.042, 1).

### 3.1.3 Power and sample size

In Examples 3.1 and 3.2 there is considerable doubt about the population value of  $p$ . In practical quality control problems a common situation is that where a purchaser would ideally like all batches of items such as the rechargeable batteries that are accepted to contain not more than 10 per cent that require recharging before 100 hours of use. However, taking marketing realities such as production cost into account, the purchaser may be prepared to accept a few batches where as many as, say, 20 per cent fail to live up to this requirement, but be reluctant to accept batches where more than 20 per cent fail. By the same token the supplier does not want the purchaser to reject batches where less than 10 per cent will fail.

If we test  $H_0: p \leq 0.10$  against the alternative  $H_1: p > 0.10$  and agree the evidence is strong enough to reject  $H_0$  if the associated  $P$ -value is  $P \leq 0.05$  this means there is a probability not greater than 0.05 that the purchaser will reject a batch that would be acceptable (i.e. that in reality contains 10 per cent or less that are defective). Be careful in your reading to distinguish between the binomial probability (lowercase  $p$ ) and the test  $P$ -value (capital  $P$ ). In acceptance sampling terms  $P = 0.05$  is called the *producer's risk* because in agreeing to accept these terms it is the maximum long run probability that an acceptable batch will be rejected. In conventional hypothesis testing terms producer's risk is the probability of an *error of the first kind*. If the consumer is prepared to accept a few batches with more than 10 but less than 20 per cent defective the consumer and producer might agree that they should devise a test procedure that ensures a certain preassigned **high** probability, say  $P^* = 0.90$  that one does not accept a batch with 20 per cent or more defective. This means there is only a probability  $1 - P^* = 0.10$  that a batch with 20 per cent or more defective will slip through. This last probability is called the *consumer's risk* because in agreeing to accept these terms it is the maximum probability that an unacceptable batch will be purchased. It is the probability of an *error of the second kind*, and  $P^*$  is the power. A practical problem is to choose the sample size  $n$  such that there is both a preassigned small probability  $P$  that batches for which  $p \leq 0.10$  will be rejected and also a preassigned large probability  $P^*$  that batches for which  $p \geq 0.20$  will also be rejected.

Here the concept of power (Section 1.3) is useful. Power depends on the value of  $P$  chosen to indicate significance, i.e. as sufficient evidence to make  $H_0$  unacceptable, and also upon the sample size  $n$ .

and upon the value of the binomial  $p$  specified for an alternative hypothesis. Two relevant questions are:

- Given a sample of size  $n$  what is the power of a test of  $H_0: p = 0.1$  against  $H_1: p = 0.2$  if we agree to reject  $H_0$  for any  $P$  less than some fixed preassigned value in a one-tail test?
- What sample size  $n$  will ensure that a test of  $H_0: p = 0.1$  against  $H_1: p = 0.2$  where we reject  $H_0$  for any  $P$  less than some fixed value (e.g.  $P = 0.05$ ) has a given power  $P^*$  for that alternative?

The rationale for these choices of  $H_0$ ,  $H_1$  is that if the true  $p$  is less than 0.1 we are less likely to reject a good batch than would be the case if  $p = 0.1$  and that if  $p > 0.2$  the power of our test will exceed that when  $p = 0.2$ . A moment's reflection shows these are desirable characteristics for both producer and consumer.

If we observe  $r$  occurrences of event  $A$  in a sample of size  $n$  and want to test  $H_0: p = p_0$  against  $H_1: p > p_0$  and agree that there is sufficient evidence to prefer  $H_1$  for all  $P$ -values less than or equal to some fixed value  $\alpha$ , then the critical region of size  $\alpha$  consists of all values of  $r \geq r_0$  where  $r_0$  is the smallest value of  $r$  such that the binomial distribution  $B(n, p_0)$  probabilities for all  $r \geq r_0$  have a sum not exceeding  $\alpha$ . In practice we either read  $r_0$  from tables or now more commonly obtain it using computer software.

Having determined  $r_0$  it is easy to find the power of this test against a single-valued alternative  $H_1: p = p_1$  because this is simply the  $\Pr(r \geq r_0)$  when  $H_1$  is true. This probability may be read from tables or obtained using appropriate computer software.

For large  $n$  asymptotic approximations may be used. These are based on the fact that under  $H_0$  the distribution of

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \quad (3.3)$$

is approximately standard normal. If, for example,  $\alpha = 0.05$  for an upper-tail test we set  $Z = 1.64$  in (3.3) and solve for  $\hat{p}$  given  $n$  and  $p_0$ . To get an asymptotic expression for the power when  $H_1: p = p_1$  is true we replace  $p_0$  by  $p_1$  in (3.3) and insert the solution  $\hat{p}$  that we have just obtained in this amended expression for  $Z$  and calculate the resulting  $Z$ , say  $Z = z_1$ . The asymptotic approximation to the power is  $\Pr(Z \geq z_1)$ .

### *Example 3.3*

*The problem.* For a sample of 20 batteries from a large batch suppose we wish to test  $H_0: p = 0.1$  against  $H_1: p = 0.2$  where  $p$  is the population proportion

of batteries requiring recharging in less than 100 hours and decide we will prefer  $H_1$  for any  $P$ -value  $P \leq 0.05$ . What is the minimum number of faulty batteries in the sample for which we prefer  $H_1$ ? Obtain both exact and asymptotic expressions for the power of this test.

*Formulation and assumptions.* Let  $X$  be the number of occurrences,  $r$ , of the event  $A$  (battery needs recharging in less than 100 hours), then under  $H_0$ ,  $X$  has a binomial (20, 0.1) distribution. We prefer  $H_1$  if our sample gives  $r_0$  or more occurrences of  $A$  where  $r_0$  is the least  $r$  such that  $\Pr(X \geq r_0) \leq 0.05$  when  $H_0$  is true. Once we know  $r_0$  the power is obtained by computing for that  $r_0$  the corresponding probability under  $H_1$ . Asymptotic results for large  $n$  use the normal approximations given above for these probabilities.

*Procedure.* If suitable tables are available these may be used although it is often easier to use software. We outline the procedure for both methods. When  $n = 20$ ,  $p = 0.10$ , Table A3 in Conover (1999) indicates that  $\Pr(X \leq 4) = 0.9568$  whence  $\Pr(X \geq 5) = 1 - 0.9568 = 0.0432$ . Thus we prefer  $H_1$  if and only if  $r \geq 5$ . To determine the power we require  $\Pr(X \geq 5)$  for a  $B(20, 0.2)$  distribution. The same tables indicate that now  $\Pr(X \geq 5) = 1 - \Pr(X \leq 4) = 1 - 0.6296 = 0.3704$ . In the asymptotic approach under  $H_0$  we have, using (3.3)

$$Z = (\hat{p} - 0.1) / \sqrt{(0.1 \times 0.9/20)} = 1.64$$

whence  $\hat{p} = 0.2100$ . Thus the relevant power is given by

$$\Pr[Z \geq (0.2100 - 0.2)/\sqrt{(0.2 \times 0.8/20)}] = \Pr(Z \geq 0.1118).$$

From standard normal distribution tables this probability is 0.4555.

Most computer software easily generates  $P$ -values corresponding to any number of successes for one- or two-tail tests. For example, the 1-proportion program in Minitab confirms that  $\Pr(X \geq 5) = 0.043$  and the same program using  $p = 0.2$  indicates  $\Pr(X \geq 5) = 0.370$  confirming our result for the power. Minitab, version 12, also gives a separate program relating power to sample size but uses an asymptotic approximation. This gives an approximate power 0.432 which differs somewhat from our 0.4555. The discrepancy is explained by the estimate being sensitive to round off. If we replace 1.64 by the more precise 1.645 the estimated power is 0.434, close to the Minitab estimate. The discrepancy between the true power and the asymptotic estimate reflects the unsatisfactory nature of the asymptotic result for even moderate  $n$  associated with small  $p$ . StatXact, version 4 gives the exact power directly for any sample size for any pair of values of  $p$  specified in  $H_0$  and  $H_1$ .

*Conclusion.* The power of our test against an alternative  $p = 0.20$  is only 0.370 if we choose  $P \leq 0.05$  (or more precisely because of discontinuities  $P \leq 0.043$ ) as our criterion for not accepting  $p = 0.10$ .

*Comments.* 1. In the discussion preceding this example we suggested a practical aim might be to reject a batch with probability (power) 0.90 if  $p \geq 0.20$ . The power 0.370 is far short of this requirement.

2. The unsatisfactory nature of asymptotic power calculations with the  $n$  and  $p$  used here is less acute with larger samples.

3. In Section 1.3 we indicated that reducing the  $P$ -value chosen as acceptable evidence against  $H_0$  generally results in reduced power against a particular

alternative. In this example tables or appropriate software indicates that if we set the requirement that  $P$  must not exceed 0.02 the power of the one-sided test against the alternative  $p = 0.2$  is reduced from 0.370 to 0.196. However, due to discontinuities in possible  $P$ -values the exact  $P$  is not 0.02 but is 0.011.

Modern software yields exact  $P$ -values for one- or two-tail tests for any  $n, p$  and  $r$  which effectively means we can obtain the values of  $\Pr(X \geq r)$ ,  $0 \leq r \leq n$  for any  $B(n, p)$  distribution. This is useful for developing tests such as those for attribute sampling considered in this section that have preset producer's risk (i.e.  $P$ -values for rejection of  $H_0$ ) and consumer's risk (i.e. power against a specified  $H_1$  that gives the maximum  $p$  the consumer considers tolerable).

### *Example 3.4*

*The problem.* Determine the power of the test  $H_0: p = 0.10$  against  $H_1: p = 0.20$  when  $P \leq 0.05$  when  $n = 50, 100, 150, 200, 300$ . Compare exact values with the asymptotic approximation. Also determine the minimum  $n$  to ensure a power of at least 0.9 for this test.

*Formulation and assumptions.* For illustrative purposes we use programs in Minitab, version 12 and in StatXact, version 4, but modifications for other packages with binomial distribution programs are generally straightforward.

*Procedure.* We describe the procedure fully for the case  $n = 100$  and only quote results for other  $n$ . We urge readers to confirm the results using available software. When  $n = 100$  and  $p = 0.10$  it is intuitively reasonable to expect a  $P$ -value not exceeding 0.05 if  $r$ , the number of batteries needing recharging before 100 hours, is close to 15. Minitab and StatXact both give the following values

$r$	15	16	17
$P$	0.073	0.040	0.021

establishing  $r = 16$  as the critical value. To obtain the power we need to know  $\Pr(X \geq 16)$  when  $p = 0.20$ . StatXact gives this probability, the exact power, to be 0.871 directly without the need to first ascertain, as we just did, that  $r = 16$  is the critical cut-off value. The sample size and power program for proportions in Minitab gives an asymptotic approximation which in this case indicates a power of 0.897 for a nominal  $P = 0.05$ . For our exact  $P$ -value 0.040 the asymptotic power approximation is 0.882. The asymptotic approximation is reasonable for this sample size. For  $n = 100$  and the remaining sample sizes of interest with the largest exact  $P$ -values meeting the requirement  $P \leq 0.05$ , the exact power and the asymptotic estimate of power based on the corresponding exact  $P$  are:

$n$	50	100	150	200	300
$P$	0.025	0.040	0.044	0.043	0.038
<i>Exact power</i>	0.556	0.872	0.963	0.989	0.999
<i>Asymptotic power</i>	0.617	0.882	0.962	0.988	0.999

How best to estimate the sample size  $n$  needed to ensure an exact power 0.9 may depend on available software. Clearly the size lies between 100 and 150. Since

asymptotic power calculations seem reasonable for samples exceeding 100 one might use an asymptotically based facility in Minitab that gives an estimate of sample size having any required power for the relevant critical  $P$ -value (here  $P = 0.05$ ). One may fine tune this using a program for exact binomial probabilities in a way we now describe. For our test the power-sample size program in Minitab indicates the required  $n = 102$ . We used this as a starting estimate for the exact result. When  $n = 100$  we found the critical region was defined by  $r \geq 16$ . When  $n = 102$  the Minitab binomial program relevant to our one-tail test shows that when  $r = 16$  the corresponding  $P$ -value is  $P = 0.047$ , while  $r = 15$  gives  $P = 0.083$ . The exact power for the alternative  $p = 0.20$  corresponding to  $P = 0.047$  is given by Minitab as 0.890. In practice this may be accepted as a suitable approximation or one might seek some insurance by choosing a slightly greater sample size such as  $n = 105$ . With that latter choice Minitab gives the exact  $P = 0.032$  with power 0.865! The apparent anomaly of reduced power with increased sample size is simply a reflection of the impact of discontinuities in  $P$ -values and possible values of the power (which remember is also a probability with associated discontinuities). Our nearest  $P$ -value above  $P = 0.032$  would have been  $P = 0.058$  and the associated power then becomes 0.914. StatXact also gives exact power for the above or other sample sizes directly. These indicate that a sample size of 109 has power 0.901 in the problem considered here.

*Conclusions.* In general for a specified single-value alternative hypothesis power increases with sample size for a fixed  $P$ -value but there are some irregularities due to discontinuities in possible exact  $P$ -values. These irregularities are generally of little practical importance.

*Comment.* Although one should be aware of the effects of discontinuities in  $P$ -values in assessing evidence for abandoning a hypothesis and in calculating power and sample sizes to achieve certain aims the importance of such effects should not be overemphasized. While it is clearly unwise, for example, to use asymptotic results for small samples, the discrepancies due to discontinuities in possible  $P$ -values are usually small in larger samples and in this context it is also worth remembering that nearly all mathematical models of real world situations are only approximations. Hopefully these will reflect key features of reality, but there will always be some discrepancies. For instance, we have already pointed out that when we take a sample of  $n$  from a large batch of  $N$  items by sampling without replacement the binomial distribution is only an approximation, albeit quite a good one.

### 3.1.4 A caution

In Section 3.1.1 we considered a situation where, in the population at large, 7 per cent of all children skin-pricked for sensitization to grass, dust mites and cats had a positive skin reaction to all three allergens. However, the presence or absence of asthma meant that each child did not have the same probability of showing three positive skin tests. We indicated there that if a random sample were taken from that larger population it is not unreasonable to expect this to reflect the population proportion of 7 per cent showing three

positive tests, because statistical theory backs an intuitive hunch that a random sample should reflect fairly closely characteristics of a population. On these grounds we indicated that it may be reasonable to take a random sample of children from a country with a desert climate and few domestic cats to see whether there is a lower proportion of positive skin tests in that population. We formalize a simple test based on the binomial assumption in Example 3.5 but warn in the *Comment* that the analysis may still not be valid.

### *Example 3.5*

*The problem.* It is known that 7 per cent of UK children give positive skin tests to grass, dust mites and cats. It is thought that a desert climate and few domestic cats may reduce the proportion of positive tests. A random sample of 54 children from such a country were given the allergen tests and only one of these had positive reactions to all three allergens. Is there sufficient evidence of a lower incidence compared with the UK?

*Formulation and assumptions.* We justify a one-tail test of  $H_0: p = 0.07$  against  $H_1: p < 0.07$  because there are good medical reasons for assuming that the proportion should be lower under such conditions. If the direction of any effect on sensitization were unknown then a two-tail test of  $H_0: p = 0.07$  against  $H_1: p \neq 0.07$  would be appropriate.

*Procedure.* As exact tables for  $n = 54$  are not widely available and a small  $p$  makes asymptotic results unreliable one should resort to appropriate software. For the one-tail test StatXact or Minitab gives an exact  $P = 0.101$ . The exact two-sided nominal 95 per cent confidence interval for  $p$  is (0.0005, 0.0989).

*Conclusion.* The evidence of a reduction is not strong. If it were important to detect even a small reduction to, say,  $p = 0.05$  a larger experiment would be needed to give reasonable power.

*Comment.* In Section 3.1.1 we pointed out possible difficulties if particular groups of patients have markedly different probabilities of showing a reaction. We stressed the need for a random, or effectively random sample, to overcome this problem. Another type of problem may arise in, for instance, the testing of a new drug. For the new drug there may be a quite different set of factors influencing the likelihood of an individual showing side-effects compared with the original drug. For instance, with one antibiotic for the treatment of pneumonia people with asthma may be more likely to exhibit side-effects, whereas for a new antibiotic the reverse may be the case. In these circumstances what could be an overall beneficial effect can sometimes be masked or distorted in our sample. The solution lies in either further experimentation or at least in using more detailed observations and tests. For example, one may require an analysis that takes into account other factors that might influence outcomes for individuals, e.g. factors such as weight, age, sex of patients. The result is a more complex analytic problem and indeed most real life statistical problems are more complicated than the simple ones discussed in this section!

## 3.2 TESTS RELATED TO THE SIGN TEST

### 3.2.1 An alternative approach to the sign-test

We described the sign test in detail in [Section 2.3](#). An alternative but equivalent approach more in the spirit of the work in [Section 3.1](#) is sometimes used and we illustrate this by an example.

*Example 3.6*

*The problem.* For the second data set in [Example 2.9](#), i.e.

5.6 6.1 6.3 6.3 6.5 6.6 7.0 7.5 7.9 8.0 8.0 8.1 8.1 8.2 8.4 8.5 8.7 9.4 14.3 26.0

if  $\theta$  is the population median test the hypothesis  $H_0: \theta = 9$  against the alternative  $H_1: \theta \neq 9$  using the sign test. Obtain a 95 per cent confidence interval for  $\theta$  using the methods developed in [Section 3.1](#).

*Formulation and assumptions.* Wide availability of programs giving exact  $P$ -values provides a useful facility for testing and estimation in a sign-test situation.

*Procedure.* For illustrative purposes we use the basic statistics 1-proportion option in Minitab. Here  $n = 20$  and the number of observations below the hypothesized median is  $r = 17$ . The sign test calls for evaluation of  $P$  for a two-tail test with  $p = 0.5$  and the Minitab program gives at the same time a confidence interval for the true  $p$  at a nominal level which we here set at 95 per cent. The program gives  $P = 0.003$  and a nominal 95 per cent confidence interval for  $p$  of (0.621, 0.967). Allowing for rounding our  $P$ -value agrees with that obtained in [Example 2.9](#) and indeed here we did nothing essentially different from what was done there. To obtain a confidence interval for  $\theta$  we now decrease  $r$  by steps of 1 until we get a 95 per cent confidence interval for  $p$  that just includes the value  $p = 0.5$ . This straightforward stepwise process indicates that when  $r = 15$  the confidence interval is (0.508, 0.913) while for  $r = 14$  it is (0.457, 0.881). Thus when  $r = 14$  we would regard  $p = 0.5$  as acceptable with  $P > 0.05$  but not when  $r = 15$ . The symmetry of the binomial distribution when  $p = 0.5$  implies that we would accept when  $r = 6$  but not when  $r = 5$ . This implies that an appropriate 95 per cent confidence interval for the median is from the sixth to the fifteenth largest observations, i.e. the open interval (6.6, 8.4), agreeing with the result in [Example 2.10](#).

*Computational aspects.* The procedures in [Examples 2.9](#) and [2.10](#) and in this example are basically equivalent. Which is preferred is largely determined by the availability of appropriate tables or computational software.

### 3.2.2 Quantile tests

The sign test is relevant to population medians. We are often interested in other quantiles of a population. For example a quantile such that three-quarters of the population values lie below it and one quarter above is known as the *upper (or third) quartile* while a

quantile such that  $r$  tenths of the observations lie below it and  $(10-r)$  tenths of the observations lie above it where  $r = 1, 2, \dots, 9$  is called the  $r$ th decile. Many tests and estimation procedures involving quantiles proceed along similar lines to those about medians by simply replacing  $p = 0.5$  by a value appropriate to the relevant quantile. For continuous distributions quantiles are unique, but for discrete distributions conventions given in most standard textbooks are needed to give unique quantiles. We have already met the usual convention for the median of a set consisting of an even number, say,  $2m$ , observations. We arrange them in ascending order and denoting the  $r$ th ordered observation by  $x_{(r)}$  we define the median as  $\frac{1}{2}(x_{(m)} + x_{(m+1)})$ . Note that the fifth decile is the same as the median.

### *Example 3.7*

*The problem.* A central examining body publishes the information that ‘three-quarters of the candidates taking a mathematics paper achieved a mark of 40 or more’ (i.e. the first population quartile is 40). One school entered 32 candidates for this paper of whom 13 scored less than 40. The president of the Parents’ Association argues that the school’s performance is below national standards. The headmaster counters by claiming that in a random sample of 32 candidates it is quite likely that 13 would score less than the lower quartile mark even though 8 out of 32 is the expected proportion. Is his assertion justified?

*Formulation and assumptions.* Denoting the first quartile by  $Q_1$  a test for the headmaster’s assertion is a test for  $H_0: Q_1 = 40$  against  $H_1: Q_1 \neq 40$ .

*Procedure.* We associate a minus with a mark below 40; thus for our ‘sample’ of 32 we have 13 minuses (and 19 pluses). If the first quartile is 40 then the probability is  $p = 0.25$  that each candidate in a random sample has a mark below 40 (and thus is scored as a minus). The distribution of minuses is therefore binomial  $B(32, \frac{1}{4})$ . Using any program that gives a 95 per cent confidence interval for  $p$  given  $r = 13$  minuses (event  $A$ ) when  $n = 32$  we find the relevant confidence interval is  $(0.24, 0.59)$ . This includes  $p = 0.25$ . The associated  $P$ -value is  $P = 0.0755$ .

*Conclusion.* If we observe 13 minus signs we would not reject the hypothesis  $H_0: \text{first quartile is } 40$  at a conventional 5 per cent significance level. Nevertheless there is some evidence against this hypothesis – enough to worry many parents and some may be reluctant to give the headmaster’s claim the benefit of the doubt until further evidence were available.

*Comments.* 1. Remember that even if we use formal significance levels, non-rejection of a hypothesis does not prove it true. It is only a statement that the evidence to date is not sufficient to reject it. This may simply be because our sample is too small. We may make an error of the second kind by non-rejection. Indeed, since the above confidence interval includes  $p = 0.5$  we would not reject the hypothesis that the median is 40.

2. We used a two-tail test. A one-tail test would not be justified unless we had information indicating the school’s performance could not be better than the

national norm. For example, if most schools devoted three periods per week to the subject but the school in question only devoted two, we might argue that lack of tuition could only depress performance. Indeed, a one-tail test at the conventional 5 per cent level would reject  $H_0$  if there are more than 12 minuses. If we observe 13 minuses and feel a one-tail test with  $P = 0.05$  as a significance indicator is appropriate and want to use the method described in this example we might determine a 90 per cent confidence interval for  $p$  and reject  $H_0$  if  $p = 0.25$  is below the lower limit for this interval. Think carefully about this to be sure you see why. Have you any reservations about the accuracy of this approach?

3. The headmaster's claim said 'if one took a random sample'. Pupils from a single school are in no sense a random sample from all examination candidates. Our test only establishes that results for this particular school are not too strongly out of line with national results in the sense that they just might arise if one took a random sample of 32 candidates from all entrants.

4. Tests for a third quartile are symmetric with those for a first quartile if we interchange plus and minus signs.

### 3.2.3 The Cox–Stuart sign test for trend

Cox and Stuart (1955) proposed a simple test for a monotonic trend, i.e. an increasing or decreasing trend. A monotonic trend need not be linear; it need only express a tendency for observations to increase or decrease subject to some local or random irregularities. Consider a set of **independent** observations  $x_1, x_2, \dots, x_n$  ordered in time. If we have an even number of observations,  $n = 2m$ , say, we take the differences  $x_{m+1} - x_1, x_{m+2} - x_2, \dots, x_{2m} - x_m$ . For an odd number of observations,  $2m+1$ , we may proceed as above omitting the middle value  $x_{m+1}$  and calculating  $x_{m+2} - x_1$ , etc. If there is an increasing trend we expect most of these differences to be positive, whereas if there were no trend and observations differed only by random fluctuations about some median these differences (in view of the independence assumption) are equally likely to be positive or negative. A preponderance of negative differences suggests a decreasing trend.

This implies that under the null hypothesis of no trend, the plus (or minus) signs have a  $B(m, 1/2)$  distribution and we are in a sign test situation.

#### *Example 3.8*

*The problem.* The US Department of Commerce publishes estimates obtained from independent samples each year of the mean annual mileages covered by various classes of vehicles in the United States. The figures for cars and trucks (in thousands of miles) are given below for each of the years 1970–83. Is there evidence of a monotonic trend in either case?

---

Cars	9.8	9.9	10.0	9.8	9.2	9.4	9.5	9.6	9.8	9.3	8.9	8.7	9.2	9.3
Trucks	11.5	11.5	12.2	11.5	10.9	10.6	11.1	11.1	11.0	10.8	11.4	12.3	11.2	11.2

---

*Formulation and assumptions.* The figures for each year are based on independent samples so we may use the Cox–Stuart test for trend. Without further information a two-tail test is appropriate as a trend may be increasing or decreasing.

*Procedure.* For cars relevant differences are  $9.6 - 9.8$ ,  $9.8 - 9.9$ ,  $9.3 - 10.0$ ,  $8.9 - 9.8$ ,  $8.7 - 9.2$ ,  $9.2 - 9.4$ ,  $9.3 - 9.5$  and all are negative. Appropriate tables or computer software immediately establishes that seven negative signs or seven positive signs when  $p = 0.5$  has an associated  $P = 0.016$ .

*Conclusion.* There is strong evidence of a downward monotonic trend.

*Comments.* 1. For trucks the corresponding differences have the signs  $-$ ,  $-$ ,  $-$ ,  $-$ ,  $+$ ,  $+$ ,  $+$ . Clearly when we only have seven data values 3 plus and 4 minus (or 3 minus and 4 plus) signs provides the strongest possible evidence to support a hypothesis of no monotonic trend. The fact that the first four differences are all negative and the last three all positive suggests the possibility of a decreasing trend followed by an increasing trend (i.e. a non-monotonic trend) rather than random fluctuations. The sample is too small to establish this, but in [Section 3.5](#) we describe a ‘runs test’ appropriate when we have larger samples for testing whether in circumstances like these fluctuations are random.

2. Periodic trends are common. For example, at many places in the northern hemisphere mean weekly temperature tends to increase from February to July and decrease from August to January. A Cox–Stuart test applied to data of this type might either miss such a trend (because it is not monotonic) or indicate a monotonic trend for records over a limited period (e.g. from February to June). Conover (1999, Examples 4 and 5, pp. 172–175) shows how in certain circumstances the Cox–Stuart test may be adapted to detect periodic trends by re-ordering the data.

3. If the same samples of cars and trucks had been used each year the independence assumption would not hold; inference would then only be valid for vehicles in that sample, for anything atypical about the sample would influence observations in all years. With independent samples for each year, anything atypical about the sample in any one year will be incorporated in the random deviation from trend in that year.

### 3.3 MATCHING SAMPLES TO DISTRIBUTIONS

Test and estimation procedures for centrality measures were covered in [Chapter 2](#) but other population characteristics such as spread or skewness may be of interest. In particular, we often want to know whether observations are consistent with their being a sample from some specified continuous distribution. Kolmogorov (1933; 1941) devised a test for this purpose.

Given observations  $x_1, x_2, \dots, x_n$ , we ask if these are consistent with our sample being from some **completely specified** distribution. This might be a uniform distribution over  $(0, 1)$  or over  $(20, 30)$ , or a normal distribution with mean 20 and standard deviation 2.7. Kolmogorov's test is distribution-free because the procedure does not depend upon which distribution is completely specified under  $H_0$ . In [Section 3.3.3](#) we look at modifications to answer questions like 'Can we suppose these data are from a normal distribution with unspecified mean and variance?'

### 3.3.1 Kolmogorov's test

The **continuous uniform distribution** (sometimes called the **rectangular distribution**) is a simple continuous distribution that arises when a random variable has the same probability of taking a value in any small interval of length  $\delta x$  within a fixed and specified finite interval  $(a, b)$ . For example, suppose pieces of thread each 6 cm in length are clamped at each end and a force is applied until they break; if each thread breaks at the weakest point and this is equally likely to be anywhere along its length, then the breaking points will be uniformly distributed over the interval  $(0, 6)$ , where the distance to the break is measured in centimetres from the left-hand clamp.

The probability density, or frequency, function has the form

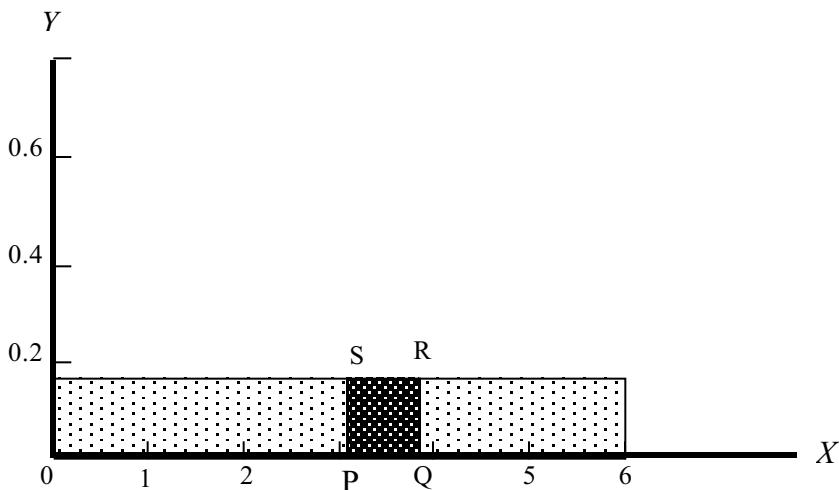
$$f(x) = 0, x \leq 0; \quad f(x) = \frac{1}{6}, \quad 0 < x \leq 6; \quad f(x) = 0, x > 6.$$

The rectangular form of  $f(x)$  over  $(0, 6)$  explains the name **rectangular distribution**. If the thread always breaks, the probability is 1 that it breaks in the interval  $(0, 6)$ , so the total area under the density curve must be 1; clearly only the density function above satisfies both this condition and that of equal probability of a break in any small segment of length  $\delta x$  no matter where that segment lies within the interval. The function is graphed in [Figure 3.1](#) and is essentially a line running from 0 to 6 at height  $\frac{1}{6}$  above the  $x$ -axis. The lightly shaded area represents the total probability of 1 associated with the complete distribution. The heavily shaded rectangle  $PQRS$  between  $x = 3.1$  and  $x = 3.8$  represents the probability that the random variable  $X$ , the distance to the break, takes a value between 3.1 and 3.8. Since  $PQ = 3.8 - 3.1 = 0.7$  and  $PS = \frac{1}{6}$ , clearly this area is  $0.7/6 \approx 0.1167$ . The probability of a break occurring in any segment of length 0.7 lying entirely in the interval  $(0, 6)$  is also 0.1167.

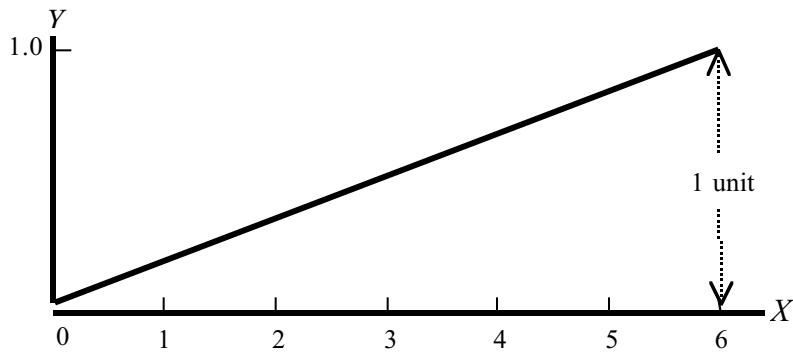
Following convention we use  $X$ ,  $Y \dots$  as names for random variables, and corresponding lowercase letters  $x$ ,  $y \dots$  (with or without subscripts) for specific values of these variables.

The Kolmogorov test uses not the probability density function, but the cumulative distribution function (cdf), a function that gives  $\Pr(X \leq x)$  for all  $x$  in  $(0, 6)$ . In our example the probability the break occurs in the first two centimetres is  $\Pr(X \leq 2)$ . Clearly this is  $1/3$ .

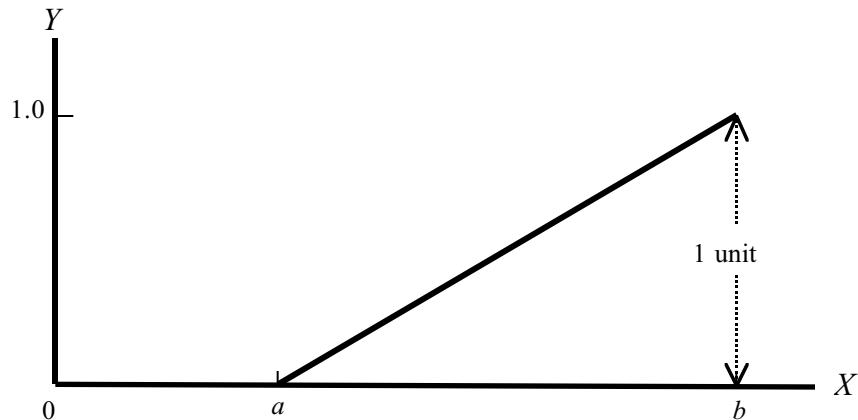
The cdf is written  $F(x)$  and for any  $x$  between 0 and 6,  $F(x) = x/6$ . It has the value 0 at  $x = 0$  and 1 at  $x = 6$ . It is graphed in [Figure 3.2](#). These notions generalize to a uniform distribution over any interval  $(a, b)$  and the cdf is  $F(x) = (x - a)/(b - a)$ ,  $a < x \leq b$ , specifying a straight line rising from zero at  $x = a$  to 1 at  $x = b$ . Clearly  $F(a) = 0$  and  $F(b) = 1$ . This is illustrated in [Figure 3.3](#).



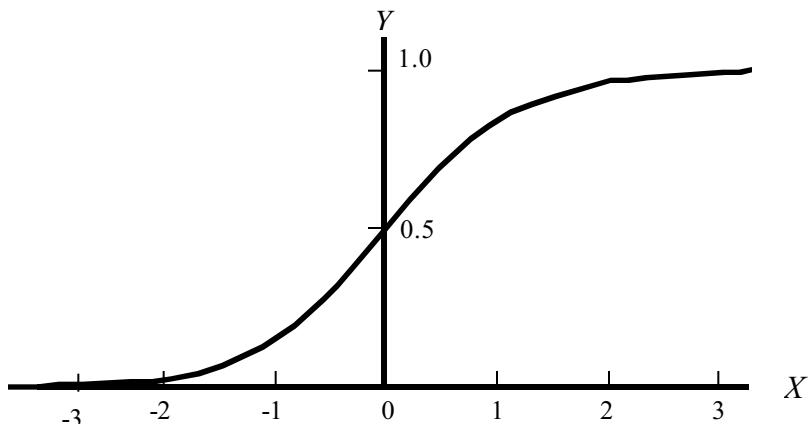
**Figure 3.1** Probability density function for a continuous uniform distribution over  $(0, 6)$ .



**Figure 3.2** Cumulative distribution function (cdf) for a uniform distribution over  $(0, 6)$ .



**Figure 3.3** Cumulative distribution function (cdf) for a uniform distribution over  $(a, b)$ .



**Figure 3.4** Cumulative distribution function (cdf) for the standard normal distribution.

For any continuous distribution the cdf is a curve starting at zero for some  $x$  and increasing as we move from left to right until it attains the value 1. It never decreases as  $x$  increases and is said to be **monotonic non-decreasing** (or **monotonic increasing**). Figure 3.4 shows the cdf for the standard normal distribution.

The Kolmogorov test compares a population cdf with a related curve  $S(x)$  based on sample values and called the **sample (or empirical) cumulative distribution function**. For a sample of  $n$  observations

$$S(x) = \frac{\text{number of sample values } \leq x}{n} \quad (3.4)$$

### Example 3.9

*The problem.* The distances from one end at which each of 20 threads 6 cm long break when subjected to strain are given below. Evaluate and graph  $S(x)$ . For convenience the distances are given in ascending order.

0.6	0.8	1.1	1.2	1.4	1.7	1.8	1.9	2.2	2.4
2.5	2.9	3.1	3.4	3.4	3.9	4.4	4.9	5.2	5.9

*Formulation and assumptions.* From (3.4) it is clear that  $S(x)$  increases by  $1/20$  at each unique  $x$  value corresponding to a break, or by  $r/20$  if  $r$  break distances coincide.

*Procedure.* When  $x = 0$ ,  $S(x) = 0$ ; it keeps this value until  $x = 0.6$ , the first break. Then  $S(0.6) = 1/20$ , and  $S(x)$  maintains this value until  $x = 0.8$  when it jumps to  $2/20$ , a value retained until  $x = 1.1$ . It jumps in steps of  $1/20$  at each break value until  $x = 3.4$ , where it increases by  $2/20$  since two breaks occur at  $x = 3.4$ .  $S(x)$  is referred to as a step function for obvious reasons. Its value at each step is given in [Table 3.1](#).

*Conclusion.* [Figure 3.5](#) shows the form of  $S(x)$  for these data.

*Comment.*  $S(x)$  is a sample estimator of the population cdf  $F(x)$ . If a sample comes from a specified distribution, the step function  $S(x)$  should not depart markedly from the population cdf  $F(x)$  unless the sample is very small, say,  $n = 10$  or less. In this problem if breaks are uniformly distributed over  $(0, 6)$  one expects, for example, about half of these to be in the interval  $(0, 3)$ , so that  $S(3)$  should not have a value very different from 0.5 and so on. In [Figure 3.5](#) we show also the cdf for a uniform distribution over  $(0, 6)$  reproduced from [Figure 3.2](#). Intuitively one may feel that this is not a good fit.

Recognizing that  $S(x)$  should not depart violently from  $F(x)$  for a sample from a distribution with that specific cdf is basic to Kolmogorov's test. The test statistic is the maximum difference in magnitude between  $F(x)$  and  $S(x)$ .

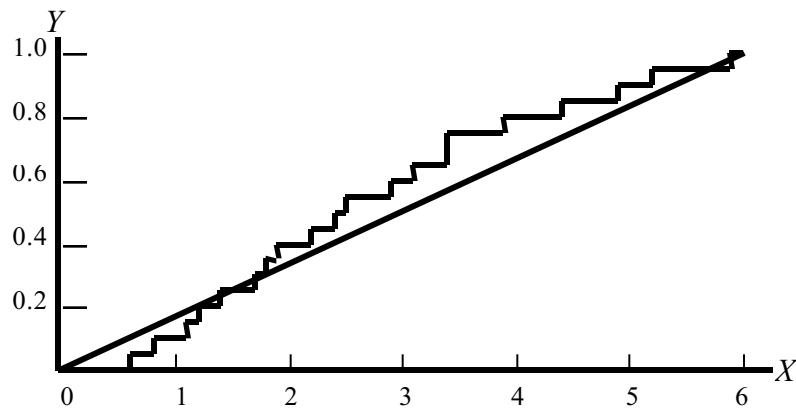
### Example 3.10

*The problem.* Given the 20 breaking points and the corresponding  $S(x)$  values in Table 3.1, is it reasonable to suppose breaking points are uniformly distributed over  $(0, 6)$ ?

**Table 3.1** Values of  $S(x)$  at step points.

$x$	0.6	0.8	1.1	1.2	1.4	1.7	1.8	1.9	2.2	2.4	2.5
$S(x)$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
$x$	2.9	3.1	3.4*	3.9	4.4	4.9	5.2	5.9			
$S(x)$	0.60	0.65	0.75	0.80	0.85	0.90	0.95	1.00			

\* Repeated sample value



**Figure 3.5** Sample cumulative distribution function (cdf) for thread breaks (step function) together with the cdf for a uniform distribution over (0,6) (straight line).

**Table 3.2** Comparison of  $F(x)$  and  $S(x)$  for thread breaks.

$x_i$	$F(x_i)$	$S(x_i)$	$F(x_i) - S(x_i)$	$F(x_i) - S(x_{i-1})$
0.6	0.10	0.05	0.05	0.10
0.8	0.13	0.10	0.03	0.08
1.1	0.18	0.15	0.03	0.08
1.2	0.20	0.20	0.00	0.05
1.4	0.23	0.25	-0.02	0.03
1.7	0.28	0.30	-0.02	0.03
1.8	0.30	0.35	-0.05	0.00
1.9	0.32	0.40	-0.08	-0.03
2.2	0.37	0.45	-0.08	-0.03
2.4	0.40	0.50	-0.10	-0.05
2.5	0.42	0.55	-0.13	-0.08
2.9	0.48	0.60	-0.12	-0.07
3.1	0.52	0.65	-0.13	-0.08
3.4	0.57	0.75	-0.18	-0.08
3.9	0.65	0.80	-0.15	-0.10
4.4	0.73	0.85	-0.12	-0.07
4.9	0.82	0.90	-0.08	-0.03
5.2	0.87	0.95	-0.08	-0.03
5.9	0.98	1.00	-0.02	0.03

*Formulation and assumptions.* The maximum difference in magnitude between  $F(x)$  and  $S(x)$  is compared with a tabulated value to assess the goodness of fit. Or better, software is used to determine the exact  $P$ -value.

*Procedure.* Table 3.2 shows values of  $F(x)$  and  $S(x)$  at each break point given in Table 3.1. It also gives at each of these points the difference  $F(x_i) - S(x_i)$ . In view of the stepwise form of  $S(x)$ , the maximum difference may not be in this set. Inspecting Figure 3.5 it is clear that a greater difference may occur immediately before such a step. There  $S(x)$  has the value attained at the previous step, so a maximum may occur among the  $F(x_i) - S(x_{i-1})$ . These differences are also recorded in Table 3.2.

Figure 3.5 shows that for much of the interval  $(0, 6)$   $S(x)$  lies above  $F(x)$ . The entry of greatest magnitude in the last two columns of Table 3.2 is  $-0.18$  when  $x_i = 3.4$ . Published tables give values of the minimum difference required for significance at nominal 5 and 1 per cent levels for various  $n$ , and in some cases also for other quantiles but modern software makes these somewhat redundant. If required, tables may be found in Gibbons and Chakriborti (1992), in Conover (1999) or in many general statistical tables. The theory for computing exact  $P$ -values is more complicated than that for situations we have met so far and Gibbons and Chakriborti (1992, Section 4.3) give a detailed account. Tables indicate that the statistic must have a value of at least 0.294 for significance at the 5 per cent level in a two-tail test. The exact test gives  $P = 0.458$ .

*Conclusion.* There is no evidence against  $H_0$ .

*Comments.* 1. A two-tail test is appropriate if the alternative hypothesis is that the observations may come from any other unspecified distribution. Situations may arise where the only alternative is that the cdf must lie either wholly at or above or wholly at or below that specified in the null hypothesis. For example, in the thread problem we may know that the test equipment places strains on the thread that, if not uniformly distributed, will be greater at the left and decrease as we move to the right. This might increase the tendency for breaks to occur close to the left end of the string. Then the cdf would always lie above the straight line representing  $F(x)$  for the uniform distribution. In that case a maximum difference with  $S(x)$  greater than  $F(x)$  for the uniform distribution would, if significant, favour the acceptable alternative, making a one-tail test appropriate.

2. In commenting on Figure 3.5 we noted that the step function did not appear to match the population cdf very well. The Kolmogorov test does not indicate a meaningful departure, a point we take up again in Section 3.3.2.

*Computational aspects.* Some general statistical packages include programs for the Kolmogorov test but may use only asymptotic results applied at nominal significance levels and these are not satisfactory even in that context unless sample sizes are at least 20, or sometimes more. The exact test program in StatXact avoids these difficulties and allows a comparison with an asymptotic approximation to the  $P$ -value, which often shows a surprising difference.

The Kolmogorov test appears to waste information by using only the difference of greatest magnitude. Tests that allow for all differences exist but tend to have few, if any, advantages. This is not as contrary as intuition may suggest, for the value of  $S(x)$  at any stage depends on how many observations are less than the current  $x$  and therefore at each stage we make the comparison on the basis of accumulated evidence.

For a discrete distribution the Kolmogorov test tends to give too few significant results and in this sense is conservative. We consider preferred tests for discrete distributions in Section 9.4.

### 3.3.2 Comparison of distributions: confidence regions

If, in [Figure 3.5](#) we had drawn step functions everywhere at distances 0.294 units above or below that representing  $S(x)$ , with adjustments at each end to stop them falling below zero or above one, we would, since 0.294 is the critical difference for significance at the 5 per cent level given by tables, have a 95 per cent confidence region for  $F(x)$  in the sense that any  $F(x)$  lying entirely between these two new step functions would be an acceptable  $F(x)$  when testing at a 5 per cent significance level in a two-tail test. This interval is not very useful in practice. A more common practical situation is one where it is reasonable to assume that our data may come from one of a (usually small) number of completely specified distributions. We might then use a Kolmogorov test for each. In each test the relevant  $S(x_i)$  will be the same for a given data set. As a result of these tests we may reject some of the hypotheses but still find more than one acceptable. It is then useful to get at least an overall eye comparison of how well the match of  $S(x)$  is to each acceptable  $F(x)$ . Ross (1990, p. 85) gives a useful procedure for such eye comparisons. Basically it consists of plots of the maximum of the two deviations given for the last two columns of each row in a table like [Table 3.2](#) against the  $S(x_i)$  for that row. Plots may easily be done for several different hypothesized population distributions on the same graph.

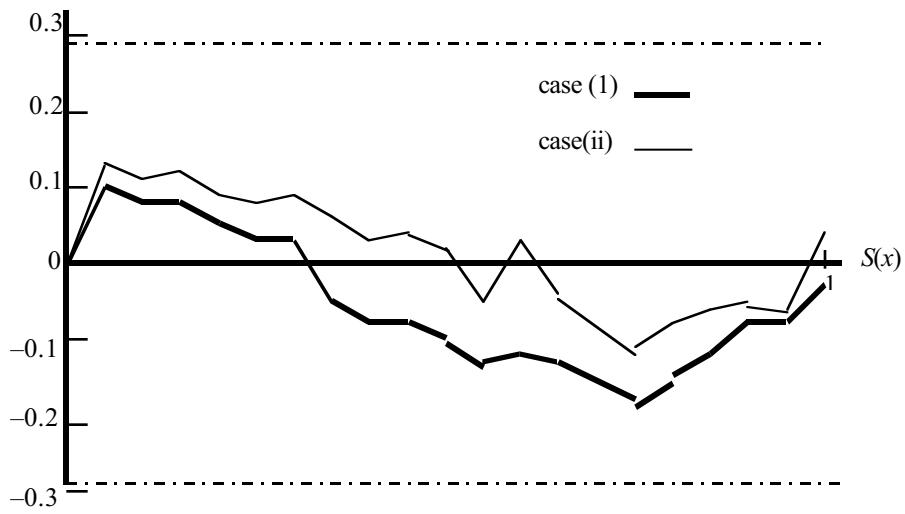
#### *Example 3.11*

*The problem.* Compare the fit of the thread break data in [Example 3.9](#) to (i) a uniform distribution over (0, 6) and (ii) the distribution with cumulative distribution function over (0, 6) given by

$$F(x) = x/5, 0 < x \leq 3; F(x) = 0.2 + 4x/30, 3 < x \leq 6. \quad (3.5)$$

*Formulation and assumptions.* All information needed for the Kolmogorov test procedure for the distribution (i) was obtained in [Example 3.10](#). We require similar information for the alternative distribution in (ii). We make a graphical comparison using the format suggested by Ross.

*Procedure.* For (i) the values of  $S(x)$  of interest are those in column 3 of [Table 3.2](#). In each case the corresponding deviation is the value of greatest magnitude in columns 4 and 5. For (ii) the values of  $S(x)$  are again those in [Table 3.2](#). However, fresh values of  $F(x)$  for each  $x_i$  must be calculated using (3.5).



**Figure 3.6** Plot of deviations  $F(x) - S(x)$  against  $S(x)$  for the two  $F(x)$  functions specified in Example 3.5.

**Table 3.3** Values of  $x_i$ ,  $S(x_i)$ ,  $d_i = \max[F(x_i) - S(x_i), F(x_i) - S(x_{i-1})]$  for population distributions specified in (i) and (ii) of Example 3.11.

$x_i$	$S(x_i)$	$d_i(i)$	$d_i(ii)$
0.6	0.05	0.10	0.12
0.8	0.10	0.08	0.11
1.1	0.15	0.08	0.12
1.2	0.20	0.05	0.09
1.4	0.25	0.03	0.08
1.7	0.30	0.03	0.09
1.8	0.35	-0.05	0.06
1.9	0.40	-0.08	0.03
2.2	0.45	-0.08	0.04
2.4	0.50	-0.10	0.03
2.5	0.55	-0.13	-0.05
2.9	0.60	-0.12	0.03
3.1	0.65	-0.13	-0.04
3.4	0.75	-0.18	-0.10
3.9	0.80	-0.15	-0.08
4.4	0.85	-0.12	-0.06
4.9	0.90	-0.08	-0.05
5.2	0.95	-0.08	-0.06
5.9	1.00	0.03	0.04

For example, when  $x_i = 1.2$ , (3.5) gives  $F(1.2) = 1.2/5 = 0.24$ . Using this and the values of  $S(1.1)$  and  $S(1.2)$  in [Table 3.2](#) we find that  $F(1.2) - S(1.2) = 0.24 - 0.20 = 0.04$  and  $F(1.2) - S(1.1) = 0.24 - 0.15 = 0.09$ . Thus, corresponding to  $x = 1.2$ , the maximum deviation is 0.09 just before the step in  $S(x)$  at  $x = 1.2$ . In [Table 3.3](#) we set out for each  $x_i$  the corresponding  $S(x_i)$  and maximum deviations for the population distributions specified in (i) and (ii). These deviations are plotted against  $S(x)$  in [Figure 3.6](#). Although it is not essential we have joined consecutive points in each case by straight line segments. The parallel broken lines at  $y = \pm 0.294$  represent the critical value for significance at the 5 per cent level. We reject  $H_0$  at a formal 5 per cent significance level only if we observe a deviation outside these lines.

*Conclusion.* There is no strong evidence against either of the population hypotheses specified in (i) and (ii).

*Comment.* The deviations, except for the early observations, are rather smaller for (ii), suggesting (but only marginally) that this may be a slightly better fit. The hypothesis (ii) implies a uniform but higher probability of breakages between 0 and 3 cm relative to a uniform distribution, but a lower probability between 3 and 6 cm. If other hypotheses are of interest relevant deviations for these may be entered in [Figure 3.6](#).

### 3.3.3 A test for normality

The Kolmogorov test is valid for any completely specified continuous distribution. If we specify in  $H_0$  only that the sample has a normal distribution and estimate the parameters  $\mu$ ,  $\sigma^2$  from the data we may still use the Kolmogorov statistic but the associated  $P$ -values differ from those for the fully specified model.

For this important practical situation a test proposed by Lilliefors (1967) is useful. We assume observations are a random sample from some unspecified continuous distribution, and test whether it is reasonable to suppose this is a member of the normal family. To this extent our test is not distribution-free, but it is nonparametric in the sense that we do not specify parameter values.

The basic idea can be extended to test compatibility with other families of distributions such as a gamma distribution with unspecified parameters, but separate tables of critical values or appropriate software to estimate  $P$ -values are needed for each family of distributions. Conover (1999, Section 6.2) describes some difficulties in obtaining critical values and indeed one may only get approximate values of the statistic corresponding to precise  $P$ -values such as  $P = 0.05$  or  $0.01$ , or alternatively Monte Carlo approximations for the exact  $P$ -value corresponding to an observed value of the statistic. StatXact provides a program for the latter. Conover

(1999, Table A14) gives approximate critical values for sample sizes up to 30 and asymptotic approximations for larger samples.

Another test for normality is the Shapiro–Wilk (1965) test which again requires special tables or appropriate software for implementation. An account of how the test works, without detail of the theory, is given by Conover (1999, Section 6.2) who also provides extensive tables that are required if relevant computer software is not available. This test has good power against a range of alternatives but the rationale is less easy to describe by intuitive arguments. StatXact includes a program for asymptotic estimates of  $P$ -values. Except for very small samples, where exact results are available, these usually prove adequate.

Tests for normality may be important in deciding whether to apply nonparametric methods in a particular problem. We describe Lilliefors' test and also quote results for analysis of the same data using the Shapiro–Wilk test.

### *Example 3.12*

*The problem.* In the Badenscallie burial ground in Wester Ross, Scotland, Sprent recorded the ages at death in years of males on all tombstones for four clans in the district and reported these in the first edition of this book. The complete list is given in the appendix. From all 117 ages recorded a random sample of 30 was taken and the ages for that sample were, in ascending order:

11	13	14	22	29	30	41	41	52	55	56	59	65	65	66
74	74	75	77	81	82	82	82	83	85	85	87	87	88	

Is it reasonable to suppose the death ages are normally distributed?

*Formulation and assumptions.* In Lilliefors' test the Kolmogorov statistic is used to compare the standard normal cdf  $\Phi(z)$  with the standardized sample cdf  $S(z)$  based on the transformation

$$z_i = \frac{x_i - m}{s}$$

where  $m$  is the sample mean and  $s$  the usual estimate of population standard deviation, i.e.

$$s^2 = \frac{\sum_i x_i^2 - (\sum_i x_i)^2 / n}{n-1}$$

*Procedure.* We find  $m = 61.43$ ,  $s = 25.04$ . Successive  $z_i$  are calculated: e.g. for  $x_1 = 11$ , we find  $z_1 = (11 - 61.43)/25.04 = -2.014$ . We treat this as a sample value from a standard normal distribution and tables (e.g. Neave, 1981, pp. 18–19) show  $\Phi(-2.014) = 0.022$ . [Table 3.4](#) is set up in analogous manner to [Table 3.2](#). Here  $\Phi(z)$ ,  $S(z)$  represent the standard normal and sample cdfs respectively. These computations are done automatically in StatXact. It is not essential, but

we have included the  $x$  values, an asterisk against a value implying it occurs more than once. The largest difference is 0.192, occurring in the final column when  $x = 74$ . Tables such as Table A14 in Conover (1999, p.548) give the 1 per cent critical value when  $n = 30$  for a one-tail test to be 0.183. StatXact computes a Monte Carlo approximation to the exact  $P$ -value for this particular data set as a worked example in the manuals for StatXact versions 3 and 4. These approximate  $P$ -values differ slightly between simulations. That in the version 3 manual is  $P = 0.0057$  while that in the version 4 manual is  $P = 0.0055$ .

*Conclusion.* The Monte Carlo approximations to exact  $P$ -values provide strong evidence that the sample is not from a normal distribution.

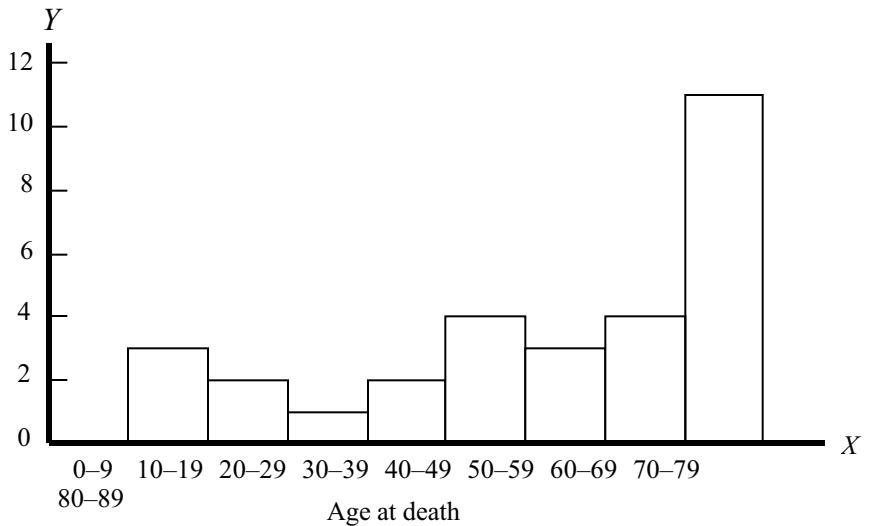
*Comments.* 1. The StatXact manuals for versions 3 and 4 also use these data to illustrate the Shapiro–Wilk test and obtain an asymptotic estimate  $P = 0.0007$ . Shapiro, Wilk and Chen (1968) carried out studies that show this test has good power against a wide range of departures from normality.

2. A glance at the sample data suggests that a few males died young and that the distribution is skew with a large number of deaths occurring after age 80. [Figure 3.7](#) shows these data on a histogram with class interval of width 10. Readers familiar with histograms for samples from a normal distribution will not

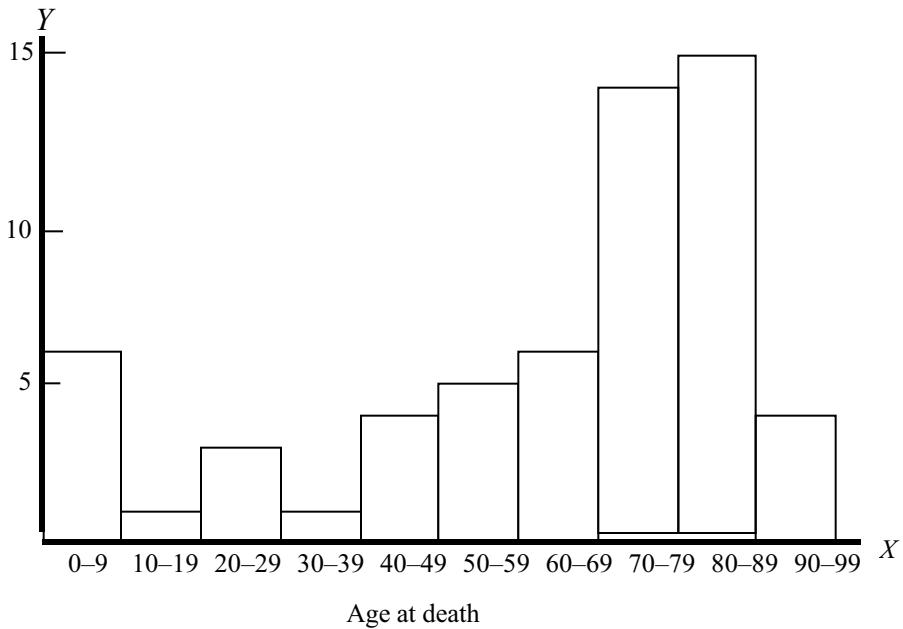
**Table 3.4** The Lilliefors normality test: Badenscallie ages at death.

$x$	$z$	$\Phi(z)$	$S(z)$	$\Phi(z_i) - S(z_i)$	$\Phi(z_i) - S(z_{i-1})$
11	-2.014	0.022	0.033	-0.011	0.022
13	-1.934	0.026	0.067	-0.044	-0.007
14	-1.894	0.029	0.100	-0.071	-0.038
22	-1.575	0.058	0.133	-0.075	-0.042
29	-1.295	0.098	0.167	-0.069	-0.035
30	-1.255	0.105	0.200	-0.095	-0.062
41*	-0.816	0.207	0.267	-0.060	-0.007
52	-0.377	0.353	0.300	0.053	0.086
55	-0.257	0.399	0.333	0.066	0.099
56	-0.217	0.414	0.367	0.047	0.081
59	-0.097	0.461	0.400	0.061	0.094
65*	0.142	0.556	0.467	0.089	0.156
66	0.183	0.572	0.500	0.072	0.105
74*	0.502	0.692	0.567	0.125	0.192
75	0.542	0.706	0.600	0.106	0.139
77	0.622	0.733	0.633	0.100	0.133
81	0.781	0.782	0.667	0.115	0.149
82*	0.821	0.794	0.800	-0.006	0.127
83	0.861	0.805	0.833	-0.028	-0.005
85*	0.942	0.827	0.900	-0.073	-0.006
87*	1.021	0.846	0.967	-0.121	-0.054
88	1.061	0.856	1.000	-0.144	-0.111

\*repeated value



**Figure 3.7** Histogram of death ages at Badenscallie.



**Figure 3.8** Histogram of death ages for clan McAlpha.

be surprised that we reject the hypothesis of normality. Indeed ‘life-span’ data, whether for man, animal or the time a machine functions without a breakdown, will often characteristically have a non-normal distribution.

*Computational aspects.* We indicated above that these data are used to illustrate Lilliefors’ test in StatXact. Because there is no analytic formula for exact  $P$ -values only a Monte Carlo approximation is available for Lilliefors’ test. These are given together with a confidence interval for that estimate and this will usually be quite short if the estimate is made on a large number of Monte Carlo

samples. In practice it is customary to use anything between 1000 (minimum) and perhaps 1 000 000 samples if high precision is required. For the Shapiro–Wilk test the distribution of the test statistic appears to be known only for sample sizes of six or less and is very nonnormally distributed for larger sample sizes. However it may be transformed to approximate normality and the transformation provides a basis for determining the asymptotic  $P$ -value given in StatXact.

### 3.3.4 Relevant inference

Situations arise where inferences about means or medians are by their nature somewhat irrelevant to sensible deductions and here tests such as Kolmogorov's or Lilliefors' test may be more useful. We comment upon one such case.

#### *Example 3.13*

For one particular clan – it would be invidious to give it the real name, so we shall call it the McAlpha clan (but the data are real) – ages at death were recorded for all males of that clan buried in the Badenscallie burial ground. For all 59 burials the ages (arranged for convenience in ascending order) were:

0	0	1	2	3	9	14	22	23	29	33	41	41	42	44	52	56	57	58	58
60	62	63	64	65	69	72	72	73	74	74	75	75	75	77	77	78	78	79	79
80	81	81	81	81	82	82	83	84	84	85	85	86	87	87	88	90	92	93	95

A histogram of these data with class interval 10 is given in [Figure 3.8](#). The mean and median are respectively 61.8 and 74 while the age range with most deaths is 80 – 89. Striking features in [Figure 3.8](#) are the skewness of the data to the left and the bimodal nature with a small peak (secondary mode) of infant deaths in the range 0–9 followed by lower death rates in 10-year intervals until age 60. The death rate then rises sharply to the primary mode at the 80–89 range. One might go through formal motions of attaching confidence intervals to the mean or median using an approach based on the  $t$ -test (obviously irrelevant because of nonnormality – Lilliefors' test indicates  $P < 0.001$ ), the Wilcoxon test (not valid because of asymmetry) or the sign test (valid but not very interesting in the light of the bimodality). Such approaches are at worst wrong and at best pointless. In this example the histogram is probably the most useful aid in interpreting the data.

What is the relevance of any formal inference procedure in a case like this? From what population is this a random sample? It is clearly not a random sample of all males buried at Badenscallie, for clans and families have differing lifespan characteristics. Perhaps the pattern of death ages is a reasonable approximation to that for the McAlpha clan in Wester Ross, Sutherland and the Western Isles (where the clan is well established). The main use we can envisage for these data would be for comparisons with patterns for other clans.

Difficulties like these may arise with data collection generally and are not confined to nonparametric methods. Analyses making various assumptions often draw attention to interpretation problems.

Using modern software it is easy to apply several tests to the same data. Indeed, for centrality tests StatXact allows one to use Wilcoxon and normal scores test and also to base tests on any scores one chooses. Do not abuse this freedom; a scoring system should be chosen on logical grounds. Remember our ethical warning in the penultimate paragraph of [Section 1.5!](#) If different choices are logical under differing assumptions and these lead to inconsistent results (as in this example for means or medians) further analysis (e.g. applying Lilliefors' test) or more simply drawing a histogram as in [Figure 3.8](#), or sometimes even just looking at the data critically, may suggest the most appropriate choice. In this example clearly a sign-test-type confidence interval for the median is less inappropriate than many others but this begs the question as to whether the median is a useful summary statistic for bimodal data.

### 3.4 ANGULAR DATA

In some investigations measurements are made on directions, e.g. wind directions in a certain town at noon on successive days, the bearings at which released pigeons disappear over the horizon and the successive stopping positions of a roulette wheel. In other studies the data, although not measured directly as angles or directions, can be represented in this way. For instance, the time of day at which babies are born in a large hospital or the days during the year in which new cases of leukaemia are diagnosed in a certain region are effectively **angular measurements**. The names **circular** or **directional measurements** are also used, the former because observations of this type are appropriately represented as points on the circumference of a circle.

Such data are sometimes analyzed as though they are distributed along a line without taking the directional aspect into account. A moment's reflection shows that this is not always a sensible course of action. As an extreme example, observations of  $1^\circ$  and  $359^\circ$  would lead to a mean and median of  $180^\circ$ . The values then appear to be very different. If, however, the two points are plotted on the circumference of a circle, the two directions are similar and a more appropriate average would be obtained from the direction bisecting the smaller arc between the two directions, in other words,  $0^\circ$ . If the data consisted of two observations of 31 December and 2 January the 'straight-line' approach leads to an average around 1 July whereas a directional analysis leads to the more reasonable average of 1 January. The book by Mardia (1972) on directional data analysis is written from a mathematical perspective. Fisher (1993) provides an introduction based around real life examples from many fields and an up-to-date treatment of theory and practice is given by

Mardia and Jupp (2000). Measures of centrality (e.g. the median) and spread (e.g. the range) may be adapted for use with angular data.

One often wants to know whether data support a hypothesis that a sample is from a population uniformly distributed on the circumference of a circle or whether they indicate clustering (Mardia 1972). The Hodges–Ajne test (Hodges, 1955; Ajne, 1968) and the range test, the latter developed from work on the circular range by Fisher (1929), address this situation.

Another question of interest is whether points are symmetrical around a particular angular direction such as the North-South line. The sign test and the Wilcoxon signed-rank test are easily adapted for this situation.

### 3.4.1 The Hodges–Ajne test

This test is used to investigate whether a sample of  $n$  observations on a circle could arise from a uniformly distributed population. The alternative is that in the population observations are more concentrated within a particular arc of the circumference; outliers may nevertheless occur well away from this arc.

To carry out this test, a straight line is drawn through the centre of the circle; this will divide the observations into two groups. The line is rotated about the centre to a position at which there is a minimum possible number of points,  $m$ , on one side of this line. If the points were regularly spaced they would lie at the vertices of a regular polygon and then either  $m = \frac{1}{2}n$  if  $n$  is even or  $m = \frac{1}{2}(n - 1)$  if  $n$  is odd. For observations around the circle from a uniform distribution there will be some variation in the angles between adjacent points and the value of  $m$  will generally be lower. The lowest values of  $m$  occur when there is a clustering of points on one side of the line. Under the assumption of uniformity, the probability that  $m$  is no more than a value  $t$  is shown by Mardia (1972) to be:

$$\Pr(m \leq t) = \frac{\binom{n}{t}(n-2t)}{2^{n-1}} \quad (3.6)$$

Mardia (1972) gives a table of critical values for this test.

#### Example 3.14

*The problem.* A midwife recorded the times of birth for twelve consecutive home deliveries. She was interested in whether births tended to occur at particular times of the day. The times (rearranged in order throughout the day) were 0100, 0300, 0420, 0500, 0540, 0620, 0640, 0700, 0940, 1100, 1200, 1720.

Since on a 24-hour circular clock one hour corresponds to  $360/24 = 15$  degrees the successive angles on the circle (assuming midnight corresponds to  $0^\circ$ ) are  $15^\circ, 45^\circ, 65^\circ, 75^\circ, 85^\circ, 95^\circ, 100^\circ, 105^\circ, 145^\circ, 165^\circ, 180^\circ, 260^\circ$ . We test the hypothesis  $H_0$  that the times of birth have a uniform distribution around the circle.

*Formulation and assumptions.* A line is drawn through the centre of the circular plot of these data and rotated until the number of points on one side of the line takes the minimum value,  $m$ . A small value for  $m$  provides evidence against the assumption of uniformity in  $H_0$ .

*Procedure.* A circular plot of the data (Figure 3.9) shows that  $m = 1$  (e.g. when the line runs from  $10^\circ$  to  $190^\circ$ ). A value of zero for  $m$  represents an even greater degree of clustering. The appropriate  $P$ -value (the probability that  $m$  is equal to zero or one, assuming that  $H_0$  is true) given by (3.6) is 0.059.

*Conclusion.* The evidence against  $H_0$  when  $P = 0.059$  casts some doubt on the assertion that births are equally likely at any time throughout the day.

*Comments.* 1. Inspection of the data shows that most of the babies are born in the morning, particularly between 0300 and 0700. If further evidence for this pattern could be obtained, the information might be useful in planning maternity services and in preparing mothers for the births of their babies.

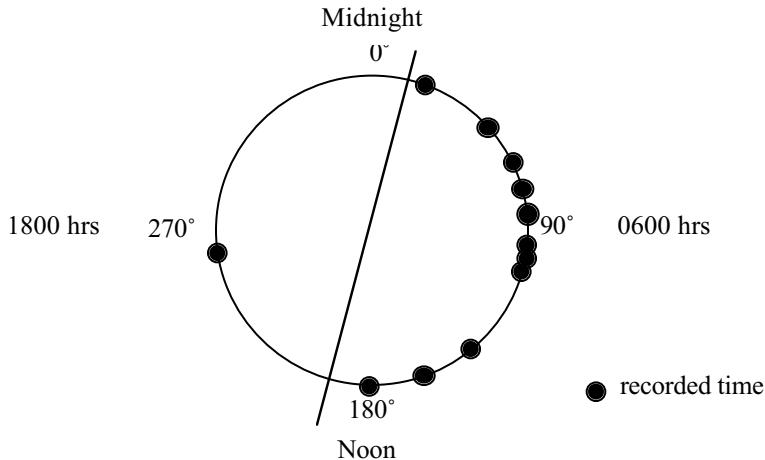
2. This test is especially useful where some observations may be clustered together but a few may be more distant, e.g. injuries from fireworks in the UK are concentrated around 5 November and deaths from drowning are more common during summer months, but such events are clearly not confined exclusively to these periods.

### 3.4.2 The range test

Like the Hodges–Ajne test, the range test is also based on the null hypothesis that a sample of  $n$  observations on the circle could arise from a uniformly distributed population, but now the alternative is that in the population all observations come from a particular arc; this test is therefore not appropriate in the presence of outliers.

To perform this test, the smallest arc which contains all of the points is found. The length of this arc,  $w$ , is the circular range. As we have already pointed out, if the points are spaced at regular intervals they will form the vertices of a regular polygon and the circular range will take the maximum possible value of  $360(n - 1)/n^\circ$ . The circular range will be small if all points occur close together. Under the assumption of uniformity, the probability that the circular range  $w$  is no more than  $r$  radians is shown by Mardia (1972) to be:

$$\Pr(w \leq r) = \sum_{k=1}^v (-1)^{k-1} \binom{n}{k} \left[ 1 - \frac{k(2\pi - r)}{2\pi} \right]^{n-1} \quad (3.7)$$



**Figure 3.9** Circular plot of delivery time data in Example 3.14 on a 24-hour clock.

where the summation is over  $k$  and  $v$  is the largest value of  $k$  such that  $1 - k\{(2\pi - r)/2\pi\}$  exceeds zero. Mardia (1972) gives a table of critical values for the range test.

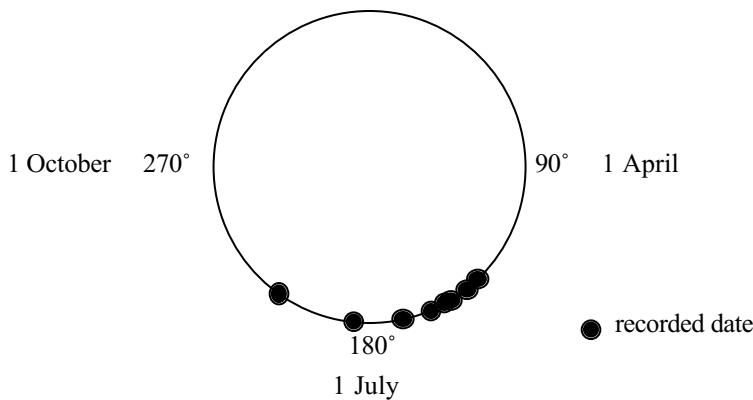
### Example 3.15

*The problem.* Smeeton and Wilkinson (1988) give data for a female psychiatric patient who repeatedly attempted to commit suicide. There was evidence to suggest that these attempts occurred during one particular part of the year. Records showed that attempts had occurred on 2 June 1980, 3 June 1980, 8 June 1980, 18 June 1980, 4 July 1980, 5 June 1981, 6 June 1981 and 31 July 1981.

The successive angles on the circle (assuming  $0^\circ$  is the start of the year) are  $151^\circ, 152^\circ, 157^\circ, 167^\circ, 182^\circ, 154^\circ, 155^\circ, 209^\circ$ . We test the null hypothesis  $H_0$  that the dates of the suicide attempts have a uniform distribution.

*Formulation and assumptions.* The circular range  $w$  is the length of the smallest arc that contains all of the dates. If the suicide attempts are uniformly distributed around the circle then the circular range will be relatively large. By contrast, if the dates form a tight cluster the circular range will be small, providing evidence against  $H_0$ .

*Procedure.* In this example, the circular range is the difference between the largest and smallest angle ( $209^\circ - 151^\circ = 58^\circ$ ). This is illustrated on the circular plot in Figure 3.10. Converting  $58^\circ$  to radians,  $r = 1.0123$  radians. The appropriate  $P$ -value given by (3.7) is  $P < 0.0001$  providing overwhelming evidence against  $H_0$ .



**Figure 3.10** Recorded dates of 8 suicide attempts on an annual clock where  $0^\circ$  represents 1 January.

*Conclusion.* In view of the strength of evidence against  $H_0$  when  $P < 0.0001$ , it is reasonable to accept that for this patient clustering of suicide attempts occurs in June and July.

*Comments.* 1. If the points within a cluster lie on both sides of  $0^\circ$  the difference between the largest and smallest values does not give the circular range; it is then most easily obtained from the circular plot.

2. Unlike the Hodges–Ajne test, the range test is only useful when all observations lie in one cluster since the circular range is highly susceptible to outliers. The occurrence of a further attempt on, say, 5 February illustrates this.

3. The detection of a single cluster of points containing all the observations is particularly relevant in psychiatry. Extra support can be targeted at the patient during the appropriate period. For the patient in this example, the data suggest that each year the anniversary of a distressing event that happened around the beginning of June could be triggering a series of suicide attempts that diminish in frequency over the subsequent weeks.

4. The circular range increases with sample size so that over subsequent years suicide attempts by this individual may be recorded outside of June and July, even if the underlying model remains the same. In fact, follow-up data for this patient (to September 1987) subsequently became available. In 1982, three attempts were recorded between 10 April and 10 June. In 1983, four attempts were recorded between 22 August and 14 December. Following this, no further attempts were recorded. Taking all of the evidence into account the most likely explanation is that any anniversary effect gradually damped out.

### 3.4.3 Median or symmetry tests for angular data

In many sports (archery, cricket, target shooting, netball, bowls, etc.) the objective – at least at some phase of the activity and for some

participants – is to hit a small target by dispatching an appropriate missile (e.g. throwing, kicking or pitching a ball, firing an arrow or bullet). In practice even the most skilled exponents will tend sometimes to go left of target and sometimes right of target (or in three dimensions perhaps also above or below target). It is often of interest to see whether there is a tendency to go more often to the left than to the right or vice versa. Clearly if we know the total number of times there is a deviation to the right or to the left in  $n$  attempts we might use a sign test to see if there is evidence that the ‘median direction’ is or is not that of the target. If there were evidence of a tendency to deviate more often to one side than the other the players or their coaches may want to take corrective action. In some cases we may also have information on the angular deviations from the target line for each attempt made by a player. We may use those angular deviations with appropriate signs (say plus to the right, and minus to the left) and after ranking absolute values use a Wilcoxon signed rank test to carry out what is usually called in this context a test of symmetry. In doing so we make an assumption that the directional distribution of the attempts is symmetric and the question of interest then is whether it is symmetric about the target direction (zero deviation).

Of course, for the serious player symmetry of shots around the target is not enough; the deviations themselves must be small. Also, in sports involving a vertical target such as archery or darts the falling of the missile above or below target has to be addressed in order to ensure an accurate ‘hit’.

These ideas are also relevant in scientific and other applications and extend to testing whether angular data indicate a symmetric distribution about some specified direction. An example of the use of the Wilcoxon test this way in a geological context is given by Fisher (1993, Example 4.17). Where these tests are appropriate they follow the usual procedure for the Wilcoxon test and the sign test given in [Chapter 2](#).

### 3.5 A RUNS TEST FOR RANDOMNESS

Many statistical inferences are valid only when the data are a sample or samples of independent (random) observations. Methods for obtaining a random sample – especially in simulation or Monte Carlo studies – often depend upon sophisticated mechanisms that purport to be equivalent to repeatedly tossing a fair coin or to repeatedly selecting one of the 10 digits between 0 and 9 inclusive,

each having a probability 0.1 of selection. One characteristic of the coin tossing situation is that in the long run there should be approximately equal numbers of heads and tails and a characteristic of random digit selection is that in a long string of such digits each digit should occur approximately the same number of times. In the coin tossing situation a sign test would be appropriate to see whether this frequency requirement is being met. We give in [Section 9.4](#) a method applicable to testing whether there is evidence that digits are not occurring with equal probability. In an ordered sequence randomness implies more than compliance with frequency criteria. For example, if the outcomes, in order, of a computer process that purports to simulate 20 tosses of a coin were

$$\text{H H H H H T T T T T T T T T H H H H H} \quad (3.8)$$

we would suspect the process did not achieve its aim. We might be equally surprised if the ordered outcomes were

$$\text{H T H T H T H T H T H T H T H T H T H T H T} \quad (3.9)$$

but reasonably happy with

$$\text{H H T H T T H T H H T H T H H H T T H} \quad (3.10)$$

A characteristic that reflects our reservations about (3.8) and (3.9) is the **number of runs**, where a run is a sequence of one or more heads or tails. In (3.9) there are three runs – a run of 5 heads, then 10 tails, then 5 heads. In (3.9) there are 20 runs, each consisting of a single head or a single tail. Intuitively we feel that (3.8) and (3.9) have respectively too few and too many runs for a truly random sequence. The sequence (3.10) has an intermediate number of runs, namely 13. Both numbers of runs and lengths of runs are relevant to tests for randomness. The distribution theory for runs was developed by Whitworth (1886) and a detailed treatment of runs tests is given by Bradley (1968, Chapters 11, 12) and a concise account by Gibbons and Chakraborti (1992, Chapter 3). We consider only a test based on the number of runs,  $r$ , in a sequence of  $N$  ordered observations of which  $m$  are of one kind (e.g. H) and  $n = N - m$  are of another kind (e.g. T). We reject the hypothesis that the outcomes are independent or random if we observe too few or too many runs. Computation of the probability of observing any given number of runs under the hypothesis of randomness is a subtle application of combinatorial mathematics and we only quote the outcome. The random variable  $R$  specifies the number of runs. We consider separately the cases  $r$  odd and  $r$  even. For  $r$  odd we set  $r = 2s + 1$  and

$$\Pr(R = 2s + 1) = \frac{\binom{m-1}{s-1} \binom{n-1}{s} + \binom{m-1}{s} \binom{n-1}{s-1}}{\binom{N}{m}}$$

If  $r$  is even we set  $r = 2s$  and

$$\Pr(R = 2s) = 2 \leftrightarrow \frac{\binom{m-1}{s-1} \binom{n-1}{s-1}}{\binom{N}{m}}$$

The test for randomness is based on the relevant tail probabilities associated with small and large numbers of runs. These were tabulated for  $n, m \leq 20$  by Swed and Eisenhart (1943) and tables based on theirs are given by Siegel and Castellan (1988, Table G). StatXact computes exact  $P$ -values. Sprent (1998, Example 6.16) shows how to compute the distribution of  $R$  for a small sample for given  $m, n$ . An asymptotic test is also available and is based on available results for the mean and variance of  $R$ . These are shown after some tedious algebra (see e.g. Gibbons and Chakraborti, 1992, Section 3.2) to be

$$E(R) = 1 + 2nm/N$$

and

$$\text{Var}(R) = 2nm(2nm - N)/[N^2(N - 1)].$$

Asymptotically

$$Z = [R - E(R)]/\sqrt{\text{Var}(R)}$$

has a standard normal distribution. The approximation is improved by the usual numerator continuity correction, i.e. adding  $1/2$  if  $R < E(R)$  and subtracting  $1/2$  if  $R > E(R)$ .

In practice a two-tail test is most often relevant, a significant result implying nonrandomness, but one-tail tests are meaningful in the sense that few runs imply clustering of like values while many runs imply an alternating pattern.

### *Example 3.16*

*The problem.* Apply a runs test for randomness to each of the sequences of heads and tails in (3.8), (3.9) and (3.10).

*Formulation and assumptions.* In each of (3.8) and (3.9)  $N = 20$  and  $m = n = 10$  while  $r = 3$  for the former and  $r = 20$  for the latter. In (3.10)  $N = 20$ ,  $m = 11$ ,  $n = 9$  and  $r = 13$ . Small or large numbers of runs indicate nonrandomness.

*Procedure.* Tables for  $N = 20$  and various values of  $m, n$  such as those referenced above indicate that in both these cases a nominal  $P = 0.05$  is associated with  $r \leq 6$  or  $r \geq 16$ . It is more satisfactory to compute exact  $P$ -values for the observed  $R$ . For these examples StatXact gives the following exact and asymptotic two-tail  $P$ -values:

	<i>Exact</i>	<i>Asymptotic</i>
<i>Data set (3.8)</i>	0.0002	0.0006
<i>Data set (3.9)</i>	$<0.0001$	0.0001
<i>Data set (3.10)</i>	0.4538	0.4575

*Conclusion.* There is strong evidence of nonrandomness in data sets (3.8) and (3.9) but no evidence in set (3.10). The low value of the statistic  $R$  for (3.8) reflects clustering while the high value in (3.9) suggests alternation. These are clear data characteristics in the respective cases.

*Comment.* The above test is not so restrictive as it might seem. For example, in sequences of supposedly random digits between 0 and 9 we may count the numbers of runs above and below the median value of 4.5 and apply the test. It might also be applied in this case to runs of odd and even digits.

## 3.6 FIELDS OF APPLICATION

Some examples given in [Section 2.7](#) for tests of centrality may be relevant to more general tests about distributions such as Kolmogorov's, Lilliefors' or the Shapiro–Wilk tests. We give here other examples where tests about distributions may be relevant.

### Biology

Heart weights are observed for a number of rats used in a drug-testing experiment. The Kolmogorov test could be used to see whether weights are consistent with a normal distribution of mean 11 g and standard deviation 3 g if these were established values for a large batch of untreated rats from the same source. This approach is appropriate if it were uncertain how the drug might affect heart weight, especially if it were felt that it might affect characteristics such as spread or symmetry.

### Forestry

The volume of usable timber per tree is obtained for 50 randomly selected trees from a mature forest. If we want to know if it is reasonable to assume volumes are normally distributed with unspecified mean and variance Lilliefors' or the Shapiro–Wilk test would be appropriate.

## Time-dependent responses

Sometimes tests or observations on individuals have to be performed one at a time and those tested or observed may be able to discuss the test with other subjects due to be tested later. This might influence the performance of later candidates. For example, in medical and other schools, borderline candidates in an examination are sometimes given a viva voce examination. Individuals tested earlier may pass hints to those due for later testing that improve their performance. The marks for candidates (taken in the order in which they were tested) may be used in a runs test to detect any trend from independent responses. With a range of scores, the values can be divided into two groups using the median (see [Exercise 3.11](#)).

## Pollution levels

It is widely believed that many pollution levels are increasing with time. Annual observations over many years at a particular point on the earth's surface or in the atmosphere of levels of ozone or other pollutants might provide data for a Cox–Stuart trend test.

## Genetics

Certain laws of simple Mendelian inheritance imply that progeny of plant crosses should occur in the ratio 3:1 with respect to some characteristics, e.g. three-quarters may produce crinkled seed and one-quarter smooth seed, or three-quarters may be tall plants and one-quarter short plants. If, in a sample of  $n$  progeny,  $r$  show the less frequent characteristic, for which in theory  $p = \frac{1}{4}$  this information may be used to test the strength of the evidence that the data are consistent with the hypothesis of a  $B(n, \frac{1}{4})$  distribution for the characteristic.

## Tasting Experiments

To test whether people can detect differences in taste between, say, two different wines  $A$  and  $B$  a set of  $n$  tasters are each asked to taste three different samples two of which are of one type of wine and the remaining one of the other and to state which of the three samples is the odd one. If they cannot discriminate on a basis of taste and are just guessing, the number of correct guesses will have a  $B(n, \frac{1}{3})$  distribution and a test of this hypothesis may be based on a  $\frac{1}{3}$  quantile ‘sign test’ along the lines developed in [Section 3.2.2](#).

## **Multiple choice examinations**

In a multiple choice examination each candidate is given  $n$  questions each containing 4 statements, one of which is true and the rest of which are false. Candidates must select the statement they believe is true. If a candidate always guesses the probability of correct selection for each question is  $p = \frac{1}{4}$  and this hypothesis may be tested using a binomial quartile-based sign test where the distribution of correct choices under the hypothesis of guessing is  $B(n, \frac{1}{4})$ .

## **Geology**

The directions of fractures or fault lines in rock structures may appear to show a preferred orientation. Tests to assess the strength of evidence that this is so may be based on those described in [Section 3.4](#).

## **Movement of animals**

Birds or animals are often believed to have preferred directions of movement under certain circumstances. For example, there have been many studies of the behaviour of homing pigeons when released to ascertain whether they show an immediate inclination to head towards home or whether they fly at random for some time before aligning to the correct flight path home. Studies have also been made to determine whether birds on a known flight path (e.g. during migration) become disoriented when passing close to electro-magnetic radiation sources and tend then to fly at random. Angular distribution tests then become appropriate.

## **Other one-sample problems**

Many one-sample problems have not been covered in detail in this or the previous chapter. One interesting one is that where ordered observations taken before a fixed time  $\tau$  have one distribution, while after that time they have a distribution differing only in centrality parameter. One problem of interest then is to determine  $\tau$ . This problem is discussed by Pettit (1979, 1981) and for the case of several change-over times by Lombard (1987).

## **3.7 SUMMARY**

**Inferences based on the binomial distribution.** Data consisting of the number of occurrences of one from two possible outcomes in

a series of trials often give rise to a binomial distribution and inferences are often required about the parameter  $p$  associated with that outcome. Inferences about quantiles of data from continuous or other measurement distributions may also be based on the value of  $p$  associated with that quantile. Studies of power and the sample sizes needed to obtain a given power against a single specified alternative are relatively straightforward when the binomial distribution is relevant to primary data. Appropriate software and extensive tables are available for binomial inference. When  $p$  is close to 0 or 1 asymptotic results may be reliable only for very large samples.

**Distribution tests.** The Kolmogorov test ([Section 3.3.1](#)) tests whether data are an acceptable fit to a completely specified continuous distribution. Often several distributions may give an adequate fit. In this case graphical methods ([Section 3.3.2](#)) are useful to get an eye comparison of the relative goodness of fit. Lilliefors' test for 'normality' with unspecified mean and variance ([Section 3.3.3](#)) uses the Kolmogorov statistic, but separate tables or relevant software are needed to test for significance or to estimate  $P$ -values. The Shapiro–Wilk test is theoretically more complicated than Lilliefors' test but is often more powerful.

**Angular data.** Directional data ([Section 3.4](#)) are often best represented by points on the circumference of a circle. Some standard general parametric and nonparametric test procedures are directly applicable to angular data but more commonly modifications are needed. Care is needed when defining concepts such as mean and median for angular data.

**Runs test.** The simplest runs test ([Section 3.5](#)) applies to numbers of runs for dichotomous outcomes (e.g. *heads* or *tails*). In a sequence of  $N$  such ordered outcomes the hypothesis of randomness or independence is rejected if there are too few or too many runs. Simple modifications to the test make possible tests for runs above and below the median for more general data; these may provide evidence for detecting certain types of departures from independence.

## EXERCISES

- 3.1 A supermarket indicates in contracts given to suppliers of oranges that not more than 1 per cent of the fruit in any consignment should show visible signs of damage. Realism forces it to agree that it will use a sampling

scheme with producer's risk set at  $P = 0.10$  for rejection of samples with less than 1 per cent blemished and consumer's risk set at  $1 - P^* = 0.05$  of accepting batches with 3 per cent or more blemished. The sampling scheme used is to select a number  $n$  from a large batch of  $N$ . For each of the sample sizes  $n = 100, 200, 300, 400, 500$  determine the maximum number of blemished fruit that would indicate the condition for the producer's risk is being met. In each case find the corresponding consumer's risk. What is the smallest value of  $n$  and for what  $r$ , the maximum number of permissible defectives in an acceptable batch of that size  $n$ , will the above conditions for producer's and consumer's risks be met?

- 3.2 A commentator on the 1987 Open Golf Championship asserted that on a good day 10 per cent of top-class players could be expected to complete a round with a score of 68 or less. On the fourth day of the championship weather conditions were poor and the commentator remarked before play started that one might expect the weather to increase scores by four strokes per round, implying that 10 per cent of players might be expected to return scores of 72 or less. In the event 26 of the 77 players competing returned scores of 72 or less. Regarding the players as a sample of 77 top-class players and assuming the commentator's assertion about scores on a good day is correct, do these fourth-day results suggest the commentator's assertion about scores in the poor weather conditions prevailing was (i) perhaps correct, (ii) optimistic or (iii) pessimistic?
- 3.3 In a pilot opinion poll 18 voters from one electorate selected at random were asked if they thought the British Prime Minister was doing a good job. Six (one-third) said 'yes' and twelve (two-thirds) said 'no'. Is this sufficient evidence to reject the hypothesis that 50 per cent of the electorate think the Prime Minister is doing a good job? The pilot results were checked by taking a larger sample of 225 voters. By coincidence 75 (one-third) answered 'yes' and 150 (two-thirds) answered 'no'. Do we draw the same conclusion about the hypothesis that 50 per cent of the electorate think the Prime Minister is doing a good job? If not, why not?
- 3.4 The journal *Biometrics* 1985, **41**, p. 830, gives data on numbers of medical papers published annually in that journal for the period 1971–81. These data are extended below to cover (in order) the period 1969–85. Is there evidence of a monotonic trend in numbers of medical papers published?

11 6 14 13 18 14 11 22 19 19 25 24 38 19 25 31 19

- 3.5 The UK Meteorological Office monthly weather summaries published by HMSO give the following annual rainfalls in mm for 15 stations in the UK during 1978. The stations are listed in order of increasing latitude. Is there evidence of a monotonic trend in rainfall from South to North?
- Margate, 443; Kew, 598; Cheltenham, 738; Cambridge, 556; Birmingham, 729; Cromer, 646; York, 654; Carlisle, 739; Newcastle, 742; Edinburgh, 699; Callander, 1596; Dundee, 867; Aberdeen, 877; Nairn, 642; Baltasound, 1142.

- 3.6 Rogerson (1987) gave the following annual mobility rates (percentage of population living in different houses at the end of the year than at the beginning) for people of all ages in the USA for 28 consecutive post-war years. Is there evidence of a monotonic trend?

18.8	18.7	21.0	19.8	20.1	18.6	19.9	20.5	19.4	19.8	19.9
19.4	20.0	19.1	19.4	19.6	20.1	19.3	18.3	18.8	18.3	18.4
17.9	17.1	16.6	16.6	16.1	16.8					

- 3.7 Are the insurance claim data in [Exercise 2.18](#) likely to have come from a normal distribution? Test using Lilliefors' test and the Shapiro–Wilk test.
- 3.8 Test whether the data in [Exercise 2.18](#) may have come from a uniform distribution over the interval (400, 7000).
- 3.9 The negative exponential distribution with mean 20 has the cumulative distribution function  $F(x) = 1 - e^{-x/20}$ ,  $0 \leq x \leq \infty$ . Use a Kolmogorov test to determine if it is reasonable to assume the excess parking times in [Exercise 2.21](#) are a sample from this distribution.
- 3.10 In the data sets in the appendix we give the ages of death for 21 members of the McGamma clan. Perform an appropriate test to determine whether it is reasonable to assume age at death is normally distributed for that clan.
- 3.11 A psychologist is testing 16 applicants for a job one at a time. Each has to perform a series of tests and the psychologist awards an overall point score to each applicant. As each applicant may discuss the tests with later applicants before the latter are tested it is suggested that those tested later may have an unfair advantage. Do the applicants' scores (in order of testing) below support this assertion?

62 69 55 71 64 68 72 75 49 74 81 83 77 79 89 42

Use an appropriate runs test. Do you consider the Cox–Stuart test ([Section 3.2.3](#)) may also be appropriate? Give reasons for your decision.

- 3.12 A circular ring road is constructed around a town. A map of the local area shows the town hall is situated at the centre of the circle. The positions of road accidents along the ring road are measured as bearings from the town hall. During a period of road works to the east of the town, the accidents that occur have bearings  $10^\circ, 35^\circ, 82^\circ, 87^\circ, 94^\circ, 108^\circ, 125^\circ$ . Is there any evidence that the accidents are linked to the road works?
- 3.13 For a children's television quiz show, a circular board with a stationary vertical pointer attached to the centre is divided into quadrants. The board is spun and contestants are asked questions on sport, popular music, current affairs and science according to the quadrant indicated when the board comes to rest. Angles are marked around the edge of the board. For ten consecutive questions, the angles indicated at rest are  $15^\circ, 46^\circ, 114^\circ, 137^\circ, 165^\circ, 183^\circ, 195^\circ, 215^\circ, 271^\circ$ , and  $328^\circ$ . Do you think that on average all types of question are equally likely?
- 3.14 The Office for National Statistics (1998) gives the numbers of deaths from railway accidents in England and Wales for 1981 to 1996 as 76, 92, 105, 86, 91, 81, 103, 92, 71, 132, 71, 57, 48, 63, 43, 60. Use an appropriate form (or forms) of the runs test to examine nonrandomness.
- 3.15 An archer fires arrows at a target that is on a bearing of  $145^\circ$  to him. The angles of fire for ten arrows are  $139^\circ, 141^\circ, 146^\circ, 148^\circ, 150^\circ, 152^\circ, 153^\circ, 155^\circ, 158^\circ, 160^\circ$ . There is concern that the arrows are tending to fall right of target so one wants to test the hypothesis  $H_0: \theta = 145$  against the alternative  $H_1: \theta \neq 145$ . Use the sign test and the Wilcoxon test to do this. Would you question the validity of using the latter test for these data?

# 4

## Methods for paired samples

### 4.1 COMPARISONS IN PAIRS

Single-sample methods illustrate basic ideas, but have limited applications. Practical problems usually involve two or more samples that may or may not be independent. We consider paired samples in this chapter, a dependent-sample situation where many problems reduce to a single-sample equivalent.

#### 4.1.1 Studies using pairing

To compare two stimuli when it is practicable it is sometimes appropriate and advantageous to apply both to each of, say,  $n$  individuals. The main reason for this is that responses to any stimulus often vary markedly between individuals. However, some important issues must be addressed before proceeding this way.

Suppose that a dentist wishes to compare two mouthwashes for the treatment of ulcerative gingivitis in the mouth. The dentist could use the number and size of the mouth ulcers in a patient to assess the severity of the problem. The patients could also indicate their perceived level of pain due to these ulcers. One way of comparing the two mouthwashes, known as a cross-over trial, is to treat a patient first with one mouthwash for a certain period, conduct an assessment of the patient's response and then transfer the patient to the other mouthwash. Following the second course of treatment the patient is again assessed. At the end of the study the two treatments are compared on the basis of findings from the patient assessments.

An advantage of this design is that patients act as their own control. In an ideal world the only difference between the two time periods would be the type of mouthwash used, which would make a comparison straightforward. Unfortunately, there may be a 'carryover' effect of the first treatment into the second period if the second mouthwash is administered before the effects of the first have completely worn off. Sometimes an interval or 'wash out' period is allowed between the two treatments in an attempt to reduce

this effect. Also, response to either mouthwash may be affected by the prior or subsequent administration of the other (interaction is the technical statistical term). In an extreme case administration of one mouthwash after the other might make a patient's condition worse, whereas if either were given separately, both may be beneficial. In passing we note that certain mouthwashes contain alcohol so the consumption of alcoholic drinks during any such treatment period may be unwise or may influence the outcome.

We do not consider cross-over trials in detail but analysis of a simple experiment of this kind is described by Koch (1972) and a more general account of such designs is given by Senn (1992).

In addition, this type of study can only be performed with chronic conditions where the first treatment received is unlikely to lead to complete recovery or death. Ulcerative gingivitis is such a condition (asthma and eczema are others) in which for many patients, improvement under treatment is swift but the condition deteriorates once the treatment is withdrawn. The assessment of ulceration by a dentist is notoriously subjective, however, as is patient assessment of pain. To allow for this a **double blind** procedure should be employed in which neither the patient nor the dentist making the assessments knows the order of presentation (which should be random). Mouthwashes can be made to have identical appearance and taste. The solutions are identified by codes available only to research workers or administrators who have no direct contact with the patient (if tablets are used they should be of the same size, colour and taste). Double blinding can also be used to allow for the placebo effect which is a psychological response shown by some patients to the knowledge that treatment is being received even though the solution or tablet (known as a placebo) does not contain an active ingredient. These aspects of a study should be carefully explained in the patient information sheet if applicable.

Ethical problems may preclude the use of the same patients to compare two treatments. For instance, in a comparison of two methods of treating oral cancer, both of which involve radiation, the estimated dose of radiation from the combined treatments may be considered unacceptably high for the patient.

If it is undesirable for ethical or practical reasons to give patients both treatments, we might use pairs of individuals chosen so that the members of each pair are as like as possible in all relevant characteristics. This process is known as matching. For example, a woman aged 30 years might be matched with a similarly aged woman of the same ethnic group, each showing a similar relevant

morbidity status. In each pair one of the women chosen at random receives the first treatment and the other the second. In this case we look at the assessment differences between members of each pair.

As mentioned in [Section 1.5](#), patients should be free to withdraw from a study at any stage. The consequence of this is that some patients may not be taking their allocated treatment by the close of the study and some may have withdrawn from the study altogether. It is recommended that the results of studies that compare two or more groups of patients be analyzed on an **intention to treat** basis in which the groups are compared as they were originally chosen. An attractive alternative might be to analyze the patients by their actual treatment at the end of the study. This would introduce bias however, as those who are unwilling or unable to continue in the study are likely to have different characteristics (e.g. greater health problems) compared with the other patients. ‘Intention to treat’ analysis also gives a fairer assessment of the impact of these strategies in the real world. These comments are relevant to planning an experiment, no matter whether we use a parametric or nonparametric method of analysis.

Returning to the study of mouth ulcer treatment, a patient’s assessment of pain may involve no more than indicating which mouthwash, if either, gave the greater relief (e.g. first mouthwash gave most relief, second mouthwash gave most relief, both equally good or both ineffective). One may score *mouthwash A gave more relief* as a **plus**, *mouthwash B gave more relief* as a **minus** and *no difference* as a **zero** or ‘no score’. Individual patient scores – plus, minus, or zero – provide the basis of a sign test of  $H_0 : \text{drugs are equally effective}$  against an appropriate one- or two-tail alternative.

### 4.1.2 Further examples

We give four specific examples.

I. Geffen, Bradshaw and Nettleton (1973) wanted to know whether certain numbers presented in random order were perceived more rapidly in the right (RVF) or left visual fields (LVF), or whether there was no consistent difference, it being a matter of chance whether an individual responded more quickly in one field or the other. For each of 12 subjects the mean response times to digital information in each field was measured. Response times varied much more between individuals than they did between fields for any individual. The data and the differences LVF – RVF for each

**Table 4.1** Mean response time (ms) to digital information.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
LVF (1)	564	521	495	564	560	481	545	478	580	484	539	467
RVF (2)	557	505	465	562	544	448	531	458	560	485	520	445
(1) – (2)	7	16	30	2	16	33	14	20	20	-1	19	22

individual are given in Table 4.1. This table shows quicker response times in the RVF for all but subject 10. No difference exceeds 33ms, whereas in either field differences between some individuals exceed 100ms. For example, it is  $580 - 467 = 113$  between subjects 9 and 12 in the LVF. Without matched pairs these differences might swamp the smaller but relatively consistent differences between fields for individuals. We explore this further in [Exercise 5.11](#).

*II.* A course organizer might compare two teaching methods such as lectures and computer assisted learning (CAL) by pairing students so that, if possible, each member of a pair is of the same gender and has the same previous knowledge of the subject. For each pair, one member is allocated to lecture classes and the other to the CAL material. At the end of the course the students take the same tests and the results are interpreted in terms of pairwise differences.

*III.* Using double blind marking (a method favoured by some institutions) one can compare consistency between two examiners. The two examiners mark the same series of essays without disclosing their assessments to one another. The differences between the marks awarded by each examiner for each essay are then compared to see whether one examiner consistently awards a higher mark, or whether differences have some other pattern or whether they appear to be purely random.

*IV.* To compare two animal diets using pairs of twin lambs the diets are fed one to each twin, and growth is measured over a period: attention is focused on growth differences between twins within each pair. Because of genetic similarity each of a pair of twin lambs fed on identical diets tends to grow at a similar rate; when fed different diets any consistent differences in growth may be attributed to the effects of diet.

In summary, the aim of pairing is to make conditions, other than the treatment or factor under investigation, as like as possible within each pair; the differences within pairs provide a measure of any treatment effect that takes the form of a ‘shift’ in the distribution.

### 4.1.3 Single-sample analysis of matched pairs

Differences between paired observations provide a single sample that can be analyzed by methods developed in Chapters 2 and 3. We must, however, consider the assumptions about observations on each member of the pair and what precautions in experimental procedure are needed to validate analyses. These points are best brought out by examples.

#### *Example 4.1*

*The problem.* Using the LVF, RVF data in [Table 4.1](#), assess the strength of the evidence for a consistent response difference between the two fields for individuals. Obtain 95 per cent confidence intervals for that difference.

*Formulation and assumptions.* We denote the observation on subject  $i$  in the RVF by  $x_i$  and that in the LVF by  $y_i$  and analyze the differences  $d_i = y_i - x_i$ ,  $i = 1, 2, \dots, n$ . The  $d_i$  are independent since each refers to a different individual. Under  $H_0$ : *the median of the differences is zero* the  $d_i$  are equally likely to be positive or negative and a sign test is justified. If we assume a symmetric distribution of the  $d_i$  under  $H_0$  we may use a Wilcoxon signed-rank (or a normal scores) test. There are several different response patterns in the two visual fields that could result in a symmetric distribution of the differences under  $H_0$ . In particular, if we assume response times are identically and independently (but not necessarily symmetrically) distributed in the two fields for any individual (but this distribution need not be the same for every individual) the difference for each individual will be symmetric about zero (since if  $X$  and  $Y$  have the same independent distributions, then  $X - Y$  and  $Y - X$  will each have the same distribution and therefore must be symmetrically distributed about zero). We also get a symmetric distribution of differences if the LVF and RVF for any individual have different symmetric distributions providing each has the same mean. It is well to be aware of such subtleties, but a Wilcoxon test is clearly justified when we assume identical distributions for LVF and RVF for any one individual under the null hypothesis; the alternative hypothesis of interest is usually that there is a shift in centrality only, as indicated by a shift in one field or the other in the median of otherwise identical distributions. Often the  $d_i$  values themselves indicate whether there is serious asymmetry that suggests this condition may not hold.

If we assume also that the differences are approximately normally distributed, a  $t$ -test is appropriate. We may test whether normality is a reasonable assumption for the  $d_i$  by Lilliefors' test or by the Shapiro–Wilk test, but these tests may have low power for small samples.

*Procedure.* We work with the single sample composed of the differences  $d_i$  and use procedures developed in [Chapter 2](#), so we only sketch details. Denoting the centrality parameter or measure of treatment effect (assuming this to be only a ‘centrality’ or ‘location’ shift) by  $\theta$ , the null hypothesis is  $H_0: \theta = 0$ , so the ordered signed deviations are simply the ordered differences obtainable from [Table 4.1](#) i.e.:

-1 2 7 14 16 16 19 20 20 22 30 33

and the corresponding signed ranks are

-1 2 3 4 5.5 5.5 7 8.5 8.5 10 11 12

Using the sign test for zero median difference we have 1 minus and 11 plus signs. Appropriate tables or software immediately give  $P = 0.0063$ , providing very strong evidence that responses differ between fields. The nominal 95 per cent confidence interval based on the sign test is (7, 22) since we reject  $H_0$  at a nominal  $P = 0.05$  level only for 2 or fewer or 10 or more minus signs. For the Wilcoxon test  $S_- = 1$ , corresponding to a two-tail  $P < 0.001$ . Indeed, you should only need a pocket calculator to show that  $P \approx 0.00098$  in this example!

For a confidence interval with coverage of at least 95 per cent tables indicate we require the 14 greatest and least Walsh averages. These averages may be obtained using Minitab, else StatXact may be used to find the interval. Many other software packages include a program giving at least an asymptotic approximation. The interval turns out to be (9.5, 23.5). The Hodges–Lehmann point estimator of the median (i.e. the median of the Walsh averages) is 17.25.

If we assume normality, a two-tail  $t$ -test gives  $P < 0.001$  and a 95 per cent confidence interval (10.1, 22.9) centred about the mean 16.5. Normal theory gives the shortest confidence interval and the fact that the interval is only slightly displaced for the sign test implies reasonable symmetry. Lilliefors' test statistic, 0.152, is well below the value required for significance and for that test StatXact gave a Monte Carlo estimate for the exact  $P = 0.617$ . For the Shapiro–Wilk test StatXact gave  $P = 0.685$ , so there is no evidence of nonnormality.

*Conclusion.* The sign, Wilcoxon and  $t$ -tests all point to strong evidence against the null hypothesis and we conclude that response times in the RVF are faster. Consistency of the confidence intervals given by these approaches suggests the mean difference is between about 10 and 22 ms.

*Comments.* 1. The set of differences we compare are independent (a necessary condition for validity of our centrality tests) because each difference is calculated for a different individual.

2. If the response rate in the LVF had been measured before that in the RVF for all subjects a difficulty in interpretation would arise. The result might then imply a learning process, people responding more quickly in the RVF because they were learning to react more rapidly; or there could be a mixture of a learning effect and an inherent faster response in the RVF. We avoid this difficulty if we decide at random which field – left or right – is to be tested first for each individual; that was done in this experiment. This should balance out and largely annul any learning factor. Another approach is to achieve balance by selecting six subjects (preferably at random) to be tested first in the LVF. The remaining six are tested first in the RVF. Such balanced designs provide a basis for separating a learning effect from an inherent difference between field responses, although a somewhat larger experiment would be needed to do this by appropriate parametric or nonparametric methods.

*Computational aspects.* Most statistical packages that have programs for the procedures discussed in this example allow one to enter either the data for each field for each individual and then computes differences automatically or else to enter the differences themselves as the raw data.

A modification to the above procedure lets us test the hypothesis that a centrality difference has a prespecified value  $\theta$ . As in Chapters 2 and 3, we simply consider deviations from that value. This effectively shifts the origin to the hypothetical median, so that for our revised data (the deviations) we test the hypothesis that the centrality parameter for that population takes the value zero.

#### *Example 4.2*

*The problem.* Eleven children are given an arithmetic test; after 3 weeks' special tuition they are given a further test of equal difficulty (we say more about this in the *Comments*). Their marks in each test (out of 90) and the individual differences are given in Table 4.2. Do these support a claim that the average improvement due to extra tuition is 10 marks?

*Formulation and assumptions.* The question essentially is whether, if we assume these children are a random sample from some hypothetical population (perhaps children of the same age trained in the same educational system or studying the same syllabus), it is reasonable to suppose the mean mark difference is 10?

*Procedure.* We consider deviations from 10 for the mark differences in the last line of Table 4.2. These deviations are:

$$-2 \quad -4 \quad 4 \quad -5 \quad 0 \quad -8 \quad -1 \quad 9 \quad 1 \quad -6 \quad -5$$

The differences arranged in order of magnitude with appropriate signs are

$$0 \quad -1 \quad 1 \quad -2 \quad -4 \quad 4 \quad -5 \quad -5 \quad -6 \quad -8 \quad 9$$

and the ranks without signs are

$$1 \quad 2.5 \quad 2.5 \quad 4 \quad 5.5 \quad 5.5 \quad 7.5 \quad 7.5 \quad 9 \quad 10 \quad 11$$

Using the rules given in [Section 2.2.3](#) for mid-ranks with ties together with the convention mentioned at the end of that section for replacing the rank 1 associated with the zero difference by 0, we get signed ranks:

$$0 \quad -2.5 \quad 2.5 \quad -4 \quad -5.5 \quad 5.5 \quad -7.5 \quad -7.5 \quad -9 \quad -10 \quad 11$$

Our statistic is the sum of the positive ranks  $S_+ = 0 + 2.5 + 5.5 + 11 = 19$ . For the exact permutation test StatXact gives  $\Pr(S_+ \leq 19.0) = 0.124$ . Doubling this value for a two-tail test we immediately see there is no strong evidence against  $H_0$ , since  $P = 0.248$ .

**Table 4.2** Marks (out of 90) in two arithmetic tests.

<i>Pupil</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>
<i>First test</i>	45	61	33	29	21	47	53	32	37	25	81
<i>Second test</i>	53	67	47	34	31	49	62	51	48	29	86
<i>Second – First</i>	8	6	14	5	10	2	9	19	11	4	5

While the sample is too small to give confidence in an asymptotic procedure for this example this gives a two-tail  $P = 0.229$ , comparing reasonably well with the exact  $P = 0.248$ . Also  $S_+ = 19$  exceeds the nominal  $P = 0.05$  value,  $S = 10$ , given in published tables for  $n = 11$  in the no-tie situation.

*Conclusion.* There is no firm evidence against the hypothesis that the mean improvement may be 10.

*Comments.* 1. How does one decide if two arithmetic tests are equally difficult? Might not the improved marks in the second test imply it was easier? The statistician should seek the assurance of the educationalist conducting the test that reasonable precautions have been taken to ensure equal difficulty. Sometimes standard tests that have been tried on large groups of students with results that show convincingly that they are for all practical purposes of equal difficulty, are used in such situations.

2. Confidence intervals for the mean or median difference may be obtained in the usual way. The nominal 95 per cent confidence interval for these data given by StatXact is (5, 12), in close agreement with the normal theory  $t$ -test interval (5.14, 11.76).

3. It may not be realistic to test simply for a centrality shift. Sometimes pupils who perform well initially benefit little from extra tuition; in this example Pupil K cannot possibly improve by more than 9 marks. Likewise very poor pupils may find the concepts of arithmetic difficult to grasp and gain little from extra tuition. Often only those in the mid-ability range show appreciable benefit. A statistician may find evidence of this simply by looking at the data – and there are indeed tests for such tendencies. Deciding what should be tested or estimated may be a topic for fruitful discussion between statistician and experimenter.

Example 4.3 is based on data for a group of 77 first-year medical students available to one of us (NCS).

#### *Example 4.3*

*The problem.* The data below are differences in *systolic blood pressure after exercise – systolic blood pressure before exercise* measured in mm Hg for a random sample of 24 from the group of 77 students. Obtain 95 per cent confidence limits for the population difference based on (i) the sign test, (ii) the Wilcoxon signed-rank test and (iii) the  $t$ -test, and comment on the appropriateness of each. For convenience we have arranged the differences in increasing order.

–5 –5 0 2 10 15 15 15 18 20 20 20 20 22 30 30 30 34 40 40 40 40 41 47  
80 85

*Formulation and assumptions.* We use standard methods developed in Chapter 2 and in this chapter.

*Procedure.* We omit computational details (see [Exercise 4.16](#)) since examples of similar calculations have already been given. The intervals quoted below were obtained using relevant statistical software.

*Conclusion.* The nominal 95 per cent confidence intervals are:

Sign test	(15, 40)
Wilcoxon	(17.5, 35)
<i>t</i> -test	(16.87, 35.96)

*Comments.* 1. Not surprisingly, the sign test interval is the longest. There is slight evidence of skewness, but none of the intervals is seriously displaced relative to the others. The Wilcoxon interval being shorter than that based on normal theory may be attributed to a slightly heavier ‘upper tail’ (especially the values 80, 85) than one would expect with a sample from a normal distribution. Lilliefors’ test for normality gives a test statistic value 0.162, and an estimated  $P = 0.1071$  based on 10 000 simulations using StatXact. For the Shapiro–Wilk test there is stronger evidence against normality, the statistic 0.8999 corresponding to  $P = 0.0205$ .

2. The only reason for taking a random sample of 24 from 77 observations was to provide a convenient illustrative example. Almost certainly one would use all observations in any detailed studies of the effects of exercise on blood pressure (and interest would be in the distribution of the differences and not just in the mean or median difference).

3. The variation in the size of the differences is unsurprising with blood pressure measurements taken by inexperienced first-year medical students. Readings are usually recorded to the nearest mm Hg but in practice they may only be recorded to the nearest 5 mm Hg by those unfamiliar with the technique. The student exercises would be performed with a range of enthusiasm. A few extremely high differences and a few small negative changes are therefore understandable.

4. Notwithstanding Comment 3, there was clearly an error on the original computer print-out we used. For one student the systolic blood pressure measurement after exercise was recorded as 15 mm Hg and gave a difference of  $15 - 118 = -103$ . A systolic blood pressure reading of less than 80 mm Hg is unlikely (unpublished data from South London). The value of 15 mm Hg for systolic blood pressure after exercise is clearly incorrect! There is a strong suspicion that the final digit has been omitted and that the true reading should be between 150 and 159. It is easy to spot such a discrepancy on a print-out and in practice one would then attempt to track down the source of error (and if possible make the needed correction). Possible sources of error are failure of a printer to reproduce a character, a mistake in entering the original data or in initial reading or recording of the data by an inattentive student. The original purpose and manner of the data collection could affect the likely accuracy of the recorded results. With increasing use of computer packages to process data such an error might go undetected. Had this value,  $-103$ , been included in our sample of 24 in place of the entry  $-5$  we would have obtained the following 95 per cent confidence intervals:

Sign test	(15, 40)
Wilcoxon	(17.5, 34.5)
<i>t</i> -test	(7.83, 36.83)

The reader who has difficulty explaining the differences between these results and those recorded above in terms of the effect of an ‘outlier’ should refer to [Section 2.3.2](#). In particular it is worth noting the Wilcoxon interval is little

changed because the introduction of a value  $-103$  has a tendency to make the sample distribution, if anything, a little more symmetric than the original sample. However, the tails become rather longer than one would expect for a normal distribution and this has the effect of elongating the  $t$ -based interval. For these amended data Lilliefors' test gave a Monte Carlo estimated  $P = 0.0083$  and a Shapiro-Wilk test gave  $P = 0.0001$ , both hinting strongly at nonnormality implicit in the tail values.

Practical experience in handling data highlights the crucial importance of detecting data errors. Many computer programs include output to help detect outliers; e.g. a print-out of maximum and minimum values often (but by no means always) highlights a glaring data error.

## 4.2 A LESS OBVIOUS USE OF THE SIGN TEST

The way the data are presented in Example 4.4 does not make it obvious that a sign test is relevant.

### *Example 4.4*

*The problem.* Members of a mountaineering club have long argued about which of two rock climbs is the more difficult. Hoping to settle the argument one member checks the club log book. This records for any climb by a member whether it is successfully completed. The log shows that 108 members have attempted both climbs with the outcomes summarized in Table 4.3. Is there evidence that one climb is more difficult?

*Formulation and assumptions.* A moment's reflection shows that a climber succeeding in both climbs, or failing in both, provides no information about relative difficulty; such cases are **ties** so far as comparing difficulty is concerned. If we had additional information, for instance about each climber's personal assessment of the difficulty, the situation would be different. As it is, our only meaningful comparators of difficulty are numbers who succeed at one climb, but fail at the other. From Table 4.3 we see that 9 succeed at the first climb but fail at the second; we may think of this as a 'plus' for the first climb. Also 14 fail on the first climb but succeed at the second; we may think of this as a 'minus' for the first climb.

**Table 4.3** Outcomes of two rock climbs.

		<i>First climb</i>	
		<i>Succeeded</i>	<i>Failed</i>
<i>Second climb</i>	<i>Succeeded</i>	73	14
	<i>Failed</i>	9	12

This is a sign test situation. If the climbs are of equal difficulty a ‘plus’ or ‘minus’ is equally likely for each climber. Thus under  $H_0$ : *climbs are equally difficult* the number of ‘plus’ signs has a  $B(n, \frac{1}{2})$  distribution where  $n$  is the total number of plus and minus signs.

*Procedure.* There are 9 plus and 14 minus signs so  $n = 23$ . StatXact gives an exact test  $P = 0.4049$ . Alternatively, using the normal approximation to the binomial distribution we find, using (2.2), that  $Z = (9.5 - 11.5)/\sqrt{(23/4)} = -0.83$  giving a two-tail  $P = 0.4065$ , so there is no evidence that one climb is easier.

*Conclusion.* We retain  $H_0$ : *climbs are equally difficult*.

*Comments.* 1. Because  $73 + 12 = 85$  from 108 pairs provide no information on the relative difficulty of the climbs this may seem wasted data, but such ‘wasted’ data give an indication of how big or small any difference might be. In some situations we need many observations because we are looking for a small difference which may not be clearly distinguishable when the only relevant criteria are success/failure or failure/success categories. In the context of these data it is likely that most of those who failed at both were less experienced or less enthusiastic than those who achieved one success and also that those who succeeded at both were the more experienced or the more enthusiastic.

2. The continuity correction of replacing 9 by 9.5 in the numerator of the asymptotic approximation has a marked effect in this example. If it is omitted the two-tail  $P$  is reduced from 0.4065 to 0.2971. Our conclusion, however, in this case would not be altered by this appreciable numerical change.

3. It would be interesting to know whether club members generally attempted the climbs in a particular order as a learning effect might then be involved.

*Computational aspects.* StatXact and also many general packages include programs for this simple test that may be used to obtain the relevant  $P$ -value for any sign test.

The test is called **McNemar’s test**, having been proposed by McNemar (1947). Conover (1999, [Section 3.5](#)) presents it more formally, but it is effectively a paired-sample sign test.

### 4.3 POWER AND SAMPLE SIZE

In [Section 3.1.3](#) we explored the power of some tests for binomial probabilities of the form  $H_0: p = p_0$  against single-valued alternatives such as  $H_1: p = p_1$  where choice of  $p_0$ ,  $p_1$  was determined by their relevance to producer’s and consumer’s risks associated with a sampling scheme. Exact power calculations were relatively straightforward and we illustrated their use in finding sample sizes to meet specified producer’s and consumer’s risks. We now consider relationships between power and sample size for single-sample centrality tests relevant both to basic single-sample situations or to differences between matched pairs. Exact results are only available

in a limited number of cases and in this context the power for a specified alternative is in general not distribution-free even when the test itself is. This restricts the overall value of exact power computations. Nevertheless, providing one exercises care in interpretation, computing approximate power for given sample sizes before beginning an experiment helps one make efficient use of resources. We first highlight some key features and limitations using simple examples for the sign test, where only relatively straightforward computations are needed. We then consider briefly the Wilcoxon signed-rank test where power and sample size determinations introduce greater difficulties both in theory and practice.

### 4.3.1 Power and sample size for the sign test

[Examples 4.5 to 4.7](#) cover three situations where we use a sign test for a median  $\theta$  of the form  $H_0: \theta = \theta_0$  against  $H_1: \theta > \theta_0$  where a specific alternative  $H_1: \theta = \theta_0 + 1$  reflects a minimal departure from  $H_0$  of interest, i.e. only a positive shift in median of at least one unit is of practical importance. Without loss of generality (see, e.g. [Example 4.2](#)) we set  $\theta_0 = 0$ . We confine attention to a one-tail test, but modifications for a two-tail test are straightforward if one uses the principle of doubling one-tail  $P$ -values for a two-tail equivalent.

If we compute the power of a test for the alternative  $H_1: \theta = 1$  then if the true shift exceeds one unit the power will in general be greater. From these examples it will emerge that for three different population distributions all having the same variance the power of the equivalent tests is different for each, confirming the statement above that power is not a distribution-free property. This should cause no surprise in the light of the well-known result that the Pitman efficiency of tests depends upon the underlying population distribution.

#### *Example 4.5*

*The problem.* For a sample of 10 from a normal distribution with unknown mean  $\theta$  and standard deviation 2 what is the power of the sign test for  $H_0: \theta = 0$  against  $H_1: \theta = 1$  when the probability of a type I error is  $\alpha = 0.05$ ? What sample size would ensure a test with power 0.9?

*Formulation and assumptions.* Under  $H_0$  the number of positive sample values (pluses) has a  $B(10, \frac{1}{2})$  distribution. If the number of plus signs is the test statistic then under  $H_1$  these will still have a binomial distribution, but the value of  $p$  is now given by  $\Pr(X > 0) = p_1$ , say, where  $X$  has a  $N(1, 4)$  distribution. Thus the test is equivalent to testing  $H_0: p = \frac{1}{2}$  against  $H_1: p = p_1$  and once  $p_1$  is calculated the power study when  $n = 10$  and that for determining the sample size

required for any specified power proceed in the way described in [Section 3.1.3](#).

*Procedure.* Since  $X$  is  $N(1, 4)$  if  $H_1$  holds it follows that  $Z = (X - 1)/2$  has a standard normal distribution, whence  $p_1 = \Pr(X > 0) = \Pr(Z > -\frac{1}{2}) = 0.6915$ , a value obtainable from tables or appropriate software. The power may then be obtained from tables or software in the way described in [Example 3.3](#), where here we test  $H_0: p = 0.5$  against  $H_1: p = 0.6915$ . Using computer software is preferable, for it is unlikely that tables will be readily available for precisely the latter value of  $p$  and an approximation such as  $p = 0.7$  may then have to be used. With these values of  $p$ , StatXact gives the power as 0.1365 for sample size 10. However, discontinuities are influential here as the smallest possible exact one-tail  $P$ -values under  $H_0$  are  $P = 0.001$ ,  $P = 0.011$  and  $P = 0.055$ . Had we replaced  $\alpha = 0.05$  by  $\alpha = 0.011$  the power would still be 0.1365 whereas if we set  $\alpha = 0.055$  the power becomes 0.3604. Minitab uses an asymptotic approximation for power, giving this as 0.3914 when  $\alpha = 0.05$ . One has reservations about asymptotic results for so small a sample, but 0.3914 is broadly in line with the exact values obtained for possible type I errors slightly above a nominal 0.05.

The asymptotic result should be more satisfactory for finding the larger  $n$  needed to ensure a power of, say, at least 0.90. The Minitab program gives this as  $n = 55$ . Finer tuning using exact results from StatXact gives  $n = 58$  as the minimum required sample size corresponding to an exact  $P = 0.0435$  and an exact power 0.9031.

*Conclusion.* For a sample of  $n = 10$  the power is only 0.1365 but this low power in part reflects the large gap between the possible  $P$  values 0.011 and 0.055. For power at least 0.90 for the one-tail test the sample size should be at least 58.

*Comments.* 1. The Pitman efficiency of the sign test relative to the  $t$ -test for samples from a normal population is 0.64. This implies that for large  $n$  the relative efficiency is less than two-thirds, suggesting that a sample of about two-thirds the size should have the same power if we apply a  $t$ -test. Many statistical packages allow power calculations for the  $t$ -test when it is optimal, and in this case these indicate that for a  $N(\theta, 4)$  distribution the necessary sample size with power 0.90 with our chosen values of  $\theta$  in  $H_0$  and  $H_1$  is  $n = 36$ , broadly in line with that suggested by the Pitman efficiency.

2. Our computed  $p_1 = 0.6915$  was based on the strong assumption that our sample was from a  $N(\theta, 4)$  distribution. We seldom know the population variance but not the mean and if we had such information (or could only assume normality with the variance unknown) clearly one should prefer the normal theory-based inference to that using the sign test. However, the results obtained here are illuminating for comparisons with the situations covered in [Examples 4.6](#) and [4.7](#) where a nonnormal population again with a variance of 4 is assumed.

#### *Example 4.6*

*The problem.* Given a sample of 10 from a double-exponential distribution with unknown mean  $\theta$  and standard deviation 2 find the power of the sign test for  $H_0: \theta = 0$  against  $H_1: \theta = 1$  if the probability of a type I error is  $\alpha = 0.05$ . What sample size would ensure a power at least 0.9?

*Formulation and assumptions.* The double-exponential distribution, also known as the Laplace distribution, with mean  $\theta$  and standard deviation 2 (variance 4) has probability density function

$$f(x) = [1/(2\sqrt{2})]\exp[-|(x-\theta)|/\sqrt{2}]. \quad (4.1)$$

The distribution is symmetric about the mean  $\theta$ . Probabilities associated with the tails are greater than those for the normal distribution with the same mean and variance and this is a classic example of a long-tail symmetric distribution.

Under  $H_0$  the number of positive sample values (pluses) again has a  $B(10, \frac{1}{2})$  distribution. If the number of plus signs is the test statistic these will also have a binomial distribution under  $H_1$  but the parameter  $p$  is now given by  $\Pr(X > 0) = p_1$ , say, where  $X$  has the double-exponential distribution (4.1) with  $\theta = 1$ . Thus the original test is equivalent to testing  $H_0: p = \frac{1}{2}$  against  $H_1: p = p_1$  and the power study for the given sample size and for determining the sample size required for any specified power proceed as described in [Section 3.1.3](#).

*Procedure.* The value of  $p_1$  is given by setting  $\theta = 1$  in (4.1) and integrating over the interval  $(0, \infty)$ . Integration is straightforward ([Exercise 4.17](#)) and gives  $p_1 = 0.7534$ . For testing  $H_0: p = \frac{1}{2}$  against  $H_1: p = 0.7534$  when  $\alpha = 0.05$  StatXact gives the exact power 0.2518. Discontinuities in possible  $p$ -values are the same as those in [Example 4.5](#) and replacing  $\alpha = 0.05$  by  $\alpha = 0.011$  does not alter the power, whereas replacing it by  $\alpha = 0.055$  gives power 0.5358. The asymptotic power computation given by Minitab when  $\alpha = 0.05$  is 0.4805. Again, here the asymptotic result is likely to be more satisfactory when seeking the larger  $n$  needed to ensure a power of at least 0.90. Minitab in this case gives  $n = 30$ . Finer tuning using exact results confirms this minimum required sample size corresponding to an exact  $P = 0.0494$  and gives an exact power 0.9023.

*Conclusion.* For a sample of  $n = 10$  the power is now 0.2518 but this low power in part reflects the large gap between the possible  $P$ -values 0.011 and 0.055. A sample of 30 gives a power at least 0.90 for the one-tail test.

*Comments.* 1. The sample size 30 is approximately half the size (58) needed for the same power when using the sign test with a normal distribution with the same variance. This is in line with the Pitman efficiency of the sign test relative to the  $t$ -test which is 2 when sampling from a double-exponential distribution. We must remember, however, that the  $t$ -test itself has lower efficiency for a sample from a double-exponential distribution than it has for a normal distribution with the same mean and variance.

2. The reason the sign test is appreciably more powerful for testing the same basic hypotheses about  $\theta$  in this test than it was for an equivalent test in the previous example is that although the same values of  $\theta$  are specified in both examples and that for  $H_0$  translates to  $p = 0.5$  in each case, that for  $H_1$  transforms to  $p = 0.6915$  and  $p = 0.7634$  respectively, so that for the second example we have a more marked departure from the value of  $p$  under  $H_0$ . That this produces a more powerful test is in line with our notion of a good test being one where its power increases as the difference between the values specified in  $H_1$  increases relative to a fixed value specified in  $H_0$ .

[Examples 4.5](#) and [4.6](#) featured samples from symmetric distributions. We now consider a skew distribution with a similar one-unit shift to the right. The case here is unusual in practice but it provides a direct comparison with [Examples 4.5](#) and [4.6](#) because we again take a case where we sample from a distribution with a variance of 4.

*Example 4.7.*

*The problem.* Given a sample of 10 from a population with an exponential distribution having probability density function

$$f(x) = \frac{1}{2} \exp\left[-\frac{1}{2}(x + 1.3862 - \theta)\right], \quad x \geq \theta - 1.3862 \quad (4.2)$$

which has median  $\theta$  and standard deviation 2 what is the power of the sign test of  $H_0: \theta = 0$  against  $H_1: \theta = 1$  if the probability of a type I error is  $\alpha = 0.05$ . Find also the sample size that will ensure a test with power at least 0.9.

*Formulation and assumptions.* The distribution specified by (4.2) is skew with a long right tail. As in the preceding example under  $H_0$  the number of positive sample values (pluses) has a  $B(10, \frac{1}{2})$  distribution. Using the number of plus signs as the test statistic this also has a binomial distribution under  $H_1$  but now  $p_1 = \Pr(X > 0)$ . Thus the original test is equivalent to testing  $H_0: p = \frac{1}{2}$  against  $H_1: p = \Pr(X > 0)$  and the power study for the given sample size and for determining the sample size required for any specified power again proceed as described in [Section 3.1.3](#).

*Procedure.* It is easily shown by straightforward integration of the probability density function over  $(0, \infty)$  that under  $H_1$   $\Pr(X > 0) = 0.8244$  ([Exercise 4.18](#)). For testing  $H_0: p = \frac{1}{2}$  against  $H_1: p = 0.8244$  when  $\alpha = 0.05$  StatXact gives the exact power 0.4539. Discontinuities in possible  $P$ -values described in [Example 4.5](#) are again relevant and if we replace  $\alpha = 0.05$  by  $\alpha = 0.011$  the test still has power 0.4539 whereas if we replace it by  $\alpha = 0.055$  the power becomes 0.7499.

The asymptotic power computation given by Minitab when  $\alpha = 0.05$  is 0.6970. Again, here the asymptotic result is likely to be more satisfactory for finding the larger  $n$  needed to ensure a power of at least 0.90. The Minitab program in this case gives  $n = 17$ . Finer tuning using exact results in StatXact shows  $n = 18$  is the minimum required sample size corresponding to an exact  $P = 0.0481$  and giving an exact power 0.9194.

*Conclusion.* For a sample of  $n = 10$  the power is 0.4539 but this is again influenced by the large gap between the possible  $P$ -values 0.011 and 0.055. To give a power at least 0.90 for the one-tail test a sample of size 18 suffices.

*Comments.* 1. The sign test is particularly useful for skew distributions and in this example where we have a long right tail the test has moderate power for quite small samples for the alternatives considered. The combination of skewness and a long tail reduces the power of a  $t$ -test appreciably as a consequence of the breakdown of normality assumptions.

2. As indicated above the example is unrealistic from the practical viewpoint because we seldom meet exponential distributions where we are interested in

simple shifts of the complete distribution that affect the mean or median alone. We are more often interested in whether a sample is from one or other of a set of exponential distributions where all positive values of  $X$  are possible but where the scale parameter  $\lambda$  is unknown. Changes in  $\lambda$  lead to changes in not only the mean or median but also in the variance and higher moments. This is more complicated than a simple centrality shift. Nevertheless the sign test is still valid for studying such changes providing we base our test on allocation of signs to observations depending upon whether they are above or below the median specified in  $H_0$ . [Exercise 4.19](#) is intended for readers interested in exploring the power and sample-size relationship for this case.

Finding an adequate sample size to give a good chance of detecting median shifts of interest is a useful exercise but one with practical limitations. [Examples 4.5 to 4.7](#) confirm that even if a test is distribution-free the power computations associated with specific hypotheses about a population mean or median are no longer distribution-free. Further, exact power calculations can only be made easily if we assume all observations are from a population with a specific distribution, whereas we have seen that one strength of the sign test is its applicability to situations where each observation may come from a different distribution providing only that all these distributions have a common median. There is also a certain irony in that we may choose a nonparametric or distribution-free test because we are uncertain about the distribution that provides our sample, yet computation of a sample size to guarantee a desired power requires knowledge of the distribution! Nevertheless reasonable approximations are often obtainable if we can make a few rational assumptions about symmetry, length of tails or other prominent characteristics.

A further limitation to the usefulness of power and sample size calculations applies both to parametric and nonparametric inference. Using a test with good power for rejection of  $H_0$  for differences of importance does not imply that if  $H_0$  is rejected then the difference is important. A confidence interval is more informative about this. To illustrate this point, consider the situations in [Examples 4.5 to 4.7](#). In each we assumed a zero median,  $\theta$ , under  $H_0$  and that only a value  $\theta \geq 1$  represented a departure of interest. If we reject  $H_0$  and calculate a 95 per cent confidence interval for  $\theta$  and this turns out to be  $(0.2, 1.3)$  there is considerable doubt about whether  $\theta \geq 1$ , so the departure from  $H_0$  may still be unimportant. We emphasize again the distinction made in [Section 1.4.2](#) between statistical significance and practical importance.

Despite their limitations approximate power computations are a useful first step in many experimental situations. They may indicate

that a proposed experiment is too small to have a realistic chance of picking up interesting departures from  $H_0$ , or more rarely may suggest we are needlessly squandering resources when a smaller experiment would give us all the information we need. Power and sample size studies provide target sizes for experiments likely to be of value, or if we have limited resources indicate whether using all of these will be enough to achieve an experimenter's aims.

Even though the power of a simple test like the sign test is not distribution-free, power studies are still useful if we can make broad assumptions about the type of population distribution we have, e.g. is it symmetric and long tailed, or something more like a uniform distribution? Is it skewed to the right or to the left? Unless it is very small the sample itself will often give some hint as to whether it comes from a distribution with one or more such characteristics. If there is evidence that the distribution is fairly symmetric and has longer tails than those associated with the normal distribution power studies that are optimal for the double-exponential distribution may provide good approximations. Example 4.7 showed that for a particular skew distribution with one long tail the sign test performs well compared to the  $t$ -test and more general studies have shown that this is broadly true for most skew distributions with a long tail. Very often a quick eye inspection of sample values aided perhaps by some exploratory data analysis involving tools like a box and whisker plot will indicate that a sample appears to have come from, say, a fairly symmetric long-tail distribution not unlike the double exponential. We might 'estimate' its variance by the sample variance and for a null hypothesis about the median,  $\theta$ ,  $H_0: \theta = \theta_0$  and a specific alternative  $H_1: \theta = \theta_1$  compute the value  $p_1 = \Pr(X > \theta_0)$  assuming the distribution of  $X$  is really a double-exponential with variance equal to the sample variance. In a situation like this, had we known the true population distribution we are effectively assuming that the correct  $p_1$  is likely to differ little from that for the double-exponential distribution assumption. Suppose that for the alternative hypothesis of interest  $p_1 = 0.82$  for a double-exponential distribution, then we might be conservative and work out the power for this situation and also that for a slightly lower  $p_1$ , say  $p_1 = 0.78$ . It is likely that the true power will lie somewhere near the values given by these approximations.

If there are problems in working out exact sample sizes to ensure a certain power for a sign test because suitable software is unavailable a good asymptotic approximation due to Noether (1987a) may be used. Denoting the probability of a type I error by  $\alpha$

and that of a type II error by  $\beta$ , so that the power is  $1 - \beta$  then the sample size required to obtain that power with a sign test is

$$n = \frac{(z_\alpha + z_\beta)^2}{4(p_1 - \frac{1}{2})^2} \quad (4.3)$$

where  $p_1$  has the meaning assigned to it throughout this section and  $z_\alpha$  is the value of a standard normal variable that must be exceeded corresponding to a  $P$ -value  $\alpha$  when  $H_0$  holds. For example, in a one-tail test with  $\alpha = 0.05$ ,  $z_\alpha = 1.645$ . A similar meaning is attached to  $z_\beta$ .

#### *Example 4.8*

*The problem.* Apply the Noether approximate formula to compute the sample size  $n$  for the test considered in [Example 4.6](#) to give a power of at least 0.9 when  $\alpha = 0.05$  in a one-tail test.

*Formulation and assumptions.* The appropriate value of  $p_1$  was shown in [Example 4.6](#) to be  $p_1 = 0.7534$  and clearly  $z_\alpha = 1.645$  and  $z_\beta = 1.282$ .

*Procedure.* Substitution of the above values in the Noether formula gives  $n \approx 33.4$ .

*Conclusion.* Conservative rounding up suggests a sample size of 34 is appropriate.

*Comments.* 1. This asymptotic result is close to the size  $n = 30$  found in [Example 4.6](#). Bearing in mind that we may frequently use such calculations when there is some uncertainty about the precise population distribution, calculations using the Noether formula will often be adequate in practice.

2. For the problem in [Example 4.5](#) the Noether formula gives  $n = 56.4$  in close agreement with the exact value  $n = 58$ , and in [Example 4.7](#) it gives  $n = 22.9$  (exact value  $n = 18$ ) suggesting the approximation is not unreasonable even when observations are from a highly skewed distribution.

### 4.3.2 Power and sample size for the Wilcoxon signed-rank test

We consider this topic in less detail than we did for the sign test partly because of additional difficulties in performing exact tests but also because the limitations of the approach indicated in our discussion for the sign test mean that approximate results are often all we can get in practice.

A simple property of the sign test under the alternatives  $H_1$  considered in [Section 4.3.1](#) is that no matter from what distribution we are sampling the test statistic, the number of plus signs, has a binomial distribution. For this statistic, for any given value of  $\theta$  in  $H_1$  all that differs between samples from different distributions is the

value of the binomial parameter  $p_1$ . This simplicity is a feature of the statistic used, the number of plus signs.

Although the corresponding statistic for the Wilcoxon signed-rank test, the sum of the positive signed ranks, has a relatively simple symmetric distribution under  $H_0$  it has generally got a rather intractable distribution under any  $H_1$  for all but a few simple population distributions. Exact power calculations have only been made for small sample sizes and for a limited number of population distributions such as the normal and a few  $t$ -distributions with small numbers of degrees of freedom, *see e.g.* Klotz (1963) and Arnold (1965). Even approximations are limited in their usefulness. An approximate formula for power for a given sample size against alternatives close to that in  $H_0$  is discussed in detail by Lehmann (1975, Section 4.2) and an example of its use is given by Hollander and Wolfe (1999, Section 3.1). The result is applicable only if all observations are from the same distribution. In addition, one needs to know the value of the population frequency function at the median or mean specified in  $H_0$  (which without loss of generality may be set at  $\theta = 0$ ) and also the value at this median of the frequency function for the sum of two independent variables having this same distribution. Except for a few distributions such as the normal where the sum also has a normal distribution, computation of the latter requires a good understanding of distribution theory and calculus.

The reason the distribution of the sum of two independent observations comes into the calculation is closely allied to the relevance of the Walsh averages in test and estimation procedures associated with signed ranks. This sum for any two sample values has the same sign as the corresponding Walsh average (which is simply that sum divided by 2) and under  $H_0$ :  $\theta = 0$  the sign is equally likely to be positive or negative, i.e.  $\Pr(x_i + x_j > 0) = \frac{1}{2}$  whereas under  $H_1$ :  $\theta = \theta_1$ , if  $\theta_1$  is positive then  $\Pr(x_i + x_j > 0) = p_1$  where  $p_1 > \frac{1}{2}$ . The value of  $p_1$  depends upon the population distribution and is not always easy to calculate.

Although the signed-rank statistic no longer has a binomial distribution the approximation due to Noether given in (4.3) may still be used to estimate the sample size having a given power. However, for a given  $\theta_1$  this may be difficult to calculate except for some simple distributions and it may also be sensitive to an incorrect choice of population distribution, again illustrating the dilemma that goes with power calculations when using nonparametric methods due to uncertainty about the population distribution.

### *Example 4.9*

*The problem.* Using the approximation based on (4.3) determine the approximate sample size needed to guarantee a power of 0.80 using a Wilcoxon test when sampling from a normal  $N(\theta, 4)$  distribution and we wish to use a one-tail test for  $H_0: \theta = 0$  against  $H_1: \theta = 1$  with probability of a type I error not exceeding  $\alpha = 0.05$ .

*Formulation and assumptions.* We require  $p_1 = \Pr(X_1 + X_2 > 0)$  where  $X_1, X_2$  are independently  $N(1, 4)$  and this together with appropriate values of  $z_\alpha, z_\beta$  are substituted in (4.3) to estimate  $n$ .

*Procedure.* Using conventional normal distribution theory we know that  $U = X_1 + X_2$  is distributed  $N(2, 8)$  under  $H_1$ . Thus  $Z = (U - 2)/(2\sqrt{2})$  has a standard normal distribution and it follows that  $\Pr(U > 0)$  implies  $p_1 = \Pr(Z > -1/\sqrt{2}) = \Pr(Z > -0.7071) = 0.7601$ . Clearly  $z_\alpha = 1.645$  when  $\alpha = 0.05$  and for power 0.8 we have  $\beta = 0.2$  and  $z_\beta = 0.842$ . Substitution of these values in (4.3) gives  $n \approx 23$ .

*Conclusion.* A sample size of 23 should nearly meet requirements.

*Comments.* Minitab, like many other packages, provides a program to determine the sample size to give required power when the optimum normal theory  $t$ -test is used in these circumstances and indicates a sample size  $n = 25$ . This suggests the asymptotic test size of 23 for the Wilcoxon test may be an underestimate since the Pitman efficiency of the Wilcoxon test relative to the optimal test in this case is  $3/\pi$ , which is slightly less than 1. However, the result is of the right order of magnitude.

## 4.4 FIELDS OF APPLICATION

In most applications if a numerical value of the difference for each matched pair is available and these do not appear too skew, the Wilcoxon test (or an analogous test using normal scores) is likely to be appropriate. The matched pairs  $t$ -test is appropriate if the differences  $d_i = y_i - x_i$  are approximately normally distributed; sometimes this may be the case even when the distributions of  $X, Y$  are each far from normal. If there is evidence of skewness in the differences a sign test is preferable.

### **Laboratory instrument calibration**

Two different brands of instrument reputedly measure the same thing (e.g. blood pressure, hormone level, sugar content of urine, bacterial content of sputum), but each is subject to some error. Samples from, say, each of 15 patients might be divided into two subsamples, the first being analyzed with one kind of instrument, the second with the other. A Wilcoxon test is appropriate to test for

any systematic difference between instruments. When purporting to measure the same thing a systematic difference from the true values in means or medians is often described as **mean** or **median bias**. The term ‘bias’ alone is usually taken to imply mean bias.

## Biology

The heartbeat rate of rabbits might be measured before and after they are fed a hormone-rich diet. The Wilcoxon test is appropriate to investigate a shift in mean. ‘Before’ and ‘after’ measurements are common in medical and biological contexts, including experiments on drugs and other stimuli, which may be either physical or biological (e.g. a rabbit’s blood pressure may be measured when on its own and again after it has shared a cage for half an hour with a rabbit of the opposite sex). Confidence intervals for the mean difference are useful both as an indication of the precision of the experiment ([Section 1.4.2](#)) and to help in reaching a decision as to whether any statistically significant difference is of practical importance.

## Occupational medicine

An instrument called a Vitalograph is used to measure lung capacity. Readings might be taken on workers at the beginning and end of a shift to study any effect on lung capacity of fumes inhaled in some industrial process, or on athletes before and after competing in a 100-metre sprint.

## Agriculture

In a pest control experiment each of 10 plots may contain 40 lettuce plants. Each plot is divided into two halves: one half chosen at random is sprayed with one insecticide, the second with another. Differences in numbers of uninfested plants in each plot can be used in a Wilcoxon test to compare effects of insecticides. Incidentally, pest control experiments are a situation where a normality assumption is often suspect.

## Psychology

Given sets of identical twins, it being known for each pair which was the first-born, for each individual in a pair the times to carry out a manual task are observed to see if there is any indication that the first-born tends to be quicker. The choice may lie between a

Wilcoxon test and a *t*-test. Confidence intervals for the mean difference will indicate the precision with which any difference is measured and whether it is of practical importance.

### Road safety

Drivers' reaction times in dangerous situations may be compared before and after each has consumed 2 pints of beer, using equipment that simulates driving conditions on the road. (This is a response to stimulus situation of the type mentioned above under the heading **Biology**.)

### Space research

Potential astronauts may have the enzyme content of their saliva determined before and after they are subjected to a zero gravitational field in a simulator. Such biochemical evidence is important in determining physiological reactions to space travel.

### Education

To decide which of two examination questions is perceived by students to be the harder, both questions could be included in a test paper in which candidates are free to choose neither, one or both of the two questions. Records are taken of the numbers who complete both, neither, only the first, only the second. Numbers in the latter two categories can be used in a McNemar sign test for evidence of unequal perceived difficulty.

### Social policy

An association of government employees wishing to find evidence to support their case that salaries in the public sector were generally below those paid for equivalent work in the private sector might obtain data for the average salaries paid in each sector for each of a number,  $n$ , of employment categories matched with respect to working conditions, responsibility, security of employment, etc. and use differences to assess evidence for their case.

## 4.5 SUMMARY

The matched pair sample tests for centrality differences considered in this chapter reduce to the analogous single-sample tests

considered in Chapter 2. See in particular the sign test ([Section 2.3](#)), the Wilcoxon signed-rank test ([Section 2.2](#)), raw data scores ([Section 2.1](#)), normal, or modified van der Waerden scores ([Section 2.4](#)). General tests for the distributions of the paired differences include the Kolmogorov test ([Section 3.3.1](#)) and Lilliefors' test and the Shapiro–Wilk test for normality ([Section 3.3.3](#)).

**McNemar's test** ([Section 4.2](#)) is relevant to paired observations to assess changes in attitude or for assessments of relative difficulty. It is equivalent to the sign test ([Section 2.3](#)).

**Power and sample size calculations** for the single sample or matched pair differences are in general not distribution-free and reasonably good approximations depend upon assumptions about the population distribution. Some power computations are relatively easy for the sign test because the statistic still has a binomial distribution under the alternative hypothesis, whereas the distribution of the signed-rank statistic often proves intractable under alternative hypotheses.

## EXERCISES

- 4.1 Verify the confidence intervals given in Comment 2 on [Example 4.2](#).
- 4.2 The blood pressures of 11 patients are measured before and after administration of a drug. The differences in systolic blood pressure (pressure before – pressure after) for each patient are:

$$7 \ 5 \ 12 \ -3 \ -5 \ 2 \ 14 \ 18 \ 19 \ 21 \ -1$$

- Use an appropriate nonparametric test to see if the sample (assumed random) contradicts the hypothesis of no systematic change in blood pressure.
- 4.3 Samples of cream from each of 10 dairies (A to J) are each divided into two portions. One portion from each is sent to Laboratory I, the other to Laboratory II, for bacterium counts. The counts (thousands bacteria  $\text{ml}^{-1}$ ) are:

Dairy	A	B	C	D	E	F	G	H	I	J
Lab I	11.7	12.1	13.3	15.1	15.9	15.3	11.9	16.2	15.1	13.6
Lab II	10.9	11.9	13.4	15.4	14.8	14.8	12.3	15.0	14.2	13.1

Use the Wilcoxon signed-rank test to assess the evidence for any consistent difference between laboratories for subsamples from the same dairy. Obtain also nominal 95 and 99 per cent confidence intervals for the mean

difference and compare these with the intervals using the optimal method when normality is assumed.

- 4.4 A hormone is added to one of otherwise identical diets given to each of 40 pairs of twin lambs. Growth differences over a 3-week period are recorded for each pair and signed ranks are allocated to the 40 differences. The lower rank sum was  $S_1 = 242$ . There was only one rank tie. Investigate the evidence that the hormone may be affecting (increasing or decreasing) growth rate.
- 4.5 A psychologist interviews both father and mother of each of 17 unrelated children with learning difficulties, asking each individually a series of questions designed to test how well they understand the problems their child is likely to face in adult life. The psychologist records whether the father (F) or mother (M) shows the better understanding of these potential problems. For the 17 families the findings are

F M M F F F F F F M F F F M F F

Is the psychologist justified in concluding that fathers show better understanding?

- 4.6 For each of nine matched pairs of students, one student is allocated to a series of lectures and the other to appropriate computer assisted learning (CAL) material. At the end of the course the students are given the same examination paper. The marks achieved (out of 100) are:

<i>Pair</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>
<i>CAL</i>	50	56	51	46	88	79	81	95	73
<i>Lectures</i>	25	58	65	38	91	32	31	13	49

Analyze these results by what you consider the most appropriate parametric or nonparametric methods to determine whether or not they provide acceptable evidence that CAL material leads to better examination results.

- 4.7 One hundred general practitioners attend a health promotion workshop. At the start of the workshop they are asked to indicate whether they are in favour of routinely asking patients about alcohol consumption. They are then shown a video on the health and social problems caused by the excessive consumption of alcoholic drinks. The video is followed by discussion in small groups. After the video and discussion they are asked the original question again. Do the results given below indicate a significant change in attitudes as a result of the video and group discussion?

		<i>Before video and discussion</i>	<i>In favour</i>	<i>Against</i>
<i>After video and discussion</i>		41	27	
		16	58	

- 4.8 A canned-soup manufacturer is experimenting with a new-formula tomato soup. A tasting panel of 70 each taste samples of the current product and the new one (without being told which is which). Of the 70, 32 prefer the new-formula product, 25 the current product and the remainder cannot distinguish between the two. Is there enough evidence to reject the hypothesis that consumer preference is equally divided?
- 4.9 Do the data in Exercise 4.8 support a claim that as many as 75 per cent of those who have a preference may prefer the new formula?
- 4.10 To produce high-quality steel one of two hardening agents A, B may be added to the molten metal. Hardness of steel varies from batch to batch, so to test the two agents batches are sub-divided into two portions, for each batch agent A being added to one portion, agent B to the other. To compare hardness, sharpened specimens for each pair are used to make scratches on each other; that making the deeper scratch on the other is the harder specimen. For 40 pairs, B is adjudged harder in 24 cases and A in 16. Is this sufficient evidence to reject the hypothesis of equal hardness?
- 4.11 For a subsample of 10 pairs from the steel batches in Exercise 4.10 a more expensive test is used to produce a hardness index. The higher the value of the index, the harder the steel. The indices recorded were:

<i>Batch no.</i>	1	2	3	4	5	6	7	8	9	10
<i>Additive A</i>	22	26	29	22	31	34	31	20	33	34
<i>Additive B</i>	27	25	31	27	29	41	32	27	32	34

Use an appropriate test to determine whether these data support the conclusion reached in Exercise 4.10.

- 4.12 On the day of the third round of the Open Golf Championship in 1987 before play started a television commentator said that conditions were such that the average scores of players were likely to be at least three higher than those for the second round. For a random sample of 10 of the 77 players participating in both rounds the scores were:

<i>Player</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>
<i>Round 2</i>	73	73	74	66	71	73	68	72	73	72
<i>Round 3</i>	72	79	79	77	83	78	70	78	78	77

Do these data support the commentator's claim? Consider carefully whether a one- or two-tail test is appropriate.

- 4.13 Pearson and Sprent (1968) gave data for hearing loss (in decibels below prescribed norms) at various frequencies. The data below show these losses for 10 individuals aged between 46 and 54 at frequencies of 0.125 and 0.25  $\text{kc s}^{-1}$ . A negative loss indicates hearing above the norm. Is there an indication of a different loss at the two frequencies?

<i>Subject</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>
$0.125 \text{ } kc \text{ } s^{-1}$	2.5	-7.5	11.25	7.5	10.0	5.0	7.5	2.5	5.0	8.75
$0.25 \text{ } kc \text{ } s^{-1}$	2.5	-5.0	6.35	6.25	7.5	3.75	1.25	0.0	2.5	5.0

- 4.14 Apply a normal scores test to the data in [Example 4.2](#).
- 4.15 Scott, Smith and Jones (1977) give a table of estimates of the percentages of UK electors predicted to vote Conservative by two opinion polling organizations, A and B, in each month in the years 1965–70. For a random sample of 15 months during that period the paired percentages were:
- |          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| <i>A</i> | 43.5 | 51.2 | 46.8 | 55.5 | 45.5 | 42.0 | 36.0 | 49.8 | 42.5 | 50.8 | 36.6 | 47.6 | 41.9 | 48.4 |
|          | 53.5 |      |      |      |      |      |      |      |      |      |      |      |      |      |
| <i>B</i> | 45.5 | 44.5 | 45.0 | 54.5 | 49.5 | 43.5 | 41.0 | 53.0 | 48.0 | 52.5 | 41.0 | 47.5 | 42.5 | 45.0 |
|          | 52.5 |      |      |      |      |      |      |      |      |      |      |      |      |      |
- Do these results indicate a significant tendency for one of the organizations to return higher percentages than the other? Obtain an appropriate 95 per cent confidence interval for any mean or median difference between predictions during the period covered.
- 4.16 Verify the correctness of the confidence intervals and the result for Lilliefors' test quoted in [Example 4.3](#).
- 4.17 Confirm the value given in [Example 4.6](#) for  $\Pr(X > 0|\theta = 1)$  for the distribution given in [\(4.1\)](#). What is the corresponding probability conditional upon  $\theta = 2$ ?
- 4.18 Confirm the value given in [Example 4.7](#) for  $\Pr(X > 0|\theta = 1)$  for the distribution given in [\(4.2\)](#). What is the corresponding probability conditional upon  $\theta = 1.2$ ?
- 4.19. Determine the sample size needed to have power at least 0.80 for the sign test that the median  $\theta$  is  $H_0: \theta = 1$  against the alternative  $H_1: \theta = 2$  with the one-tail  $P = 0.05$  if the observations are known to be a random sample from an exponential distribution with frequency function  $f(x) = \lambda e^{-\lambda x}$ ,  $x \geq 0$ . (Hint: You will need to find values of  $\lambda$  that give medians corresponding to those specified in the null and alternative hypotheses.)

# 5

## Methods for two independent samples

### 5.1 CENTRALITY TESTS AND ESTIMATES

#### 5.1.1 Extensions from single samples

We often have two independent random samples (i.e. the members of the first sample or group are independent of those in the second) and wish to make inferences about the two populations which they represent. We denote the members of the samples by  $x_1, x_2, \dots, x_m$  and  $y_1, y_2, \dots, y_n$  where for convenience and without loss of generality we assume  $n \geq m$ , i.e. that the second sample is at least as large as the first. In [Chapters 2–4](#) we saw that several distribution-free tests and estimation procedures based on permutations differed only in the scores assigned, e.g. ranks, signs, van der Waerden scores. Relevant assumptions and practical computational matters governed the choice of an appropriate procedure. These considerations extend with modifications and additions to the two-sample situation. We consider first centrality tests in the same order as we covered the single sample analogues in [Chapter 2](#).

#### 5.1.2 The Pitman permutation test

The Pitman permutation test for two independent samples has similar disadvantages to its one-sample counterpart. Like the  $t$ -test, it is not robust against certain departures from assumptions needed for its validity and there are difficulties in computing confidence intervals. Further, the conditional nature of the test makes it virtually impossible to obtain exact  $P$ -values without specialist software. For these reasons it is seldom used in practice so we omit details here although both StatXact and Testimate provide programs for exact hypothesis tests, the former using the program option for permutation tests with any chosen scores. Readers interested in this test will find an account of its application together with an example in Sprent (1998, Section 4.1).

## 5.2 RANK BASED TESTS

### 5.2.1 The Wilcoxon–Mann–Whitney test

For two independent samples, the analogue of the one sample Wilcoxon signed-rank test is the **Wilcoxon rank-sum test** proposed by Wilcoxon (1945). An equivalent test widely referred to particularly, but not only, in the medical literature, as the **Mann–Whitney  $U$  test** was developed independently by Mann and Whitney (1947). It is convenient, though not universal practice, to refer to the tests jointly as the **Wilcoxon–Mann–Whitney test**, or for brevity as the **WMW test**.

The data in [Example 1.4](#) can be viewed as two independent samples, those patients who received the new drug forming the first sample and those who did not forming the second. Information recorded on the condition of each patient is already in the form of ranks so the Wilcoxon–Mann–Whitney approach can be applied immediately. This was effectively what we did in that example.

### 5.2.2 The Wilcoxon formulation

In this form of the test, the two samples are combined and the data are ranked overall. The original two samples are then separated out with each rank being attached to the corresponding observation. The usual null hypothesis is that the two samples are from identical populations and a common alternative hypothesis is that the population distributions differ only in the mean or median. As indicated in [Example 1.4](#), if both samples come from the same population (which may be of any continuous form and need not be symmetric) we expect a mix of low, medium and high ranks in each sample. Under the alternative hypothesis we expect lower ranks to dominate in one population and higher ranks in the other. Such a shift in centrality is often referred to as an ‘additive’ treatment effect, i.e. there is a ‘constant’ difference between two treatments.

The sum of the ranks in the first sample,  $S_m$ , can be used to determine the strength of evidence against the null hypothesis (the other, perhaps larger group, could be taken instead). Taking [Example 1.4](#) as a case in point, the smaller group consists of the four patients receiving the new treatment. If their ranks are 1, 2, 3 and 5 (say), low ranks predominate and the rank-sum of 11 is small. If the ranks are 6, 7, 8 and 9, high ranks predominate and  $S_m = 30$  is relatively large. Ranks of 2, 3, 6 and 8 indicate not very different values in the two groups. The rank sum is now 19. Values of the

rank sum close to the minimum possible (10) or maximum possible (30) provide strong evidence against the null hypothesis; for intermediate values of  $S_m$  the evidence is weaker.

The test may also be used when samples are from two distributions with identical cumulative distribution functions under  $H_0$ , but under  $H_1$ , one cumulative distribution curve (see Comment 1 on [Example 3.10](#)) lies beneath the other apart from some points where the curves touch. A moment's reflection shows that under  $H_1$  low or high ranks should dominate in one sample, as opposed to a fairly even distribution of ranks under  $H_0$ . These are sometimes referred to as 'dominance' alternatives. Given the permutation distribution of rank sums under  $H_0$ ,  $P$ -values may be determined in the way described for the particular case in [Example 1.4](#).

### *Example 5.1*

*The problem.* For some models of pocket calculator the trigonometric function values are obtained by entering the number before pressing the function button (Type A models, say). Other models require the function to be selected before the number is entered (Type B models). A mathematics teacher wished to determine whether a particular Type A model of calculator allows calculations to be performed with greater speed compared with a certain Type B model. A class of 21 pupils was randomly divided into groups *A* (using the Type A model), and *B* (using the Type B model) with 10 and 11 pupils in the respective groups. The pupils were asked to carry out the same set of trigonometric calculations. The total times in minutes for each member of each group to complete the calculations were

<i>Group A</i>	23	18	17	25	22	19	31	26	29	33
<i>Group B</i>	21	28	32	30	41	24	35	34	27	39

Do the data indicate that one model of calculator is superior (i.e. leads to more rapid computations)?

*Formulation and assumptions.* A two-tail test is appropriate. We require the sum of ranks associated with the smaller sample, *Group A*, in a joint ranking of all data, or alternatively the rank sum for *Group B* could be used. The  $P$ -value associated with this sum lets us assess the strength of evidence against  $H_0$ : *the population medians are identical*, where the alternative is that one sample is from a population with greater median. We prefer to think in terms of medians rather than means because there is no need to make an assumption of symmetry. The test will also be valid for a dominance alternative, i.e. for a tendency for the computations to be done more quickly with one model of calculator, although the time difference may vary appreciably between pupils.

*Procedure.* The sample sizes are  $m = 10$  and  $n = 11$ . Most software programs for the Wilcoxon test will calculate the ranks automatically, but to obtain them manually it helps to arrange data in ascending order within each sample and then allocate ranks. We leave it as an exercise for the reader to show that if this is

done the ranks assigned to Group A are 1, 2, 3, 5, 6, 8, 9, 12, 14, 16 with a rank sum of  $S_m = 76$ . The rank sum for Group B is  $S_n = 155$ .

If software to give exact  $P$ -values is not available some relevant values for  $S_m$  are given in many tables. For this example we used the program in StatXact which indicates that for a two-tail test the relevant exact  $P = 0.0159$ . The output also confirms the value of  $S_m$  given above.

*Conclusion.* The low  $P$ -value provides fairly strong evidence against  $H_0$ . On average, pupils appear to perform calculations more speedily with the Type A model, for which the trigonometric functions are entered after the number.

*Comments.* 1. Tables for the Wilcoxon–Mann–Whitney statistic are given by Neave (1981, p. 30). Conover (1999, Table A7) and others give various quantiles for  $S_m$ ,  $S_n$ . Actual, rather than nominal, significance levels may be obtained from computer programs giving the exact permutation distribution.

2. This study is a comparison of two particular models of calculator. It would be unreasonable for the teacher to recommend that pupils purchase *any* Type A calculator on the basis of just this study. Further, calculators that are particularly useful for operating with trigonometric functions may not perform so well with other problems such as calculating a mean or standard deviation.

3. Suppose that this investigation had been conducted in a mathematics lesson of 40 minutes. The fifth pupil in Group B would then have been unable to complete all of the calculations and this observation would have been censored. In the ranking of the data, however, since this observation is the only one with a time in excess of 40 minutes a rank of 21 would still be given. A  $t$ -test could not have been validly used since an exact value is required for each observation.

*Computational aspects.* StatXact and Testimate give exact  $P$ -values corresponding to observed  $S$  providing  $m, n$  are not too large. Many general statistical packages compute  $S_m$  but leave the user to consult tables or give an asymptotic result which may be unsatisfactory if, for example, one sample is large but the other small, or if there are many tied ranks.

### 5.2.3 The Mann–Whitney formulation

A statistic  $U$ , which is a function of the rank sum  $S$ , can be calculated for either group in order to determine the strength of the evidence against the null hypothesis. For the first of the two samples this statistic is given by  $U_m = S_m - \frac{1}{2}m(m + 1)$ , with the equivalent statistic for the (perhaps larger) sample being  $U_n = S_n - \frac{1}{2}n(n+1)$ . We only need to compute one of  $S_m$  or  $S_n$ , for the sum of all the ranks from 1 to  $m + n$  is  $\frac{1}{2}(m + n)(m + n + 1) = S_m + S_n$ . Using the relations  $U_m = S_m - \frac{1}{2}m(m+1)$  and  $U_n = S_n - \frac{1}{2}n(n+1)$  one easily deduces that each has minimum value zero and that

$$U_m = mn - U_n \quad (5.1)$$

so that again only one of  $U_m$ ,  $U_n$  need be computed. Either may be used in a test, although  $U_m$  is generally given in tables.

In the Mann–Whitney approach, either  $U_m$  or  $U_n$  is calculated directly. To obtain  $U_m$  or  $U_n$  we count the number of observations in one sample exceeding each member of the other sample. Ranks are not needed, and the procedure eases calculation if no computer program is available for the test. It also forms the basis for determining a confidence interval for a difference in centrality.

### *Example 5.2*

*The problem.* Recalculate the test statistic for the data in [Example 5.1](#) using the Mann–Whitney approach.

*Formulation and assumptions.* We inspect the observations in each sample; they need not be ordered, but counting is easier when they are. It is visually easier to count the number of times each observation in Group A is exceeded by an observation in Group B. This gives  $U_n$ , from which  $U_m$  can be determined.

*Procedure.* The data from [Example 5.1](#), arranged for convenience in ascending order in each group, are:

<i>Group A</i>	17	18	19	22	23	25	26	29	31	33
<i>Group B</i>	21	24	27	28	30	32	34	35	36	39

Clearly the first observation 17 in Group A is exceeded by all 11 observations in Group B. Similarly, the observations 18, 19 are also exceeded by all observations in Group B. The observation 22 is exceeded by 10 observations in Group B. Proceeding in this way we find the numbers of observations in Group B exceeding each observation in Group A and then add these, viz.  $U_n = 11 + 11 + 11 + 10 + 10 + 9 + 9 + 7 + 6 + 5 = 89$ . Using (5.1) gives  $U_m = 110 - 89 = 21$ , easily shown to be consistent with the value of  $S_m$  found in [Example 5.1](#).

*Conclusion.* As in [Example 5.1](#).

*Comment.* Equivalence of the Wilcoxon and Mann–Whitney formulations is general.

*Computational aspects.* StatXact and Testimate give exact tail probabilities corresponding to the value of  $U$  (in this case  $P = 0.0159$ ). As with the statistic  $S$ , many general statistical packages compute  $U$  but leave the user to consult tables or give an asymptotic result based on the normal distribution (see [Section 5.6](#)). For this example, Stata or StatXact give an asymptotic  $P = 0.0167$ , which leads to the same conclusions; in other situations the asymptotic result may be unsatisfactory. Many tables only give values of the statistic that correspond to nominal conventional significance levels rather than exact  $P$ -values, but most explain what is given and how to use the tables. For instance, Neave (1981) indicates that if either  $U_m$  or  $U_n$  do not exceed 26, the two-tailed  $P$ -value for Example 5.2 is less than 0.05.

### 5.2.4 Ties

We use mid-ranks for ties as we did for the Wilcoxon signed-rank test ([Section 2.2.3](#)). If software to compute exact  $P$ -values is not

available and there are only a few ties, basing significance tests on the appropriate critical values for the ‘no-tie’ case is unlikely to be seriously misleading. If  $m, n$  are both reasonably large (say 15 or greater), a normal approximation we develop in [Section 5.6](#) may be used with reasonable confidence, adjustment being essential only when there are a moderate to large number of ties. [Example 5.4](#) illustrates a situation where ties dominate.

In the Mann–Whitney formulation, if an observation in the second sample equals an observation in the first sample it is scored as  $\frac{1}{2}$  in counting the number of observations in the second sample exceeding that observation in the first.

For small to medium-sized samples ties present no difficulty if a computer program is available to generate exact probabilities for the WMW test based on the appropriate permutation distribution.

### *Example 5.3*

*The problem.* We consider a set of data from a second experiment similar to that in [Examples 5.1](#) and [5.2](#) but where there are now some ties in the times taken by different participants. For convenience the data are given in ascending order but this is not essential, especially if suitable software is used.

<i>Group A</i>	16	18	19	22	22	25	28	28	28	31	33		
<i>Group B</i>	22	23	25	27	27	28	30	32	33	35	36	38	38

Do the data indicate that one model of calculator is superior (i.e. leads to more rapid computations)?

*Formulation and assumptions.* We use the WMW test with mid-ranks for ties.

*Procedure.* We have  $m = 11$  and  $n = 13$ . Computer programs usually assign ranks or mid-ranks automatically, but this is easily done manually since the data are ordered. The reader should verify that these are

<i>Group A</i>	1	2	3	5	5	8.5	13.5	13.5	13.5	17	19.5		
<i>Group B</i>	5	7	8.5	10.5	10.5	13.5	16	18	19.5	21	22	23.5	23.5

Unless we need  $S_m$  specifically and have no software to compute it there is no need to allocate ranks. It is easier to obtain  $U_n$  by simply counting for each first sample value the number of observations exceeding it in the second sample (scoring one half for ties) and summing these. For example, for *each* value 22 in Group A there is one tied value, scored as  $\frac{1}{2}$  and 12 values exceeding 22 in Group B giving a contribution of 12.5 to  $U_n$ . Proceeding in this way we find

$$U_n = 13 + 13 + 13 + 12.5 + 12.5 + 10.5 + 7.5 + 7.5 + 7.5 + 6 + 4.5 = 107.5$$

From (5.1),  $U_m = 11 \times 13 - 107.5 = 35.5$ . StatXact confirms this value for  $U_m$  requiring only the original data to do so, and for a two-tail test gives the exact  $P = 0.0359$ .

*Conclusion.* There is some evidence against  $H_0$  and the results suggest computations are completed more rapidly with the model tested by Group A.

*Comments.* 1. Tables such as those in Neave (1981) indicate that values of  $U_m \leq 37$  imply  $P < 0.05$  for a two-tail test. This one example suggests that a few ties do not seriously upset conclusions based on ‘no-tie’ critical values for moderate sample sizes.

2. The situation using conventional ‘no-tie’ tables is less satisfactory when there are many ties or for ties in unbalanced samples. For example, if we had a sample of 3 with values 1, 2, 2, and a sample of 13 with values 1, 1, 4, 5, 5, 5, 7, 8, 9, 9, 9, 9, 10 it is easy to show that  $U_m = 5$ . In a no-tie situation two-tail test  $\Pr(U_m \leq 5) = 0.0571$  whereas the exact permutation test allowing for ties gives  $\Pr(U_m \leq 5) = 0.0464$  as the true  $P$ -value for this specific tie pattern. Although the strength of evidence against the null hypothesis is not markedly different, use of a rigid 5 per cent significance level would lead to differing conclusions.

3. As we demonstrated in [Example 2.5](#), if ties result from rounding and greater accuracy allows us to break ties, the way the ties break may alter appreciably conclusions regarding significance.

*Computational aspects.* Software to determine exact permutation probabilities associated with the test statistics  $U$  or  $S$  is particularly valuable when there are ties. StatXact and Testimate allow this. Some general programs take ties into account by using mid-ranks but provide only an asymptotic test (which may be unreliable for small  $m + n$  or when one of  $m, n$ , is small). Alternatively, one may resort to tables appropriate to the no-tie situation but nominal significant levels can no longer be guaranteed.

A common ‘tie’ situation is one where we are not given precise measurements, but only grouped data. For example, instead of the complete sample values in [Example 5.3](#) we may be given only numbers of participants taking between 10 and 19 minutes, 20 and 29 minutes, 30 and 39 minutes. We may still calculate the  $U$  or  $S$  statistics making allowances for ties, but it may now be misleading to use tabulated critical values for these statistics.

#### *Example 5.4*

*The problem.* Instead of the data in [Example 5.3](#) suppose we are given only the numbers taking 10–19, 20–29, 30–39 minutes leading to the data given below. Perform the test in [Example 5.3](#) using this reduced information.

No. of minutes	10–19	20–29	30–39
Group A	3	6	2
Group B	0	6	7

*Formulation and assumptions.* We carry out a WMW test based on mid-ranks using, if available, a program giving exact permutation probabilities.

*Procedure.* If no suitable program is available calculating the required value of  $U_n$  is reasonably straightforward but needs care. For each of the three ties in 10–19 for Group A there are  $6 + 7 = 13$  greater values in Group B. Thus between them these three ties contribute a total of  $3 \times 13 = 39$  to  $U_n$ . Similarly each of the six Group A ties in 20–29 corresponds to six ties and seven greater values in Group B, thus each contribute a score of  $\frac{1}{2} \times 6 + 7 = 10$  so that together they contribute a score of  $6 \times 10 = 60$ . Finally by a similar argument the two group A ties in 30–39 each contribute 3.5 whence it is easily seen that  $U_n = 39 + 60 + 7 = 106$  and now from (5.1)  $U_m = 143 - 106 = 37$ . In Comment 2 on [Example 5.3](#) we noted that  $U_m = 37$  is the minimum value that implies  $P \leq 0.05$  in a two-tail test if there are no ties. The exact permutation probability that  $U_m = 37$  for these tied data in a two-tail test given by StatXact is  $P = 0.0442$ .

*Conclusion.* There is reasonably strong evidence against the hypothesis of equal medians.

*Comments.* 1. In this specific example the use of conventional no-tie tables is not seriously misleading; however this is not always so. In particular, heavy tying often leads to major discontinuities in possible  $P$ -values since the tying may reduce appreciably the number of possible values that  $U_m$  may take.

2. By pooling the data into groups we use less information in this test than we did in that in [Example 5.3](#) so it is not surprising that there is a reduction in the strength of evidence against  $H_0$  that is reflected by the slightly higher  $P$ -value.

### 5.2.5 Wilcoxon–Mann–Whitney confidence intervals

The Mann–Whitney statistic compares all differences  $d_{ij} = x_i - y_j$  between sample values. In computing  $U_m$  we allocate a score of 1 if  $d_{ij}$  is positive, zero if  $d_{ij}$  is negative and  $\frac{1}{2}$  if  $d_{ij}$  is zero. To calculate  $U_n$  we reverse these scores, i.e. score 1 if  $d_{ij}$  is negative, zero if  $d_{ij}$  is positive and  $\frac{1}{2}$  if  $d_{ij}$  is zero. To calculate a confidence interval for the difference in centrality we need the actual values of some or all  $d_{ij}$ . If  $c$  is the value of the Mann–Whitney statistic  $U$  that we would regard as providing sufficient evidence to indicate significance at the  $100\alpha$  per cent significance level, the  $100(1 - \alpha)$  per cent confidence limits are given by the  $c + 1$  smallest and  $c + 1$  largest  $d_{ij}$ . The reasoning here is not unlike that used in establishing a confidence interval using Walsh averages in the one-sample situation based on the Wilcoxon signed rank approach.

*Example 5.5*

*The problem.* Obtain nominal 95 per cent confidence limits for a population median shift based on the WMW method using the data in [Example 5.1](#).

*Formulation and assumptions.* If computer software that calculates a confidence interval directly is not available tables such as those in Neave (1981) indicate that for sample sizes of 10, 11 the critical value is  $U_m = 26$  for significance at a nominal 5 per cent level. Denoting the sample values by

$x_1, x_2, \dots, x_{10}$  and  $y_1, y_2, \dots, y_{11}$  we require the 27 largest and smallest  $d_{ij} = x_i - y_j$ . It is not essential to compute all  $d_{ij}$ , but for completeness we give these in Table 5.1. These need not be computed manually if suitable software is available either to determine the confidence limits directly, or even if only to compute the required differences.

*Procedure.* It is easier to compute differences manually if the sample values are ordered. We write those for the first sample in the top row and those for the second sample in the first (left) column. The entries in the body of Table 5.1 are the differences between the data entries at the top of that column and at the left of that row, e.g. the first entry  $-4$  is  $17 - 21 = -4$ . Note that the difference between each entry in any pair of rows is a constant equal to the difference between the corresponding entries in the left-hand data column; there is an analogous constant difference between entries in pairs of columns. The largest entries appear in the top right of the table and entries decrease and eventually change sign as we move towards the bottom left.

A count shows 21 positive and 89 negative values whence  $U_n = 89$  and  $U_m = 21$  as found in [Example 5.1](#). The critical value is  $U_m = 26$  for significance at least at a 5 per cent level so the lower limit for a 95 per cent confidence interval for the median difference  $\theta_1 - \theta_2$  is obtained by eliminating the 26 largest negative differences; the next largest negative difference is the required lower limit. Using Table 5.1 gives this lower limit as  $-13$ . Similarly, elimination of the 26 largest differences gives the upper limit  $-2$ . StatXact provides a program to perform these computations given only the original data and quickly gives these limits. Minitab will also compute the relevant differences.

*Conclusion.* A nominal 95 per cent confidence interval for the difference  $\theta_1 - \theta_2$  is  $(-13, -2)$  or equivalently, for  $\theta_2 - \theta_1$  the interval is  $(2, 13)$ .

*Comments.* 1. If we assume normality the 95 per cent confidence interval based on the relevant  $t$ -distribution is  $(-12.6, -1.9)$ . The close agreement between this and the Wilcoxon interval is heartening because there is little to

**Table 5.1** Paired differences for times taken to complete calculations.

	17	18	19	22	23	25	26	29	31	33
21	-4	-3	-2	1	2	4	5	8	10	12
24	-7	-6	-5	-2	-1	1	2	5	7	9
27	-10	-9	-8	-5	-4	-2	-1	2	4	6
28	-11	-10	-9	-6	-5	-3	-2	1	3	5
30	-13	-12	-11	-8	-7	-5	-4	-1	1	3
32	-15	-14	-13	-10	-9	-7	-6	-3	-1	1
34	-17	-16	-15	-12	-11	-9	-8	-5	-3	-1
35	-18	-17	-16	-13	-12	-10	-9	-6	-4	-2
36	-19	-18	-17	-14	-13	-11	-10	-7	-5	-3
39	-22	-21	-20	-17	-16	-14	-13	-10	-8	-6
41	-24	-23	-22	-19	-18	-16	-15	-12	-10	-8

indicate serious nonnormality or a difference between population variances in these data apart from a small indication that the Group B data might have a slightly greater spread.

2. Confidence intervals at other levels can be obtained. For example, a 99 per cent interval is obtained by rejecting the 18 extreme differences and using [Table 5.1](#) is easily seen to be  $(-16, 1)$  a result confirmed by StatXact.

3. The computational process is reminiscent of that with Walsh averages given in [Section 2.2.4](#).

*Computational aspects.* If one wishes to explore further the exact confidence level one may examine results of hypothesis testing situations at or near the end points of the interval, much as we did for the Wilcoxon signed rank test. This is achieved in the case of a 95 per cent confidence interval by an appropriate addition or subtraction to all observations in one sample and then obtaining exact  $P$ -values for the corresponding test of zero median difference. While StatXact computes exact intervals some software packages may use asymptotic approximations based on the asymptotic theory we give for hypothesis testing in [Section 5.6](#). For both samples moderate or large in size asymptotic approximations are usually quite good but they may be unreliable if the samples differ greatly in size especially when one is small.

The appropriate point estimator of the median difference based on the WMW procedure used above, known as the **Hodges–Lehmann** estimator is the median of the differences in [Table 5.1](#). It is easily verified that in the above example this is  $-7.5$ .

## 5.3 THE MEDIAN TEST

### 5.3.1 The basic test

If each population has the same median, whether or not they differ in other respects, for each sample the number of values above (or below) that common median has a binomial distribution with  $p = 0.5$ . The median test we develop here, that proposed by Mood (1954), does **not** test whether that common median,  $\theta$ , say, has a particular value, but as is also the situation in the two-sample  $t$ -test or the WMW test, only whether it is reasonable to suppose both populations have the same unknown median. Formally, if the sample of  $m$  all come from populations (they need not all be the same) each with unknown common median  $\theta_1$  and the sample of  $n$  from populations each with unknown median  $\theta_2$ , we test  $H_0: \theta_1 = \theta_2$  against alternatives like  $H_1: \theta_1 \neq \theta_2$  (two-tail) or  $H_1: \theta_1 > \theta_2$  (one-tail).

If all populations have the same median  $\theta$ , then the combined sample median,  $M$ , say, provides a point estimate of  $\theta$ . The median test examines, for each sample, how many values are above and how

many are below  $M$ . If the samples are from populations with the same median the distributions of the numbers in each above  $M$  will be approximately binomial  $B(m, 1/2)$  and  $B(n, 1/2)$ , where the relevant probabilities are now conditional upon the numbers above  $M$  in each sample adding to the number above  $M$  in the combined sample. When  $m + n$  is even and no sample value equals  $M$ , in the combined sample the numbers of values above and below  $M$  will each be  $1/2(m + n)$ . If  $m + n$  is odd at least one sample value equals  $M$ . In all cases where some values equal  $M$  we suggest omitting these and proceeding with a reduced sample. Unless many values equal  $M$  this is usually satisfactory. Conditioning on the total numbers above or below  $M$  in the combined sample leads to permutation tests that have much in common with some procedures we develop more generally in [Chapter 9](#). However, our extensive treatment of this test is largely to illustrate principles, because there is mounting evidence that is well summarized in Freidlin and Gastwirth (2001) that, especially for small or unbalanced samples, the test often has low power relative to more appropriate alternatives, so we do not recommend its use when tests such as the WMW or other tests are valid.

### *Example 5.6*

*The problem.* Suppose it is claimed that on average the salivary flow rate when chewing an unflavoured gum is greater for males than for females. To investigate this assertion, the rate of saliva production (ml/min) was ascertained for 7 female and 21 male adults. The rates for the two groups in ascending order were:

Females	0.45	0.60	0.80	0.85	0.95	1.00	1.75
Males	0.40	0.50	0.55	0.65	0.70	0.75	0.90
	1.35	1.45	1.50	1.85	1.90	2.30	2.55
						2.70	2.85
							3.85

To apply the median test, the samples are pooled and the median salivary flow rate,  $M$ , for the combined sample is determined. This can be shown to be 1.1 ml/min. None of the observed values is exactly equal to the median so there are 14 values greater and 14 less than  $M$ . For the females, there is one value greater than  $M$ , whereas for the males there are 13 such observations. If the unknown population medians are  $\theta_1$  and  $\theta_2$  respectively, an appropriate test is  $H_0: \theta_1 = \theta_2$  against  $H_1: \theta_1 < \theta_2$ .

*Formulation and assumptions.* In Sample 1, there are 6 values less than  $M$  and only 1 greater than  $M$ . This suggests that if  $H_0$  is not true it is likely that this sample is from a population with a lesser median than that suggested by the pooled sample. Similarly, the fact that Sample 2 has more values above  $M$  than it has below suggests that if  $H_0$  is not true then this sample is likely to have come from a population with a greater median than that suggested by the pooled sample. This suggests any departure from equality is in the direction specified in  $H_1$ . It is convenient to set out the information as in [Table 5.2](#).

**Table 5.2** Numbers of observations above and below  $M$  in Example 5.6.

	<i>Above M</i>	<i>Below M</i>	<i>Total</i>
<i>Sample 1</i>	1	6	7
<i>Sample 2</i>	13	8	21
	14	14	28

An important feature of this table is that the row and column totals and the grand total are fixed. There are two rows and two columns of data, so such a table is called a **2 × 2 contingency table**. The four data positions are often referred to as cells. That at the intersection of the  $i$ th row and  $j$ th column is called cell  $(i, j)$ . In Table 5.2, cell  $(2, 1)$  contains the entry 13.

The row totals are determined by the sample sizes  $m, n$  and the column totals by the rule that  $M$  is the combined sample median. The test is based on the permutation distribution of all **possible** cell entries in Table 5.2 consistent with the marginal totals. Since the marginal totals are fixed, knowledge of one cell value means that all the other cell values can be determined by subtraction. It is convenient to concentrate on the possible values in cell  $(1, 1)$  (i.e. the number of values above  $M$  in sample 1). Suppose that a value of 3 is considered. To give the correct marginal totals, by subtraction we find the cell entries in the body of a table like Table 5.2 must then be

$$\begin{matrix} 3 & 4 \\ 11 & 10 \end{matrix}$$

Such an outcome is clearly more favourable to  $H_0$  than the one observed in this example, as the proportions of observations greater than  $M$  for each sex are more similar. Indeed, the only outcome less favourable to  $H_0$  than that observed is

$$\begin{matrix} 0 & 7 \\ 14 & 7 \end{matrix} \quad (5.2)$$

This allocation and that in Table 5.2 form an appropriate ‘tail’ critical region for assessing the evidence for  $H_1$  against  $H_0$ . We need the probability associated with this tail when  $H_0$  is true. We assume temporarily that if  $H_0$  is true then  $M$  is the value of the common median for the two populations. This may not be true, but it is intuitively the ‘best’ estimate of the common median to be obtained from our data. We see below that this assumption is not critical to our argument. As in the sign test, under  $H_0$  the probability of getting one observation above the median in a sample of 7 is the binomial probability  $p_1 = \binom{7}{1} (\frac{1}{2})^7$ ; that of getting 13 above the median in a sample of 21 is  $p_2 = \binom{21}{13} (\frac{1}{2})^{21}$ . These are independent as they refer to different samples. Turning now to the column margin, in the combined sample of 28 the probability of observing 14 out of 28 observations above  $M$  is  $p_c = \binom{28}{14} (\frac{1}{2})^{28}$ . We require the probability of observing respectively 1 and 13 observations above the median in Samples 1 and 2, conditional upon observing 14 above the median in the combined sample. From the definition of conditional probability this is  $P^* = p_1 p_2 / p_c$ . Thus, for the data in Table 5.2,  $P^* = [(\binom{7}{1})(\frac{1}{2})^7 (\binom{21}{13})(\frac{1}{2})^{21}] / (\binom{28}{14})(\frac{1}{2})^{28}$ . The binomial probability  $p = 1/2$  cancels out

in the expression for  $P^*$ . This will still happen if the binomial analogues of  $p_1, p_2, p_c$  are written down for any probability  $p$ ; i.e.  $P^*$  is independent of  $p$ . Thus if  $H_0$  is true, we do not need to assume that  $M$  is the population median. That  $P^*$  is independent of the binomial parameter  $p$  when row and column marginal totals are fixed has far-reaching implications in nonparametric inference. In Chapter 9 we use this result and extensions to it in a variety of situations. To determine a relevant tail probability we calculate  $P^*$  for the configuration in Table 5.2 and for the more extreme configuration (5.2) and add the results.

*Procedure.* Evaluation of  $P^*$  is tedious. It is commonly written in an equivalent form. For a general  $2 \times 2$  contingency table with cell entries

$$\begin{matrix} a & b \\ c & d \end{matrix}$$

this form is

$$P^* = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!} \quad (5.3)$$

where  $a!$  (known as factorial  $a$ ) =  $a \times (a-1) \times (a-2) \times \dots \times 3 \times 2 \times 1$  and so on. Also  $0! = 1$ . The factorials in the numerator of (5.3) are for the marginal totals and those in the denominator for the grand total and the individual cell entries. For Table 5.2, (5.3) gives  $P^* = [7! \times 21! \times 14! \times 14!] / [28! \times 1!6! \times 13! \times 8!]$ . There is appreciable cancellation between numerator and denominator and using a pocket calculator it is easy to verify that  $P^* = 0.0355$ . Computing  $P^*$  for (5.2) gives  $P^* = 0.0029$  (Exercise 5.3). Thus the exact  $P$ -value for assessing the evidence for  $H_0$  against that for  $H_1$  is  $P = 0.0029 + 0.0355 = 0.0384$ .

*Conclusion.* There is reasonably strong evidence against  $H_0$ . In other words, when chewing unflavoured gum females have on average a lower salivary flow rate compared with males.

*Comments* 1. For a two-tail test of  $H_0: \theta_1 = \theta_2$  against  $H_1: \theta_1 \neq \theta_2$  the permutation distribution is symmetric; the upper tail is associated with the two extreme outcomes

$$\begin{matrix} 6 & 1 \\ 8 & 13 \end{matrix} \quad \text{and} \quad \begin{matrix} 7 & 0 \\ 7 & 14 \end{matrix}$$

and these have the same four cell entries (in different order) and marginal totals as those in Table 5.2 and in (5.2). It follows they have the same associated probabilities. Doubling the one-tail probability for a two-tail test, the size of the critical region is  $2 \times 0.0384 = 0.0768$ , so there is no strong evidence against  $H_0$  at a conventional 5 per cent significance level in a two-tail test, but the evidence against  $H_0$  could nevertheless be used to argue a case for further experiments.

2. Clearly there are marked discontinuities in the permutation distribution, despite there being 28 observations. Discontinuities arise partly because the observations are split into a relatively large (21) and a relatively small (7) sample. Because fixing the entry in one cell determines all others (to give correct marginal totals) we say the table has **one degree of freedom**. If we

**Table 5.3** Exact permutation distribution for median test in Example 5.6.

<i>X</i>	<i>Probability</i>
0	0.0029
1	0.0355
2	0.1539
3	0.3077
4	0.3077
5	0.1539
6	0.0355
7	0.0029

denote the number in the first cell by  $X$ , then  $X$  is a random variable which, in this example, can take only the values 0, 1, 2, 3, 4, 5, 6, 7.

3. There may be a simple biological explanation for the results in the above study. Observations during the investigation might show that males chew more vigorously than females and hence produce saliva more quickly.

*Computational aspects.* StatXact and Testimate and many general software packages have specific programs to carry out this test, which is usually referred to as the **Fisher (exact) test**, a test we discuss more fully in Section 9.2.1. StatXact generates the complete permutation distribution for all possible  $X$  values. This is given in Table 5.3, which confirms symmetry of the distribution and illustrates the marked probability discontinuities between successive  $X$  values. In Chapters 2 and 3 we saw that discontinuity decreased rapidly with increasing sample size. Because we have 28 observations the level of discontinuity here may surprise. With two or more samples discontinuities are often quite marked if one sample is small relative to the other(s). This may make some asymptotic results we discuss in Section 5.6 unreliable in such circumstances.

Markedly different sample sizes (in Example 5.6 one is three times the other) often occur in clinical trials where there may be only a few patients with a rare illness. Then comparison is often between responses for these patients and a much larger available ‘control’ group without that illness.

### 5.3.2 Confidence intervals for a median difference

Given full measurement data we may establish a confidence interval for the difference between population medians based on the median test permutation distribution. Writing the true difference  $\delta = \theta_2 - \theta_1$ , a nominal 95 per cent confidence interval for  $\delta$  is  $(d_1, d_2)$  where the end points  $d_1, d_2$  are chosen so that if we change each observation in sample 2 (or sample 1) by more than these amounts we would reject

$H_0$ : the medians are equal for population 1 and the amended population 2. This is relevant because increasing or decreasing all sample (or population) values by a constant  $k$  increases or decreases the median by  $k$ . For given sample sizes  $m, n$  it is not difficult to work out, especially if one has the appropriate computer software, the numbers above the median in the amended sample 2 that will just give significance. We explain the procedure by an example.

### Example 5.7

*The problem.* To determine whether books on management tended to be longer or shorter than books on biology, random samples of 10 books on biology and 12 on management were selected in Dundee University library and the number of pages recorded for each book sampled. The numbers (arranged in order of magnitude within each sample) were

Biology	143	173	226	233	250	287	291	303	634	637		
Management	50	164	198	221	225	302	328	335	426	534	586	618

Determine a nominal 95 per cent confidence interval for the difference between population medians based on the median (Fisher exact) test procedure.

*Formulation and assumptions.* The contingency table in this example has row totals of 10 and 12 (the sample sizes) and column totals of 11 and 11 (numbers above and below the combined sample median). The null hypothesis is that the population medians for the two types of book are equal. With two tails, the median test will indicate strong evidence against this null hypothesis at the 5 per cent significance level if the entry in row 1 column 1 (the number above the median in sample 1) is in a tail region of nominal size 0.025. Using a suitable computer program it can be shown that this is the case if the number above the median in the first sample is either less than or equal to 2 or at least 8. The actual  $P$ -value for a cell value of 2 or 8 is  $2 \times 0.0150 = 0.0300$  (3 per cent). If a computer program is not available this can be shown using formula (5.3) for the contingency tables with cell (1, 1) entries 0, 1, 2 respectively. Having established this we determine what constant adjustment is needed to each observation in sample 2 in order to just avoid rejecting  $H_0$  in a median test based on sample 1 and the adjusted sample 2 but rejecting  $H_0$  for any greater adjustment.

*Procedure.* Since we reject  $H_0$  if only 2 values in sample 1 exceed the joint sample median this implies we reject  $H_0$  if the joint sample median exceeds 303. Since the number above the median in both samples must total 11, this implies that for the adjusted sample 2 if we are not to reject  $H_0$ , it must have 8 values above the adjusted median. If the adjusted median is not to exceed 303 this means the adjustment must ensure that there are 8 values at or above 303 in the adjusted second sample, i.e. the eight largest values in the second sample must exceed 303. Clearly this occurs if we add to all values the difference between 303 and the eighth largest sample value in Sample 2. That value is 225, so we must add  $303 - 225 = 78$  to all second sample values to achieve this. This implies that if  $\theta_2 - \theta_1 > -78$  we would reject  $H_0$ . Similarly, since we reject  $H_0$  if 8 or more values exceed the combined sample median, this implies we reject  $H_0$  if

the joint sample median is less than 226. This means that for the adjusted second sample, if we are not to reject  $H_0$  we must have at least 3 values at or above 226 in that adjusted second sample. Clearly this occurs if we subtract  $426 - 226 = 200$  from all second sample values. This implies that if  $\theta_2 - \theta_1 > 200$  we reject  $H_0$ .

*Conclusion.* A nominal 95 per cent confidence interval for the median difference  $\theta_2 - \theta_1$  is  $(-78, 200)$ .

*Comments.* 1. As usual, at the ends of this interval we have ties at the combined sample medians for the first sample and the adjusted second sample. To check that we would accept  $H_0$  at the end points we must in effect perform a test with allowance for ties at the joint median. [Exercise 5.4](#) asks the reader to check that we would accept  $H_0$  in such a test.

2. The actual confidence level of the above interval is 97 per cent since the exact  $P = 0.03$  (see formulation and assumptions above).

3. Assuming normality, the  $t$ -based 95 per cent confidence interval for the mean difference is  $(-142.94, 172.04)$ . This is a wider interval than that based on median test theory. Indeed, normal theory is not appropriate for these data because the distribution of numbers of pages is skew. The WMW based interval is  $(-103, 185)$ . While a majority of books (as on most subjects) are between 150 and 350 pages in length there are a few works of 600 or more pages and a relatively small number below 150 pages – in practice books are rarely less than 50 pages if we exclude pamphlets and a few special tracts. Because only minimal restrictions about the nature of the populations are needed, the median test is reasonable when the precise population distributions are seriously in doubt. It may however have low power for small samples.

*Computational aspects.* Given a program for the Fisher exact test for a  $2 \times 2$  contingency table one can easily determine for any given  $m, n$  the number above the median in the first sample which just gives significance, and, because there is only one degree of freedom, all other entries in the  $2 \times 2$  table follow automatically. It is then, as indicated in the above example, relatively simple to determine confidence limits by appropriate additions or subtractions from all second-sample observations.

A modification of the median test proposed by Mathisen (1943) is called the **control median test**. This is described by Sprent (1998, Section 4.6) who also refers to an interesting application proposed by Gastwirth (1968) of this or the median test described above that may save resources if observations are obtained sequentially as is often the case in studies of survival times or times to failure, a topic we cover in another context in [Section 5.5](#).

## 5.4 NORMAL SCORES

As in the one-sample case transforming ranks to give scores with something like the characteristics of samples from a normal

distribution is intuitively appealing. In [Section 2.4.1](#) we referred briefly to van der Waerden and expected normal scores. For the two independent samples scenario corresponding to the WMW test van der Waerden scores replace the rank  $r$  by the  $r/(m + n + 1)$ th quantile of the standard normal distribution ( $r = 1, 2, \dots, m + n$ ). Expected normal scores may also be used but there is usually little difference between results using these and van der Waerden scores so we consider here only van der Waerden scores. Symmetry implies that the mean of the  $m + n$  van der Waerden scores is zero if there are no ties. The mean of all van der Waerden scores is not exactly zero with ties if we use quantiles based on mid-ranks; for a few ties the effect is negligible as adjacent quantile scores are replaced by scores close to their mean. For ties an alternative to the mid-rank quantile (which is the mean of the tied ranks) is to take the mean of the quantiles corresponding to each of the ranks that are tied. In practice the difference is usually slight except for large numbers of ties or ties in certain positions (see [Exercise 5.6](#)). If one has an exact permutation test program that covers the Pitman test or allows arbitrary choice of scores this program may be used to compute exact  $P$ -values for a test using van der Waerden scores by substituting these scores for the original data. StatXact and a few other packages provide a program, the normal scores test, that uses van der Waerden scores, forming these directly from the original data with the convention that for ties the mean of the quantiles corresponding to the ranks that are tied (the second of the options mentioned above) is used.

#### *Example 5.8*

*The problem.* For the data for computation times given in [Example 5.1](#), viz.

<i>Group A</i>	23	18	17	25	22	19	31	26	29	33
<i>Group B</i>	21	28	32	30	41	24	35	34	27	39

use van der Waerden scores to assess whether the data indicate that one layout is superior (i.e. leads to more rapid computations).

*Formulation and assumptions.* The data are ranked and van der Waerden scores assigned to each. Using these scores a permutation program for the Pitman test may be used with these scores in place of the raw data to compute exact  $P$ -values.

*Procedure.* It is easiest to use appropriate software to generate the scores, however they may also be obtained using tables of the standard normal distribution. For example, for this data set  $m = 10$ ,  $n = 11$  so that the van der Waerden score for the datum ranked 3 (which here is the observation 19 in Group A) is the 3/22nd quantile of the standard normal distribution, i.e. the value of  $x$  such that  $\Phi(x) = 3/22 \approx 0.1364$ . From tables of the standard normal

distribution this is found to be  $x = -1.097$ . StatXact gives a two-tail  $P = 0.0118$  for this example.

*Conclusion.* There is strong evidence against the null hypothesis of equal medians.

*Comment.* The  $P$ -value is less than that obtained using the WMW procedure and almost identical to that for a  $t$ -test. This is not surprising since the data suggest only slight departures from normality and are of a type where one would not have serious reservations about using a  $t$ -test. In these circumstances although one would expect on the basis of Pitman efficiency that the Wilcoxon test may be not quite as powerful as the  $t$ -test it would be reasonable to expect the normal scores test to perform much like the  $t$ -test as it moulds the data to a form with the characteristics of a sample from a normal distribution. Indeed when sampling from normal populations that differ only in mean the van der Waerden scores test has Pitman efficiency 1.0 relative to the  $t$ -test.

*Computational aspects.* If software for computing exact  $P$ -values with van der Waerden scores is not available an asymptotic approximation valid for large samples is given in [Section 5.6](#). In practice this is reasonable for samples of moderate size because of the ‘normalizing influence’ of the transformation which enhances the rate of convergence towards the asymptotic approximation.

## 5.5 TESTS FOR SURVIVAL DATA

Nonparametric tests are widely used in the analysis of survival and failure-time data. Two characteristics of such data have stimulated development of special tests:

- Some units are often lost to a study before the response of interest (e.g. death of a patient or complete recovery, or failure of a machine) occurs.
- Distributions of survival times of patients in clinical trials, or of times to failure for components in an engineering or industrial context are typically long-tailed or skewed to the right.

Withdrawal of a subject before the response of interest occurs is called ‘right censoring’ to indicate that we are losing information about an observation towards the right tail of a distribution. Censoring may occur at a fixed time because it is decided to terminate an experiment before all subjects have shown the response of interest or it may occur when a predetermined proportion or number of units subject to a particular treatment have responded. It may also occur because subjects withdraw or are lost to a study before the response of interest occurs. The analysis is complicated if censoring is in some way treatment dependent (e.g., if subjects are

more likely to withdraw if they are given a treatment which produces unpleasant side-effects).

The characteristic long-tailed distribution reduces the role of the normal distribution compared to the part it plays in many applications. Parametric tests for survival data often assume data are samples from exponential, Weibull or other long-tailed distributions.

In this section we discuss only briefly how these factors influence a nonparametric approach, for analysis of survival data is a subject that would itself quickly fill a book. Introductory surveys that include references to more detailed work are given by Sprent (1998, Sections 4.7–4.9) and by Hollander and Wolfe (1999, Chapter 11).

The examples in this section indicate two simple approaches used in practice. More sophisticated methods are needed for most advanced studies.

### 5.5.1 The Gehan–Wilcoxon Test

We met a sample with censored data in [Example 1.2](#). Censoring is common in medical studies that involve long-term follow up of patients after treatment as well as in some industrial contexts. Davis and Lawrence (1989) give an example involving tyre failures. Tyres that had not failed at the end of an experimental period were regarded as censored observations. In a medical context censoring may arise for one or more of the reasons discussed above. Gehan (1965a, 1965b) proposed generalized Wilcoxon-type tests for both one- and two-sample problems with various types of censoring. We consider one two-sample case.

A censored observation clearly has a longer survival time than any unit that has failed at or before the time of censoring. However, Gehan regarded any two censored observations as giving no definite information on survival times relative to each other. Nor did he regard a censored observation as providing information on whether that subject does or does not have a greater survival time than that of a later failure. He scored these *no definite information* cases as ties.

*Example 5.9.*

*The problem.* We mentioned in [Example 1.2](#) that Dinse (1982) gave survival data for 28 asymptomatic cases as well as for the 10 symptomatic cases given in that example. The complete data, an asterisk denoting censoring, are:

Symptomatic	49	58	75	110	112	132	151	276	281	362*
Asymptomatic	50	58	96	139	152	159	189	225	239	242
	257	262	292	294	300*	301	306*	329*	342*	346*
	349*	354*	359	360*	365*	378*	381*	388*		

Test  $H_0$ : survival times distributions are identical against  $H_1$ :  $G(x) \leq F(x)$  with strict inequality for some  $x$ , where  $F(u)$ ,  $G(v)$  are survival time cumulative distribution functions for symptomatic and asymptomatic cases.

*Formulation and assumptions.* A one-tail test is appropriate since  $H_1$  is one-sided. Because of censoring we use Gehan–Wilcoxon scores.

*Procedure.* The test is easily explained and carried out using analogues of the WMW statistics  $U_m$  or  $U_n$ . To obtain the modified form of  $U_n$  for each observation in the first sample (symptomatic) that is a definite failure time (not a censored observation) we count the number that survive longer in the second sample (i.e. that fail or are censored at a later time) scoring 1 for each such case as in the WMW test; we also score  $\frac{1}{2}$  for each unit in the second sample that is censored earlier. For any censored unit in the first sample we count the number of second sample units that fail after the time that item was censored and add to this count the number of second sample items that have been censored at either an earlier, the same or a later time, regarding each as a tie and scoring it as  $\frac{1}{2}$ . The statistic  $U_n$  is the sum of all such scores taken over all first sample observations. For the first observation 49 the score is 28 for clearly all second sample observations represent later failures. For the second entry 58 the score is 26.5 since there are 26 later failures and one tied value. For the final entry 362\* the score is 6 since there are no second sample values that definitely have a longer survival time but there are 12 censored units in the second sample that are each regarded as a tie and scored as  $\frac{1}{2}$ . Carrying out the process for all units in the first sample it is easily confirmed that

$$U_n = 28 + 26.5 + 26 + 25 + 25 + 25 + 24 + 16 + 16 + 6 = 217.5$$

Testimate, using an equivalent but different scoring system described in the manual, gives the exact one-tail  $P = 0.0040$ .

*Conclusion.* There is fairly strong evidence against  $H_0$ .

*Comments.* 1. We express  $H_1$  in terms of one distribution dominating the other rather than as a median shift because an efficient treatment is likely to increase life expectancy not by a constant amount for all subjects, but by amounts which vary from subject to subject depending upon whether there is slow or rapid recovery or perhaps just an arrest of the development of a disease.

2. All censored values in this example are at least 300; for the asymptomatic cases survival times are sufficiently long for many patients to survive this time, whereas only one symptomatic case does.

3. The treatment of censored data in the Gehan–Wilcoxon test is conservative. Peto and Peto (1972) and others have suggested stronger but realistic assumptions about the life expectancy of units that provide only censored observations. These often lead to more powerful tests, especially if there is substantial censoring. One such test is described in Sprent (1998, Section 4.8) where it is called the Peto–Wilcoxon test.

*Computational aspects.* StatXact includes a program for a generalized Gehan–Wilcoxon test which differs slightly from that given here, but it usually gives similar results. For this example it gives  $P = 0.0041$  for a one-tail test.

### 5.5.2 Savage scores and the log-rank transformation

The prevalence of long-tail distributions, often not unlike an exponential distribution, has given that distribution a role in survival and failure time studies almost as central as that of the normal distribution in many other types of problem. While methods of analysis based on the normal distribution are often moderately robust against many minor (and even some major) departures from assumptions, tests based on the exponential distribution are often sensitive to even small departures from that distribution. This has led to the development of distribution-free methods to overcome such problems. Savage (1956) introduced a transformation – now often referred to as the log-rank transformation – in which ranks are replaced by what are often known as Savage or exponential scores. The transformation is essentially one to expected values of the order statistics of an exponential distribution with parameter  $\lambda = 1$  (the ‘unit’ exponential distribution). If there are no ties and no censoring in two samples of  $m, n$  observations then the Savage score corresponding to the combined sample rank  $r$ , ( $r = 1, 2, \dots, N$ ) is

$$s_r' = 1/N + 1/(N-1) + \dots + 1/(N-r+1)$$

where  $N = m + n$ .

It is easily verified that the sum of all  $s_r'$  is  $N$  and an equivalent score commonly used is

$$s_r = 1 - s_r'$$

and these latter scores sum to zero. The sums of these scores for either sample are appropriate statistics in an exact permutation test which may be carried out using either specific tests for these scores or they may be substituted for the original data in a Pitman or arbitrary score permutation test program. StatXact includes a program that calculates scores automatically for a given data set making appropriate modifications for ties or censored data. We do not discuss these aspects here but they are covered in some detail in Sprent (1998, Section 4.9).

Dinse (1982) and Dinse and Lagakos (1982) have discussed estimation problems for incomplete survival data. Emerson (1982) and Brookmeyer and Cowley (1982) discuss confidence intervals for the median when data are censored. Woolson and Lachenbruch (1983) consider analysis of covariance with censored data and Hanley and Parnes (1983) consider censoring in multivariate distributions from a nonparametric viewpoint.

Other papers dealing with various refinements in analysis of censored data include Woolson and Lachenbruch (1980, 1981), Woolson (1981), Dinse (1986), Albers and Akritas (1987), Davis and Lawrence (1989), O'Quigley and Prentice (1991) and Babu, Rao and Rao (1992), a list that is indicative rather than exhaustive.

When there is uncertainty about which of several tests ranging from WMW tests to those using Savage scores is most appropriate Gastwirth (1985) proposed a test that is robust over a family of models and called them maximum efficiency robust tests (MERT). This and a related test are further considered by Freidlin, Podgor and Gastwirth (1999). StatXact has an option for MERT tests.

When we wish to test equality of centrality but make no assumption of equal spreads the situation is analogous to that of the classic Behrens–Fisher problem of testing for equality of means given samples from two normal distributions with different unknown variances. This more general problem has been considered by Pothoff (1963), Fligner and Rust (1982) and others.

## 5.6 ASYMPTOTIC APPROXIMATIONS

There are simplified formulae for asymptotic results for the WMW test without ties, but it is convenient to first introduce asymptotic results valid for more general scoring systems. These cover not only mid-ranks for ties but also apply to the rank transformations introduced in Sections 5.4 and 5.5, including cases where data are censored. In all cases we denote the score (rank, tied rank, Savage score or whatever) for the  $i$ th ordered observation in the joint sample by  $s_i$ . If the sample sizes are  $m, n$  and  $T$  denotes the sum-of-scores statistic based on the sample of  $m$ , then the general theory for the relevant permutation distribution gives, after some straightforward but tedious applications of standard results, the mean value  $E(T)$  and the variance  $\text{Var}(T)$  of  $T$  to be

$$E(T) = \frac{m}{m+n} \left( \sum s_j \right) \quad (5.4)$$

$$\text{Var}(T) = \frac{mn}{(m+n)(m+n-1)} \left( \sum s_j^2 - \frac{1}{m+n} \left( \sum s_j \right)^2 \right) \quad (5.5)$$

where summations are over all  $j$  from 1 to  $m + n$ . For large  $m, n$

$$Z = \frac{T - E(T)}{\sqrt{\text{Var}(T)}} \quad (5.6)$$

has approximately a standard normal distribution. For the WMW test with no ties setting  $T = S_m$  (5.6) reduces to

$$Z = \frac{S_m - \frac{1}{2}m(m+n+1)}{\sqrt{mn(m+n+1)/12}} \quad (5.7)$$

A continuity correction in the ‘no-tie’ case has little practical effect for large  $m$  and  $n$ , but if used the correction is to add  $\frac{1}{2}$  to  $S_m$  if  $Z$  is negative and to subtract  $\frac{1}{2}$  if  $Z$  is positive. If one prefers to use  $S_n$  the only alteration needed in (5.7) is to interchange  $m$  and  $n$  throughout (which does not alter the denominator). If one uses the Mann–Whitney statistic the numerator of (5.7) is modified giving

$$Z = \frac{U_m - \frac{1}{2}mn}{\sqrt{mn(m+n+1)/12}} \quad (5.8)$$

If there are ties and we use mid-rank scores the numerators, but not the denominators, of (5.7) and (5.8) are unaltered. Some writers (e.g. Daniel, 1990, p. 94) give rules for modifying the denominator, but with modern calculators or computers it is easy to use (5.6) directly where the  $s_j$  are mid-rank scores. Unless there are many ties the effect of these is small. As mentioned above, we may also use (5.6) with appropriate scores for large-sample tests with van der Waerden scores, Gehan–Wilcoxon scores or Savage scores. StatXact and other packages provide such asymptotic approximations for use when samples are too large for exact tests.

### *Example 5.10*

*The problem.* Ages at death for members of two clans (see appendix) in the Badenscallie burial ground (arranged in ascending order) are:

<i>McGamma</i>	13	13	22	26	33	33	59	72	72	72	77
	78	78	80	81	82	85	85	85	86	88	
<i>McBeta</i>	0	19	22	30	31	37	55	55	56		
	66	66	67	67	68	71	73	73	75		
	75	78	79	82	83	83	88	88	96		

Use the WMW large-sample approximation to test  $H_0$ : *the populations are identical* against  $H_1$ : *the population medians differ*.

*Formulation and assumptions.* With sample sizes 21, 24 a large-sample approximation should suffice. To adjust for ties we require mid-ranks.

*Procedure.* The appropriate mid-ranks are:

<i>McGamma</i>	2.5	2.5	5.5	7	10.5	10.5	15	23	23
	23	28	30	30	33	34	35.5	40	40
	40	42	43.5						
<i>McBeta</i>	1	4	5.5	8	9	12	13	14	16.5
	16.5	18.5	18.5	20	21	25	26.5	26.5	30
	32	35.5	37.5	37.5	43.5	45			

While one may use either  $S$  or  $U$ , once the ranks are recorded it is probably as easy to use  $S$ . For the McGamma's we easily find  $S_m = 518.5$ . Ignoring the effect of ties and using (5.7) gives

$$Z = (518.5 - 483)/\sqrt{1932} = 35.5/43.95 = 0.81.$$

We leave it as an exercise to show that using (5.5) to compute the variance in formula (5.6) reduces the denominator 43.95 given above only to 43.92, demonstrating the trivial effect of just a few ties.

*Conclusion.* Since  $P = \Pr(|Z| > 0.81) = 0.418$  there is no plausible evidence against the hypothesis of identical population medians.

*Comments.* 1. For both  $m, n$  greater than 20 one expects asymptotic results to agree closely with those for the exact WMW permutation test. Here the exact two-tail probability is  $P = 0.426$ .

2. As in [Example 5.9](#) we might argue that rather than  $H_1$  specifying a median shift a hypothesis of dominance of one distribution over the other may be more appropriate on genetic grounds. Inspection of the data gives a slight hint that we may have a generalized Behrens–Fisher type situation of the type described at the end of the previous section; this will reduce the power of the WMW test.

*Computational aspects.* With efficient algorithms computation of the exact permutation distribution tail probability in this case takes only seconds on most PCs but as we have pointed out above the asymptotic results are nearly as good for samples of this size and there may seem little point in calculating exact probabilities when an asymptotic result suffices. Of particular interest when using an asymptotic result is whether a reasonable fit in the tails is obtained. For the case  $m = 21, n = 24$ , if  $S_m = 641$  the exact test one-tail probability is 0.0001 and the asymptotic probability is 0.0002; if  $S_m = 557$  the corresponding probabilities are 0.0317 and 0.0311.

If an approximate 95 per cent confidence interval is required for large  $m, n$  the value of  $U$  giving the relevant number of extreme differences  $x_i - y_j$  in a table similar to [Table 5.3](#) is obtained by setting  $Z = -1.96$  and substituting the values of  $m, n$  in (5.8). In practice, for large  $m, n$  this value of  $U$  will be large and a computer program will be needed to get the required limits. Minitab and several other packages provide a program to give all differences.

Asymptotic approximations for the median test are not required if software for the Fisher exact test is available. If such software is not available an asymptotic form of the Fisher exact test may be based on the chi-squared test in a way described in [Section 9.2.2](#).

## 5.7 POWER AND SAMPLE SIZE

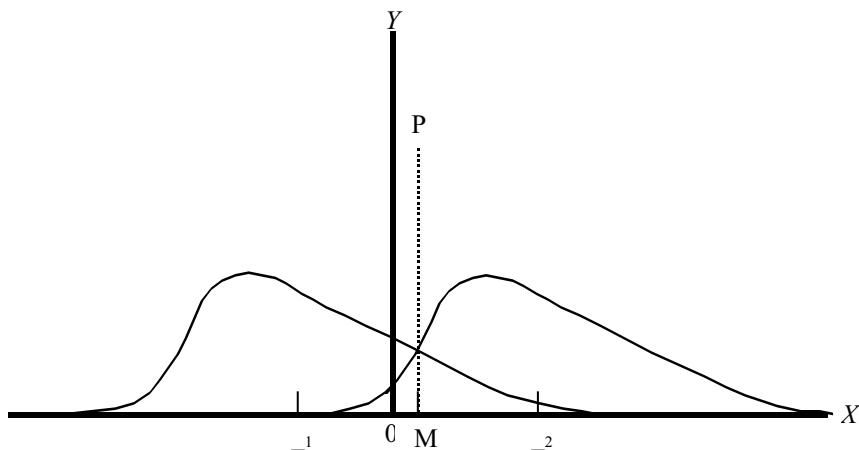
In the light of the discussion in [Section 4.3](#) for single samples it is not surprising that power and sample size results for the median test and the WMW test are not distribution-free. Not unexpectedly, the situation for the WMW test is more complicated than that for the median test because the latter is an extension from the sign test and therefore essentially only involves binomial distributions whereas the WMW test involves the distribution of ranks which may be complicated under  $H_1$ . We consider the median test first.

### 5.7.1 Power and sample size for the median test

The hypotheses only involve possible differences between unspecified population medians  $\theta_1$  and  $\theta_2$ . If we denote this difference by  $\delta = \theta_2 - \theta_1$  then the null-hypothesis is  $H_0: \delta = 0$  and the median test is equivalent to testing whether the samples support a hypothesis that the probability that a sample value is above the common median is  $p = 1/2$  for each sample. If we use the combined sample median as our estimate of the common median then under  $H_0$  the probability is only approximately  $p = 1/2$  for reasons indicated in [Section 5.3](#). Under an alternative hypothesis  $H_1$  specifying some value  $\delta \neq 0$  the probabilities  $p_1, p_2$  for sample values to lie above the combined sample median will be unequal but estimation of  $p_1, p_2$  and the interpretation of the relationship between the combined sample median and the two population medians depends on both the nature of those populations and the relative sample sizes.

The combined sample median is only easy to interpret if samples are from populations that differ, if at all, only by a median shift. The interpretation is simplest when  $m = n$ . Then if  $H_1$  specifies a fixed difference  $\delta$  between the medians, i.e.  $H_1: \theta_2 - \theta_1 = \delta$  it is easily seen that the combined sample median is an intuitively reasonable estimator of a constant midway between  $\theta_2$  and  $\theta_1$ , i.e. an estimator of  $\theta_1 + 1/2\delta = \theta_2 - 1/2\delta$ . [Figure 5.1](#) shows probability density functions for two arbitrary distributions that differ only in their medians  $\theta_1, \theta_2$ . The dotted vertical line  $PM$  meets the  $x$ -axis at the point  $M$  where  $\theta_1 + 1/2\delta = \theta_2 - 1/2\delta$ . If we now assume that under  $H_0$  the common value of the median is  $\theta_1$  and the common probability density function  $f(x)$  is completely defined we are now able to

calculate the approximate power of the median test under these restrictive assumptions by computing a probability  $p_1$  that under  $H_1$  a sample value from the specified distribution with median  $\theta_1$  takes a



**Figure 5.1** Probability density functions for two distributions that differ only by a shift in median or mean. The point  $M$  lies midway between the population medians  $\theta_1, \theta_2$  and is the median of a population consisting of an equal mixture of the two populations. The combined sample median provides an estimate of this point if both samples are of equal size  $m = n$ .

value greater than  $\theta_1 + 1/2\delta$ . Similarly,  $p_2$  is computed for a similar distribution with the median shifted to  $\theta_1 + \delta$ . Programs are now fairly readily available for power calculations for this situation when we have assigned median shifts for completely specified distributions. We illustrate the method in Examples 5.11 and 5.12. However this approach is unrealistic in many practical situations where we are likely to use the median test, for these are conditions where a WMW test may in general be more powerful.

In more complicated situations where a median shift is the only hypothesized difference a more general approach based on order statistics may be used. However, here again some knowledge of the precise form of the distribution is required. The method is outlined by Gibbons and Chakraborti (1992, Section 7.4) who also give an asymptotic approximation available for unequal sample sizes.

An alternative approach to obtaining the sample size required to obtain a given power that is probably a reasonable approximation if  $m = n$  is to use any available program to work out the sample size for a corresponding  $t$ -test with the required power and to adjust on the basis of the Pitman efficiency relative to the  $t$ -test where this is known. This is discussed briefly in the comments on Examples 5.11 and 5.12.

#### *Example 5.11*

*The problem.* For samples of size 20 from each of two normal distributions with standard deviation 2 and unknown medians  $\theta_1, \theta_2$  find the power of the

median test of  $H_0: \theta_2 = \theta_1$  against  $H_1: \theta_2 = \theta_1 + 1$  when the probability of a type I error is  $\alpha = 0.05$ . What equal sample sizes would be needed to ensure power of at least 0.80?

*Formulation and assumptions.* Under  $H_0$  the rows in the  $2 \times 2$  table for the Fisher exact test in [Section 5.3](#) may be considered as the number of positive and negative outcomes in independent samples each of size 20 when sampling from binomial distributions for which the  $p$  are equal with approximately the value  $p = 1/2$ . Under  $H_1$  the probability that any value in the sample from the first population has a value greater than  $\theta_1 + 1/2$  is given by  $p_1 = \Pr(X > \theta_1 + 1/2)$  where  $X$  has a  $N(\theta_1, 4)$  distribution and this is approximately the probability that a sample value from this distribution takes a value greater than the combined sample median. Similarly, under  $H_1$  the probability that any value in the sample from the second population has a value greater than  $\theta_1 + 1/2$  is given by  $p_2 = \Pr(X > \theta_1 + 1/2)$  where  $X$  has a  $N(\theta_1 + 1, 4)$  distribution and this is approximately the probability that a sample value from this distribution takes a value greater than the combined sample median. As a consequence of the symmetry assumptions it is easily verified that  $p_2 = 1 - p_1$ . Approximate power and sample size determinations now reduce to considering those for the problem of testing  $H_0: p_1 = p_2 = 1/2$  against  $H_1: p_1, p_2$  take the values computed when  $H_1: \theta_2 = \theta_1 + 1$  holds. Computer software exists for the latter test giving either exact or asymptotic approximations.

*Procedure.* Clearly, for the first population, since  $X$  is  $N(\theta_1, 4)$  it follows that  $Z = (X - \theta_1)/2$  has a standard normal distribution. Thus  $p_1 = \Pr(X > \theta_1 + 1/2) = \Pr(Z > 0.25) = 0.4013$ . Since  $p_2 = 1 - p_1$  this implies  $p_2 = 0.5987$ , which may be verified directly. We used the program in StatXact for power calculations for the Fisher exact test for comparing two binomials. This provides both an asymptotic estimate of power or an option for exact computation. For  $n = m = 20$  for the test considered here the program gives an exact power 0.22. Larger samples are required to achieve power 0.80. If software is not available to estimate the sample size a trial and error approach is needed. For  $m = n = 100$  StatXact gives a power 0.85. For  $n = m = 90$  the required power 0.80 is achieved.

*Conclusion.* For  $m = n = 20$  the power is 0.22. Samples of size  $m = n = 90$  are required for power 0.80.

*Comments.* 1. Since the samples are from normal distributions differing only in mean the optimal test is a  $t$ -test and for the situation in this example the relevant computations (available in many statistical packages) indicate  $m = n = 51$  are the sample sizes needed to attain a power 0.8. This is broadly in line with the expected size based on the Pitman efficiency of 0.64 for the median test when the  $t$ -test is optimal. If we assume the Pitman efficiency is reasonable for the test under consideration this would suggest the sample size  $51/0.64 = 80$  approximately. However it should be kept in mind that this is only a limiting result concerning the power for small departures independent of the choice of  $\alpha$ ,  $\beta$  and that appreciable variations from this estimate are possible for finite samples, specific alternatives and particular choices of  $\alpha$ ,  $\beta$ . However, using a sample size based on the relevant Pitman efficiency will usually give an indication of the size needed to give reasonable power in a practical context unless samples are small or unbalanced.

2. The theory behind power and sample size computations for the Fisher exact test on which the results obtainable from StatXact are based is described by Mehta and Patel (1999, Chapter 25). StatXact computes an unconditional power that ignores the restriction implicit in the median test that numbers of observations above and below the combined sample median are fixed. In practice this unconditionality may tend to overestimate the power of this particular test but in our experience this overestimation is only slight in the context here.

3. The restriction to  $m = n$  is not essential to the binomial comparison problem but it is needed to induce the property that the sample median is a reasonable estimate of  $\theta_1 + \frac{1}{2}\delta$ . This in turn depends on the assumption that samples come from identical distributions apart from possible differences between medians. Relaxation of these assumptions changes the function of  $\theta_1$  and  $\theta_2$  that is estimated by  $M$ .

4. In the light of common uncertainty about the exact nature of the populations being sampled asymptotic power calculations for comparing two binomials available in many software packages will often suffice for obtaining sample sizes needed to give substantial power since these sample sizes are often fairly large and an approximate sample size is often all that is required.

### *Example 5.12*

*The problem.* For samples of size 20 from each of two double exponential distributions with standard deviation 2 and unknown medians  $\theta_1, \theta_2$  find the power of the median test of  $H_0: \theta_2 = \theta_1$  against  $H_1: \theta_2 = \theta_1 + 1$  when the probability of a type I error is  $\alpha = 0.05$ . What equal sample sizes would be needed to ensure power of at least 0.80?

*Formulation and assumptions.* As in [Example 4.6](#) a double exponential distribution with mean  $\underline{\theta}$  and standard deviation 2 has frequency function

$$f(x) = [1/(2\sqrt{2})]\exp[-|(x - \underline{\theta})|/\sqrt{2}] \quad (5.9)$$

Using arguments similar to those in [Example 5.11](#) under  $H_0$  the rows in the  $2 \times 2$  table for the Fisher exact test in [Section 5.4](#) may be considered as the number of positive and negative outcomes in independent samples each of size 20 when sampling from binomial distributions for which the  $p$  are equal with approximately the value  $p = \frac{1}{2}$ . Under  $H_1$  the probability that any value in the sample from the first population has a value greater than  $\theta_1 + \frac{1}{2}$  is given by  $p_1 = \Pr(X > \theta_1 + \frac{1}{2})$  where  $X$  has a double exponential distribution given by (5.9) with mean  $\theta_1$  and this is approximately the probability that a sample value from this distribution takes a value greater than the combined sample median. Similarly, under  $H_1$  the probability that any value in the sample from the second population has a value greater than  $\theta_1 + \frac{1}{2}$  is given by  $p_2 = \Pr(X > \theta_1 + \frac{1}{2})$  where  $X$  has a double exponential distribution with mean  $\theta_1 + 1$  and this is approximately the probability that a sample value from this distribution takes a value greater than the combined sample median. Because of symmetry it is easily verified that as in the previous example  $p_2 = 1 - p_1$ . Approximate power and sample size determinations now reduce to a consideration of those for the problem of testing  $H_0: p_1 = p_2 = \frac{1}{2}$  against  $H_1: p_1, p_2$  take the values computed when  $\theta_2 = \theta_1 + 1$ .

*Procedure.* For the first population sampled  $p_1$  under  $H_1$  is easily computed by integrating (5.9) over the interval  $(\theta + \frac{1}{2}, \infty)$  where  $\theta = \theta_1$  (see [Exercise 5.12](#)). This gives  $p_1 = 0.3511$ . Since  $p_2 = 1 - p_1$  this implies  $p_2 = 0.6489$ , which may be verified directly by integrating (5.9) over the interval  $(\theta + \frac{1}{2}, \infty)$  where now  $\theta = \theta_1 + 1$ . The StatXact program used in [Example 5.11](#) gives for  $n = m = 20$  for the test considered here an exact power 0.45. Larger samples will be needed to achieve power 0.80. If no software is available to estimate the sample size directly a trial and error approach is needed. For  $m = n = 40$ , StatXact gives a power 0.79. For  $n = m = 41$  the required power 0.80 is achieved.

*Conclusion.* For  $m = n = 20$  the power is 0.45. Samples of size  $m = n = 41$  are required for power 0.80.

*Comments.* 1. The  $t$ -test is no longer optimal so a direct comparison is inappropriate. However, in view of its non-optimality the  $t$ -test would require a larger sample size to achieve the same power as it would have for the corresponding normal test, where, as indicated in Comment 1 on [Example 5.11](#) we require  $m = n = 51$  to attain a power 0.8, indicating, in line with the Pitman efficiency of 2.0 relative to the  $t$ -test that the median test is preferable here.

2. Comments 2, 3 and 4 on [Example 5.11](#) are also relevant here.

3. Although the median test has higher Pitman efficiency than the WMW test for double exponential distributions differing if at all only in medians several authors have indicated that for small or unbalanced samples (i.e., samples of very different sizes) the median test may not have as high a power as the WMW test for some alternatives specified in  $H_1$ .

### 5.7.2 Power and sample size for the WMW test

The added complexities of power and sample size computations for the Wilcoxon signed-rank test relative to the sign test carry over to the WMW test relative to the median test. We consider here only an approximation due to Noether (1987a) similar to that given for the signed-rank test in (4.3) for determining the equal sample sizes to ensure a given power when testing for a median shift only in otherwise identical distributions (which need not be symmetric).

The analogous formula to (4.3) is

$$n \approx \frac{\left( z_\alpha + z_\beta \right)^2}{6(p_1 - \frac{1}{2})^2} \quad (5.10)$$

where  $p_1 = \Pr[(Y - X) > 0]$  where we sample from populations of random variables  $X, Y$  that differ only in medians and  $n = m$  is the common sample size. The other terms have the same meaning as in (4.3). The distribution of  $Y - X$  is of course dependent on the nature of the distributions of  $X$  and  $Y$  and is in general obtained by standard distribution theory for change of variables. In simple situations like that when  $X, Y$  are both normally distributed  $p_1$  is easily computed as we see in the next example.

### Example 5.13

*The problem.* Samples of equal size  $n$  are to be taken from two normal distributions each with standard deviation 2 and unknown medians  $\theta_1, \theta_2$ . Use (5.10) to estimate  $n$  so that the power of the WMW test of  $H_0: \theta_2 = \theta_1$  against  $H_1: \theta_2 = \theta_1 + 1$  is at least 0.80 when the probability of a type I error is  $\alpha = 0.05$ .

*Formulation and assumptions.* Relevant values of  $z_\alpha, z_\beta$  are obtained from tables or using appropriate software. Standard normal distribution theory tells us that  $Y - X$  is in this case distributed  $N(1, 8)$  whence  $p_1$  is easily obtained. Relevant values are substituted in (5.10).

*Procedure.* We leave it to the reader (Exercise 5.13) to show that  $z_\alpha = 1.645$ ,  $z_\beta = 0.842$  and  $p_1 = 0.6381$ , whence substitution in (5.10) gives  $n \approx 54.05$ .

*Conclusion.* Rounding up the result suggests samples of size 55 are needed to attain the required power.

*Comment.* These results for the normal distribution are consistent with the known Pitman efficiency of 0.95 relative to the  $t$ -test in this case because we noted in Example 5.11 that the  $t$ -test sample sizes needed for this power are  $n = 51$ , giving a relative efficiency for the WMW test of  $51/55 \approx 0.93$ , in close agreement with the Pitman efficiency  $3/\pi$ .

Hollander and Wolfe (1999, Section 4.1) generalize (5.10) to unequal sample sizes and discuss some further results for asymptotic approximate power calculations.

## 5.8 TESTS FOR EQUALITY OF VARIANCE

Historically, tests about differences between medians or means have dominated inference for two sample problems, but there is increasing interest in differences in variation or spread which may or may not be associated with differences in means or medians. In the field of quality control, for example, this reflects consumer demand that not only should all products of a particular kind perform well *on average* but that they should also be consistently reliable.

Care is needed in testing these characteristics when we drop an assumption of normality. In the two independent sample situation when we assume normality the appropriate statistic to test equality of population variances, irrespective of whether the means are equal, when we have samples  $x_1, x_2, \dots, x_m$  and  $y_1, y_2, \dots, y_n$  is the well-known

$$F = \frac{\sum(x_i - \bar{x})^2 / (m-1)}{\sum(y_j - \bar{y})^2 / (n-1)}$$

which has an  $F$  distribution with  $m - 1$ ,  $n - 1$  degrees of freedom under the hypothesis that the population variances are equal, irrespective of whether the means are equal.

The normal distribution family is the best-known example of what are called **location-scale parameter** families. These are families of distributions of random variables  $X$ ,  $Y$  having location parameters  $\theta_1$ ,  $\theta_2$  and scale parameters  $\varphi_1$ ,  $\varphi_2$  that imply that  $U = (X - \theta_1)/\varphi_1$  and  $V = (Y - \theta_2)/\varphi_2$  are identically distributed. For the normal distribution in conventional notation we have  $\theta_i = \mu_i$ , and  $\varphi_i = \sigma_i$  ( $i = 1, 2$ ) and  $U$ ,  $V$  each have a standard normal distribution. We shall see below not only that we may be interested in spread differences between members of families of distributions that are not of the location-scale parameter form, but also that even when we have such families many of our tests for differences in dispersion assume either that  $\theta_1 = \theta_2$  or that we know the value of the difference  $\delta = \theta_2 - \theta_1$ . The conventional  $F$  statistic relevant to normal distributions is almost unique in that it provides a test for differences between the population spreads or dispersions without needing to know the values of the population location parameters.

So far in this chapter we have considered tests basically designed for shifts in means or medians, but we have pointed out that several of the tests are also relevant to tests of dominance where the alternative hypothesis is of the form  $H_1: G(x) \leq F(x)$  with strict inequality for at least some  $x$ .

This is important in the context of dispersion because for many distributions even when they belong to the same family it is impossible to have a shift in mean or median without altering other distribution characteristics such as variance. A simple example is the exponential distribution with parameter  $\lambda$ . The probability density function is  $f(x) = \lambda e^{-\lambda x}$ ,  $x \geq 0$ , and  $E(X) = 1/\lambda$ ,  $\text{Var}(X) = 1/\lambda^2$  and  $\text{med}(X) = \ln(2/\lambda)$ , thus a change in  $\lambda$  changes the mean, median and variance reflecting a dominance situation. The parameter  $\lambda$  is often referred to as a scale parameter.

Where both centrality and spread are influenced by parameter changes there are difficulties in developing completely distribution-free tests for differences in variance or more general measures of spread. As already indicated tests for differences in variance require an assumption that the means or medians are equal or that we know the difference between them. These points are discussed more fully in Sprent (1998, Section 6.5), Hollander and Wolfe (1999, Sections 5.1 and 5.2) and by Gibbons and Chakraborti (1992, Chapter 10). Most tests for a difference in spread either implicitly or explicitly

assume samples are from populations with identical medians, or that the difference between medians is known and that an adjustment has been made to align the samples by addition or subtraction of that difference to the data for one of the samples.

If the sample values, or if a test, indicate equality of medians is untenable it is often suggested that one should adjust sample values to align the sample medians. This has intuitive appeal, and is often a sensible thing to do, but examples exist to show that it may sometimes be counter-productive and in general it will tend to give rather too many small  $P$ -values.

Here we concentrate mainly on tests that **assume** population means or medians are identical or can be made so by an addition to or subtraction from one set of sample values, but in practice such assumptions are often at best approximations.

### 5.8.1 The Siegel–Tukey test

This test proposed by Siegel and Tukey (1960) is easy to carry out but like many nonparametric tests for spread or dispersion, is not very powerful. The Pitman efficiency relative to the  $F$ -test when samples are from normal distributions is only 0.61. The basic idea behind it is that if two samples come from populations differing only in variance, the sample from the population with greater variance will be more spread out. If there is a known centrality difference as well we first align the samples by subtracting the median difference from all values in one sample, but in practice this difference is seldom known. If an unknown centrality difference is assumed a common practice is to align the populations by shifting the median (or mean) of one sample to coincide with that of the other sample. This requires an appropriate addition to or subtraction of a constant for all observations in one sample. The sample variances are unaltered by this change, but as indicated above the procedure is not optimal unless the difference between sample medians happens to equal that between the population location parameters in a location-scale model situation. However, the change usually works reasonably for this test if there is no strong sample evidence of differences in skewness. If we now arrange the combined samples in order and allocate the rank 1 to the smallest observation, 2 to the largest, 3 to the next largest, 4 and 5 to the next two smallest, 6 and 7 to the next two largest, and so on, the sum of the ranks for the sample from the population having the greater variance should be smaller than if there were no difference in variance.

### Example 5.14

*The problem.* Davis and Lawrence (1989) give (as part of a larger data set collected for different purposes) the time in hours to two different types of tyre failure under similar test conditions. Failure type A is rubber chunking on shoulder and failure type B is cracking of the side wall. Do the following data suggest the population variances may differ?

Type A	177	200	227	230	232	268	272	297				
Type B	47	105	126	142	158	172	197	220	225	230	262	270

*Formulation and assumptions.* We align the medians of the two samples, then allocate ranks as described above. The lesser sum of ranks associated with one sample is calculated and tested for significance as in the WMW test. Our null hypothesis is that the variances are identical. Without further information a two-tail test is appropriate.

*Procedure.* The first sample median is  $\frac{1}{2}(230 + 232) = 231$  and the second sample median is 184.5. We align for location by subtracting  $231 - 184.5 = 46.5$  from all first-sample values and work with the adjusted samples

Type A (adjusted)	130.5	153.5	180.5	183.5	185.5	221.5	225.5	250.5
Type B	47	105	126	142	158	172		
	197	220	225	230	262	270		

Assigning ranks using the scheme outlined above is easier if we arrange all sample values in ascending order before allocation. The result of this operation is given in Table 5.4 where type A (adjusted) values are indicated in **bold**. The sum of the ranks for type A faults is  $S_m = 106$ , giving  $U_m = S_m - \frac{1}{2}m(m+1) = 70$  and  $U_n = mn - U_m = 26$ . Tables for the WMW test gives  $U = 22$  as the critical value for significance at a nominal 5 per cent level in a two-tail test for samples of 8 and 12 observations. StatXact provides a program for an exact Siegel–Tukey test and this indicates that the exact two-tail  $P$ -value corresponding to  $S_m = 106$  is  $P = 0.0979$ .

*Conclusion.* There is only weak evidence against the hypothesis of equal variance; not sufficient to reject  $H_0$  at a nominal 5 per cent significance level.

*Comments.* 1. For these data the WMW test two-tail  $P = 0.0268$ . The Siegel–Tukey program in StatXact does not adjust the sample medians to be equal. This must be done manually or by means of the program editing facility.

2. We ‘relocated’ the type A data by equating sample medians. Since variance is based on squared deviations from the mean, equating means has intuitive appeal. In practice, which of these alternatives we choose usually makes little difference. The median is easier to calculate – hardly a justification with

**Table 5.4** Allocation of ranks for Siegel–Tukey test; type A (adjusted) in **bold**.

<i>Value</i>	47	105	126	<b>130.5</b>	142	<b>153.5</b>	158	172	<b>180.5</b>	<b>183.5</b>
<i>Rank</i>	1	4	5	<b>8</b>	9	<b>12</b>	13	16	<b>17</b>	<b>20</b>
<i>Value</i>	<b>185.5</b>	197	220	<b>221.5</b>	225	<b>225.5</b>	230	<b>250.5</b>	262	270
<i>Rank</i>	<b>19</b>	18	15	<b>14</b>	11	<b>10</b>	7	<b>6</b>	3	2

modern computing methods—but it may also be more robust if there is some skewness. Lehmann (1975) recommends an alignment based on the Hodges—Lehmann estimator of the median difference.

*Computational aspects.* After ranks are allocated any program that computes exact probabilities associated with the permutation distribution for the WMW test may be used if a dedicated program such as that in StatXact is not available.

### 5.8.2 Ansari—Bradley type tests

Several almost equivalent tests based on dispersion of ranks in each sample about the combined sample mean rank were formulated about the same time. If there are  $N = m + n$  observations and these are ranked 1 to  $N$  in ascending order as in the WMW test procedure the mean rank is  $\frac{1}{2}(N + 1)$  and the deviation of rank  $i$  from this mean is  $d_i = i - \frac{1}{2}(N + 1)$ . Clearly the sum of these deviations over all  $N$  ranks is zero. If the spread in one sample is greater than that in the other low and high ranks tend to dominate in that sample. Since the  $d_i$  corresponding to low and high ranks have opposite signs the magnitude of the  $d_i$  in each sample are relevant and these are reflected by taking either absolute values of the  $d_i$  or the squares of the  $d_i$  as scores. The sum of these for one sample is an appropriate score to use in a Pitman type permutation test for a difference in spread under an assumption that the population medians are identical.

Tests that are in essence equivalent and use the sum of the  $|d_i|$  were proposed with minor variations by Freund and Ansari (1957), David and Barton (1958) and by Ansari and Bradley (1960). Gibbons and Chakraborti (1992) refer to them collectively as the Freund—Ansari—Bradley—David—Barton test. While any test program for the Pitman test may be used with the appropriate scores StatXact has a program that calculates scores given the original data and uses the name Ansari—Bradleytest. Mid-ranks are widely used for ties. This does not alter the mean rank but affects some of the scores.

**Table 5.5** Ranks and scores for an Ansari—Bradley test in [Example 5.15](#); type A (adjusted) in **bold**.

Value	47	105	126	<b>130.5</b>	142	<b>153.5</b>	158	172	<b>180.5</b>	<b>183.5</b>
Rank	1	2	3	<b>4</b>	5	<b>6</b>	7	8	<b>9</b>	<b>10</b>
Score	9.5	8.5	7.5	<b>6.5</b>	5.5	<b>4.5</b>	3.5	2.5	<b>1.5</b>	<b>0.5</b>
Value	<b>185.5</b>	197	220	<b>221.5</b>	225	<b>225.5</b>	230	<b>250.5</b>	262	270
Rank	<b>11</b>	12	13	<b>14</b>	15	<b>16</b>	17	<b>18</b>	19	20
Score	<b>0.5</b>	1.5	2.5	<b>3.5</b>	4.5	<b>5.5</b>	6.5	<b>7.5</b>	8.5	9.5

### *Example 5.15*

*The problem.* Apply the Ansari—Bradley test to the data in [Example 5.14](#).

*Formulation and assumptions.* The test assumes the medians of both populations are equal. We align the data as we did in [Example 5.14](#) by subtracting 46.5 from all first sample values and work with the adjusted sample values. The combined sample data are ranked and the  $|d_i|$  are summed over one sample for use in a Pitman type permutation test or a dedicated Ansari—Bradley test program.

*Procedure.* The original data after adjustment and the ranks and the scores  $|d_i|$  are given in [Table 5.5](#). The type A data are shown in bold.

The sum of the scores  $|d_i|$  for type A is  $S = 6.5 + 4.5 + 1.5 + 0.5 + 0.5 + 3.5 + 5.5 + 7.5 = 30$ . Using either a program for a Pitman test with these scores or a dedicated Ansari—Bradley program gives a two-tail  $P = 0.1437$ .

*Conclusion.* There is little evidence of a difference between variances.

*Comments.* 1. The  $P$ -value here is slightly greater than that for the Siegel—Tukey test.

2. The effect of lining up the samples on the basis of sample medians means that the test is not exact in the sense that there is a tendency to give too many small  $P$ -values. Simulation studies have confirmed this tendency although the exact  $P$ -values are approached for very large samples.

3. For ties it is usual to use mid-ranks.

The alternative to using the  $|d_i|$  as scores is to use the squared  $d_i$  in a test proposed by Mood (1954) otherwise analogous to the Ansari—Bradley test and called the Mood test. StatXact provides a program for this test and for the data considered in [Examples 5.14](#) and [5.15](#) the exact test indicates  $P = 0.1293$ . Asymptotic approximations valid for large samples are available for the Ansari—Bradley test and for the Mood test.

### **5.8.3 The Conover squared-rank test for variance**

If the means of  $X, Y$  are respectively  $\mu_x, \mu_y$  then equality of variance implies  $E[(X - \mu_x)^2] = E[(Y - \mu_y)^2]$ , where  $E[X]$  is the expectation of  $X$ .

Conover (1980) proposed a test for equality of variance based on the joint squared ranks of squared deviations from the means, i.e. the latter are  $(x_i - \mu_x)^2 (y_i - \mu_y)^2$ . The population means are seldom known but again we assume it is reasonable to replace them by their sample estimates  $m_x, m_y$ . We do not need to square the deviations to obtain the required rankings because the same order is achieved by ranking the absolute deviations  $|x_i - m_x|, |y_i - m_y|$ . We rank these deviations and use as scores squares of the ranks. Our test statistic  $T$

is the sum of the scores for one of the samples. For large samples  $Z$  given by (5.6) has a standard normal distribution. Conover (1980) gives quantiles of  $T$  for a range of sample sizes in a ‘no-tie’ situation. Programs generating the permutation distribution for arbitrary scores may also be used if exact tail probabilities are required and StatXact provides a specific program for the test.

### *Example 5.16*

*The problem.* For the data in Example 5.14 use the squared-rank test for equality of variance.

*Formulation and assumptions.* We require deviations of each observation from its sample mean. The absolute deviations are then ranked, the ranks squared and the statistic  $T$  is calculated.

*Procedure.* Denoting type A sample values by  $x_i$  and type B sample values by  $y_i$  we find  $m_x = 238$  and  $m_y = 180$  (it suffices to express these to the same order of accuracy as the data, here to the nearest integer). We compute the absolute deviations for each sample value, e.g. for the type A observation 227, the deviation is  $227 - 238 = -11$ , giving an absolute deviation of 11. Table 5.6 gives the ordered absolute deviations for the combined samples together with squared ranks (squared mid-ranks for ties). Type A deviations are in **bold**.

The sum of the bold squared ranks is  $T = 706.5$ . With squared ranks as scores we find from (5.6) that  $\text{Var}(T) = 78.522.49$ . Also (5.4) gives  $E(T) = 1147.6$ , whence from (5.6),  $Z = (706.5 - 1147.6)/280.219 = -1.574$ , but using an exact program such as that in StatXact, if available, is preferable to an asymptotic result. For these data StatXact gives a two-tail  $P = 0.1267$ .

*Conclusion.* The evidence against  $H_0$  is not strong.

*Comments.* 1. Calculating sample means to one decimal place would avoid data ties, but this makes little practical difference to the calculated  $P$ -value although the value of  $T$  may be somewhat different especially if high ranks are involved in ties. For example if ranks 17 and 18 are replaced by a tied value 17.5 and only one of these tied values is included in  $T$  it contributes 306.25 whereas if 17 replaces that tied value it contributes only 289. StatXact avoids ties in this example and gives a  $T = 720$  (compared to our 706.5). If ties occur StatXact in this and some other programs deals with them slightly differently from

**Table 5.6** Absolute deviations (*Adev*) and squared ranks (*Sqr*), Example 5.16; type A in **bold**.

<i>Adev</i>	<b>6</b>	<b>8</b>	8	<b>11</b>	17	22	<b>30</b>	<b>34</b>	<b>38</b>	38
<i>Sqr</i>	<b>1</b>	<b>6.25</b>	6.25	<b>16</b>	25	36	<b>49</b>	<b>64</b>	<b>90.25</b>	90.25
<i>Adev</i>	40	45	50	54	<b>59</b>	<b>61</b>	75	82	90	133
<i>Sqr</i>	121	144	169	196	<b>225</b>	<b>256</b>	289	324	361	400

the way we do when we use mid-ranks. The method is given in the StatXact manual, but in practice the two methods usually lead to similar results.

2. Whereas the Siegel–Tukey test with median alignment would establish significance at the formal 5 per cent level in a one-tail test if that were appropriate, the Conover squared-rank test just fails to do so. While the test appears in this case to be less powerful than the Siegel–Tukey test, it is reasonably robust. The results are broadly in line with those for the Ansari–Bradley test which in fact uses a scoring system not very different from that for the Siegel–Tukey test.

3. We could have used ranks of absolute deviations rather than squared ranks. In that case we essentially have a WMW test (see [Exercise 5.8](#)). Conover (1999, [Section 5.3](#)) reports that this leads to a less powerful test than the squared-rank test.

*Computational aspects.* The Conover scores may be used in any permutation test program for a Pitman-type test if a dedicated program is not available.

#### 5.8.4 The Moses test of extreme reactions

The tests we have described so far have basically considered spread or dispersion as a ‘variance’ type of characteristic. Suppose that we have two samples, A and B, and that Sample A has a smaller range of values than Sample B. One might then wish to investigate whether extreme values (e.g. for reaction or response times) are more likely to occur in the population from which Sample B was drawn.

Moses (1952) proposed a test of extreme reactions. The medians of the two associated populations are assumed to be equal. The null hypothesis is that extreme values are equally likely to occur in both populations, the alternative hypothesis being that extreme values are more likely to occur in the population from which the sample with the larger range was drawn. The two samples are combined, ranked overall and the observations put into rank order. Each observation is labelled either A or B according to the sample from which it originally came.

If both samples come from the same population then each will contain some high, some low and some intermediate values. For instance, if there are seven observations in Sample A and eight in Sample B this might be indicated by the ordering:

B B A A B A B A A B A B B B A

Under the alternative hypothesis, values for Sample A will be concentrated in one part of the series, for example:

B B B A A A A A B A A B B B B

Once the values have been ordered, the ranks of the lowest and highest scores from Sample A are noted. The number of cases contained within these values, including the extremes, is defined as the **span** of the A scores. In the first example above, the span is relatively large (13) whereas in the second case it is quite small (8). The smaller the value of the span, the stronger is the evidence against the null hypothesis. The Moses test determines the probability of observing a span of no more than that recorded, assuming that the null hypothesis is true.

If samples A and B have  $m$  and  $n$  observations respectively, then the span of A must be between  $m$  and  $m + n$  inclusive. Suppose that the observed span is  $m + k$ . Under the null hypothesis, the probability that the span  $s$  is no more than this is given by

$$\Pr(s \leq m + k) = \frac{\sum_{i=0}^k \binom{i+m-2}{i} \binom{n+1-i}{n-i}}{\binom{m+n}{m}} \quad (5.11)$$

Probabilities for the upper tail of the scan length distribution can be calculated in a similar manner. Exact  $P$ -values for the one-tail Moses test can be found using SPSS, although tail probabilities are simple (if tedious) to compute using tabulated values of binomial coefficients.

### *Example 5.17*

*The problem.* Nineteen women who had received results from cervical cancer screening completed a questionnaire about their interest in receiving detailed information about cervical cancer, a high score indicating a great degree of interest. Those women who had received a negative screening result (Group A) were compared with those with positive result (Group B). The questionnaire scores (out of 25) were as follows:

<i>Group A:</i>	8	9	10	12	13	15	18	20	21
<i>Group B:</i>	1	3	5	6	11	14	19	22	23

We test the null hypothesis that extreme values are equally likely to occur in the *positive test* and *negative test* populations of women.

*Formulation and assumptions.* The two groups are combined and the observations ranked. The ranks are written in ascending order. The span of Group A is then obtained by recording the number of cases, including the extremes, contained within the lowest and highest ranks relating to this group. A one-tail test is appropriate, as there are grounds for believing that the scores for women with a positive test result might be more variable (*see Comment 3*).

*Procedure.* The ranks for the combined observations are obtained. These are shown in [Table 5.7](#). Ranks relating to Group A are given in bold.

**Table 5.7** Allocation of ranks for Moses test; Group A in **bold**.

<i>Value</i>	1	3	5	6	<b>8</b>	<b>9</b>	<b>10</b>	11	<b>12</b>	<b>13</b>
<i>Rank</i>	1	2	3	4	<b>5</b>	<b>6</b>	<b>7</b>	8	<b>9</b>	<b>10</b>
<i>Value</i>	14	<b>15</b>	<b>18</b>	19	<b>20</b>	<b>21</b>	22	23	25	
<i>Rank</i>	11	<b>12</b>	<b>13</b>	14	<b>15</b>	<b>16</b>	17	18	19	

We have  $m = 9$ ,  $n = 10$ . The lowest rank for a Group A observations is 5, the highest is 16, giving a span of 12. The minimum possible value for the span is  $m = 9$ , thus  $k = 3$ . Using (5.11) we find  $P = 0.0149$ . The SPSS program confirms this.

*Conclusions.* There is considerable evidence against the null hypothesis. One may reasonably conclude that extreme questionnaire scores occur more frequently for patients with a positive screening result.

*Comments.* 1. The span of the Group A observations is often referred to as the range, and it is a measure highly susceptible to outliers. One way of taking this into account is to decide before the investigation to remove (or ‘trim’) one or more Group A observations from the upper and lower extremes. The span is then based on the values that remain. For the above data, SPSS gives a  $P$ -value of 0.128 when one observation is trimmed from each end. There is little reason to use a trimmed span in this example, as there are no obvious outliers. Trimming when there are outliers in other contexts is discussed briefly in [Section 11.4](#).

2. If ties occur between two or more members of the same group the tied values are adjacent in the ordering and the span is not affected. In addition, if ties occur between members of the two different groups the span will generally not be affected. In the absence of trimming, the only exception is where a tie involves either the highest or lowest observation from Group A; the value for the span is then unclear. In SPSS the span is calculated by assuming that the Group A value in the tie is actually the larger. This can overestimate the value of the span and hence the  $P$ -value. In the above example, if the highest value in Group A is changed to 22, the resulting tie leads to a  $P$ -value 0.04.

3. The findings above are in line with the belief of many psychologists that individuals who receive a positive diagnosis for a serious illness tend to be either ‘blunters’ whose reaction to the bad news is denial, or ‘monitors’ who wish to be informed about every detail regarding their illness. Individuals who do not have to face up to such bad news are assumed to have an intermediate degree of interest in that illness. Thus the ‘blunters’ would have low scores on the cervical cancer questionnaire with the ‘monitors’ having relatively high scores. This could explain why Group B has more extreme values than Group A.

*Computational aspects.* Since  $n = m + 1$ , there is considerable simplification in the numerator of (5.11) for this example.

### 5.8.5 Some related procedures

Klotz (1962) proposed using van der Waerden scores in a test analogous to the Mood test, while Capon (1961) suggested a similar test using expected normal scores. Sukhatme (1957) proposed a test having analogies with the Mann–Whitney formulation of the WMW test but the scoring is more complicated and there are practical limitations to its use. It is described by Gibbons and Chakraborti (1992, Section 10.7) who point out that it can be adapted to construct confidence intervals for a scale parameter.

Gastwirth (1965) proposed statistics for detecting location and scale differences based on simple scores assigned to top and bottom non-overlapping fractions of the combined sample ordered data, with zero scores given to any data with intermediate ranks. The motivation for his approach was that tests for centrality or location based on transformation of ranks such as those using van der Waerden scores that have high Pitman efficiency relative to optimal normal theory tests when the latter are appropriate give greater weights to scores associated with extreme ranks. Gastwirth was especially interested in tests that were more efficient for dispersion than the tests we have discussed here. A general account of his approach is given in Sprent (1998, Section 6.6).

## 5.9 TESTS FOR A COMMON DISTRIBUTION

In [Section 3.3](#) we developed tests of the hypothesis that a single sample was drawn from some specified distribution. For two independent samples we may want to know if it is reasonable to suppose each comes from the same unspecified distribution. The Smirnov test (Smirnov, 1939, 1948) has similarities to the Kolmogorov test developed in [Section 3.3.1](#).

### 5.9.1 Smirnov test for a common distribution

The hypothesis  $H_0$ : *two samples come from the same distribution* may be tested against  $H_1$ : *the distributions have different cumulative distribution functions (cdfs)*. We do not specify further the nature of any differences: the distributions may have the same mean but different variances; one may be skew, the other symmetric, etc. We compare the sample cdfs; the test statistic is the difference of greatest magnitude between these two functions.

*Example 5.18*

*The problem.* Use the Smirnov test to decide if it is reasonable to conclude that the samples in [Example 5.14](#) are from the same populations.

*Formulation and assumptions.* We compute the sample cumulative distribution functions  $S_1(x)$ ,  $S_2(y)$  at each sample value and at each of these we also compute and record the difference  $S_1(x) - S_2(y)$ . For samples of size  $m$ ,  $n$  respectively  $S_1(x)$ ,  $S_2(y)$  are step functions with steps  $1/m$ ,  $1/n$  respectively at each sample value (or with multiple steps at ties).

*Procedure.* [Table 5.8](#) gives in successive columns the sample values and corresponding values of  $S_1(x)$ ,  $S_2(y)$  and  $S_1(x) - S_2(y)$  at each sample point. The difference of greatest magnitude (final column) is 0.5, occurring twice. Here a one-tail test is essentially a test of whether the function for type B failures is above that for type A failures for at least one common  $x$ ,  $y$  value, against the alternative that it is everywhere at or below. However, a two-tail test is appropriate for unspecified general alternatives. An exact test is preferable if a suitable program is available. One is provided in StatXact and in this example gives an exact one-tail  $P = 0.0748$  and an exact two-tail  $P = 0.1496$ .

**Table 5.8** Calculation of the Smirnov test statistic,

Type A $x_i$	Type B $y_j$	$S_1(x)$	$S_2(y)$	$S_1(x) - S_2(y)$
47	0	0.083		-0.083
105	0	0.167		-0.167
126	0	0.250		-0.250
142	0	0.333		-0.333
158	0	0.416		-0.416
172	0	0.500		-0.500
177		0.125	0.500	-0.375
	192	0.125	0.583	-0.458
200		0.250	0.583	-0.333
	220	0.250	0.667	-0.417
	225	0.250	0.750	-0.500
227		0.375	0.750	-0.375
230	230	0.500	0.833	-0.333
232		0.625	0.833	-0.208
	262	0.625	0.917	-0.292
268		0.750	0.917	-0.167
	270	0.750	1.000	-0.250
272		0.875	1.000	-0.125
297		1.000	1.000	0.000

*Conclusion.* There is a slight indication that the cumulative distribution for type B failures may be strictly above that for type A for at least some common  $x$ ,  $y$ .

*Comments.* 1. Sometimes a test for location or variance establishes a difference whereas a Smirnov test indicates no overall distributional difference. This is because Kolmogorov–Smirnov type tests are often less powerful than tests for specific characteristics such as differences in location.

2. Like the Kolmogorov test, the Smirnov test may appear not to be making full use of the data because it uses only the maximum difference. As we pointed out in [Section 3.3.1](#), the statistic uses cumulative information. In [Section 5.9.2](#) we see that power may be increased by considering all differences.

3. Tables of nominal significance levels may be used if no program for an exact test is available. Asymptotic approximations are available for large samples. Many published tables for the Smirnov test give not critical values but quantiles which must be exceeded for significance at a given level. Neave (1981, p. 31) gives values for the equivalent  $mn[\max|S_1(x) - S_2(y)|]$  for significance for a wide range of sample sizes.

4. The rationale behind this test is sketched by Sprent (1998, Example 6.14).

*Computational aspects.* Several standard packages give programs for the Smirnov test (often referred to as the Kolmogorov–Smirnov test), but care should be taken to see whether each test uses only an asymptotic approximation or a more appropriate approach for small samples.

Chandra, Singpurwalla and Stephens (1981) developed a test of the Kolmogorov–Smirnov type for the Weibull distribution, often met with survival data. Modifications of the Smirnov test for discrete data are discussed by Eplett (1982), while Saunders and Laud (1980) devised a multivariate Kolmogorov goodness-of-fit test.

## 5.9.2 The Cramér–von Mises test for identical populations

The differences  $S_1(x) - S_2(y)$  at each sample value are not independent; this makes it difficult to work out the distribution for statistics that take account of all differences. However, for one such test, the Cramér–von Mises test (Cramér, 1928; von Mises, 1931), an approximate theory gives simple significance tests that, except for very small samples, are virtually independent of sample size, and these are useful if one only wants broad measures of evidence for or against identity rather than more precise  $P$ -values.

The test statistic is a function of the sum of squares of the differences  $S_1(x) - S_2(y)$  at all sample points. Denoting this sum of squares by  $S_d^2$  the test statistic is  $T = mnS_d^2/(m + n)^2$ . In a two-tail test for significance at the 5 per cent level  $T$  must exceed 0.461; for significance at the 1 per cent level  $T$  must exceed 0.743.

### *Example 5.19*

*The problem.* Perform the Cramér–Von Mises test for the data in [Example 5.18](#).

*Formulation and assumptions.* We square and add the differences in the last column of [Table 5.8](#), then form the statistic  $T$ .

*Procedure.* From the last column of [Table 5.8](#) we find the sum of squares of the differences is 2.0024, whence  $T = 96 \times 2.0024 / (20 \times 20) = 0.48$ .

*Conclusion.* Since  $T > 0.461$  we conclude that the population cumulative distribution functions differ for at least some  $x$ .

*Comment.* The Cramér–Von Mises test is often more powerful than the Smirnov test and is easy to use because of the approximation. Accurate tables exist for some values of  $m, n$ . The only labour additional to that for the Smirnov test is calculation of the sums of squares of differences.

### **5.9.3 The Wald–Wolfowitz run test**

Wald and Wolfowitz (1940) proposed a test based on runs for a population difference when we are given two independent samples. We omit details here but the test is described briefly and an example given in Sprent (1998, Section 6.9) and StatXact provides a dedicated program for an exact test. Difficulties arise if there are tied values in different samples when the test is often less satisfactory than the Smirnov test. For the data used in [Examples 5.18](#) and [5.19](#) a program in StatXact for this test gives the exact two-sided  $P = 0.8174$ . Even though there are no problems with ties in this example the result is so much at variance with that for the Smirnov test as to reflect the often low power of this test.

### **5.9.4 The two sample runs test on a circle**

Suppose we have two independent samples of angular data, A and B. In this test, the points are plotted on a circle and labelled A and B according to the sample from which they came. The number of runs around the perimeter of the circle is noted. In proceeding around the circle once, the final run and initial run must be from different groups. The number of runs has therefore to be even (if the extreme points are from the same group when arranged on a line, the number of runs is reduced by one when the points are arranged on a circle). Where there is a tendency for points from the same group to cluster, the number of runs is relatively small; for alternation the number of runs is relatively large (see [Section 3.5](#)).

If  $R$  and  $R_C$  represent the number of runs when the data are arranged on a line and on a circle respectively, then using the notation of [Section 3.5](#) the lower tail probabilities for the number of runs on the circle,  $2s$ , are given by:

$$\Pr(R_C \leq 2s) = \Pr(R \leq 2s+1)$$

with upper tail probabilities calculated in a similar manner.

### *Example 5.20*

*The problem.* The midwife in [Example 3.13](#) also recorded the sex of the 12 home deliveries. She was interested in whether or not the times of births for boys and girls occurred in clusters. The data were as follows (M = male, F = female)

Time	0100	0300	0420	0500	0540	0620	0640	0700	0940	1100	1200	1720
Sex	M	F	F	F	M	M	F	F	F	F	M	M

*Formulation and assumptions.* We test the null hypothesis  $H_0$  that for the populations of male and female births the times of delivery are randomly ordered with respect to gender. Since the midwife is interested only in clustering and not in alternation, a one-tail test is appropriate. A small value for the number of runs,  $2s$ , indicates possible clustering.

*Procedure.* The data are plotted on a circle. The numbers of boys and girls are  $m = 5$  and  $n = 7$  respectively. The count of runs starts from the beginning of a particular run. We therefore count runs from the female birth that occurred at 0300. The number of runs,  $2s$ , is equal to four. As reasoned above, we could not have had five runs on a circle for these two groups. The relevant  $P$ -value for the angular case is therefore given by  $\Pr(R \leq 5)$  where  $R$  is the number of runs for such data arranged along a line. This  $P$ -value is 0.197.

*Conclusion.* There is no strong evidence against  $H_0$ . On the basis of these data, the times of delivery for boys and girls do not seem to cluster with respect to gender.

*Comment.* Recall that if the number of runs for the data on a line is odd, the number of runs is one fewer when the data are plotted on a circle. This implies that low  $P$ -values are only obtained from small studies in the most extreme situations. For instance, unless both  $m$  and  $n$  are at least eight there will only be strong evidence against  $H_0$  (formally  $P < 0.05$ ) if the number of runs is two.

## 5.10 FIELDS OF APPLICATION

Little imagination is needed to think of realistic situations where one might wish to compare medians or means of two populations (i.e. look for centrality or ‘treatment’ differences) on the basis of two independent samples. Here are a few relevant situations.

## **Medicine**

For comparing the efficacy of two drugs for reducing hypertension, blood cholesterol levels, relief of headaches or other conditions, independent samples are often needed because of ‘interaction’ or ‘hangover’ effects if each drug is given to the same patients ([Section 4.1](#)); for this or ethical reasons it may be inappropriate to give both drugs to any one person even after a considerable time lapse.

## **Sociology**

To explore the possibility that town and country children may attain different levels of physical fitness, samples of each might be scored on an appropriate scale in a fitness test and the results compared nonparametrically.

## **Mineral exploration**

A mining company has options on two sites but only wishes to develop one. Sample test borings are taken at each and the percentage of the mineral of interest in each boring is determined; these are the basis for a test for population differences in mean or median levels; if there is evidence that one site is richer the company may want to estimate the difference because development costs and other factors may well differ between sites. This would call for confidence intervals and power considerations may come into play to determine whether a larger experiment is need to provide a sound basis for a decision. Results may need to be interpreted with caution because borings close together may not be independent.

## **Manufacturing**

New and cheaper production methods are often tried. Manufacturers may compare products using a new process or raw material with existing ones to assess quality and durability. Interest here is often not only in ‘average’ quality, but also in differences in variability.

## **Psychology**

Children with learning difficulties may be given a treatment that it is hoped will encourage them to respond to commands. Sixty commands are given to a sample of 10 treated children and to a further sample of 12 untreated children. The number of favourable

responses is recorded for each child. Interest will lie in whether there is a response level difference and a confidence interval for any response shift is likely to be informative.

## 5.11 SUMMARY

**Wilcoxon–Mann–Whitney test.** A rank test for differences in centrality or domination of one distribution over the other. The test statistic is the rank sum associated with either sample. An alternative formulation was given by Mann and Whitney. The two formulations are described in Sections 5.2.2 and 5.2.3. Adjustments for ties are given in Section 5.2.4 and confidence intervals for measurement data are obtained in Section 5.2.5. Asymptotic results are given by (5.7) and (5.8) in Section 5.6.

**The median test.** A test for centrality based on numbers in each of two samples above and below the combined sample median. It is equivalent (Section 5.3) to a case of the more general Fisher exact test. Confidence intervals (Section 5.3.2) are available when we have full measurement data. The only assumption about the population distributions is that of identical medians under  $H_0$ . The test is not recommended if assumptions allow more powerful tests such as the WMW test.

**Van der Waerden scores.** Among tests based on transformation of ranks that using van der Waerden scores (Section 5.4) is easy to use, but results are often similar to those given by the WMW test (Section 5.4).

**Survival data.** The long-tail distributions associated with much survival data, together with censoring has triggered the development of many special methods. These include the **Gehan–Wilcoxon test** (Section 5.5.1) which is a modification of the WMW test. Savage scores and the **log-rank transformation** (Section 5.5.2) provide another approach.

**Power and sample size.** Studies of power and sample size are reasonably straightforward only for the median test providing samples of the same size are taken from distributions that differ only in mean or median, a situation where that test is often not the most appropriate! In similar situations and some that are slightly more

general reasonable asymptotic approximations are available for the WMW test. A brief introduction is given in [Section 5.7](#).

**Spread or dispersion.** The **Siegel–Tukey test** ([Section 5.8.1](#)) and **Ansari–Bradley type tests** ([Section 5.8.2](#)) are often less powerful and less robust than the **Conover squared-rank test** ([Section 5.8.3](#)) or a similar test using absolute deviations. The **Moses test** ([Section 5.8.4](#)) for extreme reactions is relatively simple, but being based on a range of ranks it is susceptible to outliers. Many dispersion tests require for strict validity that both populations medians or means are the same or that the difference, if any, between them is known.

**Comparison of distributions.** The **Smirnov test** ([Section 5.9.1](#)) has analogies with the one-sample Kolmogorov test ([Section 3.3.1](#)). The alternative **Cramér–Von Mises test** ([Section 5.9.2](#)) is easy to use in formal hypothesis tests as the significance levels are virtually independent of sample size except for very small samples. A less satisfactory test is the **Wald–Wolfowitz run test** briefly referred to in [Section 5.9.3](#). A two-sample runs test can be applied to angular data ([Section 5.9.4](#)).

## EXERCISES

- 5.1 For the data in [Example 5.5](#) carry out the procedures suggested in the remarks under *Computational aspects* using the most appropriate computer software available to determine if the end points of the confidence interval can validly be included in a nominal 95 per cent confidence interval.
- 5.2 Use (5.5) to compute the exact variance using mid-ranks in [Example 5.10](#) and compare it with the result given in that example.
- 5.3 Using (5.3) calculate  $P^*$  for the data in (5.2).
- 5.4 In [Example 5.7](#) verify that we accept  $H_0$  at the end points of the nominal 95 per cent confidence interval computed in that example for the population median difference. Verify also that for any point outside that interval we would reject  $H_0$ .
- 5.5 An alloy is composed of zinc, copper and tin. It may be made at one of two temperatures H (higher) or L (lower). We wish to know if one temperature produces a harder alloy. A sample is taken from each of 9 batches at L and 7 at H. To arrange them in ascending order of hardness, all specimens are scraped against one another to see which makes a deeper scratch (a deeper scratch indicates a softer specimen). On this basis the specimens are ranked 1 (softest) to 16 (hardest) with the results given below. Should we reject the hypothesis that hardness is unaffected by temperature? State any assumptions needed for validity of the test you use.

<i>Temperature</i>	H	L	H	H	H	L	H	L	L	H	H	L	L	L	L	L
<i>Rank</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

- 5.6 If we have samples of  $m, n$  where  $m + n = 20$  and the only tie is one between ranks 2 and 3 what would be the differences between the van der Waerden scores if the scores for these ties are based on the mid-rank 2.5 and the van der Waerden scores if these are based on the mean of the van der Waerden scores corresponding to rank 2 and to rank 3. Carry out a similar comparison if there were ties for ranks 10, 11 and 12.
- 5.7 Perform a Siegel–Tukey test on the data in [Example 5.14](#) after shifting one set of sample values to align the sample means.
- 5.8 Using the data in [Example 5.14](#) carry out a test analogous to the Conover squared rank test but using absolute ranks instead of squared ranks.
- 5.9 Hotpot stoves use a standard oven insulation. To test its effectiveness they take random samples from the production line and heat the ovens selected to 400°C, noting the time taken to cool to 350°C after switching off. For a sample of 8 ovens the times in minutes are:

15.7 14.8 14.2 16.1 15.3 13.9 17.2 14.9

They decide to explore a cheaper insulation, and using this on a sample of 9 the times taken for the same temperature drop are:

13.7 14.1 14.7 15.4 15.6 14.4 12.9 15.1 14.0

Are the firm justified in asserting there is no real evidence of a different rate of heat loss? Obtain a 95 per cent confidence limit for the difference in median heat loss (a) with and (b) without a normality assumption. Comment critically on any differences between your conclusions.

- 5.10 A psychologist wants to know whether men or women are more upset by delays in being admitted to hospital for routine surgery. He devises an anxiety index measured on patients 1 week before scheduled admission and records it for 17 men and 23 women. These are ranked 1 to 40 on a scale of increasing anxiety. The sum of the ranks for the 17 men is 428. Is there evidence that anxiety is sex-dependent? If there is, which sex appears to show the greater anxiety?
- 5.11 Suppose we are given the data for response times in LVF and RVF in [Table 4.1](#), but the information that they are paired is omitted. In these circumstances we might analyze them as independent samples. Would we then conclude that responses in the two fields differed? Does your conclusion agree with that found in [Example 4.1](#)? If not, why not?
- 5.12 Confirm the value of  $p_1$  under  $H_1$  for the double exponential distribution introduced in [Example 5.12](#) that is quoted in the *Procedure* section of that example.
- 5.13 Verify the correctness of the required sample sizes given in [Example 5.13](#) for the WMW test in the circumstances given there.
- 5.14 A psychologist notes total time (in seconds) needed to perform a series of simple manual tasks for each of eight children with learning difficulties and seven children without learning difficulties. The times are:

---

<i>Without difficulties</i>	204	218	197	183	227	233	191
<i>With difficulties</i>	243	228	261	202	343	242	220

---

Use a Smirnov test to find whether the psychologist is justified in asserting these samples are likely to be from different populations. Do you consider a one- or a two-tail test appropriate? Perform also a Cramér–Von Mises test. Does it lead you to a different conclusion? If you think the psychologist should have tested more specific aspects of any difference, perform the appropriate tests.

- 5.15 Apply the Smirnov test for different population distributions to the oven cooling data in [Exercise 5.9](#).
- 5.16 The numbers of words in the first complete sentence on each of 10 pages selected at random is counted in each of the books by Conover (1980) and Bradley (1968). The results were:

---

<i>Conover</i>	21	20	17	25	29	21	32	18	32	31
<i>Bradley</i>	45	14	13	31	35	20	58	41	64	25

---

Perform tests to determine whether there is evidence that in these books

- (i) sentence lengths show a difference in centrality;
  - (ii) the variances of sentence lengths differ between authors;
  - (iii) the distributions of sentence lengths differ in an unspecified way;
  - (iv) the sentence lengths for either author are not normally distributed.
- 5.17 Lindsey, Herzberg and Watts (1987) give data for widths of first joint of the second tarsus for two species of the insect *Chaetocnema*. Do these indicate population differences between the width distributions for the two species?

---

<i>Species A</i>	131	134	137	127	128	118	134	129	131	115
<i>Species B</i>	107	122	144	131	108	118	122	127	125	124

---

- 5.18 Carter and Hubert (1985) give data for percentage variation in blood sugar over 1-hour periods for rabbits given two different dose levels of a drug. Is there evidence of a response difference between levels?

---

<i>Dose I</i>	0.21	-16.20	-10.10	-8.67	-11.13	1.96	-10.19	-15.87	-12.81
<i>Dose II</i>	1.59	2.66	-6.27	-2.32	-10.87	7.23	-3.76	3.02	15.01

---

- 5.19 The journal *Biometrics* published data on the numbers of completed months between receipt of a manuscript for publication and the first reply to the authors for each of the years 1979 and 1983. The data are summarized below. Is there evidence of a difference in average waiting times between 1979 and 1983?

<i>Completed months</i>	<i>0</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	$\geq 6$	
<i>Number of authors</i>	1979	26	28	34	48	21	22	34
	1983	28	27	42	44	17	6	16

- 5.20 Hill and Padmanabhan (1984) give body weights (g) of diabetic and normal mice. Is there evidence of a significant difference in mean body weight? Obtain the Hodges–Lehmann estimate of the difference together with a 95 per cent confidence interval. Compare this interval with that based on the *t*-distribution.

<i>Diabetic</i>	42	44	38	52	48	46	34	44	38
<i>Normal</i>	34	43	35	33	34	26	30	31	27
	32	32	36	32	32	38	42	36	44
							44	33	38

- 5.21 The data below are numbers of words with various numbers of letters in 200-word sample passages from the presidential addresses to the Royal Statistical Society by W.F. Bodmer (1985) and J. Durbin (1987). Is there acceptable evidence of a difference between the average lengths of words used by the two presidents?

<i>Number of letters</i>	<i>1–3</i>	<i>4–6</i>	<i>7–9</i>	<i>10 or more</i>
<i>Bodmer</i>	91	61	24	24
<i>Durbin</i>	87	49	32	32

- 5.22 The following data are DMF scores for 34 male and 54 female first-year dental students. The DMF score is the total of the numbers of decayed + missing + filled teeth.

<i>Males</i>	8	6	4	2	10	5	6	6	19	4	10	4	10	12	7	2	5
	1	8	2	0	7	6	4	4	11	2	16	8	7	8	4	0	2
<i>Females</i>	4	7	13	4	8	8	4	14	5	6	4	12	9	9	9	8	12
	4	8	8	4	11	6	15	9	8	14	9	8	9	7	12	11	7
	4	10	7	8	8	7	9	10	16	14	15	10	4	6	3	9	3
	10	3	8														

Use an asymptotic WMW test to determine whether the DMF score differs significantly between males and females. Do tied ranks have much influence on the appropriate test statistic?

- 5.23 Records from a maternity hospital show that during a particular year, identical and non-identical twins were born on the following days:

---

<i>Identical</i>	3 January, 17 February, 14 April, 31 May, 6 August, 2 October, 7 December
<i>Non-identical</i>	30 January, 28 February, 5 March, 13 April, 2 June, 29 June, 8 July, 17 July, 4 August, 17 September, 9 October, 26 October, 17 November, 11 December

---

Is there evidence of possible clustering of births of identical and non-identical twins?

# 6

## Three or more samples

### 6.1 COMPARISONS WITH PARAMETRIC METHODS

Parametric one- or two-sample analyses of continuous data using test and estimation procedures based on normality are modified appreciably for three or more samples. Emphasis shifts from the  $t$ -distribution and related test and estimation procedures to the analysis of variance (ANOVA) and tests based on the  $F$ -distribution (although the  $t$ -distribution is related to a particular  $F$ -distribution). The concept of experimental design becomes more important. Readers unfamiliar with these topics need only to skim through this section briefly, or they may prefer to proceed directly to [Section 6.2](#).

Given  $k$  independent samples from normal distributions all with the same (not necessarily known) variance and means  $\mu_1, \mu_2, \dots, \mu_k$ , the basic overall significance test is that of  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  against  $H_1: \text{not all } \mu_i \text{ are equal}$ . In analysis of variance terminology large values of the statistic

$$F = (\text{between samples mean square}/\text{within samples mean square})$$

indicate evidence against  $H_0$ . Under  $H_0$ ,  $F$  has an  $F$ -distribution with  $k - 1$  and  $N - k$  degrees of freedom, where  $N$  is the total number of observations. When it is appropriate,  $H_0$  is sometimes expressed as the hypothesis of *no difference between treatments*.

The above  $F$ -test is usually a preliminary to more specific tests and estimation procedures concerning possible differences within chosen subsets, sometimes (but by no means always) pairs of the  $\mu_i$ ,  $i = 1, 2, \dots, k$ . New concepts such as least significant differences or multiple comparison tests are then sometimes introduced even in the independent sample or ‘one-way classification’ situation.

The two-dependent-samples situation in Chapter 4 generalizes to designed experiments. For the overall parametric test of  $H_0: \text{no difference between treatments}$  against  $H_1: \text{some treatments differ}$  (in centrality) testing and estimation procedures now depend on the experimental design. A well-known design is that of randomized

blocks. The nature of the treatment structure also influences the analysis; factorial experiments not only allow us to study the basic effect of two or more factors, but also whether factors ‘interact’ with one another.

In normal theory inference **linear models** provide a framework that links analysis of variance and linear regression analysis.

Nonparametric analyses parallel some aspects of the normal theory linear model, but relaxing assumptions means that some linear model techniques have no direct nonparametric analogue.

We develop first overall tests for treatment differences analogous to those used in the analysis of variance. These generalize some of the methods in [Chapters 4 and 5](#). Detailed testing of certain aspects of data subsets is deferred to [Section 6.4](#).

Having more than two samples often means there are sufficiently many observations for asymptotic results to be reasonable but this may not be so if some samples are very small even when that total number of data is large. Over-reliance on asymptotic results when some samples are small has in the past reflected a paucity of tables giving even nominal significance levels for permutation tests based on ranks, etc. Programs like StatXact and Testimate have eased this problem.

In this and the following four chapters we shall find that the same nonparametric technique may often be applied to problems that appear at first sight to be different. We have already met examples of such equivalences; e.g. the sign test and McNemar’s test ([Section 4.2](#)), also that between Wilcoxon’s signed-rank test in a specific highly tied situation and the sign test ([Exercise 2.6](#)). We also introduced the median test ([Section 5.3](#)) as a special case of the widely applicable Fisher exact test ([Section 9.2.1](#)).

## 6.2 CENTRALITY TESTS FOR INDEPENDENT SAMPLES

### 6.2.1 The Pitman permutation test

A permutation test based on raw data that extends the two-independent sample test mentioned briefly in [Section 5.1.2](#) was developed by Pitman (1938). For reasons similar to those given in that section the test is seldom used in practice. Exact  $P$ -values for the test corresponding to the overall  $F$ -test in the one-way classification ANOVA mentioned in [Section 6.1](#) are only obtainable with present software for fairly small samples so one has to resort to

Monte Carlo approximations or asymptotic results if the samples are not all small. We omit details here but the procedure is outlined together with an example in Sprent (1998, Section 7.2). StatXact includes a program under the title *One-way ANOVA with general scores* that may be used for the test by taking the raw data as scores.

### 6.2.2 The Kruskal–Wallis test

Kruskal and Wallis (1952) extended the WMW test using the Wilcoxon formulation to three or more samples. As well as being applicable to data consisting only of ranks it is relevant as an overall test for equality of population means or medians when samples are from otherwise identical and continuous distributions. It also has reasonable power for testing for identical population cumulative distribution functions against an alternative that one or more of these cumulative distributions are distinct in the sense of dominating the others. As in the two-sample case, a shift in mean or median is often referred to as an *additive treatment effect*, a term used in parametric (normal theory) linear models.

Suppose we have  $k$  random samples, the  $i$ th sample ( $i = 1, 2, \dots, k$ ) consisting of  $n_i$  observations, the  $j$ th of these being  $x_{ij}$ , ( $j = 1, 2, \dots, n_i$ ). The total number of observations is  $N = \sum_i n_i$ . We assume there are no ties and rank the  $N$  observations from smallest (rank 1) to largest (rank  $N$ ). Let  $r_{ij}$  be the rank allotted to  $x_{ij}$  and  $s_i = \sum_j r_{ij}$  be the sum of the ranks for the  $i$ th sample. We compute  $S_k = \sum_i (s_i^2/n_i)$ . The test statistic is then

$$T = \frac{12S_k}{N(N+1)} - 3(N+1) \quad (6.1)$$

If all the samples are from the same population we expect a mixture of small, medium and high ranks in each sample, whereas under the alternative hypothesis high (or low) ranks may dominate in one or more samples. Consequently, under the alternative hypothesis  $S_k$  will contain the square of at least one relatively large rank sum, leading to larger values of  $T$ .

Computing exact permutation probabilities under  $H_0$  by hand is impractical except for small samples. For  $N$  moderate or large,  $T$  has a chi-squared distribution with  $k - 1$  degrees of freedom under  $H_0$ .

#### *Example 6.1*

*The problem.* An estate agency sells properties in villages  $A$ ,  $B$  and  $C$ . The homes for sale in these villages are at the following prices (£K):

<i>A</i>	39	45	71	
<i>B</i>	51	63	88	97
<i>C</i>	99	150	260	

Use the Kruskal–Wallis test to examine the validity of the hypothesis that the house prices in these samples come from identical populations.

*Formulation and assumptions.* A suitable null hypothesis is  $H_0$ : *the populations from which the samples were drawn have identical medians* while the alternative is  $H_1$ : *the population medians are not all equal*.

*Procedure.* The  $N$  ( $= 10$ ) house prices are ranked overall and the ranks are assigned to the appropriate groups. These are shown in Table 6.1.

The sum of ranks is  $1 + 2 + 5 = 8$  for the houses in Village *A*,  $3 + 4 + 6 + 7 = 20$  for Village *B* and  $8 + 9 + 10 = 27$  for Village *C*, whence  $S_k = (8)^2/3 + (20)^2/4 + (27)^2/3 = 364.33$ . Hence (6.1) gives

$$T = \frac{123 \leftrightarrow 6433}{10 \leftrightarrow 11} - 3 \leftrightarrow 11 = 6745$$

corresponding to an exact  $P = 0.010$ .

*Conclusion.* There is strong evidence that house prices differ between villages.

*Comments.* 1. Properties in Village *C* seem to be more expensive. This may be because the village has more ‘executive-style’ homes. One could only make an informed comparison of house prices in the three villages by looking at similar types of property. Remember too that properties offered for sale may not be representative of the local housing stock. We have a small non-random sample which may not be representative of the ‘population’ of village properties.

2. We could use  $S_k$  as a test statistic rather than  $T$  since all other quantities in  $T$  are invariant in the exact permutation distribution (i.e. are the same for any permutation). The use of  $T$  is preferred for practical reasons (e.g. availability of tables and for asymptotic approximations).

3. Calculating the  $P$ -value by hand is tedious. StatXact computes an exact  $P$  directly from the above data. Tables giving critical values of  $T$  for small sample sizes (strictly relevant to the ‘no-tie’ situation) are given by Neave (1981, pp. 32–34) while Hollander and Wolfe (1999, Table A12) give upper-tail  $P$ -values for large  $T$  for a range of sample sizes. Good estimates of exact probabilities are also available using the Monte Carlo facility in StatXact. Using Stata or other general statistical packages the chi-squared approximation with  $T = 6.745$  gives

**Table 6.1** House prices (£K) and ranks in three villages.

<i>Village A</i>	<i>Price</i>	39	45	71	
	<i>Rank</i>	1	2	5	
<i>Village B</i>	<i>Price</i>	51	63	88	97
	<i>Rank</i>	3	4	6	7
<i>Village C</i>	<i>Price</i>	99	150	260	
	<i>Rank</i>	8	9	10	

$P = 0.0343$ , considerably different from the exact value indicating, not surprisingly, that the asymptotic result can be misleading for small samples.

4. The test is scale invariant and exactly the same ranks and  $T$  value would be obtained if the prices were converted to US\$, euros or any other currency at the ruling exchange rate.

### 6.2.3 The Kruskal–Wallis test with ties

As in the WMW test we give mid-rank values to tied data. With ties a more general formula for  $T$  is needed. In addition to the terms calculated above, we need also to calculate  $S_r = \sum_{i,j} r_{ij}^2$ , where some of the  $r_{ij}$  will now be mid-ranks. Readers familiar with the analysis of variance will recognize  $S_k$  and  $S_r$  as uncorrected **treatment** and **total** sums of squares for ranks. An appropriate correction for the mean is subtracted from each, namely  $C = \frac{1}{4}N(N+1)^2$ . The test statistic is

$$T = \frac{(N-1)(S_k - C)}{S_r - C} \quad (6.2)$$

If there are no ties, this can be shown to be equivalent to (6.1), since then  $S_r = N(N+1)(2N+1)/6$ .

#### Example 6.2

*The problem.* Uniform editions by each of three writers of detective fiction are selected. The numbers of sentences per page, on randomly selected pages in a work by each are

<i>C.E. Vulliamy</i>	13	27	26	22	26
<i>Ellery Queen</i>	43	35	47	32	31
<i>Helen McCloy</i>	33	37	33	26	44
	33	37	33	26	44
	33	37	33	26	44
	33	37	33	26	44

Use the Kruskal–Wallis test to examine the validity of the hypothesis that these may be samples from identical populations.

*Formulation and assumptions.* The null hypothesis is that the populations are identical. We consider the alternative hypothesis that the samples are from populations that are not identically located.

*Procedure.* Manual calculation of the required statistic is outlined for illustrative purposes, but one only need do this if no program is available to compute an exact  $P$ -value. Ranks are shown in Table 6.2. From that table we find that  $s_1 = 1 + 2 + 4 + 4 + 6 = 17$ ,  $s_2 = 72.5$  and  $s_3 = 81.5$ , whence

$$S_k = (17)^2/5 + (72.5)^2/6 + (81.5)^2/7 = 1882.73.$$

The sum of squares of all the allocated ranks is  $S_r = 2104.5$  and  $C = 18 \times (19)^2/4 = 1624.5$  since  $N = 18$ . We compute

$$T = 17(1882.73 - 1624.5)/(2104.5 - 1624.5) = 9.146.$$

**Table 6.2** Sentences per page and ranks in samples for three authors.

	<i>Number</i>	13	22	26	26	27		
		<i>Rank</i>	1	2	4	4	6	
<i>Queen</i>	<i>Number</i>	31	32	35	37	43	47	
		<i>Rank</i>	7	8	12	13.5	15	17
<i>McCloy</i>	<i>Number</i>	26	33	33	33	37	44	54
		<i>Rank</i>	4	10	10	10	13.5	16

Neave's tables show that  $T \geq 8.157$  implies strong evidence against  $H_0$  (significant at the 1 per cent level). StatXact confirms the above value of  $T$  and gives an exact  $P = 0.0047$ .

*Conclusion.* There is strong evidence to support a difference in sentence length. Inspection of the data indicates that Vulliamy uses longer sentences (fewer sentences per page).

*Comments.* 1. Using uniform editions avoids difficulties that may arise with different page sizes or type fonts.

2. The total number of observations is small ( $N = 18$ ). However, if we perform an asymptotic test  $T = 9.146$  is just short of the chi-squared value 9.21 required for significance at the 1 per cent level. Indeed  $P = 0.0103$  for this chi-squared approximation.

*Computational aspects.* Monte Carlo estimates are provided by StatXact and these may be useful for sample sizes a little larger than those in this example. The computational burden for exact  $P$ -values rapidly increases as  $N$  increases. Many standard packages such as Minitab and Stata include the Kruskal–Wallis test, usually computing  $T$  but giving only an asymptotic  $P$ -value.

For the above example most statisticians would consider a parametric analysis of variance to be suitable. This leads to basically similar conclusions ([Exercise 6.1](#)).

Boos (1986) developed more comprehensive tests for  $k$  samples using linear rank statistics that test specifically for location, scale, skewness and kurtosis. Willemain (1980) tackles the interesting problem of estimating the mean of a population given only the largest observation in each of  $n$  samples. This scenario may arise if each of  $n$  consultants is given one dose of a new or expensive drug and asked to give this to the patient of greatest need on his or her list. A record may be taken of the change, say, in some level of a blood constituent and the aim may be to estimate the mean or median change that could be expected if the drug were used for all patients suffering from the relevant disease.

Shirley (1977) proposed a nonparametric test to determine the lowest dose level at which there is a detectable treatment effect compared with the level for untreated control subjects. An improved version of this test was given by Williams (1986) and House (1986) extended the idea to a randomized block design.

#### 6.2.4 Tests based on transformation of ranks

A modification of the Kruskal—Wallis test is to use van der Waerden scores in place of ranks. We replace the rank or mid-rank  $r$  by the  $r/(N+1)$ th quantile of a standard normal distribution. Ignoring a slight discrepancy in the case of ties, the mean of these scores is zero. If we evaluate quantities corresponding to  $S_k$  and  $S_r$  in [Section 6.2.2](#) with ranks replaced by van der Waerden scores and set  $C = 0$ , we may compute a  $T$  analogous to (6.2) for the van der Waerden scores. Again, asymptotically  $T$  has a chi-squared distribution with  $p - 1$  degrees of freedom. The asymptotic result is usually reasonable for all but very small samples although some packages including StatXact now include a program giving exact  $P$ -values for samples that are not too large and provide Monte Carlo estimates for larger samples. The exact  $P$ -value given by StatXact using van der Waerden scores for the data in [Example 6.2](#) is again 0.047. In [Exercise 6.2](#) we ask you to reanalyze the data in [Example 6.2](#) using van der Waerden scores with whatever programming facilities are available. Expected normal scores may also be used.

Other transformations are possible, e.g. to Savage scores if these are appropriate. There is no call for their use in [Examples 6.1](#) or [6.2](#) since there is little evidence of asymmetry with a long right tail.

#### 6.2.5 The Jonckheere—Terpstra test

The Kruskal—Wallis test is an omnibus test for differences in centrality or for differences of a dominance nature. If treatments represent, for example, steadily increasing doses of a stimulant we may want to test hypotheses about means or medians,  $\theta_i$ , of the form  $H_0: \text{all } \theta_i \text{ are equal}$  against  $H_1: \theta_1 \leq \theta_2 \leq \theta_3 \leq \dots \leq \theta_k$ , where at least one of the inequalities is strict, or against an ordered dominance alternative  $H_1: F_1(x) \geq F_2(x) \geq \dots \geq F_k(x)$ , again at least one inequality being strict at least for some  $x$ . This is a one-tail test. Reversal of all inequalities gives an analogous test in the opposite tail. A test for such ordered alternatives was devised by Jonckheere (1954), but it had been conceived independently by Terpstra (1952).

In essence it is an extension of the WMW test using the Mann—Whitney formulation. The samples **must** be ordered in the sequences specified in  $H_1$ . Exact permutation tests are more readily available for small samples than they are for the Kruskal—Wallis test, and an asymptotic test is also available. We discuss this after an example. The asymptotic test at least is available in many standard statistical software packages.

### *Example 6.3*

*The problem.* Hinkley (1989) gives braking distances taken by motorists to stop when travelling at various speeds. A subset of his data is:

<i>Speed (mph)</i>	<i>Breaking distances (feet)</i>
20	48
25	33    59    48    56
30	60    101    67
35	85    107

Use the Jonckheere—Terpstratetest to assess the evidence for a tendency for braking distance to increase as speed increases.

*Formulation and assumptions.* We test  $H_0$ : *braking distance is independent of initial speed* against  $H_1$ : *braking distance increases with speed*. The samples are already arranged in order of increasing speed implicit as the natural order under  $H_1$ .

*Procedure.* If there are  $k$  samples we calculate the sum,  $U$ , of all Mann—Whitney statistics  $U_{rs}$  relevant to the  $r$ th sample ( $r = 1, 2, \dots, k - 1$ ) and any sample  $s$  for which  $s > r$ . Thus for the four samples above we calculate  $U_{12}$ ,  $U_{13}$ ,  $U_{14}$ ,  $U_{23}$ ,  $U_{24}$  and  $U_{34}$ . For example,  $U_{12}$  is the sum of the number of sample 2 values that exceeds each sample 1 value. Clearly here  $U_{12} = 2.5$  since there is only one sample 1 value (48) and this is equal to one value in sample 2 and is exceeded by two others (56, 59); ties are scored as  $\frac{1}{2}$  as in the Mann—Whitney statistic. Similarly  $U_{13} = 3$ ,  $U_{14} = 2$ ,  $U_{23} = 12$ ,  $U_{24} = 8$  and  $U_{34} = 5$ . Adding all  $U_{rs}$  gives a total  $U = 32.5$ . Obtaining the exact permutation distribution of ranks is tedious without a suitable computer program. However the exact test in StatXact indicates  $P = \Pr(U \leq 32.5) = 0.0011$ , for the one-tail test relevant here.

*Conclusion.* There is clear evidence (not surprisingly!) that increasing speed increases braking distance.

*Comments.* 1. When it is relevant, the Jonckheere—Terpstratetest is generally more powerful than Kruskal—Wallistest. If a Kruskal—Wallistest is applied to these data, the exact  $P = 0.0133$  while the asymptotic test gives  $P = 0.0641$ . It is not surprising that the asymptotic result is unreliable for such small samples. A parametric analysis of variance using these data gives a variance ratio  $F = 4.45$  with 3, 6 degrees of freedom which again would not indicate significance at the 5 per cent level ( $P = 0.057$  for this  $F$ -value).

2. Older editions of the UK Highway Code used the formula  $d = v + v^2/20$  for minimum stopping distances  $d$  feet, when travelling at  $v$  mph. The mean values

in the above data for each speed are fairly close to those obtained from this formula, but there is considerable variation between drivers at any one speed.

*Computational aspects.* StatXact computes exact  $P$ -values for somewhat larger samples than it does for the Kruskal–Wallis test. It is illuminating to compare these with tail probabilities for the asymptotic test given below. In this example the asymptotic result gives  $P = 0.0022$ ; although twice as large as the exact test level, the discrepancy is not so great as that indicated above for the Kruskal–Wallis test.

The asymptotic Jonckheere–Terpstra test is based on the fact that  $U$  as defined in [Example 6.3](#) has a mean  $E(U) = \frac{1}{2}(N^2 - \sum_i n_i^2)$  and variance  $\text{Var}(U) = \{N^2(2N + 3) - \sum_i [n_i^2(2n_i + 3)]\}/72$ . For large  $N$  and the individual  $n_i$  not too small the distribution of

$$Z = \frac{U - E(U)}{\sqrt{\text{Var}(U)}}$$

is approximately standard normal. The sample sizes in [Example 6.4](#) are too small for an asymptotic result to inspire confidence, but calculations sought in [Exercise 6.3](#) give, for this example,  $Z = 2.85$ , corresponding, as indicated above, to a one-tail  $P = 0.0022$ .

The above expression for  $\text{Var}(U)$  needs adjustment for ties. The adjustment is trivial for relatively few ties but in [Section 9.3.2](#) we show that the existence of many ties may have a dramatic effect. A formula for adjusting  $\text{Var}(U)$  for ties is given by Lehmann (1975, p. 235) and a modified version is given by (9.11) in [Section 9.3.2](#).

[Exercise 6.21](#) covers a situation where  $H_1: F_1(x) \geq F_2(x) \geq \dots \geq F_k(x)$  is an appropriate alternative hypothesis.

Tests involving ordered treatment differences that are sometimes of interest include that where we wish to test  $H_0: \text{all } \mu_i \text{ are equal}$  against  $H_1: \theta_1 \leq \theta_2 \leq \theta_3 \leq \dots \leq \theta_{q-1} \leq \theta_q \geq \theta_{q+1} \geq \dots \geq \mu_k$  where at least one inequality is strict. These tests are called **umbrella tests** and were proposed by Mack and Wolfe (1981) and are discussed in some detail by Hollander and Wolfe (1999, Section 6.3).

## 6.2.6 The median test for several samples

The median test in [Section 5.3](#) generalizes easily to three or more samples. The alternative hypothesis is now one of a difference between population medians without specifying which populations differ in location, how many differences there are, or their direction. As in the two-sample case, each sample may come from **any** unspecified population (they need not all have the same distribution). Given  $k$  samples where the unknown population

**Table 6.3** A general contingency table for the median test.

<i>Above M</i>	<i>Below M</i>	<i>Total</i>
$a_1$	$b_1$	$n_1$
$a_2$	$b_2$	$n_2$
.	.	.
.	.	.
$a_k$	$b_k$	$n_k$
$\overline{A}$	$\overline{B}$	$\overline{N}$

medians are  $\theta_1, \theta_2, \dots, \theta_k$ , we test  $H_0: \theta_1 = \theta_2 = \dots = \theta_k$  against  $H_1: \text{not all } \theta_i \text{ are equal}$ . The procedure generalizes that in [Section 5.3.1](#). If  $M$  is the combined sample median for all observations, in each sample we count the numbers of observations above and below  $M$ . As in [Section 5.3.1](#), we reject sample values equal to  $M$  and then work with reduced samples. Assuming sample values that equal  $M$  have already been dropped, suppose that of the  $n_i$  observations in sample  $i$  there are  $a_i > M$  and  $b_i < M$ ,  $i = 1, 2, \dots, p$ . We record these numbers above and below  $M$  in a  $k \times 2$  contingency table of  $k$  rows and 2 columns. The values  $a_i, b_i$  are constrained for each sample so that  $a_i + b_i = n_i$  the number in that sample. When no observations equal  $M$ , the column totals  $A$  and  $B$  both equal  $\frac{1}{2}N$ . Table 6.3 illustrates this for a general contingency table having  $k$  rows and 2 columns.

The probability  $P^*$  of the above cell values with the row and column totals fixed is given by a generalization of [\(5.3\)](#):

$$P^* = \frac{A! B! \prod_i (n_i!) }{N! \prod_i (a_i!) \prod_i (b_i!)} \quad (6.3)$$

where  $\prod_i (x_i!)$  is the product  $(x_1!) \times (x_2!) \times \dots \times (x_k!).$  Other  $k \times 2$  contingency tables with the same marginal totals must be considered to obtain the  $P$ -value. This is illustrated by an example.

#### *Example 6.4*

*The problem.* Six dental surgeons (*I*, *II*, *III*, *IV*, *V* and *VI*) perform the removal of third molar (or wisdom) teeth from adult patients. To investigate the assertion that some surgeons extract these teeth more quickly, the following data were obtained. These relate to 28 patients each of whom had just one third molar teeth removed. The times (in minutes) taken for the six dental surgeons to operate were:

<i>I</i>	23	25	34	45			
<i>II</i>	7	11	14	15	17	24	40
<i>III</i>	5	8	16	20	26		
<i>IV</i>	12	21	31	38			
<i>V</i>	30	43					
<i>VI</i>	4	9	10	18	19	35	

Inspecting the data we find that the median extraction time,  $M$ , for all 28 patients is 19.5 minutes, so no sample value is equal to the median. Table 6.4 shows the numbers of extraction times above and below  $M$  by each surgeon. Use a median test to assess the strength of any evidence against  $H_0$ : *all six samples come from populations with the same median extraction time.*

*Formulation and assumptions.* The test is an extension of the Fisher exact test developed in [Section 5.3](#) for the two-sample situation to the  $k \times 2$  table having the form in [Table 6.3](#). We omit details but the formula (6.3) is obtainable by extending arguments similar to those leading to (5.3). For the data in Table 6.4 a  $P$ -value is calculated based on a critical region consisting of all  $6 \times 2$  contingency tables having the given marginal totals for which, under  $H_0$ ,  $P^*$  does not exceed that observed. For the  $2 \times 2$  tables in [Example 5.6](#), the more extreme configurations were obvious. For larger contingency tables this is not the case, nor can we attribute these extreme configurations to a particular tail; i.e. our alternative hypothesis  $H_1$ : *not all medians are equal* is essentially two-tail, not unlike the situation in the Kruskal–Wallis test ([Section 6.2.2](#)), or indeed the  $F$ -test in parametric ANOVA.

*Procedure.* For Table 6.4 the calculation using (6.3) reduces to

$$P^* = \frac{(14!)(14!)(4!)(7!)(5!)(4!)(2!)(6!)}{(28!)(4!)(2!)(2!)(3!)(2!)(1!)(0!)(5!)(3!)(1!)(0!)(5!)}$$

where, by definition,  $0! = 1$ . Despite cancellations between numerator and denominator, it is tedious to verify that  $P^* = 0.00013$  using only a pocket calculator.

**Table 6.4** Numbers of observations above and below the combined sample median,  $M$ , in [Example 6.3](#).

Dental surgeon	Above $M$	Below $M$	Total
<i>I</i>	4	0	4
<i>II</i>	2	5	7
<i>III</i>	2	3	5
<i>IV</i>	3	1	4
<i>V</i>	2	0	2
<i>VI</i>	1	5	6
	14	14	28

The  $P^*$  for all other contingency tables with the same marginal totals that have the same or lower values of  $P^*$  can be found similarly. Realistically, a computer program or some suitable approximation is required. An appropriate program gives a  $P$ -value of 0.046 for the comparison of third molar extraction times.

*Conclusion.* There is moderately strong evidence against  $H_0$ , indicating some variation in median extraction times between the dental surgeons.

*Comments.* 1. Individual sample sizes here are all small (particularly for dental surgeon V). The main differences in [Table 6.4](#) seem to be that surgeon I probably has a median above  $M$  and surgeon VI (possibly also surgeon II) has a median below  $M$ . If we were to perform separate sign tests of  $H_0: \theta_i = M$  for each individual sample we would not be able to demonstrate strong evidence against  $H_0$  (even in a one-tail test) since for any particular surgeon the number of patients operated on is very small. In practice much larger samples of patients would be selected for such a study.

2. The results from a comparison of dental surgeons need to be interpreted carefully. Most extractions are completed in around 15 to 20 minutes, but complicated cases can take 1 hour or more. These can be identified in advance from a radiograph, so consultants may deal with serious cases, with less experienced staff operating on the straightforward patients. Higher extraction times for surgeon I could indicate greater responsibility rather than inexperience.

3. The difficulty in carrying out an exact test without a suitable computer program makes an asymptotic test of interest. We discuss one below.

4. A situation where we may want to test for a common median when samples come from distributions differing in other respects (e.g. spread, skewness, etc.) arises if we compare alternative components that may be used in an industrial process. We may measure a characteristic such as time to failure for a number of replicates of each type of component. A preliminary step may then be to test whether it is reasonable to suppose all may have the same median time to failure. If there is clear evidence they do not, one might reject from further consideration those that appear to have a lower median time. Further choice may depend on preference for a component with little variability in its failure time distribution, or on non-statistical factors, such as cost.

*Computational aspects.* The StatXact, Testimate or Stata programs for the Fisher exact test ([Section 9.2.1](#)) may be used for the median test.

In the notation used in [Table 6.3](#) an asymptotic test makes use of the statistic

$$T = \sum_i \frac{(a_i - n_i A / N)^2}{n_i A / N} + \sum_i \frac{(b_i - n_i B / N)^2}{n_i B / N} \quad (6.4)$$

The term  $n_i A/N$  is equal to the *expected number* of cases above  $M$  for the  $i$ th surgeon, assuming that the null hypothesis is true. Readers familiar with the chi-squared test for independence in contingency tables will recognize (6.4) as the sum of

$$[(\text{observed} - \text{expected value})^2/\text{expected value}]$$

for all cells, which we discuss in [Section 9.2.2](#), where an alternative form easier to compute manually is given by (9.8). For a  $k \times 2$  table  $T$  has asymptotically a chi-squared distribution with  $k - 1$  degrees of freedom. Chi-squared tables (e.g. Neave, 1981, p. 21) giving minimal critical values for significance at conventional levels are widely available and nearly every statistical software package has a program for this test.

### *Example 6.5*

*The problem.* For the data in [Example 6.4](#) test for equality of population medians using the asymptotic test based on (6.4).

*Formulation and assumptions.* We calculate all quantities in (6.4) from [Table 6.4](#) and compare the resulting  $T$  with the tabulated critical value.

*Procedure.* The expected numbers ( $n_i A/N$ , etc.) are calculated for each cell in [Table 6.4](#) using (6.4). For instance, with surgeon  $I$  the expected number of cases above  $M$  is  $n_i A/N$  or  $4 \times 14/28 = 2$ . The expected values, in the order given in [Table 6.4](#) are:

2	2
3.5	3.5
2.5	2.5
2	2
1	1
3	3

Substituting these and the observed cell values from [Table 6.4](#) in (6.4), we find  $T = 2 \times (2^2/2 + 1.5^2/3.5 + 0.5^2/2.5 + 1^2/2 + 1^2/1 + 2^2/3) = 11.15$ . Tables indicate  $T \geq 11.07$  is the critical value of chi-squared with  $k - 1 = 5$  degrees of freedom for significance at the 5 per cent level.

*Conclusion.* There is moderately strong evidence against  $H_0$ .

*Comments.* 1. All samples are small, but exact and asymptotic results agree well. For the chi-squared distribution with 5 degrees of freedom  $\Pr(T \geq 11.15) = 0.048$ , close to the exact  $P = 0.046$  obtained in [Example 6.4](#).

2. The alternative hypothesis that not all medians are equal is essentially two-tailed, but the asymptotic test uses a single tail of the chi-squared distribution. This is because we use the squares of the discrepancies *observed* – *expected* and these are necessarily positive, no matter what is the sign of the actual discrepancy.

*Computational aspects.* If StatXact is used for the Fisher exact test it gives an asymptotic test statistic that differs from  $T$  given by (6.4). This is because it uses a statistic, called the Fisher statistic, which also has an asymptotic chi-squared distribution. For this particular example the Fisher statistic takes the value 10.23 and  $\Pr(\gamma^2 \geq 10.23) = 0.069$ . The reason for this discrepancy is explored further in Section 9.2. For larger samples the two asymptotic statistics are usually in reasonable agreement.

In practice, we may be interested in whether there are differences in medians for some subset of the  $k$  populations. One might take pairs of samples and compute, say, 95 per cent confidence limits for the median differences for each pair using the method given in [Section 5.3.2](#). Since these confidence intervals will generally be determined with varying precision for different pairs (due to differences in sample size and perhaps differences in the population distributions) direct comparisons involving these intervals sometimes lead to what at first sight appear bizarre consequences. For example, if zero is included in the 95 per cent confidence interval for the difference  $\theta_2 - \theta_1$  and also in the 95 per cent confidence interval for the difference  $\theta_3 - \theta_1$  there is no strong evidence against the null hypotheses that  $\theta_2 = \theta_1$  or that  $\theta_3 = \theta_1$ . It is tempting to conclude that this implies a lack of evidence against the hypothesis  $\theta_3 = \theta_2$ . This does not follow. For instance, if most of the confidence interval for  $\theta_2 - \theta_1$  is positive and most of that for  $\theta_3 - \theta_1$  is negative, the likely differences with respect to  $\theta_1$  are in opposite directions. As a consequence, differences between possible values for  $\theta_2$  and  $\theta_3$  may in fact be quite large and the 95 per cent confidence interval for  $\theta_2 - \theta_3$  might not include zero. Similar considerations apply for pairwise comparisons in parametric analyses and the practice is not recommended (*see also* [Section 6.4.1](#) for further comments).

If assumptions needed for, say, a Kruskal–Wallis test appear to hold that test should be preferred to the median test, the latter then usually being appreciably less powerful. For the data in [Example 6.4](#), however, one may have reservations about the Kruskal–Wallis test because the sample values suggest appreciable differences in the spread of times taken by different surgeons.

### 6.3 CENTRALITY TESTS FOR RELATED SAMPLES

In [Example 4.3](#), we analyzed systolic blood pressure measured before and after exercise for a group of students. The data from such

a study from two related groups. The sign test gives an individual a positive sign if the systolic blood pressure increases on exercise and a negative sign if it decreases. The numbers of positive and negative signs are then compared. This procedure amounts to ranking the two blood pressure measurements for each individual and comparing them before and after exercise. In a general repeated-measures study, measurements on individuals are made at  $t$  points where  $t$  can be three or more. The data collected from  $t$  related samples.

Another type of study that generates several related samples is the **randomized block** design. In its simplest form, each patient receives two treatments in random order. Alternatively, matched patients can be used to compare the two treatments ([Section 4.1.1](#)). A pair of linked observations is known as a block. Using matched pairs allows us to make treatment comparisons by analyzing differences between responses within blocks, which effectively eliminate differences **between** blocks. It is advantageous to do this when we expect results for each member of a pair to be more homogeneous under a null hypothesis of no treatment effect than would be the case if treatments were applied randomly to large sets of less homogeneous units.

Generalizing this type of study to compare  $t$  ( $>2$ ) treatments we might replace our homogeneous pairs by blocks of  $t$  units, blocks being chosen so that units within each block are as homogeneous as possible for characteristics other than the applied treatments. For example, to compare the effects of five different diet regimes on pigs, each block may be a litter of five pigs. In each litter the five diets are allocated, one to each pig, chosen at random.

To compare three different cake recipes (treatments) we might make batches of cakes using each recipe and divide each batch into four equal parts and cook one part of each batch in four different ovens (each oven is a block). We do this because each oven may operate at a slightly different temperature and they may have varying heat efficiencies. To assess the merits of the recipes we might ask experts to rank them in order of preference for taste. Comparison between cakes baked in the same oven is desirable because, although all mixtures may produce poor cakes (all slightly burnt for example) in a particular oven and all pleasant tasting cakes in another, one hopes that the relative preference ordering for products from each oven may be reasonably consistent from oven to oven. We have so far not considered randomization within ovens. If all mixtures are cooked at the same time it would be wise to allocate the three different recipes to shelf positions at random as these may

affect the cakes produced. If the cakes in each oven are to be cooked one after another the time order could sensibly be randomized separately for each oven as the end product might be affected by the time the mixture stands before cooking. Alternatively, there may be some carry-over effect on flavour depending upon which cake is cooked first. These are matters reflecting principles of good experimental design and are relevant irrespective of whether the method of analysis is parametric or nonparametric.

### 6.3.1 Pitman type tests

When data consist of measurements or other sample values from continuous distributions, it is again possible to use permutation tests of the type developed by Pitman (1938) for tests equivalent to the overall  $F$ -test in the analysis of variance. Once again for reasons given in Sections 5.1.1 and 6.2.1 these are seldom used in practice although the procedure is described and illustrated in simple terms by Edgington (1995, Chapter 5) and Sprent (1998, Section 8.1). The former reference gives a program for Monte Carlo tests for this permutation ANOVA. At the time of writing we know of no software in major statistical packages for calculating exact permutation  $P$ -values for repeated measures or randomized block designs using raw measurement data. We do not discuss the test further because of these practical limitations.

### 6.3.2 The Friedman test

Friedman (1937) extended the sign test to the case of several related samples. As we have noted, in the sign test the two scores (– or +) can be thought of as corresponding to ranks 1 and 2. In the Friedman test, the scores for each individual are ranked (e.g. if measurements are made at five points in time the five readings will be ranked from 1 to 5). Each reading is then replaced by the appropriate rank.

The null hypothesis is that the population distributions at each time point are the same, so that for instance, the population medians are equal. If the null hypothesis is true, inspection of the readings for a particular time point will reveal a mixture of ranks. If the population medians differ, then for at least one of the time points the ranks should be mainly high or mainly low.

In the randomized block design, instead of several discrete times there are an equivalent number of treatments. The number of units in

each block equals the number of treatments (each treatment is applied to exactly one unit in each block).

In parametric analysis of variance for continuously distributed observations such as measurements we remove differences between patients (or blocks) as a source of variability before making time point (or treatment) comparisons. Do not worry about how this is done if you are unfamiliar with the analysis of variance. For Friedman's test differences between patients (or blocks) are removed automatically by the process of replacing observations by their ranks when we do this separately within each block. The test is also immediately applicable if the basic data are only ranks within each block (see [Exercise 6.11](#)). Friedman's test in the context of that and similar examples examines the consistency of ranks rather than acting as a test for centrality. This approach, which we discuss further in [Section 7.2](#), was developed independently by M.G. Kendall.

This is how the test works. Suppose that in a study  $b$  patients are measured at  $t$  time points (or in a randomized block design  $t$  treatments are applied each to 1 of  $t$  units in each of  $b$  blocks). If  $x_{ij}$  denotes the measurement at time  $i$  for patient  $j$  where  $i$  runs from 1 to  $t$  and  $j$  from 1 to  $b$ , we replace the  $x_{ij}$  for each patient by ranks 1 to  $t$ , this ranking being carried out separately for each patient (block). We assume that there are no tied ranks. The sum of the ranks for time point  $i$  is denoted by  $s_i$ ,  $i = 1, 2, \dots, t$ . The Friedman statistic can then be written

$$T = \frac{12 \sum s_i^2}{bt(t+1)} - 3b(t+1) \quad (6.5)$$

If  $b$  and  $t$  are not too small,  $T$  has approximately a chi-squared distribution with  $t-1$  degrees of freedom. For this no-tie situation, some tables are available for conventional critical values of  $T$ , e.g. Neave (1981, p. 34). Modern statistical software provides many programs for exact or asymptotic tests.

### *Example 6.6*

*The problem.* For a group of seven students, the pulse rate (per minute) was measured before exercise (I), immediately after exercise (II), and 5 minutes after exercise (III). The data are given in [Table 6.5](#). Use the Friedman statistic to test for differences between pulse rates on the three occasions.

*Formulation and assumptions.* Separately for each student we replace each observation by its rank and calculate the Friedman statistic for these ranks to test  $H_0$ : *population distributions of pulse rates are identical against  $H_1$ : at least one*

**Table 6.5** Pulse rate values for students before and after exercise.

Time	I	II	III
Student			
A	72	120	76
B	96	120	95
C	88	132	104
D	92	120	96
E	74	101	84
F	76	96	72
G	82	112	76

**Table 6.6** Pulse rate values, ranks within students.

Time	I	II	III
A	1	3	2
B	2	3	1
C	1	3	2
D	1	3	2
E	1	3	2
F	2	3	1
G	2	3	1
Total	10	21	11

time point has a different distribution of pulse rates reflected by a median shift or dominance. As there are no ties within students, formula (6.5) may be used.

*Procedure.* Relevant ranks are given in Table 6.6 along with the rank total for each time point. To use (6.5), we calculate  $\sum_i (s_i^2)$ . The column totals are squared and summed to give  $10^2 + 21^2 + 11^2 = 662$ . We have  $b = 7$  students and  $t = 3$  time points, whence

$$T = \frac{12 \times 662}{7 \times 3 \times (3+1)} - 3 \times 7 \times (3+1) = 10.57$$

This gives an exact  $P = 0.0027$ . By contrast, the chi-squared approximation with  $T = 10.57$  (2 degrees of freedom) gives  $P = 0.0051$ .

*Conclusion.* There is strong evidence of a difference in pulse rates between the three times.

*Comments.* 1. The difference between the  $P$ -values given by the exact and approximate methods is a cause for concern. For such small data sets it is unwise to rely on the chi-squared approximation.

2. Inspection of the raw data shows that the strength of evidence against the null hypothesis is not really surprising. From a physiological point of view, the pulse rate should increase on exercise and fall off to the resting value once the exercise has been completed. Note that some individuals appear to take longer than others to return to the resting pulse rate.

3. While the physiological explanation of the differences is sensible for this example care must be taken with measurements repeated in time because, unlike the randomized block situation, time order is fixed, not randomized. This means that evidence against  $H_0$  may be due to some factor other than that under study. In the present study if all individuals recorded their pulse rates using a clock in the exercise room and because of some unnoticed mechanical defect the clock mechanism slowed only during the time of the second reading and this was not detected this could explain higher readings on occasion II. Alternatively, if for some reason the air temperature at the time of the third reading had dropped dramatically from that at the time of the second reading this might explain the relatively lower rates at occasion III.

### 6.3.3 The Friedman test with ties

Tied ranks within individuals (blocks) are given mid-rank values. If there are such ties we use a more general formula for  $T$ . In addition to the above terms we need to calculate  $S_r = \sum_{i,j} r_{ij}^2$  where  $r_{ij}$  is the rank corresponding to  $x_{ij}$  (with no ties  $S_r = bt(t + 1)(2t + 1)/6$ ). It is also convenient to write  $S_t = \sum_i (s_i^2)/b$ .

As with the Kruskal–Wallis test, a correction factor is required and this is  $C = 1/4bt(t + 1)^2$ . The Friedman statistic can be written

$$T = \frac{b(t-1)(S_t - C)}{S_r - C} \quad (6.6)$$

If there are no ties this is equivalent to (6.5). As with the no-ties case, if  $b, t$  are not too small  $T$  has an approximate chi-squared distribution with  $t - 1$  degrees of freedom. Iman and Davenport (1980) suggest that a better approximation than (6.6) is

$$T_1 = \frac{(b-1)(S_r - C)}{S_r - C} \quad (6.7)$$

which, under a null hypothesis of no treatment difference, has approximately an  $F$ -distribution with  $t - 1$  and  $(b - 1)(t - 1)$  degrees of freedom. If the ranking is identical in all blocks the denominator of  $T_1$  is zero. Iman and Davenport show that in formal significance testing terms this may be interpreted as a result significant at a  $P = (1/t)^{b-1}$  significance level.

**Table 6.7** Nodes to first flower, total for four plants.

Block	I	II	III	IV
<i>Treatment</i>				
<i>Control</i>	60	62	61	60
<i>Gibberellic acid</i>	65	65	68	65
<i>Kinetin</i>	63	61	61	60
<i>Indole acetic acid</i>	64	67	63	61
<i>Adenine sulphate</i>	62	65	62	64
<i>Maelic hydrazide</i>	61	62	62	65

**Table 6.8** Nodes to first flower, ranks within blocks.

Block	I	II	III	IV	Total
<i>Treatment</i>					
<i>Control</i>	1	2.5	1.5	1.5	6.5
<i>Gibberellic acid</i>	6	4.5	6	5.5	22
<i>Kinetin</i>	4	1	1.5	1.5	8
<i>Indole acetic acid</i>	5	6	5	3	19
<i>Adenine sulphate</i>	3	4.5	3.5	4	15
<i>Maelic hydrazide</i>	2	2.5	3.5	5.5	13.5

### Example 6.7

*The problem.* Pearce (1965, p. 37) quoted results of a greenhouse experiment carried out by J. I. Sprent (unpublished). The data, given in Table 6.7, are the numbers of nodes to first initiated flower summed over four plants in each experimental unit (pot) for the pea variety Greenfeast subjected to six treatments – one an untreated control, the others various growth substances. There were four blocks allowing for differences in light intensities and temperature gradients depending on proximity to greenhouse glass. The blocks were arranged to make these conditions as like as possible for all units (pots of four plants) in any one block. Use the Friedman statistic to test for differences between treatments in node of flower initiation.

*Formulation and assumptions.* Within each block we replace each observation by its rank and calculate the Friedman statistic for these ranks to test  $H_0$ : no difference between treatments against  $H_1$ : at least one treatment has a different centrality parameter from the others. We may test using (6.6) or (6.7).

*Procedure.* Relevant ranks are given in Table 6.8 where we add a column giving rank totals  $s_i$  for each treatment. Squaring each rank and adding we get  $S_r = 361$ . The uncorrected treatment sum of squares is

$$S_t = (6.5^2 + 22^2 + 8^2 + 19^2 + 15^2 + 13.5^2)/4 = 339.625$$

and  $C = 4 \times 6 \times 7^2/4 = 294$ ; using (6.6)  $T = 4 \times 5(339.625 - 294)/(361 - 294) = 13.62$ , and (6.7) gives  $T_1 = 3(339.625 - 294)/(361 - 339.625) = 6.40$ .

Using  $T_1$ , as recommended by Iman and Davenport, we find that with 5, 15 degrees of freedom critical values for significance at the 5, 1 and 0.1 per cent significance levels are respectively 2.90, 4.56 and 7.57 so we would judge our result significant at the 1 per cent level. If instead we compare  $T$  with the critical values of the chi-squared distribution with 5 degrees of freedom the result is significant at the 5 per cent but not the 1 per cent level. For the chi-squared distribution with 5 degrees of freedom  $\Pr(T \geq 13.620) = 0.0182$ . Comparing  $T$  with the critical values given by Neave for six treatments and four blocks we find significance is indicated at the 1 per cent level if  $T \geq 12.71$ , so again significance is indicated at that level, but remember that Neave's tables apply strictly to the 'no-tie' situation. The Friedman test program in StatXact confirms and refines the above findings based on  $T$ . The exact  $P$ -value based on the relevant permutation test is  $P = 0.0045$ , while the asymptotic  $P$ -value based on the chi-squared distribution is  $P = 0.0182$  as already stated.

*Conclusion.* There is strong evidence of a difference between treatments.

*Comments.* 1. While for practical reasons we use  $T$  or  $T_1$  as our test statistic we could use  $S_t$  because both  $S_t$  and  $C$  remain unaltered in the permutation distribution which involves only permuting ranks **within** blocks.

2. A parametric randomized blocks analysis of variance of the original data gives  $F = 4.56$ , which corresponds almost exactly to  $P = 0.01$ .

3. Whereas analysis of variance introduces a sum of squares reflecting difference between block totals there is no such term in the Friedman analysis because the sum of ranks in all blocks is the same, namely  $\frac{1}{2} t(t + 1)$ .

*Computational aspects.* For larger samples exact  $P$ -value computations like those in StatXact may be slow or even impossible. The Monte Carlo option will then give good estimates of  $P$ . The test statistic is calculated in many statistical software packages, but often only asymptotic tail probabilities are quoted.

Ranks within blocks might be replaced by normal scores. Experience suggests that such rescaling has few advantages. Indeed, if there are ties, the block differences removed by ranking may be reintroduced, though usually not dramatically.

Ranking within blocks is robust against many forms of heterogeneity of variance, in that it removes any inequalities of variance between blocks (see [Exercise 6.14](#)).

Durbin (1951) gives a rank-based Friedman-type test for data in incomplete blocks.

### 6.3.4 The Quade test

Quade (1979) proposed a test that is often more powerful than the Friedman test. It also eliminates block differences but weights the

rank scores to give greater weight in those blocks where the raw data indicate possibly more marked treatment effects. Details, together with a numerical example, are given by Sprent (1998, Section 8.2) and by Conover (1999, Section 5.8). StatXact includes a program for this test. Whereas the Friedman test is basically an extension of the sign test, the Quade test is effectively an extension of the Wilcoxon signed-rank test and is equivalent to it when  $t = 2$ .

If  $x_{ij}$  denotes the observed value for treatment  $i$  in block  $j$  the range,  $D_j$  of values in block  $j$  is the difference between the greatest and least  $x_{ij}$ , i.e.

$$D_j = \max_i x_{ij} - \min_i x_{ij}.$$

For the Quade test the  $D_j$  are ranked in ascending order from 1 to  $b$ , using mid-ranks for ties if needed. If we denote the rank of  $D_j$  for block  $j$  by  $q_j$  and the Friedman rank of  $x_{ij}$  by  $r_{ij}$ , then the Quade scores are defined as  $s_{ij} = q_j[r_{ij} - 1/2(t+1)]$ . These scores are used in StatXact for an exact permutation test but if that program is not available the Iman and Davenport (1980) analogue of (6.7) may be used in an asymptotic test and it takes the form

$$T_1 = \frac{(b-1)S_t}{S_s - S_t}$$

where  $S_t = \sum_i[(\Sigma_j s_{ij})^2]/b$  and  $S_s = \sum_{i,j} s_{ij}^2$ . Asymptotically  $T_1$  has an  $F$ -distribution with  $t - 1$  and  $(b - 1)(t - 1)$  degrees of freedom. As with the Friedman test, if the denominator is zero this may be interpreted as a result significant at the  $P = (1/t)^{b-1}$  significance level.

For the data in Example 6.7 for the Quade test StatXact gives an exact  $P = 0.0019$ .

### 6.3.5 An alternative extension of Wilcoxon-type tests

Hora and Conover (1984) proposed ranking all observations simultaneously without regard to treatments or blocks and carrying out an analysis of variance on the ranks (or normal scores derived from these). The procedure for ranks is described by Iman, Hora and Conover (1984).

When applied to the data in Example 6.7 the relevant  $F$ -statistic for treatment difference is  $F = 4.96$  (compared with 4.56 for a parametric test and 6.40 using (6.7)). The reader familiar with standard analysis of variance may wish to verify these results (see Exercises 6.4 and 6.13).

### 6.3.6 The Page test for ordered alternatives

Page (1963) proposed an analogue to the Jonckheere–Terpstra test for ordered alternatives applicable to blocked data. If the treatments are arranged in the order specified in the alternative hypothesis and  $s_i$  is the sum of ranks for treatment  $i$  (as in the Friedman test) the Page statistic is  $P = s_1 + 2s_2 + 3s_3 + \dots + ts_t$ . Asymptotically, if  $H_0$ : *no treatment difference* holds, then  $P$  has a normal distribution with mean  $tb(t+1)^2/4$  and variance  $b(t^3 - t^2)/[144(t-1)]$ . Tables (Daniel, 1990, Table A17) give critical values of  $P$  for small  $b$ ,  $t$ . The test is discussed in more detail by Daniel (1990, Section 7.3), by Marascuilo and McSweeney (1977, Section 14.12), by Sprent (1998, Section 8.2) and by Hollander and Wolfe (1999, Section 7.2). StatXact includes a program for calculating exact  $P$ -values for this test with the usual provision for a Monte Carlo approximation if the sample size is too large. An asymptotic approximation is also given. Hollander and Wolfe (1999, Table A23) give tail probabilities for the Page statistic  $P$  given above for a range of treatment and block sizes for which the asymptotic approximation may prove unreliable.

## 6.4 MORE DETAILED TREATMENT COMPARISONS

In parametric analysis of variance after a preliminary overall test interest usually shifts to specific treatment comparisons. In factorial experiments, for example, interest may centre on main effects and interactions. In other contexts one may be interested in pairwise comparisons for a selected subgroup of treatments or in comparing one specified treatment (often called a control) with each of the remaining treatments. One may also want to make comparisons between groups of treatments; for example, if three of a set of seven fertilizer treatments include a nitrate component while the remaining four do not one may want to make inferences about differences between the overall means or medians for the nitrate fertilizers and the corresponding measure for the nitrate-free fertilizers.

In parametric ANOVA such comparisons are referred to as *contrasts in a linear model*. The reader familiar with the notion of factorial experiments (factorial treatment structures) will know that in a  $2 \times 2$  factorial experiment with factors  $A$  and  $B$  each at two levels the *main effect* for  $A$  is defined as the difference between the mean for all experimental units receiving the higher level of factor  $A$  and the mean for all units receiving the lower level of factor  $A$ .

For nonparametric methods there are analytical procedures for assessing such contrasts covering both hypothesis tests and estimation (e.g. obtaining confidence intervals for differences). A detailed treatment is beyond the scope of this elementary text, but an account of the basic theory illustrated with a range of examples is given by Hettmansperger and McKean (1998, Chapter 4). Many contrasts of practical interest are considered by Hollander and Wolfe (1999, Chapters 6 and 7) and by Conover (1999, Chapter 5).

#### 6.4.1 Multiple comparisons

Careful thought should be given to which of many possible contrasts are relevant and to the implications of the fact that not all contrasts are independent of one other. We drew attention to a possible anomaly that may arise when comparisons are not independent in the penultimate paragraph of [Section 6.2.6](#). In an experiment in which one treatment may reasonably be regarded as a *control* it is often both relevant and sensible to compare the mean or median for this treatment with the corresponding measure for each remaining treatment. Such contrasts may be of interest, for example, when comparing the efficacy of three different drugs in treating a disease and the experiment includes also a group consisting of individuals not receiving any drug. If an experiment is conducted where the different treatments consist of administration of the same drug at increasing dose levels the main contrast of interest will be in how response changes with dose level and overall tests for increasing or decreasing responses with dose are provided by the Jonckheere–Terpstra or the Page test. In some circumstances umbrella tests of the type mentioned at the end of [Section 6.2.5](#) may also be appropriate. Further tests using specific contrasts may also be relevant to find, for example, if there is a threshold dose which must be exceeded before there is evidence of a real treatment effect.

For both parametric and nonparametric analyses many more-or-less automatic procedures known as **multiple comparison tests** have been devised. They are designed to safeguard against making unwarranted inferences when many contrasts that are not independent are examined. These aim to restrict the number of unwarranted conclusions that differences are real that may arise as a result of carrying out all possible tests. For example, if 20 treatments are involved in an experiment and an overall parametric test such as the *F*-test in the analysis of variance or a Kruskal–Wallis test indicates no acceptable evidence of treatment differences, a comparison of the largest observed treatment mean with the smallest observed

treatment mean will very often produce a  $P$ -value sufficiently small to indicate strong evidence that these means differ. This is an ever-present danger if one decides to carry out a test because the data look as though that test might tell us something interesting.

Many statisticians advise strongly against using multiple comparison procedures blindly especially when there is no obvious treatment structure such as that implied in factorial experiments (where multiple comparison procedures are irrelevant) or in an experiment where comparisons with a control are of prime interest.

We shall not go into the sometimes deep and often contentious logical arguments for and against multiple comparison methods here. We stress rather that if used at all they should be used with caution. A more constructive approach is one where an experimenter (often in association with a statistician) elects (or nominates) comparisons that are of interest before the data are obtained and makes relevant assessments of evidence about these on the basis of what information is later provided by the data. This is in the spirit of our overall attitude in this book towards interpretation of  $P$ -values as a tool for weighing evidence rather than a decision tool implicit in the assignment of pre-fixed formal significance levels.

We indicate one approach that is widely used for nominated pairwise comparisons. This is to compute what are known as **least significant differences**. Logically they can only be justified if applied to prechosen comparisons (i.e. those chosen before the data are collected, or at latest before they are examined) and even then they are applied only if an overall Kruskal–Wallis, Friedman or other relevant test indicates strong evidence for differences. The critical  $P$ -value or ‘significance level’ used in multiple comparison tests should be no less stringent than that in the overall test.

When using the Kruskal–Wallis test the criteria for accepting a centrality difference between the  $i$ th and  $j$ th sample are that

1. the overall test indicates significance and
2. if  $m_i = s_i/n_i$ ,  $m_j = s_j/n_j$  are the mean ranks for these samples, then

$$|m_i - m_j| > t_{N-t, \alpha} \sqrt{[(S_r - C)(N-1-T)(n_i + n_j)/\{n_i n_j (N-t)(N-1)\}]} \quad (6.8)$$

where  $T$  is given by (6.2) and  $t_{N-t, \alpha}$  is the  $t$ -value required for significance at level  $\alpha$  per cent in a  $t$ -test with  $N-t$  degrees of freedom, other quantities being defined as in Section 6.2.2 or 6.2.3.

An analogous result holds for van der Waerden scores with sample rank means replaced by sample score means and  $C$  now zero and  $T$  being the form of the statistic appropriate for these scores.

Nominated comparisons for the Friedman test may be based on either treatment rank means or totals since we have equal replication. The algebraic expression in terms of rank totals is simpler. An analogue of (6.8) gives the requirement for a least significant difference between the treatment rank totals  $s_i, s_j$ :

$$|s_i - s_j| > t_{(b-1)(t-1), \alpha} \sqrt{[2b(S_r - S_t)/(b-1)(t-1)]} \quad (6.9)$$

in the notation of [Section 6.3.2](#).

Most statisticians agree with Pearce (1965, pp. 21–2) that it is reasonable to compare a subset of all samples (or treatments) if that subset is selected before the data are obtained (or at least before they are inspected) using formulae like (6.8) or (6.9) if one could sensibly anticipate that the specified pair are likely to show a treatment difference that may be of practical interest or importance.

#### *Example 6.8*

*The problem.* In [Example 6.2](#) special interest attaches to differences in sentence length between Vulliamy and Queen because the former is an English author with an academic background and limited fictional output and Queen is a popular and prolific American writer. Use a least significant difference test to determine whether the difference is significant.

*Formulation and assumptions.* Relevant quantities are mostly available from the solution to [Example 6.2](#), and these are substituted in (6.8).

*Procedure.* Relevant rank means are  $m_1 = 17/5 = 3.4$  (Vulliamy),  $m_2 = 72.5/6 = 12.1$  (Queen). Also  $S_r - C = 2104.5 - 1624.5 = 480.0$ ,  $T = 9.146$ ,  $N = 18$ ,  $n_1 = 5$ ,  $n_2 = 6$ ,  $t = 3$ ; tables give the critical  $t$ -value at the 1 per cent significance level with 15 degrees of freedom to be 2.95. The left-hand side of (6.7) is  $12.1 - 3.4 = 8.7$  and the right-hand side reduces to

$$2.95 \times \sqrt{[480 \times (17 - 9.146) \times 11 / (30 \times 17 \times 15)]} = 6.86.$$

*Conclusion.* Since  $8.7 > 6.86$  the difference is formally significant at the 1 per cent level, i.e. there is strong evidence of a real difference.

*Comment.* We use the 1 per cent level since the overall test indicated significance at this level.

The description of the experiment and the data in the next example were kindly provided by Chris Theobald.

#### *Example 6.9*

*The problem.* As part of a study of the feeding habits of the larvae of the Blue-tailed Damsel Fly, 6 of the larvae were collected along with 6 members of each of seven species of prey on which they usually fed. Each larva was placed on a cocktail stick in a glass beaker that contained water and one each of the seven types of prey. Records were kept of the order in which each larva ate the prey. The results are shown in [Table 6.9](#). Some ties denoted by mid-ranks

**Table 6.9** Preferences of Blue-tailed Damsel Fly larvae for various prey.

Larva	1	2	3	4	5	6	Total
<i>Prey</i>							
<i>Anopheles</i>	1	2	1	1	3	1	9
<i>Cyclops</i>	7	7	7	6.5	6.5	7	41
<i>Ostracod</i>	5.5	5.5	5	5	5	2.5	28.5
<i>Simocephalus (wild)</i>	3	4	4	6.5	1.5	4	23
<i>Simocephalus (domestic)</i>	2	1	6	4	6.5	6	25.5
<i>Daphnia magna</i>	4	3	2	2	1.5	2.5	15
<i>Daphnia longspina</i>	5.5	5.5	3	3	4	5	26

arose due to failure to observe the order when a larva ate two prey in quick succession. Is there evidence that the larvae prefer to eat some species of prey before others?

The cyclops was usually the last to be eaten. If the experimenter had prior reason to anticipate such an outcome, how might you appropriately examine whether the tendency was statistically significant?

*Formulation and assumptions.* Evidence of preference requires a Friedman test analogous to that in [Example 6.6](#). A multiple comparison test using (6.9) is one approach to examining the situation regarding cyclops.

*Procedure.* The Friedman test follows the lines in [Example 6.6](#). If not using StatXact or some other program for a Friedman test the reader should confirm ([Exercise 6.20](#)) that  $t = 7$ ,  $b = 6$ ,  $S_r = 837$ ,  $S_t = 775.75$ ,  $C = 672$ ,  $T = 22.64$  and that  $T_1 = 8.47$ , indicating a  $P$ -value less than 0.01, or probably less than 0.001. Indeed the Monte Carlo estimate of the exact  $P$ -value in StatXact indicates almost certainly that  $P < 0.0005$ . Setting  $\alpha = 0.01$  we find the appropriate  $t$ -value with  $(b - 1)(t - 1) = 30$  degrees of freedom is 2.75, so when using (6.9) the least significant difference between treatment rank totals given in the last column of Table 6.9 is  $2.75 \times \sqrt{[2 \times 6(837 - 775.75)/30]} = 13.61$ .

*Conclusion.* The difference 13.61 is exceeded between totals for cyclops and all other prey except ostracod, so there is strong evidence that there is a low preference for cyclops.

*Comment.* We have used the only relevant test given in this book to clarify the situation regarding cyclops but it is possible to devise a test comparing cyclops with the average rank response for all other prey which would confirm the conclusion that cyclops was less favoured. A more sophisticated approach to multiple comparisons is given by Leach (1979, Section 6.2). A useful paper on this topic is that by Shirley (1987), and an important earlier paper is Rosenthal and Ferguson (1965). References given in the preliminary discussion in this section are also relevant.

### 6.4.2 Factorial treatment structures

A lot of work has been done in recent years on nonparametric analyses of factorial treatment structures. The detail is beyond the scope of this book and a fuller discussion would only be useful to readers already familiar with the equivalent analysis of variance. Some key references to work in this area include Grizzle, Starmer and Koch (1969), Iman (1974), Conover and Iman (1976), Scheirer, Ray and Hare (1976), Mack and Skillings (1980), de Kroon and van der Laan (1981), Groggel and Skillings (1986) and the paper by Shirley (1987) referred to in [Section 6.4.1](#). Further references, together with a critical review of some problems in analysis of data expressed as scores, are given by Thomas and Kiwanga (1993) and Sprent (1998, Section 7.6) gives a simple exposition of some of the basic ideas. A more sophisticated approach will be found in Hettmansperger and McKean (1998, Section 4.4).

### 6.4.3 Commonsense analysis

We now show how simple nonparametric methods may be adapted to deal with a practical complication. The example indicates the versatility that goes with common-sense applications of nonparametric methods in what is often termed **exploratory data analysis**.

#### *Example 6.10*

*The problem.* Suppliers of word-processing programs are testing an updated version designed to make it easier to prepare technical reports containing graphs, mathematical formulae, etc. To see whether the updated version (package A) shows an advantage over the current version (package B) four people chosen at random from nine are asked to prepare a specimen report using package A and the remaining five to prepare the same report using package B. All participants had similar experience in word processing. The numbers of mistakes made by each operator in their first drafts of the report were:

Package A	2	11	24	26
Package B	17	25	28	31

A WMW test just fails to shows a significant difference in a one-tail test (that test is appropriate because improvements in package A should not tend to increase mistakes) at a conventional 5 per cent level. The relevant Wilcoxon statistic ([Exercise 6.5](#)) is  $S_m = 13$  and the corresponding one-tail  $P = 0.0556$ .

It is now disclosed that some participants in this test had a training in the technology which was the subject of the report whilst others did not. The results could now be displayed separately for each group:

With knowledge of technology (K)	Package A	2	11
	Package B	17	25

<i>Without knowledge of technology (NK)</i>	<i>Package A</i>	24	26
	<i>Package B</i>	28	31    63

Clearly, whichever package is used, those with no knowledge of the technology make more mistakes than those using the same package who have a knowledge. Explore the use of this new information in establishing whether there is evidence that package *A* reduces mistakes, independent of prior knowledge.

*Formulation and assumptions.* A simplistic approach would be to apply the WMW test separately to the two groups *K* and *NK* but the samples are now too small to show significance no matter what outcome! We might regard the results as representing samples from four populations and apply a Kruskal–Wallis test. If this gave a significant result we would need to follow it by multiple comparison tests comparing packages *A* and *B* with technical knowledge (*K*), and packages *A* and *B* with no technical knowledge (*NK*). For the Kruskal–Wallis test ([Exercise 6.6](#))  $T = 7.1333$ . In an asymptotic test this is not significant at the 5 per cent level, but it is significant in an exact test at a nominal 1 per cent level (Neave, 1981, p. 33), a result confirmed by StatXact where  $P = 0.0079$ . The asymptotic test is clearly unreliable for such small samples. Tests based on least significant differences ([Exercise 6.6](#)) indicate significance at the 5 per cent level between packages *A* and *B* for *K* but not for *NK*.

Alternatively, we might look upon knowledge (*K*) as a factor (often called a covariate) likely to reduce errors in a similar way for each package and allow for its effect by considering separately permutation distributions for *K* and *NK*, and then combine these results in an appropriate manner.

*Procedure.* We rank all observations as for the Kruskal–Wallis test, i.e.

<i>With knowledge of technology (K)</i>	<i>Package A</i>	1	2
	<i>Package B</i>	3	5
<i>Without knowledge of technology (NK)</i>	<i>Package A</i>	4	6
	<i>Package B</i>	7	8    9

but now we carry out separate permutation distribution calculations for *K*, *NK* respectively using these scores. With *K* the possible scores for package *A* are clearly (1, 2) (1, 3) (1, 5) (2, 3) (2, 5) (3, 5). The associated sums are  $S_1 = 3, 4, 6, 5, 7, 8$  each with probability 1/6. Similarly, for *NK* possible scores for package *A* are (4, 6) (4, 7) (4, 8) (4, 9) (6, 7) (6, 8) (6, 9) (7, 8) (7, 9) (8, 9) with sums  $S_2 = 10, 11, 12, 13, 13, 14, 15, 15, 16, 17$  each with probability 1/10 under  $H_0$ . The samples are independent, so each sum  $S = S_1 + S_2$  has associated probability 1/60 and the tail probability relevant to a one-tail test is  $\Pr(S \leq k)$ , where  $k$  is the value we observe. We double this probability for a two-tail test. To work out the relevant probabilities we proceed, much as we did in [Example 1.4](#), to work out the number of ways each sum can be obtained. In [Table 6.10](#) each entry is the sum of the  $S_1$  value at the top and the  $S_2$  value at the left. We repeat the values 13 and 15 for  $S_2$  as each occurs twice. This ensures equal probabilities of 1/60 for all 60 sums in the body of the table. If a particular sum occurs  $r$  times the associated probability is  $r/60$ . We easily deduce the probabilities associated with each sum: these are given in [Table 6.11](#).

In this example adding the ranks associated with package  $A$  for  $K$  and  $NK$  we find  $S = 1 + 2 + 4 + 6 = 13$ . This is the same as the WMW statistic. Can you see why? From [Table 6.10](#) or [Table 6.11](#),  $\Pr(S \leq 13) = 1/60$ .

*Conclusion.* Since for a one-tail test  $P = 0.0167$ , there is quite strong evidence that package A is superior.

*Comments.* 1. In this *ad hoc* analysis the assumption that if one package is superior then this should be evident for both  $K$  and  $NK$  is critical. In practice, one might have found that one group preferred one package and one the other – an effect known as an **interaction** in the terminology of factorial experiments.

2. A parametric analysis of variance of the raw data when broken into the four samples does not indicate a significant difference (*see Exercise 6.7*).

3. Those familiar with factorial treatment structures will realize that this experiment might be looked upon as a  $2 \times 2$  factorial experiment with knowledge status as one factor and type of package as the other.

**Table 6.10** Equiprobable sums under separate permutation of sub-samples in [Example 6.10](#).

$S_2$	$S_1$	3	4	5	6	7	8
10		13	14	15	16	17	18
11		14	15	16	17	18	19
12		15	16	17	18	19	20
13		16	17	18	19	20	21
13		16	17	18	19	20	21
14		17	18	19	20	21	22
15		18	19	20	21	22	23
15		18	19	20	21	22	23
16		19	20	21	22	23	24
17		20	21	22	23	24	25

**Table 6.11** Probability of each rank-sum in Table 6.10.

<i>Value k of S</i>	$\Pr(S = k)$	<i>Value k of S</i>	$\Pr(S = k)$
13	1/60	20	8/60
14	2/60	21	7/60
15	3/60	22	5/60
16	5/60	23	4/60
17	6/60	24	2/60
18	8/60	25	1/60
19	8/60		

*Computational aspects.* The method given under *Procedure* is only feasible for larger data sets with a suitable computer program. StatXact provides one for

what are termed WMW tests with stratification. The method extends to more than two covariate levels, e.g. we might have tested the packages using groups with (a) no knowledge (b) elementary knowledge and (c) advanced knowledge.

#### 6.4.4 Binary responses

In many situations each experimental unit may show one of two responses – win or lose, succeed or fail, live or die, a newborn child is male or female. For analytic purposes we usually score such responses as 0 or 1. For example, generalizing from [Example 4.4](#), suppose five members A, B, C, D, E of a mountaineering club each attempt three rock climbs at each of which they either succeed or fail. If a success is recorded as 1 and a failure as 0, the outcomes may be summarized as follows:

<i>Member</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>Climb 1</i>	1	1	0	0	1
<i>Climb 2</i>	1	0	0	1	0
<i>Climb 3</i>	0	1	1	1	1

Cochran (1950) proposed a method applicable to such situations to test the hypothesis  $H_0$ : *all climbs are equally difficult* against  $H_1$ : *the climbs vary in difficulty*. In conventional terms climbs are ‘treatments’ and climbers are blocks. If we have  $t$  treatments in  $b$  blocks and binary (i.e. 0, 1) responses the appropriate test statistic is

$$Q = \frac{t(t-1) \sum_i T_i^2 - (t-1)N^2}{tN - \sum_j B_j^2} \quad (6.10)$$

where  $T_i$  is the total (of 1s and 0s) for treatment  $i$ ,  $B_j$  is the total for block  $j$  and  $N$  is the grand total. The exact permutation distribution of  $Q$  is not easily obtainable but for large samples  $Q$  has approximately a chi-squared distribution with  $t - 1$  degrees of freedom. Although not immediately obvious, for two treatments the test reduces to McNemar’s test and in Section 9.5 we give an alternative form of the McNemar test, which is exactly equivalent to  $Q$  given above.

Cochran’s test is discussed more fully by Conover (1999, Section 4.6) and by Sprent (1998, Section 8.4).

## 6.5 TESTS FOR HETEROGENEITY OF VARIANCE

Most tests for heterogeneity of variance discussed in [Section 5.8](#) extend to several samples. We illustrate this for the squared-rank test given in [Section 5.8.3](#). The procedure parallels that for the Kruskal–Wallis test with squared ranks of absolute deviations replacing data ranks as scores. The absolute deviations of all sample values from their sample mean are ranked over all samples and these ranks are squared. The statistic  $T$  is identical to (6.2) except that now  $s_i$  is the sum of the squared rank deviations for sample  $i$  and  $C$  is the square of the sum of these squared ranks divided by  $N$ : symbolically, denoting such a squared rank by  $r_{ij}^2$ ,  $C = [\sum_{i,j} r_{ij}^2]^2/N$ .  $S_r$  is here the sum of squares of the squared ranks, i.e.  $S_r = \sum_{i,j} r_{ij}^4$ . If all population variances are equal, for a reasonably large sample  $T$  has a chi-squared distribution with  $t - 1$  degrees of freedom. If there are no rank ties the sum of squared ranks is  $N(N+1)(2N+1)/6$ . Also, for no ties the denominator of  $T$  reduces to

$$S_r - C = (N - 1)N(N + 1)(2N + 1)(8N + 11)/180$$

### Example 6.11

*The problem.* As well as increasing with speed it is thought possible that braking distance may be more variable from driver to driver as speed increases. Use the squared rank test to test for heterogeneity of population variance on the basis of the following sub-sample from Hinkley (1989) of initial speeds (mph) and braking distance (feet).

Speed	Braking distance			
5	2	8	8	4
10	8	7	14	
25	33	59	48	56
30	60	101	67	

*Formulation and assumptions.* We obtain the absolute deviations  $|x_{ij} - m_i|$  of each observation  $x_{ij}$  from its sample mean  $m_i$ . These deviations are then ranked over combined samples and the statistic  $T$  is calculated and compared to the relevant chi-squared critical value.

*Procedure.* The means at each speed are  $m_1 = 5.5$ ,  $m_2 = 9.67$ ,  $m_3 = 49$ ,  $m_4 = 76$ . The absolute deviations together with their ranks (1 for least, 14 for greatest, with mid-ranks for ties) are as follows:

5 mph	Absolute deviation	3.5	2.5	2.5	1.5
	Overall rank	7	4.5	4.5	2
10 mph	Absolute deviation	1.67	2.67	4.33	
	Overall rank	3	6	8	
25 mph	Absolute deviation	16	10	1	7
	Overall rank	12.5	11	1	9

<i>30 mph</i>	<i>Absolute deviation</i>	16	25	9
	<i>Overall rank</i>	12.5	14	10

The sum of squared ranks at 5 mph is  $s_1 = 7^2 + 4.5^2 + 4.5^2 + 2^2 = 93.5$ . Similarly,  $s_2 = 109$ ,  $s_3 = 359.25$ ,  $s_4 = 452.25$ , so the sum of all squared ranks is 1014 and  $S_p = (93.5)^2/4 + (109)^2/3 + (359.25)^2/4 + (452.25)^2/3 = 106\ 587.724$ . It is left as an exercise for the reader to verify that  $C = 73\ 442.571$  and that

$$S_r = 7^4 + 4.5^4 + \dots + 10^4 = 127\ 157.25,$$

whence  $T = 8.02$ . Tables of the chi-squared distribution indicate that the critical value for significance at the 5 per cent level with 3 degrees of freedom is  $T = 7.815$ .

*Conclusion.* There is evidence of heterogeneity of variance in the light of significant at a nominal 5 per cent level.

*Comments.* 1. One has reservations about an asymptotic result with such small samples, but we can do little better except in extremely small samples where we might calculate exact permutation distributions (*see Computational aspects* below). For the original data there may of course be differences in centrality whether or not we accept  $H_0$ : no variance difference in a squared rank test.

2. An alternative would be a Kruskal–Wallis test using overall ranks of absolute deviations.

*Computational aspects.* StatXact provides a program for an exact permutation test. The squares of the rank deviations are submitted as data to the program for one-way ANOVA with general scores and this gives  $P = 0.0240$  broadly in line with the asymptotic result based on  $T$ . Indeed if  $T = 8.02$  the exact chi-squared  $P = 0.046$ , a discrepancy that should not surprise one with such small samples.

## 6.6 SOME MISCELLANEOUS CONSIDERATIONS

### 6.6.1 Analogues of other parametric situations

Readers familiar with the design of experiments and the analysis of data using ANOVA and related techniques will appreciate that the nonparametric analogues introduced in this chapter correspond to only a small portion of this methodology. We have only touched upon the concept of factorial treatment structures and have not dealt at all with the analogues of many experimental designs such as balanced incomplete blocks. Many of these and other aspects are covered in considerable detail by Hollander and Wolfe (1999, Chapters 6 and 7).

### 6.6.2 Power sample size and efficiency

In the light of the problems hinted at in our discussions about power calculations in the much simpler one- and two-sample situations in

earlier chapters it will hardly surprise readers to learn that exact power calculations for most of the procedures considered in this chapter are only available for a few very restricted, and often unrealistic, situations. Exploration of the circumstances in which a specific method is likely to be beneficial and any indication of sample sizes needed to meet specific objectives may be aided by information about Pitman efficiency. Pitman efficiencies for some of the methods developed in this chapter as well as in other situations discussed by them are considered by Hollander and Wolfe (1999, Sections 6.10 and 7.16).

Bijnning and Kessler (1999) give some asymptotic power results for the Jonckheere–Terpstra test and some closely related tests and report that simulation studies indicate that their asymptotic results provide good approximations even for moderate sample sizes.

### 6.6.3 Runs test for three or more samples

If we have three or more groups and clustering or alternation is suspected it is possible to derive the relevant tail probabilities for the numbers of runs, *see e.g.*, Mood (1940). The algebra involved is challenging, however, and Mood's work appears not to have been widely used. This may reflect the relatively low power of such tests in many circumstances.

### 6.6.4 A quantile-based procedure for minimum dispersion

Sometimes samples come from several populations for each of which the  $k$ th quantile is known to be the same. We may then be interested in selecting the population with the smallest variance or some other measure of spread. For example, if several different methods of producing an item all lead to the same proportion of defectives (*e.g.* falling below a minimum permitted weight) we may want to select the production method with the smallest variability of weight among the items produced. A nonparametric approach to this problem is given by Gill and Mehta (1989, 1991).

## 6.7 FIELDS OF APPLICATION

Parametric analysis of variance of designed experiments has historical origins in agriculture but soon spread to the life and natural sciences, medicine, industry and, more recently, to business and the social sciences. Development of nonparametric analogues

was stimulated by a realization that data often clearly violated normal theory assumptions and, more importantly, to provide a tool for analysis of ordinal data expressed only as ranks or preferences (see e.g. [Example 6.9](#)). Ranking or preference scores often combine assessment, either deliberately or subconsciously, of a number of factors that are given different weights by individuals. It is thus of practical interest to see if there is still consistency between the way individuals rank the same objects despite the fact that they may not give the same weight to each factor. Our first example illustrates this point.

### Preferences for washing machines

Consumers' preferences for washing machines are influenced by their assessment of several factors, e.g. price, reliability, power consumption, washing time, load capacity, ease of operation and clarity of instructions. Individuals weigh such factors differently; a farmer's wife offering bed and breakfast to tourists will rate ability to wash bed linen highly; a parent with a large young family the ability to remove sundry stains from children's clothing; for many a low price and running economy may be key factors. No machine is likely to get top ratings on all points, but manufacturers will be keen to achieve a high overall rating for their product from a wide range of consumers. A number of people may be asked to state preferences (e.g. ranks) for each manufacturer's machine and for competitors; each manufacturer wants to know if there is consistency in rankings – whether most people give a particular machine a preferred rating – or whether there is inconsistency or general dislike for some machine. Each consumer is a block, each machine a treatment, in the context of a Friedman test. The hypothesis under test is  $H_0$ : *no consistency in rankings* against  $H_1$ : *some consistency in rankings*. We discuss this type of situation further in [Section 7.2](#).

### Literary discrimination

A professor of English asserts that short stories by a certain writer (A) are excellent, those by a second writer (B) are good, and those by a third (C) are inferior. To test his claim and judgement he is given 20 short stories to read on typescripts that do not identify the authors and asked to rank them 1 to 20 (1 for best, 20 for worst). In fact 6 are by A, 7 by B and 7 by C. Rankings given by the professor when checked against authors are:

<i>Excellent author</i>	1	2	4	8	11	17
<i>Good author</i>	3	5	6	12	16	18
<i>Inferior author</i>	7	9	10	13	14	15
						20

Do these results justify his claim of discriminatory ability? A Kruskal–Wallis test could be used to test the hypothesis that the ranks indicate the samples are all from the same population (i.e. no discriminatory ability) but as the samples have a natural ordering – excellent, good, inferior – a Jonckheere–Terpstra test is more appropriate. *See Exercise 6.9.*

### Assimilation and recall

A list of 12 names is read out to students. It contains in random order, 4 names of well-known sporting personalities, 4 of national and international political figures, and 4 of people prominent in local affairs. The students are later asked which names they can recall and a record is made of how many names each student recalls in each of the three categories. By ranking the results we may test whether recall ability differs systematically between categories, e.g. do people recall names of sporting personalities more easily than those of people prominent in local affairs? *See Exercise 6.8.*

### Tasting tests

A panel of tasters may be asked to rank different varieties of raspberry in order of preference. A Friedman test is useful to detect any pattern in taste preference. *See Exercise 6.11.*

### Quantal responses

Four doses of a drug are given to batches of rats, groups of four rats from the same litter forming a block. The only observation is whether each rat is dead or alive 24 hours later. Cochran's test is appropriate to test for different survival rates at the four doses.

## 6.8 SUMMARY

Centrality tests for several independent samples include the following:

**The Kruskal–Wallis test** ([Section 6.2.2](#)) is a rank analogue of the one-way classification analysis of variance. The test statistics commonly used are [\(6.1\)](#) or, if there are ties, [\(6.2\)](#). If  $p$  is the number of samples [\(6.1\)](#) and [\(6.2\)](#) have asymptotically a chi-squared

distribution with  $p - 1$  degrees of freedom. An alternative is to transform ranks to van der Waerden scores.

The **Jonckheere–Terpstra test** ([Section 6.2.5](#)) is appropriate if  $H_1$  orders the treatments. The test statistic is a sum of certain Mann–Whitney statistics. An asymptotic approximation is widely used, but care is needed with ties.

The **median test** ([Section 6.2.6](#)) is applicable to  $k$  samples from any populations (these may differ from one another in aspects other than the median). The test is essentially a special case of the Fisher exact test, or asymptotically a Pearson chi-squared test applied to a  $k \times 2$  contingency table calculating the statistic using [\(6.4\)](#). The test often has low power relative to many of its competitors.

For related samples the **Friedman test** ([Section 6.3.2](#)) is applicable to randomized block designs where ranks are allocated within blocks. One test statistic is given by [\(6.5\)](#) or [\(6.6\)](#) and asymptotically it has a chi-squared distribution. Another statistic is given by [\(6.7\)](#). An alternative test is the **Quade test** ([Section 6.3.4](#)). The **Page test** ([Section 6.3.6](#)) is an analogue of the Jonckheere–Terpstra test applicable in a randomized block context if  $H_1$  orders treatments.

**General treatment comparisons** ([Section 6.4](#)) need care. Analogies to least significant differences in parametric analyses for both the Kruskal–Wallis and Friedman test situations are given in [\(6.8\)](#) and [\(6.9\)](#).

The **Cochran test** ([Section 6.4.4](#)) is applicable to blocked binary response data and the test statistic [\(6.10\)](#) has an asymptotic chi-squared distribution.

The **squared-rank test** and some other tests for heterogeneity of variance ([Section 6.5](#)) extends from those for two samples.

## EXERCISES

Readers unfamiliar with (parametric) analysis of variance may ignore questions on that topic.

- 6.1 Perform a parametric analysis of variance on the data in [Example 6.2](#), comparing your result with those for the Kruskal–Wallis test.
- 6.2 Reanalyze the data in [Example 6.2](#) using van der Waerden scores.

- 6.3 Carry out the asymptotic Jonckheere–Terpstra test for the data in [Example 6.3](#).
- 6.4 Analyze the data in [Example 6.7](#) using a parametric analysis of variance.
- 6.5 Verify for [Example 6.10](#) that the relevant statistic for the WMW test is 13.
- 6.6 Perform the Kruskal–Wallis test on the four-sample data in [Example 6.10](#) to confirm the results quoted in that example, including computation of least significant differences between package A and package B (i) for those with technical knowledge and (ii) for those without technical knowledge.
- 6.7 Perform a parametric analysis of variance (ANOVA) of the data on numbers of mistakes in [Example 6.10](#).
- 6.8 At the beginning of a session 12 names are read out in random order to 10 students. Four are names of prominent sporting personalities (Group A), four of national and international politicians (Group B) and four of local dignitaries (Group C). At the end of the session students are asked to recall as many of the names as possible. The numbers recalled were:

<i>Student</i>	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	<i>VI</i>	<i>VII</i>	<i>VIII</i>	<i>IX</i>	<i>X</i>
<i>Group A</i>	3	1	2	4	3	1	3	3	2	4
<i>Group B</i>	2	1	3	3	2	0	2	2	2	3
<i>Group C</i>	0	0	1	2	2	0	4	1	0	2

Rank the data within each block (student) and use a Friedman test to assess evidence of a difference between recall rates for the three groups. In particular, is the recall rate for group B and/or group C significantly lower than that for group A? Carry out an ANOVA on the given data. Do the conclusions agree with the Friedman test? If not, why not?

- 6.9 Use the ranks in the literary discrimination example in [Section 6.7](#) to assess validity of the professor's claim to discriminate between the authors.
- 6.10 A sergeant major orders 34 men to parade tallest on the right, shortest on the left, numbered 1 (tallest) to 34 (shortest). Each man is then asked whether he smokes or drinks alcoholic beverages and the rank numbers of men in the various categories are as follows:

<i>Drinker and smoker</i>	3	8	11	13	14	19	21	22	26	27	28	31	33
<i>Smoker, non-drinker</i>	2	12	25	32	34								
<i>Drinker, non-smoker</i>	1	7	15	20	23	24	30						
<i>Non-smoker, non-drinker</i>	4	5	6	9	10	16	17	18	29				

Is there evidence of association between height, smoking and drinking habits? Would you reach the same conclusion if ranks were replaced by van der Waerden scores? (In this basic analysis ignore the ‘factorial’ nature of the treatment structure although in practice this should be taken into account in a more sophisticated analysis.)

- 6.11 Five tasters rank four varieties of raspberry in order of preference. Do the results indicate a consistent taste preference? Mid-ranks are given for ties.

<i>Taster</i>	1	2	3	4	5
<i>Variety</i>					
<i>Malling Enterprise</i>	3	3	1	3	4
<i>Malling Jewel</i>	2	1.5	4	2	2
<i>Glen Clova</i>	1	1.5	2	1	2
<i>Norfolk Giant</i>	4	4	3	4	2

- 6.12 Four share tipsters are each asked to predict on 10 randomly selected days whether the London FTSE Index (commonly known as Footsie) will rise or fall on the following day. If they predict correctly this is scored as 1, if incorrectly as 0. Do the scores below indicate differences in tipsters' ability to predict accurately?

<i>Day</i>	1	2	3	4	5	6	7	8	9	10
<i>Tipster 1</i>	1	0	0	1	1	1	1	0	1	1
<i>Tipster 2</i>	1	1	1	1	0	1	1	0	0	0
<i>Tipster 3</i>	1	1	0	1	1	1	1	1	0	1
<i>Tipster 4</i>	1	1	0	0	0	1	1	1	0	1

- 6.13 Replace the data in [Table 6.7](#) by ranks 1 to 24 (using mid-ranks for ties where appropriate) and carry out an ordinary randomized block analysis of variance of these ranks to confirm the *F*-value quoted in [Section 6.3.4](#).
- 6.14 Berry (1987) gives the following data for numbers of premature ventricular contractions per hour for 12 patients with cardiac arrhythmias when each is treated with 3 drugs A, B, C.

<i>Patient</i>	1	2	3	4	5	6	7	8	9	10	11	12
A	170	19	187	10	216	49	7	474	0.4	1.4	27	29
B	7	1.4	205	0.3	0.2	33	37	9	0.6	63	145	0
C	0	6	18	1	22	30	3	5	0	36	26	0

Use a Friedman test to investigate differences in response between drugs. In particular, is there evidence of a difference between drug A and drug B? Note the obvious heterogeneity of variance between drugs. Carry out an ordinary randomized block analysis of variance on these data. Do you consider it to be valid? Is the Friedman analysis to be preferred? Why?

- 6.15 Cohen (1983) gives data for numbers of births in Israel for each day in 1975. We give below data for numbers of births on each day in the 10th, 20th, 30th and 40th weeks of the year.

<i>Day</i>	<i>Mon</i>	<i>Tue</i>	<i>Wed</i>	<i>Thu</i>	<i>Fri</i>	<i>Sat</i>	<i>Sun</i>
<i>Week</i>							
10	108	106	100	85	85	92	96
20	82	99	89	125	74	85	100
30	96	101	108	103	108	96	110
40	124	106	111	115	99	96	111

Perform Friedman analyses to determine whether the data indicate (i) a difference in birth rate between days of the week that shows consistency over the four selected weeks and (ii) any differences between rates in the 10th, 20th, 30th and 40th weeks.

- 6.16 Snee (1985) gives data on average liver weights per bird for chicks given three levels of growth promoter (none, low, high). Blocks correspond to different bird houses. Use a Friedman test to see if there is evidence of an effect of growth promoter.

<i>Block</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>
<i>None</i>	3.93	3.78	3.88	3.93	3.84	3.75	3.98	3.84
<i>Low dose</i>	3.99	3.96	3.96	4.03	4.10	4.02	4.06	3.92
<i>High dose</i>	4.08	3.94	4.02	4.06	3.94	4.09	4.17	4.12

Since dose levels are ordered, a Page test is appropriate. Try this also.

- 6.17 Lubischew (1962) gives measurements of maximum head width in units of 0.01 mm for three species of *Chaetocnema*. Part of his data is given below. Use a Kruskal–Wallis test to see if there is a species difference in head widths.

<i>Species 1</i>	53	50	52	50	49	47	54	51	52	57
<i>Species 2</i>	49	49	47	54	43	51	49	51	50	46
<i>Species 3</i>	58	51	45	53	49	51	50	51		

- 6.18 Biggins, Loynes and Walker (1987) considered various ways of combining examination marks where all candidates sat the same number of papers but different candidates selected different options from all those available. The data below are the marks awarded by four different methods of combining results for each of 12 candidates. Do the schemes give consistent ranking of the candidates? Is there any evidence that any one scheme treats some candidates strikingly differently than the way they are treated by other schemes so far as rank order is concerned? Is there any evidence of a consistent difference between the marks awarded by the various schemes?

<i>Cand.</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>
<i>A</i>	54.3	30.7	36.0	55.7	36.7	52.0	54.3	46.3	40.7	43.7	46.0	48.3
<i>B</i>	60.6	35.1	34.1	55.1	38.0	47.8	51.5	44.8	39.8	43.2	44.9	47.6
<i>C</i>	59.5	33.7	34.3	55.8	37.0	49.0	51.6	45.6	40.3	43.7	45.5	48.2
<i>D</i>	61.6	35.7	34.0	55.1	38.3	46.9	51.3	44.8	39.7	43.2	44.8	47.5

- 6.19 The pea node data in [Example 6.7](#) include a control treatment with no growth substance because the experimenter wished to compare all other treatments with this as a base. Regarding each such as a nominated comparison check whether any exceed the least significant difference.
- 6.20 Confirm the numerical values quoted in [Example 6.9](#) that are relevant for the Friedman test for the data in that example.
- 6.21 Chris Theobald supplied the following data from a study of 40 patients suffering from a form of cirrhosis of the liver. One purpose was to examine whether there was evidence of association between spleen size and blood platelet count. Blood platelets form in bone marrow and are destroyed in the spleen, so it was thought that an enlarged spleen might lead to more platelets being eliminated and hence to a lower platelet count. The spleen size of each patient was found using a scan and scored from 0 to 3 on an arbitrary scale, 0 representing a normal spleen and 3 a grossly enlarged spleen. The platelet count was determined as the number in a fixed volume of blood. Do these data indicate an association between spleen size and platelet count in the direction anticipated by the experimenter?

<i>Spleen size</i>	<i>Platelet count</i>												
<i>0</i>	156	181	220	238	295	334	342	359	365	374	391	395	481
<i>1</i>	65	105	121	150	158	170	214	235	238	255	265	390	
<i>2</i>	33	70	87	109	114	132	150	179	184	241	323		
<i>3</i>	79	84	94	259									

# 7

## Correlation and concordance

### 7.1 CORRELATION IN BIVARIATE DATA

We frequently want measures that summarize the **strength** of a relationship between two variables. Interpreted sensibly, such measures are often appropriate even when both variables are outcome variables, neither being considered as ‘explanatory’ in the sense of it indicating the cause of the other variable taking the values it does. Correlation is one measure of the strength of association or dependence between two variables. In the parametric context, the Pearson **product moment correlation coefficient** estimates the degree of *linear* association between two variables. The coefficient takes values between  $-1$  and  $+1$ ; these extreme values are attained only when points lie exactly on a straight line (with negative and positive slopes or gradients respectively). If one of the variables tends to be large when the other is large and small when the other is small, the correlation is positive. If large values of one variable occur with small values of the other the correlation is negative. If the two variables are independent of each other, the value of the correlation is zero. A non-linear relationship can also produce a correlation value close to zero; it is prudent to plot a scatter diagram before calculating any correlation coefficient.

If we have measurements on two variables (such as height and weight) for a sample of  $n$  individuals, these paired observations can be written  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where the population distributions are  $X$  and  $Y$ . The sample Pearson product moment correlation coefficient,  $r$ , is defined as

$$r = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (7.1)$$

which for computational purposes is usually rearranged and written

$$r = c_{xy} / [\sqrt{(c_{xx}c_{yy})}]$$

where

$$c_{xy} = \sum_i (x_i y_i) - (\sum_i x_i)(\sum_i y_i)/n, \quad c_{xx} = \sum_i x_i^2 - (\sum_i x_i)^2/n,$$

$$c_{yy} = \sum_i y_i^2 - (\sum_i y_i)^2/n.$$

It is easy to verify from (7.1) that  $r$  is unaltered by what are called location and scale changes (*see also* location and scale parameters, [Section 5.8](#)), i.e. if all  $x_i$  are replaced by  $(x_i - k)/s$  and all  $y_i$  are replaced by  $(y_i - m)/t$  where  $k, m$  are any constants that represent location or centrality changes and  $s$  and  $t$  are any positive constants that represent scale changes. Another way of putting this is that  $r$  is invariant under such linear transformations of the  $x$  and  $y$ . This property extends to the alternative correlation coefficients described in [Sections 7.1.3](#) to [7.1.5](#) and also has important practical implications in some regression methods considered in [Chapter 8](#).

In parametric inference the Pearson coefficient is particularly relevant to a bivariate normal distribution where the sample coefficient  $r$  is an appropriate estimate of the population correlation coefficient  $\rho$ . If  $\rho = 0$  for a bivariate normal distribution this implies  $X$  and  $Y$  are independent and values of  $r$  close to zero support that hypothesis. Even when the sample  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  is assumed to come from a bivariate normal distribution, inference about an unknown population correlation coefficient  $\rho$  based upon the sample coefficient  $r$  given by (7.1) is more complicated than that for inferences about the means or variances of the distributions of  $X$  or  $Y$ . If the joint distribution of  $X$  and  $Y$  is not bivariate normal, parametric inference about  $\rho$  is even more difficult.

In contrast, basic nonparametric correlation inference does not require the assumption of bivariate normality and can be applied to both paired observations of continuous data and to data consisting of ranks. These ranks may be the original data or they may be derived from continuous measurements.

### 7.1.1 A Pitman test for zero correlation

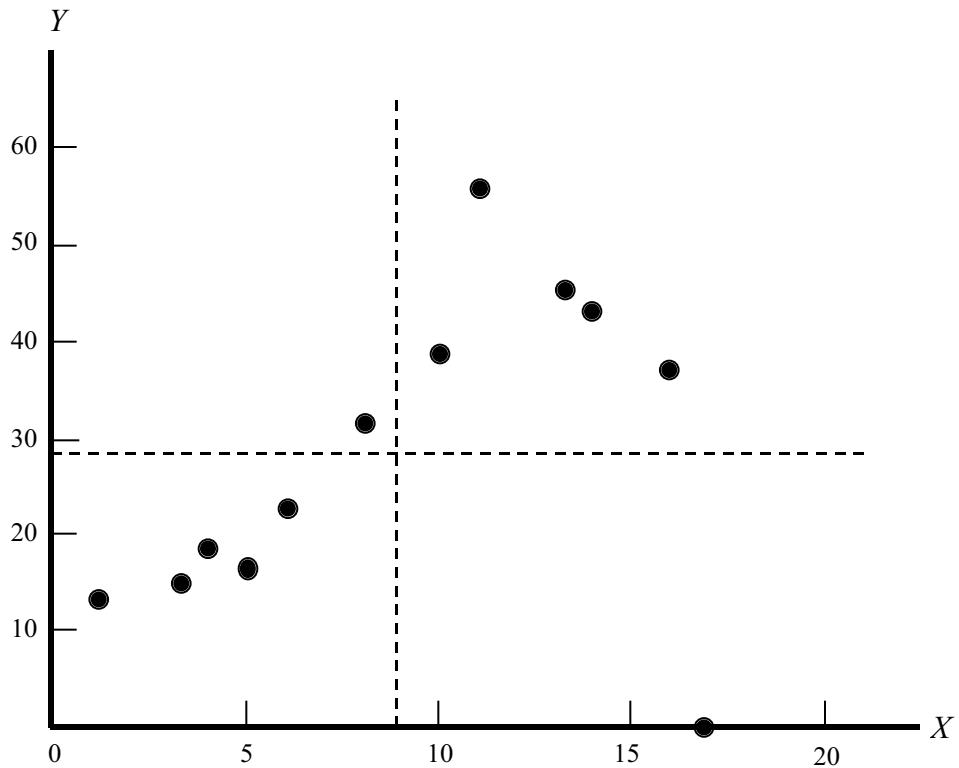
Pitman (1937b) introduced a permutation test based on the Pearson coefficient for continuous data that may be used for inference. Like other permutation tests for continuous data such as those for the one- and two-sample problems considered in [Chapters 2](#) and [5](#) the test is a conditional test and fresh computation of the permutation distribution is required for each data set. For testing  $H_0: \rho = 0$  against a one-sided alternative,  $H_1: \rho > 0$ , say, or a two-sided alternative  $H_1: \rho \neq 0$ , an exact permutation  $P$ -value may be

calculated. Modern computer software makes the computation feasible for small samples. For larger samples Monte Carlo or asymptotic approximations are available in, for example, StatXact. The appropriate test is based on the fact that if we fix the order of the  $x$  values then all  $n!$  permutations of the  $y$  values are equally likely under an hypothesis of independence or lack of association between  $X$  and  $Y$ . Permutations giving values of  $r$  near to  $-1$  or  $+1$  ( $\pm 1$ ) indicate strong evidence against the hypothesis  $H_0: \rho = 0$ . This test has similar disadvantages to other permutation tests based on raw scores or continuous data; not only are fresh computations of the permutation distribution, or at least of relevant tail probabilities, required for each data set but also the test lacks robustness. However, when normality assumptions hold, the Pitman efficiency of the test is 1. It is easily seen from (7.1) that the permutation distribution of  $r$  corresponds to that of  $\sum_i(x_iy_i)$ , since the other quantities in (7.1) are all invariant under permutation of the  $y_i$ . The test is seldom used in practice so we once again omit details but the interested reader will find a simple numerical example in Sprent (1998, Section 9.1).

### 7.1.2 Other measures of bivariate association

Desirable properties of any correlation coefficient are that its values should be confined to the interval  $(-1,1)$  and that lack of association implies a value near zero. Values near  $+1$  should imply a strong positive association and values near  $-1$  imply a strong negative association. For the Pearson coefficient,  $r = \pm 1$  implies linearity, but for the rank coefficients we introduce below values of  $\pm 1$  need not, and usually do not, imply linearity in continuous data from which the ranks may have been derived. Rather, we are interested in what for continuous data is known as **monotonicity**. If  $x$  and  $y$  increase together, this is a monotonic increasing relationship whereas if  $y$  decreases as  $x$  increases the relationship is monotonic decreasing. For rank correlation the value  $+1$  implies strictly increasing monotonicity, the value  $-1$  strictly decreasing monotonicity.

Rank correlation coefficients are also relevant to, and indeed were originally developed for use in, situations where there is no underlying continuous measurement scale, but where the ranks simply indicated order of preference expressed by two assessors for a group of objects, e.g. contestants in a diving contest or different brands of tomato soup in a tasting trial.



**Figure 7.1** Scatter diagram for data in Table 7.1. The broken lines at right angles represent axes when the origin is transferred to the  $x$ ,  $y$  sample medians.

**Table 7.1** A data set showing scores  $x$ ,  $y$  for two examination questions,  $Q_1$  and  $Q_2$ , on the same paper.

$Q1$	$x$	1	3	4	5	6	8	10	11	13	14	16	17
$Q2$	$y$	13	15	18	16	23	31	39	56	45	43	37	0

We illustrate some basic concepts in nonparametric correlation using the artificial but realistic data in Table 7.1. This gives for 12 candidates the scores achieved in an examination paper consisting of a short question marked out of 20 and a long question marked out of 60. To a certain extent the two marks are positively associated. However, some students concentrated their revision on the material for Question 1 ( $Q_1$ ), which they answered well, and performed less well on Question 2 ( $Q_2$ ). One student only revised the material for  $Q_1$  and did not even attempt  $Q_2$ . Figure 7.1 is a scatter diagram for these data.

Figure 7.1 shows the tendency for the  $Q_2$  marks to increase with the  $Q_1$  marks up to a certain point, beyond which the  $Q_2$  marks tend to decrease. Other examples of this type of relationship can be found

in a biological context. A measured response often increases broadly in accord with increasing level of a stimulus, up to a maximum beyond which the rate of increase may fall off, perhaps due to a toxic effect associated with an overdose of a potentially beneficial treatment. For the data in [Table 7.1](#), the Pearson coefficient is  $r = 0.373$ . This is not significantly different from zero at the 5 per cent level even in a one-tail test (exact one-tail  $P \approx 0.116$ ) if we assume the sample is from a bivariate normal distribution, but that assumption is hardly realistic for these data. This lack of significance of the Pearson coefficient is counter to an intuitive feeling that there is evidence of a reasonably high degree of association in the data. Not unexpectedly, because it lacks robustness, the Pitman permutation test for these data gives an exact one-tail  $P = 0.112$ , close to the value assuming normality.

In [Sections 7.1.3](#) to [7.1.5](#) we consider three nonparametric measures of correlation and see how each behaves with the data in [Table 7.1](#). We present the coefficients in a way that shows relationships between these measures of correlation and some concepts introduced in earlier chapters.

Exact tests are based on appropriate permutations of data, ranks or counts. For tests based on ranks the simplest permutation procedure is to fix the rank order associated with one variable, conventionally  $x$ , and to calculate the value of the chosen coefficient under all equally likely permutations of values of ranks of the other variable,  $y$ . The probabilities of the observed or more extreme values of the chosen coefficient (or some linear function of it) determine the size of the relevant critical region for testing for zero correlation.

### **7.1.3 The Spearman rank correlation coefficient**

The rank correlation coefficient that bears his name was proposed by Spearman (1904) and is computationally equivalent to the Pearson coefficient calculated for ranks (in place of the original continuous data if these are given). The coefficient is often denoted by the Greek letter  $\rho$  (rho), and referred to as **Spearman's rho**. Because  $\rho$  is also used for the correlation coefficient in bivariate normal distributions (estimated by the Pearson coefficient  $r$ ) to avoid confusion we denote an estimate of the **Spearman coefficient** by  $r_s$  and the corresponding population value by  $\rho_s$ . The formula [\(7.1\)](#) may be used to calculate  $r_s$  if we replace  $(x_i, y_i)$  by their ranks  $(r_i, s_i)$ . If there are no ties a simpler formula can be obtained by

straightforward algebraic manipulation using properties of sums of ranks and sums of squares of ranks. The formula is

$$r_s = 1 - \frac{6T}{n(n^2 - 1)} \quad (7.2)$$

where  $T = \sum_i (r_i - s_i)^2$ , i.e. the sum of the squares of the difference between the ranks for each sample pair. If the  $x$  and  $y$  ranks are all equal for each individual, i.e.  $r_i = s_i$  for all  $i$ , then clearly  $T = 0$  and  $r_s = 1$ . If there is a complete reversal of ranks, tedious elementary algebra establishes that  $r_s = -1$ . If there is no correlation between ranks it can be shown that  $E(T) = n(n^2 - 1)/6$ , so that  $r_s$  has expected value zero. If the observations are a random sample from a bivariate distribution with  $X, Y$  independent we expect near-zero values for  $r_s$ .

Tables giving critical values at nominal 5 and 1 per cent significance levels for testing  $H_0: \rho_s = 0$  against one- and two-sided alternatives are widely available [see e.g. Hollander and Wolfe, (1999, Table A31) or Neave (1981, p. 40)]. For not too large samples StatXact provides exact  $P$ -values for this test with the choice of an asymptotic or a Monte Carlo estimate of  $P$  for larger samples. Since  $r_s$  is a monotonic function of  $T$  when  $n$  is fixed we may use  $T$  itself as a test statistic and critical values of  $T$  have been tabulated for small  $n$ . We prefer to use  $r_s$  because it gives a more easily appreciated indication of the level of association; it complies with the convention that possible values of a correlation coefficient should lie in the interval  $[-1, 1]$ .

Both this coefficient and one we develop in [Section 7.1.4](#) are often used to test whether there is broad equivalence between ranks or orderings assigned by different assessors. Do two examiners concur in their ranking of candidates? Do two judges agree in the placings of  $n$  competitors in a diving contest? Do job applicants' ranks for manual skill based on a psychological test show a relationship to their rankings in a further test for mathematical skills?

These coefficients are also appropriate for tests of trend. Does  $Y$  increase (or decrease) as  $X$  increases? Such tests are often more powerful than the Cox–Stuart test described in [Section 3.2.3](#).

### *Example 7.1*

*The problem.* Compute the Spearman coefficient for the data in [Table 7.1](#) and use your result to test  $H_0: \rho_s = 0$  against the one-sided alternative  $H_1: \rho_s > 0$ .

*Formulation and assumptions.* The coefficient is computed using (7.2) and tables or appropriate software are used to assess the evidence.

**Table 7.2** Paired ranks and their differences,  $d$ , for the data in [Table 7.1](#).

<i>x-rank</i>	1	2	3	4	5	6	7	8	9	10	11	12
<i>y-rank</i>	2	3	5	4	6	7	9	12	11	10	8	1
$d$	-1	-1	-2	0	-1	-1	-2	-4	-2	0	3	11

*Procedure.* To compute  $T$  in (7.2) we replace the data in [Table 7.1](#) by the paired ranks in Table 7.2.  $T$  is calculated by taking the difference between ranks,  $d$ , in each column, squaring that difference, and adding the squares. Thus

$$T = (-1)^2 + (-1)^2 + (-2)^2 + \dots + (11)^2 = 1 + 1 + 4 + \dots + 121 = 162$$

whence  $r_s = 1 - (6 \times 162)/(12 \times 143) = 0.4336$ . One-tail critical values for significance at nominal 5 and 1 per cent levels given by tables are 0.503 and 0.678. For this example StatXact gives an exact one-sided  $P = 0.081$  and using an asymptotic approximation (see [Section 7.1.6](#))  $P = 0.080$ .

*Conclusion.* Since  $r_s = 0.4336$  is less than the value 0.503 required for significance at a nominal 5 per cent level and corresponds to an exact  $P \approx 0.08$  there is not strong evidence against a hypothesis of lack of monotonic association.

*Comment.* We indicated in [Section 7.1.2](#) that the exact test based on the Pearson coefficient had an associated  $P = 0.112$  which is slightly greater than that obtained here. This again reflects the influence of the curved nature of the relationship for high Q1 marks. The conceptual relationship between the Spearman and the Pearson coefficients often results in both coefficients lacking robustness when some observations are markedly out of line with the general pattern.

*Computational aspects.* Many general statistical packages have programs that compute  $r_s$  or  $T$  or an equivalent statistic, but often either leave one to look up tables or else give only an asymptotic result to establish whether there is strong evidence against  $H_0$ . If no specific program is included for the Spearman coefficient one might use the paired ranks as ‘data’ for a Pearson test. This will give the correct numerical value of the coefficient, but if such a program gives only an asymptotic  $P$ -value this will usually be calculated on the assumption that the data are a sample from a bivariate normal distribution and so it is not relevant.

The exact permutation distribution of  $r_s$  for small samples is not hard to compute and indeed StatXact will do this. The basic idea is similar to that outlined in [Section 7.1.1](#) for the permutation distribution of the Pearson coefficient. In practice one would be likely to use a facility like that in StatXact to compute the distribution, but if it were needed we illustrate the procedure for four paired ranks in [Example 7.2](#).

**Table 7.3** Possible arrangements of  $y$ -ranks for four data pairs corresponding to the  $x$ -ranks 1, 2, 3, 4 for a set of four paired observations.

$x$ -ranks	1	2	3	4	
Case Number	$y$ -ranks				$r_s$
1	1	2	3	4	1.0
2	1	2	4	3	0.8
3	1	3	4	2	0.4
4	1	3	2	4	0.8
5	1	4	2	3	0.4
6	1	4	3	2	0.2
7	2	1	3	4	0.8
8	2	1	4	3	0.6
9	2	3	1	4	0.4
10	2	3	4	1	-0.2
11	2	4	1	3	0.0
12	2	4	3	1	-0.4
13	3	1	2	4	0.4
14	3	1	4	2	0.0
15	3	2	1	4	0.2
16	3	2	4	1	-0.4
17	3	4	1	2	-0.6
18	3	4	2	1	-0.8
19	4	1	2	3	-0.2
20	4	1	3	2	-0.4
21	4	2	1	3	-0.4
22	4	2	3	1	-0.8
23	4	3	1	2	-0.8
24	4	3	2	1	-1.0

### Example 7.2

*The problem.* Given four paired observations calculate the exact permutation distribution of the Spearman correlation coefficient  $r_s$ . Assume there are no ties.

*Formulation and assumptions.* It is easily seen that there are  $4! = 24$  possible data pairings. For each such pairing we may compute the value of  $r_s$  and since all pairings are equally likely under a hypothesis of no association each has an associated probability of  $1/24$ .

*Procedure.* The number of possible pairings is obtained by noting that the  $y$ -rank associated with the  $x$ -data value that is ranked 1 may take any of the four values 1, 2, 3, or 4. Once we note which  $y$ -rank is associated with the  $x$ -rank 1 there are three remaining  $y$ -ranks that may be associated with the  $x$ -rank 2. This leaves 2 possible  $y$ -ranks for association with the  $x$ -rank 3 and only 1  $y$ -rank for

association with the  $x$ -rank 4. Thus, as already stated, there are  $4 \times 3 \times 2 \times 1 = 4! = 24$  possible arrangements. These are given in [Table 7.3](#) together with the value of  $r_s$  computed for each.

From the last column of [Table 7.3](#) we easily confirm that the possible values of  $r_s$  and the numbers of times each occurs are those given below. Since each of the 24 cases is equally likely to occur division of these numbers by 24 gives  $Pr$ , the probability of occurrence of each possible value of  $r_s$  under the hypothesis  $H_0: \rho_s = 0$ .

$r_s$	-1	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0
<i>Occurs</i>	1	3	1	4	2	2	2	4	1	3	1
<i>Pr</i>	1/24	1/8	1/24	1/6	1/12	1/12	1/12	1/6	1/24	1/8	1/24

*Conclusion.* The distribution of  $r_s$  under the null hypothesis is discrete and takes only 11 possible values in the interval  $[-1, 1]$  with the given probabilities.

*Comment.* The distribution is clearly symmetric about zero, and this property holds for all values of  $n$ .

*Computational aspects.* The program for the Spearman coefficient in StatXact will generate the complete distribution for any specified value of  $n$  that is not too large. For  $n = 10$  the discontinuities are appreciably smaller than those in this example, and the statistic takes 166 distinct values in the interval  $[-1, 1]$ .

Ties in data are replaced by mid-ranks. To obtain exact  $P$ -values with ties requires a program like that in StatXact although the value of  $r_s$  may be computed by applying the Pearson coefficient formula (7.1) to these mid-ranks. Thomas (1989) showed that if (7.2) is used to compute  $r_s$  when there are ties represented by mid-ranks it gives a value that is always greater than or equal to the correct value given by the Pearson formula. We consider an asymptotic approximation in [Section 7.1.6](#).

#### 7.1.4 The Kendall correlation coefficient

We use methods reminiscent of the Mann–Whitney scoring procedure for the WMW test to obtain another measure for monotonic association. Inferences involve a coefficient known as the **Kendall correlation coefficient**, or **Kendall's tau**, which is closely related to the Jonckheere–Terpstra test statistic.

In practice, as in the Jonckheere–Terpstra test, we need not replace continuous data by ranks, although the test may also be carried out using ranks. Computation is simplified if we assume the  $x$  data are arranged in ascending order. We may use either data in a form like that in [Table 7.1](#) where the  $x$  are already in ascending order

or the corresponding ranks in [Table 7.2](#). It is often convenient to use the labels  $x$ -ranks and  $y$ -ranks to refer respectively to the ranks associated with the first and second members of a bivariate pair  $(x, y)$ , even when our basic data are themselves ranks only and there is no specific underlying continuous variable (e.g. in the case of taste preferences or two judges ranking a number of competitors in a sporting event like ice skating or high diving).

Denoting the ranks of  $x_i, y_i$  by  $r_i, s_i$  respectively, the Kendall coefficient is based on the principle that if there is association between the ranks of  $x$  and  $y$ , then if we arrange the  $x$ -ranks in ascending order (i.e. so that  $r_i = i$ ) the  $s_i$  should show an increasing trend if there is positive association and a decreasing trend if there is negative association. Kendall (1938) therefore proposed that after arranging observations in increasing order of  $x$ -ranks, we score each paired difference  $s_j - s_i$  for  $i = 1, 2, \dots, n - 1$  and  $j > i$  as +1 if this difference is positive and as -1 if negative. Kendall called positive and negative differences **concordances** and **discordances** respectively. Denoting the numbers of concordances and discordances by  $n_c, n_d$  respectively, Kendall's tau, which we shall denote by  $t_k$ , although historically it has usually been denoted by the Greek letter  $\tau$ , is

$$t_k = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (7.3)$$

We distinguish between this sample estimate  $t_k$  and the corresponding population value that we shall denote by  $\tau_k$ . If we are comparing ranks assigned by two judges to a finite set of objects such as ten different brands of tomato soup this is essentially the entire population of interest, so our calculated coefficient is a measure of agreement between the judges in relation to that population. On the other hand, if we had a sample of  $n$  bivariate observations from some continuous distribution (of measurements, say) and calculate  $t_k$  using (7.3) this is an estimate of some underlying  $\tau_k$  that is a measure of the degree, if any, of monotonic association, or dependence between the variables  $X, Y$  in the population. Since there are  $\frac{1}{2}n(n-1)$  pairs  $s_j - s_i$ , if all are concordances  $n_c = \frac{1}{2}n(n-1)$  and  $n_d = 0$ , whence  $t_k = 1$ . Similarly, if all are discordances, then  $t_k = -1$ . If the rankings of  $x$  and  $y$  are independent we expect a fair mix of concordances and discordances and  $t_k$  should be close to zero. When there are no ties in the rankings  $n_c + n_d = \frac{1}{2}n(n-1)$ . Although the coefficient is often referred to as a

**rank** correlation coefficient it is clear that to compute it we only need to know the order of the relevant  $x$  and  $y$  and not the actual ranks as we did for the Spearman coefficient.

### Example 7.3

*The problem.* Compute Kendall's tau for the data in [Table 7.1](#).

*Formulation and assumptions.* We may compute  $n_c$  from the  $y$ -ranks in [Table 7.2](#). We may also obtain  $n_d$  this way, or deduce it from  $n_d = \frac{1}{2}n(n - 1) - n_c$ .

*Procedure.* To count the number of concordances we inspect the  $y$ -ranks in [Table 7.2](#), noting for each successive rank the number of succeeding ranks that are greater. The first  $y$ -rank is 2 and this is succeeded by 10 greater ranks, namely 3, 5, 4, 6, 7, 9, 12, 11, 10, 8. Similarly, the next  $y$ -rank, 3, gives 9 concordances. Proceeding this way it is easily verified that the total number of concordances is

$$n_c = 10 + 9 + 7 + 7 + 6 + 5 + 3 + 0 + 0 + 0 + 0 = 47.$$

Since  $n = 12$  we easily deduce that  $n_d = \frac{1}{2} \times 12 \times 11 - 47 = 19$ . This may be verified by counting discordances directly from [Table 7.2](#).

*Conclusion.* For the given data  $t_k = (47 - 19)/66 = 0.4242$ .

*Comments.* 1. Since the  $x$  are ordered in [Table 7.1](#), the order of the corresponding  $y$  values is a one-to-one ordered transformation of the ranks in [Table 7.2](#). Thus we could equally well have computed the number of concordances and discordances by noting the signs of all differences  $y_j - y_i$  in [Table 7.1](#) as these signs correspond to those of the  $s_j - s_i$  above.

2. As noted above the counting procedure for concordances is reminiscent of the Mann—Whitney counts for the WMW test, and we show below equivalence to the Jonckheere—Terpstra test.

3. We discuss a test of significance for  $H_0: \tau_k = 0$  in [Example 7.4](#).

The  $x_i$  are in ascending order in [Table 7.1](#); though not essential, this simplifies computation. More generally, the number of concordances is the number of positive  $b_{ij}$  among all  $\frac{1}{2}n(n - 1)$  of the  $b_{ij} = (y_j - y_i)/(x_j - x_i)$ ,  $i = 1, 2, \dots, n - 1$  and  $j > i$ . The number of negative  $b_{ij}$  is the number of discordances.

Equivalence to a Jonckheere—Terpstra test is evident if we arrange the  $x$  in ascending order as in [Table 7.1](#) and regard the  $x_i$  (or the associated ranks) as indexing  $n$  ordered samples each of 1 observation where that observation is the corresponding  $y$  value. Clearly the number of concordances is the sum of the Mann—Whitney sample pairwise statistics used in the Jonckheere—Terpstra test in [Section 6.2.5](#). Thus, in theory, any program for the exact Jonckheere—Terpstra test may be used for Kendalls coefficient to test  $H_0: \text{no association between } x \text{ and } y \text{ ranks}$  against either a one-

or two-sided alternative. In practice, however, most programs for the Jonckheere—Terpstra test will, in this situation, only give exact tail probabilities for very small  $n$ . Asymptotic results, or Monte Carlo estimates, of the exact tail probability must be used with this approach if the samples are moderate to large. We consider more specific asymptotic results in [Section 7.1.6](#). Programs in general statistical packages often calculate  $t_k$  directly, some indicating significance levels (usually asymptotic), others leaving the user to determine these at a nominal level from tables. Extensive tables are given, for example, by Neave (1981, p. 40) and by Hollander and Wolfe (1999, Table A30). Some tables give equivalent critical values of  $n_c - n_d$  for various  $n$ . However, calculating  $t_k$  gives a better feel for the degree of association implied by any correlation coefficient. When there are no ties tests may be based on  $n_c$  only, for it is clear from (7.3) that if  $n$  is fixed,  $t_k$  is a linear function of  $n_c$ . StatXact includes a program that computes exact  $P$ -values for the test of no association based on this coefficient for small to moderate sample sizes.

In passing it is worth noting one difference between the usual scoring systems for the Kendall coefficient and that used for Jonckheere—Terpstra, namely that whereas the former uses 1, 0, —1 for concordances, ties and discordances, the latter uses 1,  $\frac{1}{2}$ , 0. This is not a fundamental difference because providing certain obvious adjustments are made to relevant formulae the scoring systems may be interchanged without altering our conclusions.

#### *Example 7.4.*

*The problem.* For the data given in [Table 7.1](#) use Kendall's tau to test  $H_0$ : no association against  $H_1$ : positive association between  $x$  and  $y$ .

*Formulation and assumptions.* The appropriate test is a one-tail test of  $H_0: \tau_k = 0$  against  $H_1: \tau_k > 0$ .

*Procedure.* In [Example 7.3](#) we found  $t_k = 0.4242$ , also showing that  $n_c - n_d = 47 - 19 = 28$ . Tables give nominal 5 and 1 per cent critical values for significance when  $n = 12$  in a one-tail test as 0.3929 and 0.5455. Corresponding values for  $n_c - n_d$  are 26 and 36. More precisely, StatXact for these data gave an approximate Monte Carlo estimate of the exact one-tail  $P$  of 0.032.

*Conclusion.* There is moderately strong evidence against the hypothesis of no association. In a formal significance test one would reject  $H_0$  at a nominal 5 per cent significance level and accept a positive association between the ranks of  $x$  and  $y$ .

*Comments.* 1. Had the alternative hypothesis been  $H_1: \tau_k \neq 0$ , requiring a two-tail test we would not have rejected  $H_0$  at the 5 per cent significance level.

2. The evidence of association provided by Kendall's tau is stronger than that provided by Spearman's rho. We give a likely explanation in [Section 7.1.7](#).

*Computational aspects.* As indicated above, most general statistical programs do not give exact  $P$ -values or significance levels for relevant hypothesis tests. That in StatXact gives exact results for small samples, but a Monte-Carlo approximation is needed for moderate to large samples.

For small samples we may calculate the exact distribution of  $t_k$  in a way like that for the Spearman coefficient in [Example 7.2](#).

### *Example 7.5*

*The problem.* Calculate the exact permutation distribution of Kendall's  $t_k$  when  $n = 4$ , under the null hypothesis of no association.

*Formulation and assumptions.* If the  $x$ -ranks are in ascending order, then for each of the  $4! = 24$  equally likely permutations of the  $y$ -ranks we count the number of concordances,  $n_c$ . When  $n = 4$ ,  $n_c + n_d = 6$  whence, given  $n_c$ , we easily find  $n_d$ . We then compute  $t_k$  using (7.3) and count the number of times each value of  $t_k$  occurs. Division of each number of occurrences by 24 gives the probability of observing each value.

*Procedure.* In [Exercise 7.1](#) we ask for computation of  $t_k$  for each of the 24 possible permutations of 1, 2, 3, 4. For example, for the permutation 4, 2, 1, 3 we easily see that  $n_c = 2$ , whence (7.3) gives  $t_k = (2 - 4)/6 = -0.333$ . It is easy to verify that 4 other permutations give the same  $n_c$ , whence  $\Pr(t_k = -0.333) = 5/24$ .

*Conclusion.* Completing [Exercise 7.1](#) gives the following probabilities associated with each of the 7 possible values of  $t_k$ :

$t_k$	1	0.67	0.33	0	-0.33	-0.67	-1
Probability	1/24	1/8	5/24	1/4	5/24	1/8	1/24

*Comments.* 1. The distribution of  $t_k$  is symmetric.

2. When  $n = 4$ , a  $P$ -value less than 0.05 can only be obtained in a one-tail test corresponding to  $t_k = \pm 1$  for which  $P = 1/24 \approx 0.042$ . Samples of this size are of little use in practice for inference.

*Computational aspects.* StatXact will calculate complete distributions for small samples. The Monte Carlo option in that program may be used to estimate  $P$ -values rapidly and with good accuracy for quite large samples.

Modifications are needed to calculate and use  $t_k$  when there are rank ties. With ties (7.3) no longer takes values +1 or -1 even when all mid-rank pairs lie on a straight line. Ties among the  $x$  or  $y$  are both scored zero in Kendall's coefficient because they are neither concordances nor discordances. Now (7.3) can no longer take the value  $\pm 1$  because  $|n_c + n_d|$  is necessarily less than  $1/2n(n - 1)$ . It can be shown (see e.g. Kendall and Gibbons, 1990, Chapter 3) that we

obtain a coefficient that takes values in the interval  $[-1, 1]$  if we replace the denominator in (7.3) by  $\sqrt{[(D - U)(D - V)]}$  where

$$D = \frac{1}{2}n(n - 1), U = \frac{1}{2}\sum u(u - 1), V = \frac{1}{2}\sum v(v - 1)$$

and  $u, v$  are the number of consecutive ranks in a tie within the  $x$ - and the  $y$ -ranks respectively, and the summations in  $U, V$  are over all sets of tied ranks. Clearly if there are no tied ranks  $U = V = 0$  and the modified denominator reduces to that in (7.3). We denote this modified statistic by  $t_b$  and it is often called **Kendall's tau-b**. If there is an exact linear relationship between mid-ranks  $t_b$  takes either the value +1 or -1. A simple example with heavy tying illustrates some of the above points.

#### *Example 7.6*

Two judges are each asked to arrange 12 Cabernet-Sauvignon wines in order of preference. Both agree on which wine they rank 1 (or best). There are 4 wines that they both agree are almost as good but neither can separate within that group, so each gives these a tied rank 3.5, while there are a further 5 wines that they both classify as fairly good but neither judge can separate the wines within that group. Each of these wines is thus given a mid-rank 8. Both also agree the remaining 2 wines are equally bad and thus are ranked 11.5. Thus the mid-ranks allocated are those given below

Judge A	1	3.5	3.5	3.5	3.5	8	8	8	8	11.5	11.5
Judge B	1	3.5	3.5	3.5	3.5	8	8	8	8	11.5	11.5

Clearly there is a linear relationship between these mid-ranks pairs since they all lie on a straight line through the origin with unit slope. Since the ranks awarded by Judge A (i.e.  $x$ -ranks) are ordered,  $n_c$ , the number of concordances, is obtained by counting the number of positive  $s_j - s_i$ ,  $j > i$ . This is easily found by a direct (but careful) counting process, (see [Exercise 7.12](#)) to be 49, whence (7.3) gives  $t_k = 0.7424$  since there are no discordances. Also in [Exercise 7.12](#) we ask you to verify that  $U = V = \frac{1}{2}(4 \times 3 + 5 \times 4 + 2 \times 1) = 17$ , whence

$$t_b = 49/\sqrt{(66 - 17)^2} = 1.$$

Here  $t_b$  exhibits the desirable property of a correlation coefficient, i.e.  $t_b = 1$ , when there is complete ranking agreement, whereas  $t_k$  does not.

The next example covers a less extreme situation involving ties.

#### *Example 7.7*

*The problem.* Life expectancy showed a general tendency to increase during the nineteenth and twentieth centuries as standards of health care and hygiene improved. The extra life expectancy varies between countries, communities and even families. [Table 7.4](#) gives the year of death and age at death for 13 males in a clan we call the McDelts buried in the Baden scallie burial ground in Wester

**Table 7.4** Year of death and ages of 13 McDeltas.

Year	1827	1884	1895	1908	1914	1918	1924	1928	1936	1941	1964	1965	1977
Age	13	83	34	1	11	16	68	13	77	74	87	65	83

Ross, Scotland (see [Appendix](#)). Is there an indication that life expectancy is increasing for this clan in more recent years?

*Formulation and assumptions.* If there is an increase in life expectancy those dying later will tend to be older than those dying earlier. Years of death,  $x$ , are already arranged in ascending order so we may count the number of concordances and discordances by examining all pairs of  $y$  and scoring each as appropriate (1 for concordance, -1 for discordance, 0 for a tie).

*Procedure.* There are no ties in the  $x$ . The first  $y$  entry is 13. In the following entries this is exceeded in nine cases (concordances); there are two discordances and one tie. Similarly for the second  $y$  entry (83) there is one concordance, nine discordances and one tie. Proceeding this way we find that

$$n_c = 9 + 1 + 6 + 9 + 8 + 6 + 4 + 5 + 2 + 2 + 0 + 1 = 53$$

and

$$n_d = 2 + 9 + 4 + 0 + 0 + 1 + 2 + 0 + 2 + 1 + 2 + 0 = 23.$$

Since  $n = 13$ , (7.3) gives  $t_k = (53 - 23)/(1/2 \times 13 \times 12) = 0.3846$ . There are two tied values at each of  $y = 13$  and at  $y = 83$ . Using the formula for the denominator adjustment we have  $D = 78$ ,  $U = 0$  and  $V = 1/2 \times 2 \times 1 + 1/2 \times 2 \times 1 = 2$ , hence

$$t_b = (53 - 23)/\sqrt{[78 \times (78 - 2)]} = 0.3896.$$

This small difference between  $t_k$  and  $t_b$  does not affect our conclusion. StatXact gives Monte Carlo estimates of the exact  $P$  and in three simulations of 10 000 samples gave estimated  $P = 0.0359, 0.0399$  and  $0.0381$ , suggesting  $P < 0.04$ .

*Conclusion.* There is reasonable evidence against the hypothesis  $H_0$ : *no increasing trend of life expectancy* and favouring the one-tail alternative that expectancy has increased during the period covered.

*Comments.* 1. If there are only a few ties and we ignore them and use (7.3) the test is conservative, for clearly if we allow for ties by using  $t_b$  the denominator is less than that in (7.3).

2. A one-tail test is justified because we are interested in whether data for the McDelta clan follow established trends in most developed countries.

3. Care is needed in counting concordances if there are ties in both  $x$  and  $y$ . For example, for the data set:

$x$	2	5	5	7	9
$y$	5	1	3	8	1

the first pair (2, 5) gives one concordance (7, 8) and three discordances; the second pair (5, 1) gives one concordance (7, 8) and no discordances because of the tied values in  $x$  at (5, 3) and in  $y$  at (9, 1); the next pair (5, 3) gives one concordance (7, 8) and one discordance (9, 1).

*Computational aspects.* With more than a few ties a program that gives at least a Monte Carlo estimate of the relevant tail probabilities is especially useful.

For moderately large samples with only a few ties, use of (7.3) and tables of critical values for the no-tie situation or the use of asymptotic results given in [Section 7.1.6](#) should not be seriously misleading.

We have so far established only that it is intuitively reasonable to expect Kendall's  $t_k$  to take values near zero when the  $(x_i, y_i)$  are sample values from a bivariate population where the random variables  $X$  and  $Y$  associated with that population are independent, implying a population  $\tau_k = 0$  (though the converse may not be true).

This leaves open the question of what precisely the sample Kendall coefficient is estimating when  $\tau_k \neq 0$ . We now give a probabilistic interpretation of  $\tau_k$  which leads naturally to  $t_k$  as a reasonable sample estimator for it whether or not  $X$  and  $Y$  are independent and establishes the ‘parameter’  $\tau_k$  as a reasonable measure of a certain type of dependence between  $X$  and  $Y$ . Further, the measure is distribution-free as it makes only the assumption that we are sampling from some unspecified bivariate distribution.

We base our argument on the fact that if  $(X_s, Y_s)$  and  $(X_t, Y_t)$  are independent samples from a population with some joint cumulative distribution function  $F_{XY}(x, y)$ , then each has that same distribution.

We consider some properties of a variable  $B_{ts} = (Y_s - Y_t)/(X_s - X_t)$ . If  $B_{ts} = (Y_s - Y_t)/(X_s - X_t) > 0$  this implies that either the joint event  $(Y_s > Y_t \text{ and } X_s > X_t)$  or the joint event  $(Y_s < Y_t \text{ and } X_s < X_t)$  has occurred and since these are mutually exclusive this in turn implies that

$$\Pr(B_{ts} > 0) = \Pr[(Y_s - Y_t)/(X_s - X_t) > 0] = \frac{\Pr(Y_s > Y_t \text{ and } X_s > X_t)}{\Pr(Y_s > Y_t \text{ and } X_s > X_t) + \Pr(Y_s < Y_t \text{ and } X_s < X_t)}.$$

If also  $X$  and  $Y$  are independent we have

$$\Pr(Y_s > Y_t \text{ and } X_s > X_t) = \Pr(Y_s > Y_t) \Pr(X_s > X_t). \quad (7.4)$$

Because  $Y_s$  and  $Y_t$  both have the marginal distribution of  $Y$  it follows that  $\Pr(Y_s > Y_t) = 1/2$ . Similarly  $\Pr(X_s > X_t) = 1/2$ , whence from (7.4) independence implies  $\Pr(Y_s > Y_t \text{ and } X_s > X_t) = 1/2 \times 1/2 = 1/4$ . Similarly under independence  $\Pr(Y_s < Y_t \text{ and } X_s < X_t) = 1/4$  which in turn implies that under independence

$$\Pr(B_{ts} > 0) = \Pr[(Y_s - Y_t)/(X_s - X_t) > 0] = 1/4 + 1/4 = 1/2.$$

If we now define the population Kendall  $\tau_k$  to be

$$\tau_k = 2\Pr[B_{ts} > 0] - 1 \quad (7.5)$$

it follows that under independence  $\tau_k = 2 \times \frac{1}{2} - 1 = 0$ . A moment's reflection indicates that if the departures from independence are such that high values of  $Y$  tend to be associated with high values of  $X$  and low values of  $Y$  with low values of  $X$  then

$$\Pr(B_{ts} > 0) = \Pr[(Y_s - Y_t)/(X_s - X_t) > 0]$$

will be greater than  $\frac{1}{2}$ , and will equal 1 if  $Y_s > Y_t$  always implies  $X_s > X_t$  in which case from (7.5)  $\tau_k = +1$ . A similar argument shows that if high values of  $Y$  tend to be associated with low values of  $X$  and low values of  $Y$  with high values of  $X$  then  $\tau_k$  will be negative and for a complete reversal of rank orders  $\tau_k = -1$ .

In the discussion following Example 7.3 we showed that if for all pairs of  $n$  sample values  $(x_i, y_i), (x_j, y_j)$  we formed the quotients  $b_{ij} = (y_j - y_i)/(x_j - x_i)$ ,  $i = 1, 2, \dots, n-1$ , for all  $j > i$ , then a positive value of this quotient corresponds to a concordance and a negative value to a discordance and that in the notation used in (7.3)  $n_c/(n_c + n_d)$  provides a sensible estimate of  $\Pr[(Y_s - Y_t)/(X_s - X_t) > 0]$  in (7.5). Further in the no-tie case since  $n_c + n_d = n(n-1)$  it is easily verified that substitution of  $n_c/(n_c + n_d)$  in (7.5) leads to (7.3). Kerridge (1975) gives an interesting example of an estimation problem with this probabilistic interpretation.

It should be realized that the type of dependence described above is by no means the only possible kind. For example, it is not uncommon for high values of  $Y$  to go with both high values of  $X$  and low values of  $X$  while intermediate values of  $Y$  go with intermediate values of  $X$ . At its simplest this might represent a near quadratic relationship between  $X$  and  $Y$ . The correlation measures described in this chapter are generally unsuitable for detecting this or many other possible kinds of nonlinear dependence.

### 7.1.5 A median test for correlation

Blomqvist (1950; 1951) proposed a test that embodies similar concepts to the median test considered in Section 5.3. Suppose the medians of the marginal distribution of  $X, Y$  are respectively  $\theta_x, \theta_y$ . In a sample we expect about half the observed  $x$  to be below  $\theta_x$  and about half the  $y$  to be below  $\theta_y$ . If  $X$  and  $Y$  are independent we expect a good mix of high, medium and low values of  $Y$  to be associated with the various values of  $X$ . If we knew  $\theta_x$  and  $\theta_y$  we could shift the origin in a scatter diagram of the observed  $(x_i, y_i)$  to  $(\theta_x, \theta_y)$ . Then under  $H_0: X, Y$  are independent we expect about one quarter of all

points to be in each of the four quadrants determined by the new axes. Usually we do not know  $\theta_x$ ,  $\theta_y$ , so Blomqvist proposed replacing them by the sample medians  $M_x$ ,  $M_y$  as reasonable estimates. In [Figure 7.1](#) axes with a new origin at  $(M_x, M_y)$  are shown by broken lines, and working anticlockwise from the first (or top right) quadrant the numbers of points in the respective quadrants are 5, 1, 5, 1. The concentration of points in the first and third quadrants suggests dependence between  $X$  and  $Y$ . We may formalize a test of  $H_0$  against a one-sided alternative of either *positive* or *negative* association, or a two-sided alternative of *some association*, following closely the procedure in [Section 5.3.1](#). In general, if  $n$  is even, denoting the sample medians for  $X$ ,  $Y$  by  $M_x$ ,  $M_y$  and assuming no values coincide with these medians, we count the numbers of pairs  $a$ ,  $b$ ,  $c$ ,  $d$  in each of the four quadrants determined after this shift of the scatter plot origin to  $(M_x, M_y)$ . Because of sample median properties we get a  $2 \times 2$  contingency table of the form shown in [Table 7.5](#).

The marginal totals follow from the properties of the sample median and restriction to even  $n$  with no observation at either  $M_x$  or  $M_y$ . If  $n$  is odd at least one sample value for  $x$  and one sample value for  $y$  will coincide with the medians, resulting in points at or on axes through the sample medians; such points are omitted from the count. This necessitates modifications that we give at the end of this section.

In the case envisaged in [Table 7.5](#), if  $X$ ,  $Y$  are independent then clearly  $a$ ,  $b$ ,  $c$  and  $d$  each has expected value  $1/4n$ . Marked departures from this value indicate association between  $X$ ,  $Y$ . The appropriate test procedure then uses the Fisher exact test. We illustrate this in [Example 7.8](#), but first we consider what is an appropriate correlation measure having the desirable properties that it takes values in the interval  $(-1, 1)$  with values near  $\pm 1$  indicating near-monotonic dependence, and independence resulting in values near zero (although the converse may not be true).

**Table 7.5** Contingency table for the median test for correlation.

	<i>Above</i> $M_x$	<i>Below</i> $M_x$	<i>Row total</i>
<i>Above</i> $M_y$	$a$	$b$	$1/2$
<i>Below</i> $M_y$	$c$	$d$	$1/2n$
<i>Column total</i>	$1/2n$	$1/2n$	$n$

Two possible coefficients are

$$r_d = (a + d - b - c)/n \quad (7.6)$$

and

$$r_m = 4(ad - bc)/n^2 \quad (7.7)$$

It is easily verified that in the special circumstances of [Table 7.5](#), where the marginal totals imply that  $a = d$ ,  $b = c$  and  $a + b = 1/2n$ , both (7.6) and (7.7) have the desirable properties for a correlation coefficient and are equivalent, each taking the same set of  $1/2n+1$  possible values with the same associated probabilities under the null hypothesis of independence. While (7.6) has arithmetic simplicity, (7.7) generalizes more readily to other contexts, including the case  $n$  odd, so it is usually preferred.

### *Example 7.8*

*The problem.* Calculate  $r_m$  (or  $r_d$ ) for the data in [Table 7.1](#) and test the hypothesis of independence against that of positive association between  $X$  and  $Y$ .

*Formulation and assumptions.* We find the sample medians  $M_x$ ,  $M_y$  and deduce the entries in [Table 7.5](#). The relevant coefficient values are computed using (7.6) or (7.7) and the test of significance is based on the Fisher exact test as in [Example 5.6](#). In this case a one-tail test is appropriate since the alternative is in a specified (positive) direction.

*Procedure.* For the data in [Table 7.1](#) the sample medians are  $M_x = 9$ ,  $M_y = 27$ , from which we deduce the values of  $a$ ,  $b$ ,  $c$ ,  $d$  in [Table 7.5](#) (most easily done by inspection of [Figure 7.1](#)) to be  $a = d = 5$  and  $b = c = 1$ . Since  $n = 12$  we easily calculate  $r_m = 2/3 = r_d$ . We use the Fisher exact test, the relevant tail probabilities being associated with the observed value  $a = 5$  and the more extreme  $a = 6$ . Calculating these probabilities using (5.3) is easy even if an appropriate computer program is not available. The relevant probability is  $P = 0.0400$ .

*Conclusion.* There is fairly strong evidence against the hypothesis of independence and favouring that of positive association. The numerical value of the appropriate median correlation coefficient is  $r_m = 0.667$ .

*Comments.* 1. The Pearson correlation coefficient was not significant for these data; it is not robust against departures from normality of the type implicit in the observation  $(17, 0)$ . However,  $r_m$  is robust and indeed we would reach the same conclusion if the point  $(17, 0)$  were replaced by any point in the fourth quadrant (positive  $x$ , negative  $y$ ) relative to axes with origin at the sample medians if there were no other data changes. In general, the median correlation test is robust against a few major departures from a monotonic or near-monotonic trend. Unfortunately, it has low Pitman efficiency compared to the Pearson coefficient test when the relevant normality assumption holds.

2. The Blomqvist test is applicable no matter what the marginal distributions of  $X$ ,  $Y$  may be.

*Computational aspects.* For all but very small samples a computer program such as that in StatXact or Testimate is desirable for the Fisher exact test. The asymptotic result based on the chi-squared test is generally quite reliable for large samples.

If  $n$  is odd, or generally if there are values equal to one or both sample medians, the marginal totals in a table like [Table 7.5](#) will no longer be  $1/2n$ . It is then appropriate to use the Fisher exact test with the marginal totals actually observed, and if a correlation coefficient is required to use a modification of  $r_m$ , namely

$$r_m = (ad - bc) / \{\sqrt{[(a + b)(c + d)(a + c)(b + d)]}\}$$

where  $a, b, c, d$  are the observed cell values. This reduces to [\(7.7\)](#) when all marginal totals are  $1/2n$ .

The coefficient  $r_m$  is easy to calculate, but in practice it is used less often than the Spearman or Kendall correlation coefficients.

Several other nonparametric tests for correlation have been suggested. Gideon and Hollister (1987) proposed a measure that is also relatively easy to calculate, although this advantage is partly negated for small  $n$  by the fact that the coefficient takes only relatively few possible values and the permutation distribution under the null hypothesis is less easily established than that for the Kendall or Spearman coefficients. The authors give fairly extensive tables of nominal critical values. They also compare the power of the test for monotonic association when using their coefficient to those using other coefficients, with generally favourable results for the former.

Dietz and Killen (1981) extend the concept of Kendall's tau to a multivariate test for trend applicable to a pharmaceutical problem.

### 7.1.6 Asymptotic results

For large  $n$ , tests using the Spearman and Kendall coefficients may be based on the distribution of functions of  $r_s$  and  $t_k$  that have an asymptotic standard normal distribution.

Detailed discussions of the distribution of  $r_s$  under  $H_0: \rho_s = 0$  are given by Gibbons and Chakraborti (1992, Section 12.3) and by Kendall and Gibbons (1990, Section 4.14). Convergence to normality of the distribution tends to be slow so the asymptotic result should not be used for values of  $n < 30$ , and some writers suggest an even higher value. For large  $n$  it is usually assumed that  $z = r_s\sqrt{(n - 1)}$  has approximately a standard normal distribution under  $H_0$ . This approximation is identical in form to one often used for the Pearson coefficient. A better approximation that is often

reasonable for values of  $n$  as low as 10 suggested in Kendall and Gibbons and elsewhere is

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

and this has approximately a  $t$ -distribution with  $n - 2$  degrees of freedom under  $H_0$ . This approximation is used in StatXact for asymptotic estimation of a  $P$ -value for both the Pearson and Spearman coefficients.

For Kendall's tau the exact distribution of  $t_k$  under  $H_0$ :  $\tau_k = 0$  is difficult to obtain. Again a full discussion is given in Gibbons and Chakraborti (1992, Section 12.2) and by Kendall and Gibbons (1990, Section 4.8) who show under  $H_0$  that  $E(t_k) = 0$  and  $\text{Var}(t_k) = 2(2n + 5)/[9n(n - 1)]$  and that for reasonably large  $n$

$$z = \frac{3t_k \sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \quad (7.8)$$

has approximately a standard normal distribution. This is the basis of the asymptotic  $P$ -value determination used in StatXact and in many general statistical programs, not all of which give exact  $P$ -values. Modifications are needed for ties and Gibbons and Chakraborti discuss these, although a few ties may make little difference. Some writers suggest that the asymptotic result for the Kendall coefficient should not be used if  $n < 15$  so it would not be recommended for the data in [Example 7.7](#). However, in that example where  $t_k = 0.3846$  we find  $z = 1.83$  implying that for a one-tail test  $P = 0.0336$ , broadly in line with the Monte Carlo approximations to the exact  $P$  in [Example 7.7](#) despite the small sample size.

A difficulty in obtaining confidence intervals for the population coefficients,  $\rho_s$  and  $\tau_k$  using asymptotic results is that the limits may lie outside the closed interval  $[-1, 1]$ . Another is that the distributions of the sample statistics are not simple for non-zero  $\rho_s$  or  $\tau_k$ . In particular, the variances are not the same as those under a zero value hypothesis. The problem is discussed further by Kendall and Gibbons (1990, Chapters 4 and 5) who give useful approximations.

### 7.1.7 A comparison of the Spearman and Kendall coefficients

Statisticians are often asked which of the coefficients – Kendall or Spearman – is to be preferred. There is no clear-cut answer. They seldom lead to markedly different conclusions, though [Examples 7.1](#)

and 7.4 show this is possible when one ranked pair is clearly out of line with the general trend. For instance, in Example 7.1 one pair of ranks, i.e. (12, 1), contributes  $(12 - 1)^2 = 121$  to a total  $T$  value of 162, suggesting  $r_s$  may be sensitive to such an outlier.

An alternative way of calculating the coefficients clarifies the relationship between them. We assume there are no ties and that the pairs of ranks are arranged in ascending order of  $x$ -ranks as in Tables 7.1 and 7.2. Writing  $s_k$  for the rank of  $y_k$  we define  $s_{ij} = s_j - s_i$ ,  $i = 1, 2, 3, \dots, n - 1$  for all  $j > i$ . Clearly, the number of concordances in Kendall's  $t_k$  equals the number of positive  $s_{ij}$ , and the number of discordances equals the number of negative  $s_{ij}$ .

We do not prove it, but tedious algebra shows that if we denote by  $n_{cs}$  the sum of the values of the positive  $s_{ij}$  (the values of differences that are concordant) and by  $n_{ds}$  the sum of the values of the negative  $s_{ij}$  (the values of differences that are discordant), then in the ‘no-tie’ case  $n_{cs} + n_{ds} = n(n^2 - 1)/6$  and  $r_s$  is given by

$$r_s = \frac{n_{cs} - n_{ds}}{n(n^2 - 1)/6} \quad (7.9)$$

which has a formal similarity to (7.3). We may write the  $s_{ij}$  as an upper triangular matrix

$$\begin{matrix} s_2 - s_1 & s_3 - s_1 & s_4 - s_1 & \cdot & \cdot & \cdot & s_n - s_1 \\ s_3 - s_2 & s_4 - s_2 & \cdot & \cdot & \cdot & s_n - s_2 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & & & & & \\ s_{n-1} - s_{n-2} & & & & & & s_n - s_{n-2} \\ & & & & & & s_n - s_{n-1} \end{matrix}$$

For the data in Table 7.2 the matrix (see Exercise 7.3) is

$$\begin{matrix} 1 & 3 & 2 & 4 & 5 & 7 & 10 & 9 & 8 & 6 & -1 \\ 2 & 1 & 3 & 4 & 6 & 9 & 8 & 7 & 5 & 5 & -2 \\ -1 & 1 & 2 & 4 & 7 & 6 & 5 & 3 & 3 & -4 \\ 2 & 3 & 5 & 8 & 7 & 6 & 4 & 2 & -3 & \\ 1 & 3 & 6 & 5 & 4 & 2 & -5 & \\ 2 & 5 & 4 & 3 & 1 & -6 & \\ 3 & 2 & 1 & -1 & -8 & \\ -1 & -2 & -4 & -11 & \\ -1 & -3 & -10 & \\ -2 & -9 & \\ -7 & \end{matrix}$$

To compute  $t_k$  using this tableau we count the number of positive entries and obtain  $n_c = 47$  and the number of negative entries,  $n_d = 19$ .

These were the values obtained in [Example 7.3](#). To obtain  $n_{cs}$  we add all positive entries in the matrix. In [Exercise 7.3](#) we ask you to verify that  $n_{cs} = 1 + 3 + 2 + 4 + \dots + 3 + 2 + 1 = 205$ . Similarly, adding all negative entries gives  $n_{ds} = 1 + 2 + 1 + 4 + \dots + 10 + 2 + 9 + 7 = 81$ . Substitution in (7.9) gives  $r_s = 6 \times (205 - 81)/(12 \times 143) = 0.4336$ , agreeing with the value in [Example 7.1](#).

Examining the triangular matrix helps explain why, in this example,  $r_s$  plays down any correlation relative to that indicated by  $t_k$ . A few negative  $s_{ij}$  in the last column make relatively large contributions to the sum  $n_{ds}$ . This is clearly attributable to the fact that the low rank value of the 12th and last observation,  $s_{12} = 1$ , is out of line with the high ranks assigned to its near neighbours.

Many other aspects of parametric correlation have nonparametric equivalents, including that of partial correlation. An introductory treatment to that topic is given in Sprent (1998, Section 9.6) and it is more fully discussed by Kendall and Gibbons (1990, Chapter 8), and by Siegel and Castellan (1988, Section 9.5). Care is needed to interpret partial correlation coefficients for ranks correctly.

### 7.1.8 Efficiency power and sample size

Various results concerning efficiency, power and sample sizes needed to ensure a required power when specific alternative hypotheses hold for both the Kendall and Spearman coefficients are scattered throughout the literature. One important result is that when sampling from the bivariate normal distribution both the Kendall and Spearman coefficients have the same Pitman efficiency of 0.912 relative to the Pearson coefficient. The Kendall coefficient tends to do rather better than the Pearson coefficient when sampling from long-tailed symmetric distributions and has Pitman efficiency of 1.266 relative to the Pearson coefficient for a bivariate double exponential distribution. Noether (1987a) gives the following asymptotic formula for the sample size needed to ensure power  $1 - \beta$  in a one-tail test of  $H_0: \tau_k = 0$  against  $H_1: \tau_k = \tau_1$  where  $\tau_1$  has some fixed nonzero value:

$$n \approx \frac{4(z_\alpha + z_\beta)^2}{9\tau_1^2}$$

Here  $z_\alpha, z_\beta$  have the fairly obvious standard normal tail probability meanings assigned to them in (4.3). Some further results on power

for the Kendall coefficient are given by Hollander and Wolfe (1999, pp. 375—376).

## 7.2 RANKED DATA FOR SEVERAL VARIABLES

If more than two observations are ranked for each of a set of  $n$  experimental units we often want to test for evidence of concordance between rankings of the units.

We indicated in [Section 6.3.2](#) that the Friedman test may be applied to rankings of objects (preferences for different varieties of raspberry, placings in a gymnastics contest by different judges, ranking of candidates by different examiners) to test whether there is evidence of consistency between those making the rankings. Kendall, independently of Friedman, proposed the use of a function of the Friedman statistic which is often referred to as the **Kendall coefficient of concordance** and tabulated some small-sample critical values relevant to testing the hypothesis that the rankings were essentially random against the alternative of evidence of consistency. Kendall regarded his coefficient of concordance as an extension of the concept of correlation to more than two sets of rankings. Whereas in the bivariate case we may use the Kendall correlation coefficient to measure both agreement (positive association) and disagreement (negative association), concordance, whether measured by Kendall's original statistic or the Friedman modification, is one-sided in the sense that rejection of the null hypothesis indicates positive association. For example, if four judges A, B, C, D rank five objects in the order given in [Table 7.6](#) we would not with this test reject the hypothesis of no association, for the Friedman statistic ([6.5](#)) would take the value zero.

**Table 7.6** Rankings of five objects by four judges.

Object	Judge			
	A	B	C	D
I	1	5	1	5
II	2	4	2	4
III	3	3	3	3
IV	4	2	4	2
V	5	1	5	1

There is here complete agreement between judges A and C and between judges B and D: but the latter pair are completely at odds with judges A and C. The Kendall and the Friedman statistics do not detect such patterns.

An example indicates Kendall's approach in developing his concordance test.

### *Example 7.9*

*The problem.* There are six contestants in a diving competition and three judges each independently rank their performance in order of merit (1 for best, 6 for least satisfactory). The rankings allocated by each judge are given in [Table 7.7](#). Assess whether there is evidence of consistency between judges.

*Formulation and assumptions.* It is clear that the judges do not agree completely but there is a reasonable consensus that competitor III is a good performer and there is fairly strong support for competitor I while competitors IV and VI are thought to perform poorly. A moment's reflection indicates that a low rank total (last column) indicates a performance considered by the judges overall to be better than that of a competitor who attains a high rank total. Had the judges been inconsistent in their judgement (effectively just allocating ranks at random) one would expect all rank sums to be nearer to the average sum of these totals over all six competitors which is clearly  $63/6 = 10.5$ . Had there been complete agreement between judges it is easily verified that the six competitor rank sums would be some permutation of 3, 6, 9, 12, 15, 18. Kendall's coefficient of concordance is based on the sum of squares of deviations of the competitors' rank sums from their mean or expectation. If we denote this sum of squares by  $S$  then it is intuitively reasonable, and can indeed be shown that  $S$  has a maximum when the judges are in complete agreement and a minimum of zero when all rank sums equal the mean of 10.5 (which would imply some tied ranks in the situation considered in [Table 7.7](#)). Arguing along these lines Kendall proposed the statistic

$$W = \frac{S}{\text{Max}(S)}$$

**Table 7.7** Ranks awarded to six competitors by three judges in a diving competition.

Competitor	Judge			Rank total
	A	B	C	
I	2	2	4	8
II	4	3	3	10
III	1	1	2	4
IV	6	5	5	16
V	3	6	1	10
VI	5	4	6	15

which he called a **coefficient of concordance**. Clearly this takes the maximum value 1 when there is complete agreement and the minimum value zero when there is no agreement. (It will of course also take the value zero in a situation like that in [Table 7.6](#) where there are contrary opinions held by pairs of judges.) Tables of critical values are available but any program that gives exact  $P$ -values for the Friedman test may be used also to obtain  $P$  for the Kendall statistic.

*Procedure.* Clearly from [Table 7.7](#),

$$S = (8 - 10.5)^2 + (10 - 10.5)^2 + \dots + (15 - 10.5)^2 = 99.5.$$

If there are no ties it is not difficult to show that  $\text{Max}(S) = m^2 n(n^2 - 1)/12$  if there are  $m$  judges and  $n$  competitors, which in this case gives  $\text{Max}(S) = 157.5$ , whence  $W = 99.5/157.5 = 0.632$ . StatXact has a program for this test (although the Friedman test program will give the same exact  $P$ -value) and this confirms the value of  $W$  and gives the exact  $P = 0.062$ .

*Conclusion.* There is not very strong evidence of concordance between the judges.

*Comment.* The weakness of the evidence indicates the fairly low power of the test against alternatives other than a very high level of agreement. The judges are particularly erratic in their assessment of competitor V.

*Computational aspects.* As indicated above  $P$ -values are obtainable from a program for the Friedman test, but this will not calculate the value of  $W$  directly.

We do not prove it but it can be shown that the statistic  $T$  for the Friedman test given by [\(6.5\)](#) and  $W$  satisfy the relationship

$$W = T/[m(n - 1)]. \quad (7.10)$$

Some modifications to the procedure in the above example are needed for ties. These are basically similar to those for ties in the Friedman test. If we compute  $T$  for the tied Friedman case as described in [Section 6.3.3](#) we may derive the correct  $W$  using [\(7.10\)](#).

An intuitively reasonable measure of concordance between rankings of competitors by several judges is one based on the mean of pairwise rank correlation coefficients and it can be shown that for  $m$  judges the mean,  $R_s$ , of all  $\frac{1}{2}m(m - 1)$  such Spearman coefficients is such that  $R_s = (mW - 1)/(m - 1)$ . It follows that when  $R_s = 1$  (i.e. when all pairwise rankings are in complete agreement) then  $W = 1$ . When  $W = 0$  then  $R_s = -1/(m - 1)$  which can be shown to be the least possible value for  $R_s$ . Examination of the overall pattern of pairwise rank correlation coefficients is useful for detecting patterns of the type illustrated for an extreme case in [Table 7.6](#) where the pairwise Spearman coefficients are easily shown to take the value +1 twice and the value -1 four times exhibiting a definite pattern but where  $W = 0$ . This indicates that, like rank correlation in the bivariate

situation,  $W = 0$  when there is no association pattern, but the converse may not hold.

### 7.3 AGREEMENT

All measures so far described in this chapter relate to the concept of association between variables. With positive association between two variables if one variable takes a large value then so will the other. This does not necessarily mean, however, that the scores on some scale for both variables are identical. If, in addition to positive association, scores on both variables tend to be the same then we say that there is a high level of **agreement**. Agreement is not the same as association. It is possible for a correlation coefficient to be very high whilst at the same time the agreement is very low (for instance, if scores for one variable are consistently higher than the score for the other). Agreement in this sense can be assessed for categorical variables.

This concept of agreement is particularly important in assessing observer variation. Suppose that two dentists examine the X-ray of a particular tooth for the presence or absence of dental caries on two separate occasions. Although the information available from the X-ray remains the same, there are two possible types of variation with such examinations. A dentist re-examining the X-ray of a tooth some time later may make a different judgment (e.g. caries absent rather than caries present) —this is known as **intra-observer** variation. Also, a colleague may make a different decision about the same tooth, leading to **inter-observer** variation. For both types of variation, the same question can be asked — how well do the two sets of data **agree** with each other?

If there are relatively few possible categories (ordered or nominal), the most obvious measure of agreement between two sets of data is the proportion of cases in which agreement occurs. This has the clear drawback that a substantial amount of agreement can occur by chance alone. We need a chance-corrected measure of agreement. A simple and widely used measure is the **kappa statistic**, developed by Cohen (1960).

#### 7.3.1 The kappa statistic for agreement between two assessors

Suppose that two observers allocate each of  $n$  patients to one of  $k$  possible categories. We denote the number of patients allocated to

**Table 7.8** A general contingency table for assessing agreement between two observers.

	Observer B				
	1	2	...	k	Total
1	$n_{11}$	$n_{12}$	...	$n_{1k}$	$n_{1+}$
2	$n_{21}$	$n_{22}$	...	$n_{2k}$	$n_{2+}$
.	.	.	...	.	.
<i>Observer A</i>	.	.	...	.	.
.	.	.	...	.	.
k	$n_{k1}$	$n_{k2}$	...	$n_{kk}$	$n_{k+}$
	$n_{+1}$	$n_{+2}$	...	$n_{+k}$	$n$

category  $i$  by Observer A and to category  $j$  by Observer B by  $n_{ij}$ , the total number of patients allocated to category  $i$  by Observer A by  $n_{i+}$  and the total number of patients allocated to category  $j$  by Observer B by  $n_{+j}$ . Table 7.8 shows the data in a  $k \times k$  contingency table.

Dividing the number of observations for the cell in the  $i$ th row and  $j$ th column ( $n_{ij}$ ) by the total number of patients,  $n$ , gives the proportion  $p_{ij}$  of patients for that cell. The observed proportion of agreement is the proportion of patients for which  $i$  and  $j$  are equal (on the diagonal of the table) which is given by

$$p_o = (n_{11} + n_{22} + \dots + n_{kk}) / n = \sum_i p_{ii}$$

The expected proportion of agreement,  $p_e$  is determined using the appropriate row and column totals in a similar manner to that used for expected numbers in the familiar chi-squared test (Sections 9.1.1 and 9.2.2) so that

$$p_e = \sum_i p_{i+} p_{+i}$$

where  $p_{i+} = n_{i+}/n$  and  $p_{+i} = n_{+i}/n$ .

The kappa statistic is then defined as:

$$\kappa = \frac{\text{Observed proportion} - \text{Expected proportion}}{1 - \text{Expected proportion}}$$

or symbolically as

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

The kappa coefficient takes values between  $-1$  and  $1$ , with the value  $1$  for perfect agreement, zero for the level of agreement that would be expected by chance and negative values for less than chance agreement, i.e. apparent disagreement. Landis and Koch (1977) suggested benchmarks for kappa, e.g. a score over  $0.8$  indicates good agreement,  $0.6$  to  $0.8$  indicates substantial agreement and  $0.4$  to  $0.6$  moderate agreement. Assuming that the patients are assessed independently of each other and that the assessors operate independently we can test the null hypothesis that for the population  $\kappa = 0$ . An exact test for small samples is available in StatXact together with a Monte Carlo approximation for moderate sized samples and an asymptotic result for large samples. Fleiss, Lee and Landis (1979) give an asymptotic formula for the standard error of kappa assuming  $H_0$  holds, namely

$$se(\kappa) = \frac{\sqrt{p_e + p_e^2 - \sum_i p_{i+}p_{+i}(p_{i+} + p_{+i})}}{(1-p_e)\sqrt{n}}$$

The null hypothesis is tested by assuming that if it holds then  $Z = \kappa/[se(\kappa)]$  has a standard normal distribution.

A rather involved asymptotic formula for a confidence interval for kappa based on the estimated standard error of the maximum likelihood estimate of  $\kappa$  is given by Fleiss (1981).

#### *Example 7.10*

*The problem.* Two dentists inspected 100 patients and classified them as either *requiring treatment* or *not requiring treatment*. Table 7.9 shows the decisions by the two dentists as to whether or not they thought treatment was required. Use the kappa statistic to calculate the chance-corrected agreement for this sample and test the null hypothesis that  $\kappa = 0$ .

**Table 7.9** Need of treatment for a series of patients as assessed by two dentists.

		<i>Dentist B</i>		<i>Total</i>
		<i>Treatment needed</i>	<i>Treatment not needed</i>	
<i>Dentist A</i>	<i>Treatment needed</i>	40	5	45
	<i>Treatment not needed</i>	25	30	55
<i>Total</i>		65	35	100

*Formulation and assumptions.* It is assumed that patients are classified independently of each other and that the dentists do not confer. Cohen's kappa is calculated using the formulae for observed and expected proportions of agreement given above. The null hypothesis may be tested using an exact test if relevant software is available or the asymptotic formula for the standard error.

*Procedure.* We first display the data in terms of observed proportions:

		<i>Dentist B</i>		<i>Total</i>
		<i>Treatment needed</i>		
<i>Dentist A</i>	<i>Treatment needed</i>	0.40	0.05	0.45
	<i>Treatment not needed</i>	0.25	0.30	0.55
<i>Total</i>		0.65	0.35	1.00

As there are only two assessors and two categories the formulae for the observed and expected proportions of agreement are straightforward. The observed proportion of agreement is  $0.4 + 0.3 = 0.7$ . However, some of this agreement could have been expected by chance, and this is calculated from the row and column totals. Under chance agreement, the expected proportion of cases in the "yes/yes" cell of the table is given by  $0.65 \times 0.45 = 0.2925$ . The expected proportion of cases in the "no/no" cell is given by  $0.35 \times 0.55 = 0.1925$ .

The total proportion of **expected** agreeing cases is therefore  $0.2925 + 0.1925 = 0.485$ . Hence

$$\kappa = (0.7 - 0.485) / (1 - 0.485) = 0.417.$$

The asymptotic formula for the standard error of kappa under the null hypothesis gives:

$$\begin{aligned} \text{se}(\kappa) &= \sqrt{\frac{0.485 + 0.485^2 - 0.45 \times 0.65(0.45 + 0.65) - 0.55 \times 0.35(0.55 + 0.35)}{1 - 0.485}} \sqrt{100} \\ &= 0.09215, \end{aligned}$$

whence

$$Z = \kappa / \text{se}(\kappa) = 0.417 / 0.09215 = 4.53,$$

and reference to a standard normal distribution indicates  $P < 0.0001$ . StatXact confirms this and indicates that the same is true for an exact  $P$ .

*Conclusion.* There is moderate chance-corrected agreement ( $\kappa = 0.417$ ) between the two dentists. The very small  $P$ -value shows overwhelming evidence against the null hypothesis.

*Comments.* 1. Superficially, the observed agreement appears satisfactory but once chance agreement is taken into account the level of agreement is less impressive.

2. The strong evidence against the null hypothesis should come as no surprise. One would expect dentists, who receive several years of training, to have considerable agreement with each other. However, the fact that kappa is **only** 0.417 should be a cause for concern.

3. An asymptotic 95 per cent confidence interval for kappa is (0.256, 0.578). This is very wide, indicating that even a sample of 100 patients does not provide a great deal of information about the population value for kappa. Studies of observer agreement frequently involve complex and time-consuming methods for assessing patients or specimens. As a consequence, samples are generally small (typically no more than 50) and the benchmarks given by Landis and Koch (1977) are in practice therefore not always helpful.

*Computational aspects.* One can calculate kappa along with the asymptotic *P*-value for the null hypothesis  $\kappa = 0$  using SPSS or Stata. SPSS additionally gives the standard error for the calculated kappa value, from which the 95 per cent confidence interval for kappa can be obtained. Similar information, together with an exact *P*-value or a Monte Carlo estimate for large samples is also given by StatXact. The asymptotic standard error given by StatXact is slightly different from that given above, being based on the maximum likelihood estimate of  $\kappa$  rather than on the null hypothesis value.

### 7.3.2 Extensions of Cohen's kappa statistic

If there are more than two categories, some types of disagreement between the two assessors might be particularly serious. For instance, in assessing a patient's state of health, disagreeing responses of well and poor are more serious than well and moderately well. In the calculation of **weighted kappa** (Cohen, 1968), different types of disagreement are given different weights.

The kappa coefficient has been given in a conditional form (Light, 1971) and has been extended to allow for more than two assessors (Conger, 1980; Posner et al., 1990). A straightforward method of calculating multiple kappa is to take the arithmetic mean of the kappa values obtained by taking pairs of observers in turn; however, the formulae for hypothesis testing and the calculation of confidence intervals are involved (Fleiss, 1981). Stata can be used to calculate kappa for several observers and to test the null hypothesis  $\kappa = 0$ .

Kappa has also been developed for ordinal data (Fleiss, 1978) and continuous data (Rae, 1988). In certain situations kappa is equivalent to the intraclass correlation coefficient (Fleiss and Cohen, 1973; Rae, 1988). Sample size calculation for the case of two assessors has been investigated (Cantor, 1996).

## 7.4 FIELDS OF APPLICATION

### Political science

Leaders of political parties may be asked to rank issues such as the economy, health, education, transport and current affairs in order of

importance. In comparing orderings given by leaders of two parties, rank correlations may be of interest. If leaders of more than two parties are involved the coefficient of concordance may be appropriate, or we may prefer to look at pairwise rank correlations, because leaders of parties at different ends of the political spectrum may tend to reverse, or partly reverse, rankings.

## **Psychology**

A psychologist might show prints of 12 different paintings separately to a twin brother and sister and ask each to rank them in order of preference. We might use Blomqvist's, Spearman's or Kendall's coefficient to test for consistency in the rankings made by brother and sister.

## **Business studies**

Market research consultants list a number of factors that may stimulate sales, e.g. consistent quality, reasonable guarantees, keen pricing, efficient after-sales service, clear operating instructions, etc. They ask a manufacturers association and a consumers association each to rank these characteristics in order of importance. A rank correlation coefficient will indicate the level of agreement between manufacturers and consumers views on relative importance.

## **Personnel management**

A personnel officer ranks 15 sales representatives on the basis of total sales by each during one year. His boss suggests he should also rank them by numbers of customer complaints received about each. A rank correlation could be used to see how well the rankings relate. A Pearson correlation coefficient might be appropriate if we had for each sales representative figures both for sales and precise numbers of customer complaints. Note that a positive correlation between numbers of sales and numbers of complaints need not imply that a high rate of complaints stimulates sales.

## **Horticulture**

Leaf samples may be taken from each of 20 trees and magnesium and calcium content determined by chemical analysis. A Pearson coefficient might be used to see if levels of the two substances are related, but this coefficient can be distorted if one or two trees have levels of these chemicals very different from the others; such

influential observations are not uncommon in practice. A rank correlation coefficient may give a better picture of the correlation. If a third chemical, say cobalt, is also of interest, a coefficient of concordance might be appropriate.

## Medicine

For many forms of cancer alternative forms of primary treatment are surgery, chemotherapy or radiotherapy. Which is preferred for a particular patient depends on factors like age, general health status, location and stage of development of the tumour, etc. Two consultants may or may not agree on which option is to be preferred for individual patients. If each of two consultants make independent assessments of the appropriate treatment for each of a group of  $N$  patients then Cohen's kappa may be used as a measure of agreement.

## 7.5 SUMMARY

The **Pearson product moment correlation coefficient** is traditionally used as a measure of association for samples from continuous distributions. When normality is not assumed inferences may be based on a permutation test but the procedure is not robust.

The most widely used measures of rank correlation are the **Spearman rank correlation coefficient (rho)** ([Section 7.1.3](#)) and the **Kendall rank correlation coefficient (tau)** ([Section 7.1.4](#)). For the former ([7.2](#)) and for the latter ([7.3](#)) apply for the no-tie case; modifications to both are needed with ties and asymptotic results for large samples are given in [Section 7.1.6](#).

The **Blomqvist median coefficient** ([Section 7.1.5](#)) is usually estimated by ([7.6](#)) or preferably by ([7.7](#)) with appropriate modification for ties.

For multivariate ranked data **Kendal's coefficient of concordance** ([Section 7.2](#)) is equivalent to the Friedman test statistic for ranked data in randomized blocks given in [Section 6.3.2](#).

The **Cohen kappa** statistic ([Section 7.3.1](#)) is a widely applicable useful measure of agreement between independent assessors on possible courses of action. Some caution is needed in interpreting the statistic. Rejection of the null hypothesis that any indication of

agreement is due to chance does not necessarily imply that the level of agreement is strong.

## EXERCISES

- 7.1 Compute the probabilities associated with each possible value of Kendall's  $t_k$  under the hypothesis of no association when  $n = 4$  to verify the results quoted in [Example 7.5](#).
- 7.2 Compute the probabilities analogous to those sought in Exercise 7.1 for Spearman's  $r_s$  statistic.
- 7.3 Verify the numerical values in the triangular matrix of  $s_{ij}$  given on p. 259 for the data in [Table 7.2](#) and also the values of  $n_{cs}$  and  $n_{ds}$ .
- 7.4 Reanalyse the data in [Table 7.4](#) for McDelta clan deaths using the Spearman's rho. How do your conclusions compare with those based on Kendall's tau?
- 7.5 In [Table 7.4](#) ages at death are ordered by year of death. Use the Cox–Stuart trend test ([Section 3.2.3](#)) to test for a time trend in life spans. Do the results agree with those based on the Kendall and on the Spearman coefficients? If not, why not?
- 7.6 A china manufacturer is investigating market response to seven designs of dinner set. The main markets are the British and American. To get some idea of preferences in the two markets a survey of 100 British and 100 American women is carried out and each woman is asked to rank the designs in order of preference from 1 for favourite to 7 for least acceptable. For each country the 100 rank scores for each design is totalled. The design with the lowest total is assigned rank 1, that with the next lowest total rank 2, and so on. Overall rankings for each country are:

<i>Design</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
<i>British rank</i>	1	2	3	4	5	6	7
<i>American rank</i>	3	4	1	5	2	7	6

Calculate the Spearman and Kendall correlation coefficients. Is there evidence of a positive association between orders of preference?

- 7.7 The manufacturer in Exercise 7.6 later decides to assess preferences in the Canadian and Australian markets by a similar method and the rankings obtained are:

<i>Design</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
<i>Canadian rank</i>	5	3	2	4	1	6	7
<i>Australian rank</i>	3	1	4	2	7	6	5

Calculate the Spearman and Kendall correlation coefficients. Is there evidence of a positive association between orders of preference?

- 7.8 Perform an appropriate analysis of the ranked data for all four countries in [Exercises 7.6](#) and [7.7](#) to assess the evidence for any overall concordance. Comment on the practical implications of your result.
- 7.9 In a pharmacological experiment involving  $\beta$ -blocking agents, Sweeting (1982) recorded for a control group of dogs, cardiac oxygen consumption (MVO) and left ventricular pressure (LVP). Calculate the Kendall and Spearman correlation coefficients. Is there evidence of correlation?

Dog	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>
<i>MVO</i>	78	92	116	90	106	78	99
<i>LVP</i>	32	33	45	30	38	24	44

- 7.10 Bardsley and Chambers (1984) gave numbers of beef cattle and sheep on 19 large farms in a region. Is there evidence of correlation?

<i>Cattle</i>	41	0	42	15	47	0	0	0	56	67	707
<i>Sheep</i>	4716	4605	4951	2745	6592	8934	9165	5917	2618	1105	150
<i>Cattle</i>	368	231	104	132	200	172	146	0			
<i>Sheep</i>	2005	3222	7150	8658	6304	1800	5270	1537			

- 7.11 Paul (1979) discusses marks awarded by 85 different examiners to each of 10 scripts. The marks awarded by 6 of these examiners were:

	<i>Script</i>									
<i>Examiner</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
<i>1</i>	22	30	27	30	28	28	28	28	36	29
<i>2</i>	20	28	25	29	28	25	29	34	40	30
<i>3</i>	22	28	29	28	25	29	33	29	33	27
<i>4</i>	24	29	30	28	29	27	30	30	34	30
<i>5</i>	30	41	37	41	34	32	35	29	42	34
<i>6</i>	27	27	32	33	33	23	36	22	42	29

Use rank tests to determine (i) whether the examiners show reasonable agreement on ordering the scripts by merit and (ii) whether some examiners tend to give consistently higher or lower marks than others.

- 7.12 For the tied data in [Example 7.6](#) confirm that correct values have been computed for all terms required to calculate  $t_b$ .
- 7.13 For the data in [Example 7.7](#) explore the use of the asymptotic approximations for the Spearman coefficient given in [Section 7.1.6](#).

- 7.14 Timber veneer panels produced by a factory are classified after visual inspection by experts as either de luxe (*DL*), standard (*S*) or reject (*R*) depending upon the number and type of faults an observer detects. The management checks for consistency of classification between two observers by asking each to independently allocate each of a set of 150 panels to the categories they consider relevant. Use Cohen's kappa statistic to assess whether the following findings for a pair of observers indicate reasonable agreement between the two observers. Comment critically on the interpretation of your results.

		<i>Observer A</i>			<i>Total</i>
<i>Grade</i>		<i>DL</i>	<i>S</i>	<i>R</i>	
<i>Observer B</i>	<i>DL</i>	7	2	1	10
	<i>S</i>	4	96	18	118
	<i>R</i>	1	10	11	22
<i>Total</i>		12	108	30	150

# 8

## Regression

### 8.1 BIVARIATE LINEAR REGRESSION

Correlation and regression are closely related. Correlation is mainly concerned with qualitative aspects of possible relationships. For example, in looking at measurements of height and weight we might be interested in the possibility of positive association between the two variables and if there is association whether in broad terms the relationship is linear, monotonic, linear in ranks, etc. Regression is concerned with the quantitative aspects of relationships such as determining the slope and intercept of a straight line that in some sense provides a ‘best’ fit to given data. If a polynomial of degree  $p$  is needed to give a good fit regression provides values for the  $p + 1$  constants that determine a best-fitting polynomial. There is equivalence between some aspects of the two approaches, e.g. in straight-line regression a test of zero slope is equivalent to a test of zero correlation. Values of +1 or -1 for the Pearson product moment correlation coefficient tell us that all the observed points lie on a straight line. Other relationships between correlation and regression emerge in this chapter. In particular in straight-line regression we shall be interested in correlations between what are usually referred to as residuals (departures of the observed  $y_i$  from their values predicted by the fitted regression equation) and the  $x_i$ . This concept, which we explain more fully below, is basic to both classic least squares regression based on the Pearson product moment correlation coefficient and to nonparametric regression which may be based on that coefficient or on the Spearman or Kendall coefficient.

Least squares is the classic method of fitting a straight line to bivariate data. The method has optimal properties subject to well-known independence and homogeneity assumptions. If we add certain normality assumptions there are well-established procedures for hypothesis testing and estimation. Regression provides useful tools for forecasting and prediction. However, given a set of bivariate observations  $(x_i, y_i)$  there are many cases where the assumptions needed to validate least squares procedures do not hold. This may

result in misleading or invalid inferences. While least squares is relatively insensitive to some types of departures from basic assumptions it is strongly influenced by others. This has led to the development of a group of techniques known as regression diagnostics. We do not pursue that approach in this book but an elementary account of the basic ideas is given in Sprent (1998, Section 11.3) and both Atkinson (1985) and Cook and Wiseberg (1982) provide comprehensive treatments. McKean, Sheather and Hettmansperger (1990) describe regression diagnostics for rank-based methods. An alternative to using classic least squares with regression diagnostics is to use distribution-free methods.

Our treatment follows closely an approach developed by Maritz (1995, Chapter 5) but we omit much of the theory given there.

### 8.1.1 Least squares and raw-data permutation tests

For estimating slope in bivariate linear regression there is a Pitman-type permutation test procedure for continuously distributed ‘raw’ data analogous to those in Sections 2.1.1, 5.1.1, 6.2.1 and 7.1.1. It is seldom used in practice largely because its disadvantages outweigh its advantages. In particular, like its counterparts, it often yields inferences similar to normal-theory based equivalents even when these are misleading due to a breakdown in assumptions. We outline the approach not because of its practical importance but because it introduces notions and highlights difficulties relevant to, and often overcome by more practical alternatives based on the Spearman and Kendall coefficients that we consider later in the chapter.

Before moving to the permutation procedure based on the Pearson coefficient we look briefly at a classic bivariate least squares regression model. Many readers will already be familiar with this model but we emphasize here those aspects that help one understand the rationale behind many distribution-free approaches. In the classic parametric approach it is assumed that for each of a set of  $n$  given observed  $x_i$  – which may be either random variables or a set of fixed values that may or may not be chosen in advance – we observe some value  $y_i$  of a random variable,  $Y_i$ , which has the properties that its mean depends upon (i.e., is conditional upon) the value of  $x_i$  in such a way that

$$E(Y_i|x_i) = \alpha + \beta x_i \quad (8.1)$$

while the variance of  $Y_i$  is independent of  $x$  and for all  $x_i$  has the value

$$\text{Var}(Y_i) = \sigma^2$$

where  $\alpha$ ,  $\beta$ ,  $\sigma^2$  are unknown. The straight-line relationship between  $E(Y|x)$  and  $x$  of the form  $E(Y|x) = \alpha + \beta x$  between the conditional mean of  $Y$  and a given  $x$  defines the regression of  $Y$  on  $x$ . The line has slope  $\beta$  and intercept  $\alpha$  on the  $y$ -axis. The notation  $Y_i|x_i$  is the conventional notation for an event or variable  $Y_i$  having some property conditional upon a specified  $x_i$ .

For many inference purposes it is assumed also that the conditional distribution of  $Y_i|x_i$  is  $N(\alpha + \beta x_i, \sigma^2)$ . A classic regression problem is to estimate  $\alpha$ ,  $\beta$  and sometimes also  $\sigma^2$  given a set of  $n$  paired observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where for a given  $x_i$  the  $y_i$  is an observed value of the random variable  $Y_i$  featured in (8.1). These conditions hold if each  $(x_i, y_i)$  satisfies a relationship

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where each  $\varepsilon_i$  is an unobserved value of a  $N(0, \sigma^2)$  random variable and the  $\varepsilon_i$  are independent of each other and also of  $x_i$ . The portion ' $\alpha + \beta x_i$ ' is the systematic or **deterministic** part of the model for  $y_i$  and the  $\varepsilon_i$  is the **random** element. Least squares estimation seeks values  $a, b$  that minimize

$$S = \sum_i (y_i - a - bx_i)^2 \quad (8.2)$$

for estimates of  $\alpha, \beta$ . Denoting these estimates by  $\hat{\alpha}, \hat{\beta}$  the straight line  $y = \hat{\alpha} + \hat{\beta}x$  is called the least squares regression of  $y$  on  $x$ .

If, for any line  $y = a + bx$  we denote the  $y$ -coordinate corresponding to  $x = x_i$  by  $\hat{y}_i$  (i.e.  $\hat{y}_i = a + bx_i$ ) then the differences between the observed and predicted values, i.e.  $e_i = y_i - \hat{y}_i$ ,  $i = 1, 2, \dots, n$ , are called the **residuals** with respect to that line. Equation (8.2) may be written  $S = \sum_i e_i^2$  and the least squares estimators of  $\alpha, \beta$  are so-called because they minimize the sum of squares of residuals.

To obtain  $\hat{\alpha}, \hat{\beta}$ , we differentiate  $S$  separately with respect to  $a$  and  $b$ . These derivatives are then set equal to zero to obtain the so-called **normal equations** that are solved for  $a, b$ . These equations are

$$\begin{aligned} \sum_i x_i (y_i - a - bx_i) &= 0, \\ \sum_i (y_i - a - bx_i) &= 0. \end{aligned}$$

The solutions are  $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$  where  $\bar{y} = (\sum_i y_i)/n$ ,  $\bar{x} = (\sum_i x_i)/n$  and

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad (8.3)$$

With the assumptions above classic least squares regression is a generalization of classic tests for comparison of treatment means. In particular a test of  $H_0: \beta = 0$  is equivalent to a test of equality of a set of sample means. This follows from (8.1) if we regard the  $x_i$  as ‘indicators’ or labels attached to samples. If there are  $n$  paired observations  $(x_i, y_i)$  each  $y_i$  corresponding to a particular  $x_i$  may be looked upon as an observed value from the sample labelled by that  $x_i$ . The total number of samples  $m$  may be any number between 2 and  $n$  and the numbers of observations,  $n_j$ , in sample  $j$ ,  $j = 1, 2, \dots, m$  are subject to the constraint  $n_1 + n_2 + \dots + n_m = n$ . In particular if no two  $x_i$  are equal there are  $n$  samples each of one observation and at the other extreme if there are only two distinct  $x_i$  there are two samples with  $n_1$  and  $n_2 = n - n_1$  observations respectively. Suppose in this last case we label the first sample by  $x = 0$  and the second sample by  $x = 1$  and the first sample values are

$$y_{11}, y_{12}, \dots, y_{1n_1}$$

with mean  $m_0$  and the second sample values are

$$y_{21}, y_{22}, \dots, y_{2n_2}$$

with mean  $m_1$ . It can then be shown from (8.3) that  $\hat{\beta} = m_1 - m_0$ , the sample mean difference used in the familiar  $t$ -test for equality of two treatment means (Exercise 8.1). Thus the test of equality of means is in this case identical to testing  $H_0: \beta = 0$ . In the more general case of  $m$  samples a test for  $\beta = 0$  is a test for identity of all  $m$  population means.

It is well-known that for the classic least squares model the estimator of  $\beta$  does not depend upon  $\alpha$ , but that of  $\alpha$  depends upon  $\beta$  through  $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$ . From this expression for  $\hat{\alpha}$  it is easily seen that the fitted equation may be written without reference to the intercept in the form

$$y = \bar{y} - \hat{\beta} (x - \bar{x}) \quad (8.4)$$

implying that the line passes through the point  $(\bar{x}, \bar{y})$ .

A graphical interpretation of this is that the slope of the fitted line is unaltered by a change of origin. In particular if we shift the origin to the bivariate mean  $(\bar{x}, \bar{y})$  and write  $x' = x - \bar{x}$ ,  $y' = y - \bar{y}$  equation (8.4) becomes

$$y' = \hat{\beta} x'$$

and (8.3) reduces to

$$\hat{\beta} = (\sum_i x'_i y'_i) / (\sum_i x'^2_i) \quad (8.5)$$

The form (8.5) still holds if we only shift the origin to the mean of the  $x_i$  and use the original  $y_i$ . In Section 7.1 we pointed out that the Pearson correlation coefficient was unaltered by linear transformations of  $(x, y)$  of the form  $x' = (x - k)/s$  and  $y' = (y - m)/t$  where  $s, t$  are both positive. We have just seen that in regression although the estimate of  $\alpha$  is affected by a change in origin, that of  $\beta$  is not, i.e. transformations of the form  $x' = (x - k)$ ,  $y' = (y - m)$  do not affect the estimate of  $\beta$ . However, it is easy to establish using (8.3) that the transformation  $x' = (x - k)/s$  and  $y' = (y - m)/t$  alters both the true value of  $\beta$  and its least squares estimate by the same scale factor  $s/t$ , changing  $\beta$  to  $\beta' = s\beta/t$ . For this reason we do not make scale changes in this chapter.

We shall however make use of another property of the correlation coefficient in permutation tests that we drew attention to in Section 7.1.1, namely that because all other quantities in (7.1) remain constant under permutation we may base permutation tests on the statistic  $T = \sum x_i y_i$  in the case of the Pearson coefficient as an alternative to using the sample correlation coefficient  $r$ . A corresponding property carries over to the Spearman coefficient with ranks replacing the  $x_i, y_i$ .

Dropping the normality assumption for the  $Y_i$  makes joint inferences about  $\alpha$  and  $\beta$  difficult so we consider first the estimation of  $\beta$  only. The problem then reduces to one of making inferences about differences in medians or means of samples labelled by the different  $x_i$ . For any independent  $Y_i, Y_j$  associated with distinct  $x_i, x_j$  we drop the normality assumptions on the conditional distributions of  $Y|x$  and assume now only that for any  $x_i, x_j$  they have distributions  $F_i(Y_i|x_i), F_j(Y_j|x_j)$  that differ only in their centrality measure which will be taken in general to be the median (but which will coincide with the mean for conditional distributions that are symmetric providing that the mean exists). For the straight line regression model the median of  $F_i(Y_i|x_i)$  is  $\text{Med}(Y_i|x_i) = \alpha + \beta x_i$  and that of  $F_j(Y_j|x_j)$  is  $\text{Med}(Y_j|x_j) = \alpha + \beta x_j$ , these being the analogues of (8.1) and the difference between the medians is clearly

$$\text{Med}[(Y_j - Y_i)|x_j, x_i] = \beta(x_j - x_i).$$

Assuming that differences between distributions are confined to a median difference implies that for all  $i$

$$D_i(\beta) = Y_i - \beta x_i$$

are identically and independently distributed with median  $\alpha$ . Since the  $D_i$  are therefore independent of the  $x_i$  it follows that they are

uncorrelated with the  $x_i$ . As we have pointed out the  $x_i$  can be altered by addition or subtraction of a constant without affecting either the slope or its estimate. In particular, it will often simplify the algebra if we adjust the  $x_i$  by adding an appropriate constant to make  $\sum x_i = 0$ , which in graphical terms implies shifting the origin to a point on the  $x$ -axis corresponding to the mean  $\bar{x}$ . Writing  $d_i = y_i - \beta_0 x_i$  an intuitively reasonable test of the hypothesis  $H_0: \beta = \beta_0$  against a one- or two-sided alternative is a test for zero correlation between the  $x_i$  and the  $d_i$  since we know that if  $H_0$  holds then  $D_i(\beta_0)$  is uncorrelated with the  $x_i$  and the  $d_i$  are then observed values of the variable  $D_i(\beta_0)$ . Whether the most appropriate test is based on the Pearson, Spearman, Kendall or some other coefficient will depend upon what assumptions are made about characteristics of the distribution  $F(Y|x)$ , (e.g. whether long-tailed, symmetric or asymmetric, etc.). We call the  $d_i$  residuals for convenience, but this is an unorthodox use of the term which more conventionally refers to the  $e_i = y_i - \alpha_0 - \beta_0 x_i$  where  $\alpha_0$  is some hypothesized value of  $\alpha$ . An appropriate statistic based on the Pearson coefficient for the test of zero correlation is

$$T(\beta_0) = \sum_i x_i d_i = \sum_i x_i (y_i - \beta_0 x_i). \quad (8.6)$$

To estimate  $\beta$  an intuitively reasonable estimation procedure takes the form of solving for  $b$  the equation

$$T(b) - \sum_i x_i \sum_i d_i / n = 0. \quad (8.7)$$

This follows from the form of  $c_{xy}$  given below (7.1) since equating  $c_{xy}$  to zero ensures a zero value for the sample correlation coefficient since it makes the numerator in (7.1) zero.

Recalling that we may adjust the  $x_i$  to have mean zero, i.e. so that  $\sum_i x_i = 0$  without affecting our estimate of  $\beta$ , we assume this has been done so that (8.7) simplifies to

$$T(b) = \sum_i x_i (y_i - bx_i) = 0 \quad (8.8)$$

with solution  $\hat{\beta} = (\sum_i x_i y_i) / (\sum_i x_i^2)$ . It is easily verified that this is identical to the solution given by (8.3) with the constraint  $\bar{x} = 0$ .

We have made no assumptions about the distribution of the  $D_i$  except that they are identical for all  $i$ , so that for hypothesis testing a permutation test for a zero Pearson correlation coefficient based on the sample values  $(x_i, d_i)$  is appropriate. As we have indicated in other cases a difficulty with inferences based on raw data using a Pitman permutation test is that the results tend to be similar to those based on equivalent theory assuming normality even when the

normality assumption is clearly violated. In other words the method lacks robustness.

This is illustrated for a small data set in Example 8.1. There are many situations where data like these arise, although usually in larger data sets where the complexity of computation may mask a clear understanding of what is happening.

*Example 8.1*

*The problem.* The water flow in cubic metres per second ( $y$ ) at a fixed point in a mountain stream is recorded at hourly intervals ( $x$ ) following a snow-thaw starting at time  $x = 0$ .

Hours from start of thaw (x)	0	1	2	3	4	5	6
Flow in cubic metres/sec (y)	2.5	3.1	3.4	4.0	4.6	5.1	11.1

During previous thaws there has often been a near straight-line relationship between time and flow. Use the method of least squares to fit a straight line to the data and a permutation test based on the Pearson correlation coefficient to test the null hypothesis  $H_0: \beta = 1$  (implying that on average flow increases by 1 cubic metre per second over a period of 1 hour) against the alternative  $H_1: \beta \neq 1$ .

*Formulation and assumptions.* We have seen that the estimate of  $\beta$  for classic least squares and permutation based least squares are identical and may be obtained using (8.3), or from (8.5) after replacing  $x$  by  $x' = x - 3$ , or from any least squares regression program. The requested hypothesis test is performed by computing the  $P$ -value associated with the Pearson coefficient between the  $x_i$  and  $d_i = y_i - \beta x_i$  after setting  $\beta = 1$ .

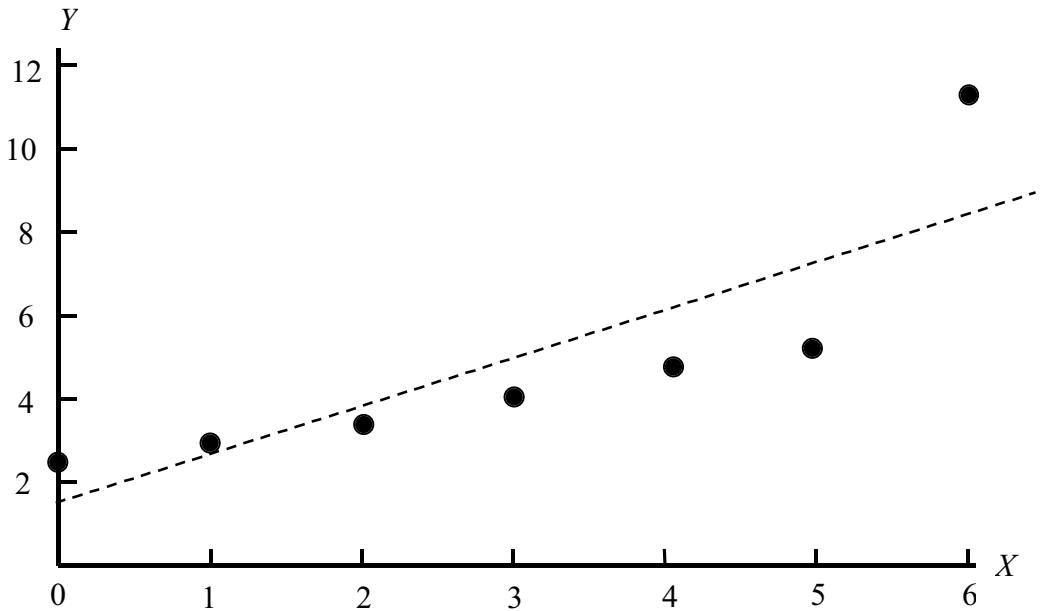
*Procedure.* Direct substitution of the values of  $x$  and  $y$  in the appropriate formula gives  $\hat{\beta} = 1.1070$ . It is also easily verified that the classic least squares estimate of  $\alpha$  is  $\hat{\alpha} = 1.508$ . To test the hypothesis  $H_0: \beta = 1$  against a two tail alternative we calculate for each data pair  $d_i = y_i - x_i$  and compute the correlation coefficient between the  $x$  and  $d$  values. Ideally for such a small sample an exact permutation test program should be used to compute the  $P$ -value appropriate to the hypothesis  $H_0: \rho = 0$  implying that the  $x_i$  and the  $d_i$  are uncorrelated. Relevant values of the  $x_i$  and  $d_i$  are:

$x_i$	0	1	2	3	4	5	6
$d_i = y_i - x_i$	2.5	2.1	1.4	1.0	0.6	0.1	5.1

For these data StatXact gives  $r = 0.1392$  for the sample Pearson coefficient and the exact two-tail  $P = 0.7925$ .

*Conclusion.* The least squares line of best fit is  $y = 1.508 + 1.107x$ . Since  $H_0: \beta = 1$  has an associated permutation test  $P = 0.7925$  the data certainly do not provide evidence against  $H_0: \beta = 1$ .

*Comments.* 1. Figure 8.1 shows the data points and fitted line. The fit is unsatisfactory largely because the point  $(6, 11.1)$  seems to be an outlier relative to the other points. If we omit the point  $(6, 11.1)$  it is easily verified that the other points are well fitted by an amended least squares regression line which is now  $y = 2.491 + 0.517x$ .



**Figure 8.1** Scatter diagram and classic least squares line of best fit for the data in Example 8.1.

In passing we also mention that if the point  $(6, 11.1)$  is omitted the  $P$ -value associated with the test of  $H_0: \beta = 1$  against  $H_1: \beta \neq 1$  reduces to  $P = 0.0028$  providing strong evidence (as one would expect from an inspection of Figure 8.1) against  $H_0$ . A larger sample, perhaps including readings between 5 and 6 hours and after six hours would give a clearer indication of the degree of association and the shape of the relationship after the first 5 hours.

2. Unfortunately both the theory and practice of obtaining exact permutation-based confidence intervals for  $\beta$  present difficulties. Possible procedures are discussed by Maritz (1995, Section 5.2) and by Sprent (1998, Section 10.2). Using the asymptotic results given there a trial and error hypothesis testing approach may then be used to refine the asymptotic limits in programs such as that for the Pearson coefficient in StatXact. For these data (including the outlying pair  $(6, 11.1)$ ) the 95 per cent confidence interval for  $\beta$  is  $(0.495, 1.984)$  which is not very different from the interval  $(0.231, 1.984)$  given by classic parametric least squares. Both clearly demonstrate the tendency for the upper limit (as is the point estimate of  $\beta$  itself) to be pulled upward by this last point.

3. The sudden departure from the apparent near-linear relationship noted at the 6-hour reading might result from a rapidly increasing rate of thaw after 5 hours, perhaps one caused by a temperature or wind change or other meteorological factor. If so, this may have important practical implications because readings like these are often used as a basis for warnings of likely flooding at points downstream from the recording station.

4. If there are outlying values and there is no reasonable explanation for them, such values may have been incorrectly recorded. In the context considered here it is unlikely that the last observation is an error unless there had been some mishap with the meter used to measure flow. Another possibility might be that the recorded  $y$  values are the mean of two readings taken, say, 1 minute apart at

each hour to average out any sudden surges and that for the last reading somebody added the two but forgot to divide by 2 to get the mean. In practice the observers should know if such an error were possible. In other contexts such as biological experiments it is not uncommon for recorded  $y$  to be the mean of two or more observations. A further possible explanation for the last reading is that a lead digit has been omitted and that the final point should be (16, 11.1); this lies much closer to the straight-line relationship suggested by the remaining points. In the context of this example this might occur if, for instance, readings were suspended overnight and the final reading was the first on the following morning. Again, in practice it should be possible to check if this were the source of an error.

5. Data like these are common in environmental studies, especially following incidents, natural or accidental, that have an impact over an extended period. For example, each  $y$  reading could be parts per million of a toxic gas in the atmosphere  $x$  hours after an explosion at a chemical plant or the  $y$  could be some measure of radioactivity recorded at daily (or weekly) intervals after an accident at a nuclear power station.

6. The assumption that consecutive  $\varepsilon_i$  are independent might be queried in this example, but unless some physical condition intervenes, such as a block in the waterway restricting flow over an appreciable period, it is probably reasonable to assume that over intervals of 1 hour the random elements in flow are for practical purposes effectively independent.

*Computational aspects.* Even simple tests based on the Pearson coefficient require suitable computing facilities. Asymptotic tests may be unreliable if normality assumptions break down seriously. Bootstrapping methods (a topic discussed in [Section 11.3](#)) may sometimes be useful in this context.

Better procedures for dealing with data like that in [Example 8.1](#) are often provided by rank-based estimation procedures.

### 8.1.2 Regression associated with the Spearman rho\*

In practice procedures we give in [Section 8.1.3](#) are easier to apply and are more widely used than those in this section but the method introduced here illustrates some key features of distribution-free regression and helps to clarify the relationship between procedures.

The obvious analogue to the Wilcoxon rank sum test procedure is to replace the  $d_i$  in [\(8.6\)](#) by  $\text{rank}(d_i)$ . This leads to simplifications of the exact test and estimation procedures when the  $x_i$  are equally spaced, for inferences may then be based on the distribution of Spearman's rho, since, if the  $x_i$  are equally spaced they are linear functions of the ranks and hence the statistic

$$T_s(\beta_0) = \sum_i x_i \text{rank}(d_i) = \sum_i x_i \text{rank}(y_i - \beta_0 x_i) \quad (8.9)$$

---

\* This section may be omitted at first reading because the method is not often used in practice.

is appropriate for a test of zero Spearman correlation. It is convenient for estimation problems, though not essential, to assume also that any origin change needed to make the mean of the  $x_i$  zero has been made. The test is not conditional on either the observed  $x$  or  $y$  because for any hypothesized  $\beta_0$  and fixed  $n$  only the paired set of ranks (or linear transformations of these ranks) from 1 to  $n$  are needed for the permutation reference set.

[Example 8.2](#) shows that any programs providing an exact test for Spearman's  $\rho_s = 0$  may be used to test  $H_0: \beta = \beta_0$  against a one- or two-sided alternative.

Point estimation of  $\beta$  is now less straightforward than it is for least squares but it is still possible and confidence intervals based on the permutation test are relatively easy to obtain if appropriate software is available.

Estimation is based on an analogue of (8.8) derived from (8.9) and we seek  $b$  satisfying the equation

$$T_s(b) = \sum_i x_i \text{rank}(y_i - bx_i) = 0 \quad (8.10)$$

A difficulty is that there is in general no value of  $b$  for which (8.10) holds exactly because  $T_s(b)$  is a step function in  $b$ . This is because the ranking of the  $d_i$  will only change as  $b$  passes through a value where two or more  $d_i, d_j$  are equal, i.e. for  $b$  such that for some  $i, j, y_i - bx_i = y_j - bx_j$ . Denoting this value of  $b$  by  $b_{ij}$  it follows that  $b_{ij} = (y_j - y_i)/(x_j - x_i)$ . Thus  $T_s(b)$  does not change in value for any  $b$  lying between successive values of  $b_{ij}$  for  $i = 1, 2, \dots, n-1, j > i$ . It is easy to show [see e.g. Sprent (1998, Section 10.3)] that  $T_s(b)$  is a nonincreasing step function in  $b$  and that if for different  $(i, j)$  the corresponding  $b_{ij}$  are equal this implies three or more observations are either collinear or the lines joining the relevant pairs of points have identical slopes. The steps will then not all be of equal height.

Some implications of the properties of  $T_s(b)$  for hypothesis testing and estimation, including confidence intervals, are best demonstrated by an example. The discussion in this example is somewhat open-ended so we have not divided it under headings of *Problem*, *Formulation and assumptions*, *Procedure*, etc.

### *Example 8.2*

The data in [Example 8.1](#) with the  $x_i$  adjusted to have mean zero become

$x$	-3	-2	-1	0	1	2	3
$y$	2.5	3.1	3.4	4.0	4.6	5.1	11.1

It is convenient to record the pairwise slopes  $b_{ij}$  as an upper triangular matrix similar to that used for the  $s_{ij}$  in [Section 7.1.7](#). A typical entry is

$$b_{26} = (5.1 - 3.1)/[2 - (-2)] = 0.500$$

shown in bold in the matrix below. Statistical software to calculate the  $b_{ij}$  is becoming increasingly common. Minitab will produce pairwise slopes in the order of the columns in the upper triangular matrix. For our data these are

0.600	0.450	0.500	0.525	0.520	1.433
	0.300	0.450	0.500	<b>0.500</b>	1.600
		0.600	0.600	0.567	1.925
			0.600	0.550	2.367
				0.500	3.250
					6.000

Repeated values among the  $b_{ij}$  imply collinearity or parallelism of some pairwise joins. [Table 8.1](#) is easily if tediously formed using (8.9) although the complete table is not needed for most inferences. We use the property that  $T_s(b)$  remains constant between successive values of  $b_{ij}$  so that we only need to evaluate (8.9) for one value  $\beta_0$  of  $b$  in each interval between successive values. When  $b$  coincides with some  $b_{ij}$  there are ties in the ranks of the  $d_i$  and using mid-ranks for these gives a value of the corresponding  $T_s(b)$  which is the mean of its values for  $b$  immediately greater than and immediately less than that  $b_{ij}$ . As an illustration of the computation of  $T_s(b)$  for a given  $b$  consider that for the interval between  $b_{16} = 0.520$  and  $b_{15} = 0.525$ . It suffices to set  $b = 0.523$ , say and to evaluate  $T_s(0.523) = \sum_i x_i \text{rank}(y_i - 0.523x_i)$ . For the given data we obtain

$x_i$	-3	-2	-1	0	1	2	3
$y_i$	2.5	3.1	3.4	4.0	4.6	5.1	11.1
$d_i = y_i - 0.523x_i$	4.07	4.15	3.92	4.0	4.08	4.05	9.53
$\text{rank}(d_i)$	4	6	1	2	5	3	7

whence

$$T_s(0.53) = (-3) \times 4 + (-2) \times 6 + (-1) \times 1 + 0 \times 2 + 1 \times 5 + 2 \times 3 + 3 \times 7 = 7.$$

The remaining values of  $T_s(b)$  in column 3 of [Table 8.1](#) are obtained in a similar manner. Since there is no unique  $b$  corresponding to each possible value of  $T_s(b)$  it is sometimes convenient to associate with each such value the **mid- $b$  value** which is the mid-point of the interval of  $b$  values associated with that  $T_s(b)$ . For example, we have just shown that  $T_s(b) = 7$  is associated with  $b$  in the interval (0.520, 0.525) so the associated mid- $b$  value is 0.5225. These mid- $b$  values are given in column 2 of [Table 8.1](#).

The maximum value, 28, of  $T_s(b)$  corresponds to a Spearman sample coefficient  $r_s = 1$  and since  $T_s$  is a linear function of  $r_s$  if we divide any  $T_s(b)$  by 28 we get a corresponding value of the equivalent statistic  $r_s$ . The relevant values are given in column 4 of [Table 8.1](#). We show below that this is useful for determining confidence intervals for  $\beta$ , but first we consider point estimation of  $\beta$ . Maritz proposes an estimate obtained by regarding the step function  $T_s(b)$  as an approximation to a continuous function and then using linear interpolation between the closest mid- $b$  values on either side of zero. In this case we interpolate between  $T_s = 1$ ,  $b = 0.5583$  and  $T_s = -2$ ,  $b = 0.5835$  to obtain a point

**Table 8.1** Intervals for  $b$ -values corresponding to each possible  $T_s(b)$  arising from the data in [Example 8.2](#). The values of  $T_s(b)$  when  $b = b_{ij}$  are the means of the values in adjacent  $b$  intervals. Values of  $r_s$  corresponding to each  $T_s(b)$  are also given.

<i>b</i> -interval	<i>mid-b value</i>	$T_s(b)$	$r_s$
$-\infty$ to 0.300		28	1.000
0.300+ to 0.450-	0.375	27	0.964
0.450+ to 0.500-	0.475	23	0.821
0.500+ to 0.520-	0.510	12	0.429
0.520+ to 0.525-	0.5225	7	0.250
0.525+ to 0.550-	0.5375	3	0.107
0.550+ to 0.567-	0.5583	1	0.036
0.567+ to 0.600-	0.5835	-2	-0.071
0.600+ to 1.433-	1.0165	-7	-0.250
1.433+ to 1.600-	1.5165	-13	-0.464
1.600+ to 1.925-	1.7625	-18	-0.643
1.925+ to 2.367-	2.146	-22	-0.786
2.367+ to 3.250-	2.8085	-25	-0.893
3.250+ to 6.000-	4.625	-27	-0.964
6.000+ to $\infty$		-28	-1.000

estimate corresponding to  $T_s = 0$  and this is

$$b = 0.5583 + (0.5835 - 0.5583)/3 = 0.567.$$

It is fortuitous that this is identical with the median of the  $b_{ij}$  although an estimate obtained in this way will usually be close to that median. In practice if only a point estimate of  $\beta$  is required one need calculate only a few entries in Table 8.1 for values of  $b$  in intervals close to  $\text{med}(b_{ij})$  and then use linear interpolation.

If an exact test program for the Spearman coefficient is available hypothesis tests about  $\beta$  are straightforward. For example, to test  $H_0: \beta = 1$  against  $H_0: \beta \neq 1$  we may use the values of  $x_i$  and  $d_i$  computed in [Example 8.1](#) and insert these in a Spearman test program. Not surprisingly (as can be verified from Table 8.1) this gives  $r_s = -0.250$  and the corresponding two-tail  $P = 0.5948$  so there is clearly no substantial evidence against  $H_0$ . If no program is available for an exact test, tables like Table A31 in Hollander and Wolfe (1999) clearly indicate that only values of  $r_s$  greater than about 0.7 in magnitude provide any acceptable evidence against  $H_0$ .

Given a facility like that in StatXact for generating the complete distribution of  $r_s$  under  $H_0: \rho_s = 0$  it is relatively easy to obtain confidence intervals for  $\beta$ . Using this distribution one finds that for  $n = 7$  the one-tail  $P$  associated with  $r_s \geq 0.786$  is  $P = 0.024$ . From Table 8.1 we note that because tied ranks occur for certain values of  $b$  there is no upper tail  $b$  corresponding exactly to  $r_s = 0.786$ , but in the lower tail when  $r_s = -0.786$  the corresponding mid- $b$  value is 2.146. The smallest value of  $b$  that will just produce an  $r_s = -0.786$  is a value slightly

above 1.925, say 1.926. In the upper tail of  $r_s$  values we see that the largest value of  $b$  that will produce a value of  $r_s$  of at least 0.786 is  $b$  slightly less than 0.500, say 0.499. Thus the  $b$ -interval (0.500, 1.925) is an appropriate  $(1 - 2 \times 0.024)100 = 95.2$  per cent confidence interval.

If no program giving exact  $P$ -values is available the values of  $r_s$  corresponding to ‘significance’ at many conventional levels are available from tables like Table A31 in Hollander and Wolfe which, for a sample of 7, gives  $\Pr(r_s \geq 0.786) = 0.025$ . Using this value leads to the same conclusions for an approximate 95 per cent confidence interval.

The interval obtained here is slightly shorter than that given in Comment 2 on [Example 8.1](#). More interestingly the point estimate  $\beta = 0.567$  is much smaller than the full-data least squares estimate and reasonably close to the least squares estimate 0.517 obtained if the ‘rogue’ point (6, 11.1) is omitted. This indicates that the suspect point is less influential on the point estimate of  $\beta$  although it still exerts an upward pressure on the confidence limit. Here the confidence limits, unlike those in classic parametric least squares estimation, are no longer symmetric about the point estimator, but they reflect more realistically the upward pressure of the point (6, 11.1) on estimates of  $\beta$ .

Clearly if confidence intervals are required at or near conventional levels only a few values of  $T_s(b)$  corresponding to large values of  $|T_s(b)|$  need be calculated.

There is an asymptotic procedure for approximate confidence intervals that usually works well for larger  $n$ . This is discussed briefly in Sprent (1998, Section 10.3). Since inference procedures discussed in this section are generalizations of the WMW test it is not surprising that they have the same Pitman efficiencies.

If the  $x_i$  are not equally spaced  $T_s(b)$  given by (8.9) is no longer a linear function of Spearman’s rho and inferences are less straightforward. A brief outline is given by Sprent (1998, Section 10.5) but in this situation the method given in the next section is generally preferable. Some writers have suggested that inferences about  $\beta$  might be based on a statistic that replaces  $x_i$  by  $\text{rank}(x_i)$  in (8.9), i.e.

$$S(b) = \sum_i \text{rank}(x_i) \text{rank}(y_i - bx_i). \quad (8.11)$$

This is clearly equivalent to using  $T_s(b)$  for equally spaced  $x$  for hypothesis testing, although care is needed for estimation since the mean of the ranks of the  $x_i$  are no longer zero, but adjustments can be made for this. However, if the points are not equally spaced the transformation to ranks is not a linear transformation of the  $x_i$  and this precludes the use of  $S(b)$  for making reliable inferences about a true  $\beta$  for the original data.

### 8.1.3 Regression associated with Kendall’s tau

In the previous section we saw that regression inference procedures based on Spearman’s rho are possible but a major restriction is that

the  $x_i$  be equally spaced. This restriction is not needed for inferences based on Kendall's tau, since  $\tau_k$  depends only upon the order of the observations and not on magnitudes of the differences between data values. Further there is no advantage in adjusting the  $x_i$  to have mean zero. It simplifies presentation without loss of generality, if, when the  $x_i$  are all different we assume that  $x_1 < x_2 < \dots < x_n$ . We consider briefly the case when some  $x_i$  are equal in Section 8.1.4.

The statistic used for inference about  $\beta$  is

$$T_t(b) = \sum_i \text{sgn}[d_{ij}(b)] \quad (8.12)$$

where

$$\begin{aligned} d_{ij}(b) &= (y_j - a - bx_j) - (y_i - a - bx_i) = (y_j - bx_j) - (y_i - bx_i) = d_j - d_i = \\ &= (y_j - y_i) - b(x_j - x_i) \end{aligned}$$

and summation of the signs of the  $d_{ij}$  is over all  $i = 1, 2, \dots, n-1$  and  $j > i$ . Since the  $x_i$  are all different and in ascending order it is clear that  $T_t(b)$  is the numerator in the expression (7.3) for Kendall's tau composed of the numbers of concordances minus the number of discordances in the data pairs  $(x_i, d_i)$ .

Use of a statistic equivalent to  $T_t(b)$  was first proposed by Theil (1950) and the procedure is widely known as **Theil's method**. Sen (1968) highlighted the relationship to Kendall's tau so we refer to it as the **Theil–Kendall method**.

Clearly  $T_t(b)$  is a linear function of the sample estimator of Kendall's tau and using arguments similar to those in Sections 8.1.1 and 8.1.2 an appropriate point estimator of  $\beta$  is obtained by setting  $T_t(b) = 0$ . Since the  $x_i$  are in ascending order it is easy to see that  $T_t(b)$  is unaltered if we replace  $d_{ij}(b)$  by  $b_{ij} - b$  where, as in Section 8.1.2  $b_{ij} = (y_j - y_i)/(x_j - x_i)$ . Clearly then  $T_t(b) = 0$  if we choose  $b = \text{med}(b_{ij})$  for then the number of positive and the number of negative  $d_{ij}$  will be equal. As was the case for  $T_s(b)$  it is easy to see that  $T_t(b)$  only changes in value when  $b$  passes through a value of  $b_{ij}$  and since the statistic is the numerator term in Kendall's tau it follows that if all  $b_{ij}$  are distinct as  $b$  increases from  $-\infty$  to  $\infty$   $T_t(b)$  is a step function decreasing by steps of 2 from  $\frac{1}{2}n(n-1)$  to  $-\frac{1}{2}n(n-1)$ . Clearly division of  $T_t(b)$  by  $\frac{1}{2}n(n-1)$  leads to Kendall's  $t_k$  which may be used as an equivalent statistic. If the  $b_{ij}$  are not all distinct some of the steps will be multiples of 2. In particular if a  $b_{ij}$  occurs  $r$  times it induces a step of height  $2r$ . The value of  $T_t(b)$  at any  $b_{ij}$  may be regarded as the mean of the values immediately above and below that  $b_{ij}$ .

Tests of hypotheses about  $\beta$  are straightforward using either Kendall's  $t_k$  or the equivalent  $T_t(b)$  as the test statistic, and if we

know the exact distribution of the sample Kendall coefficient when  $\tau_k = 0$ , a confidence interval for  $\beta$  may be obtained. The procedures are illustrated in Example 8.3.

### Example 8.3

*The problem.* Given the data in Example 8.1, i.e.

$x$	0	1	2	3	4	5	6
$y$	2.5	3.1	3.4	4.0	4.6	5.1	11.1

use the Theil–Kendall method (i) to test the hypothesis  $H_0: \beta = 1$  against the alternative  $H_1: \beta \neq 1$ ; (ii) to obtain a point estimate of  $\beta$  and a confidence interval giving at least 95 per cent coverage.

*Formulation and assumptions.* To test the hypothesis in (i) we compute  $T_t(1)$  and use an exact test program to assess evidence for or against  $H_0$ . To obtain a point estimate of  $\beta$  we determine the median of the  $b_{ij}$  defined above and for a confidence interval we then compute  $T_t(b)$  and the corresponding  $t_k$  for all  $b$ . Using the exact distribution of  $t_k$  when  $\tau_k = 0$  or appropriate tables if the exact distribution when  $n = 7$  is not available we compute the relevant confidence interval for  $\beta$ .

*Procedure.* The  $b_{ij}$  may conveniently be arranged in a triangular matrix with the first row elements  $b_{12} \ b_{13} \ \dots \ b_{1n}$ , second row elements  $b_{23} \ b_{24} \ \dots \ b_{2n}$  and so on to give

0.600	0.450	0.500	0.525	0.520	1.433
	0.300	0.450	0.500	0.500	1.600
		0.600	0.600	0.567	1.925
			0.600	0.550	2.367
				0.500	3.250
					6.000

identical to the matrix introduced in Example 8.2. Inspection shows that the point estimate of  $\beta$ , i.e.  $\text{med}(b_{ij})$  is 0.567 because there are ten greater and ten lesser  $b_{ij}$ . To test the hypothesis  $\beta = 1$  we subtract  $b = 1$  from each  $b_{ij}$  and find this implies 15 discordances (negative values) and 6 concordances whence  $T_t(1) = 6 - 15 = -9$  and when  $n = 7$  this implies  $t_k = -9/21 = 0.4286$ . Corresponding to this value of  $t_k$  StatXact gives an exact one-sided  $P = 0.1194$  which is doubled for a two-tail test so there is no convincing evidence against  $H_0$ .

Using the arguments outlined before this example and noting that there are respectively 2, 4 and 4 tied values of  $b_{ij}$  at 0.450, 0.500, 0.600 we can establish the values of  $T_t$  for all values of  $b$ . These are given in Table 8.2 together with corresponding values of  $t_k$  obtained by dividing each  $T_t(b)$  by 21.

To obtain an approximate 95 per cent confidence interval for  $\beta$  we must first determine a value of  $|T_t|$  with a one tail  $P$  as close as possible to but not exceeding  $P = 0.025$ . The exact distribution when  $n = 7$  given by StatXact indicates that  $P = 0.015$  when  $T_t = 15$  while  $P = 0.035$  when  $T_t = 13$ . If the StatXact program is not available these values may also be obtained from tables such as those in Kendall and Gibbons (1990, Appendix Table 1) or Hollander

and Wolfe (1999, Table A30). The latter give the relevant probabilities for the statistic  $T_t$ . From Table 8.2 we deduce that the shortest  $(1 - 2 \times 0.15)100 = 97$  per cent interval based on  $|T_t| = 15$  will be  $(0.500-, 1.925+)$ , or the open interval  $(0.500, 1.925)$ . See also Comment 3 below.

*Conclusion.* The point estimate of  $\beta$  is 0.567 and a nominal 95 (actual 97) per cent confidence interval for  $\beta$  is  $(0.500, 1.925)$ .

*Comments.* 1. The point estimate of  $\beta$  is identical to that obtained in Example 8.2 using Spearman's  $r_s$  and the nominal 95 per cent confidence interval is virtually the same despite minor differences in the exact coverage.

2. Comparing the methods used in this and the previous section we see that the Theil–Kendall procedure is easier to compute and more importantly in practice (though not relevant to this specific example) the restriction to equally spaced  $x_i$  is no longer needed.

3. To find a confidence interval we do not need to form Table 8.2 since the observation  $T_t = 15$  implies  $n_c - n_d = 15$  and since  $n_c + n_d = 21$  this implies  $n_c = 18$  and  $n_d = 3$ . Thus we should reject values of  $\beta$  greater than or equal to the three largest  $b_{ij}$  or less than or equal to the three smallest. Inspection of the matrix of  $b_{ij}$  given under *Procedure* shows that the third largest and third smallest  $b_{ij}$  are respectively 2.367 and 0.450. Rejecting these three implies the limits 0.500 and 1.925 given above.

*Computational aspects.* For reasons that are not clear to us more extensive tables are available for Kendall's tau than is the case for most statistics relevant

**Table 8.2** Intervals for  $b$  values corresponding to each  $T_t(b)$  or corresponding  $t_k$  for the data in [Example 8.3](#). The values of  $T_t(b)$  when  $b = b_{ij}$  are the means of its values in adjacent  $b$  intervals.

<i>b</i> -interval	$T_t(b)$	$t_k$
$-\infty$ to 0.300	21	1.000
0.300+ to 0.450-	19	0.905
0.450+ to 0.500-	15	0.714
0.500+ to 0.520-	7	0.333
0.520+ to 0.525-	5	0.238
0.525+ to 0.550-	3	0.143
0.550+ to 0.567-	1	0.048
0.567+ to 0.600-	-1	-0.048
0.600+ to 1.433-	-9	-0.428
1.433+ to 1.600-	-11	-0.524
1.600+ to 1.925-	-13	-0.619
1.925+ to 2.367-	-15	-0.714
2.367+ to 3.250-	-17	-0.810
3.250+ to 6.000-	-19	-0.905
6.000+ to $\infty$	-21	-1.000

to nonparametric inference. This to some extent relieves the need for readily available software either for the Theil–Kendall method or indeed for inferences about Kendall’s tau.

The close relationship between Kendall’s  $t_k$  and  $T_t$  makes it easy to carry over asymptotic approximations between them. Using results for  $\tau_k$  it is easily shown (see, e.g. Maritz, 1995, Section 5.2.5) that  $E(T_t) = 0$  and that  $\text{Var}(T_t) = n(n - 1)(2n + 5)/18$ . For large  $n$  the distribution of  $Z = T_t/\sqrt{\text{Var}(T_t)}$  is approximately standard normal and asymptotic inferences can then be made in the usual way. Even for  $n = 7$  the approximation is sometimes not seriously misleading ([Exercise 8.4](#)) but caution is advisable with so small a sample.

We showed in [Section 7.1.4](#) that the Kendall procedure is a special case of the Jonckheere–Terpstra method which in turn is a generalization of the Mann–Whitney formulation of the WMW procedure, whereas the method using Spearman’s rho is a generalization of the Wilcoxon formulation of the WMW procedure. Although the generalizations are not exactly equivalent it is not unreasonable when applying them to regression problems to expect them to lead to broadly similar conclusions as we saw in Examples [8.2](#) and [8.3](#).

### 8.1.4 Some alternative approaches

Having noted that the Theil–Kendall procedure is a generalization of the WMW method it is interesting to explore the possibility of procedures that generalize those of a sign test. Theil proposed one which is usually called the **abbreviated Theil method**. It uses a small independent subset of the  $b_{ij}$  and is not recommended for sample sizes smaller than about 14. Even for larger  $n$  the full Theil–Kendall method is preferable if adequate computing facilities or tables are available.

If  $n$  is even the only  $b_{ij}$  used in the abbreviated method are  $b_{i, i + 1/2 n}$  where  $i = 1, 2, \dots, 1/2 n$  while if  $n$  is odd we use only the  $b_{i, i + 1/2(n + 1)}$  where  $i = 1, 2, \dots, 1/2(n - 1)$ . These estimators all involve different data pairs and hence are independent so test and estimation procedures based on the sign test may be used. The point estimator of  $\beta$  is the median of the reduced set of  $b_{ij}$  indicated above. Tests and confidence intervals are easily derived from the corresponding sign-test procedures.

### Example 8.4

*The problem.* We give below the modal length ( $y$  cm) of samples of Greenland turbot of various ages ( $x$  years) based on data given by Kimura and Chikuni (1987). Fit a straight line using the abbreviated Theil method and obtain a 95 per cent confidence interval for  $\beta$ .

Age (x)	4	5	6	7	8	9	10	11	12
Length (y)	40	45	51	55	60	67	68	65	71
Age (x)	13	14	15	16	17	18	19	20	
Length (y)	74	76	76	78	83	82	85	89	

*Formulation and assumptions.* For the 17 points we calculate the sub-set of pairwise slopes  $b_{10,1}, b_{11,2}, \dots, b_{17,8}$  and obtain their median. A confidence interval for  $\beta$  is obtained in a manner analogous to that for a median based on the sign test.

*Procedure.* Since the  $x_i$  are equally spaced in this example the denominators in each of the required  $b_{ij}$  are all equal to  $x_{10} - x_1 = 13 - 4 = 9$  so we need only compute the relevant  $y_j - y_i$ , obtain their median and divide that by 9 to obtain an estimate of  $\beta$ . Thus  $y_{10} - y_1 = 76 - 40 = 34$ ,  $y_{11} - y_2 = 76 - 45 = 31$ , etc. The complete set of differences is 34, 31, 25, 23, 23, 15, 17, 24. The median of these is 23.5 so  $\hat{\beta} = 23.5/9 = 2.61$ . Using the argument in [Section 2.3.2](#) for a nominal 95 per cent confidence interval for the above set of 8 differences between the  $y$ -values we use appropriate  $B(8, \frac{1}{2})$  distribution probabilities ([Exercise 8.5](#)) to find that an exact 93 per cent confidence interval for the median of these differences is the interval (17, 31) and dividing the limits by 9 gives the corresponding interval for  $\beta$ , i.e. (1.89, 3.44). Similarly the interval (1.67, 3.78) is a 99.2 per cent interval. With so small a sample the discontinuities in coverage of possible intervals are quite marked.

*Conclusion.* An appropriate estimate for  $\beta$  is  $\hat{\beta} = 2.61$  and a 93 per cent confidence interval for  $\beta$  is (1.89, 3.44).

*Comment.* Using the program in Minitab for generating all pairwise slopes for these data one can show that the full Theil–Kendall procedure gives an estimator  $\hat{\beta} = \text{med}(b_{ij}) = 2.65$ . We may also establish that a 95.8 per cent confidence interval is (2.2, 3.1) following the procedure outlined in Comment 3 in [Example 8.3](#). This follows because the one-tail  $P = 0.021$  when  $n = 17$  corresponds to  $T_t = 50$ , implying  $n_c - n_d = 50$ . We then deduce that to establish the relevant confidence interval we reject the 43 smallest and 43 largest  $b_{ij}$ . Inspection of the paired values establishes the above interval. If one has a facility to generate all  $b_{ij}$  we recommend the full rather than the abbreviated Theil–Kendall method for samples of this size. A comparison of the lengths of the confidence intervals given by the two methods indicates an appreciable loss of efficiency when using the abbreviated method. This is not surprising because the abbreviated method only uses 8 of the 136 items of information (the  $b_{ij}$ ) used in the full method. It does however use an independent set, whereas there are correlations in the complete set.

Many variants of procedures considered in this chapter for slope estimation appear in the literature. Another sign-test analogue related to the Spearman coefficient rather than the Kendall coefficient is briefly discussed in Sprent (1998, Section 10.5). As alternatives to procedures involving correlation between  $x_i$  and rank ( $d_i$ ) such as that discussed in [Section 8.1.2](#), Adichie (1967) considered possible transformations of the rank( $d_i$ ) such as that to van der Waerden or to normal scores, but so far as we know these have not been widely used in practice and one might expect their performance to show little improvement on the direct use of ranks.

It will be clear that rank-based methods, especially those such as the Theil–Kendall method depend heavily on the  $b_{ij}$ . Clearly the use of median estimators based on these pairwise slope estimators introduces robustness to point estimators in the presence of rogue observations or observations that might indicate inadequacy of the model. In [Example 8.3](#), for instance, the observation (6, 11.1) which is so obviously out of line with the other observations influences only the  $b_{ij}$  that involve that point, i.e. those in the last column of the triangular matrix of  $b_{ij}$ , and these have little influence on  $\text{med}(b_{ij})$  as a point estimator of  $\beta$ . They do however exert an influence on the confidence limits and raise the upper limit above what it would be with no such out-of-line observation.

When there are no rogue or suspect observations one might feel that more weight ought to be given to those  $b_{ij}$  associated with larger values of  $x_j - x_i$ . This is what classic least squares does, for it can be shown that that estimator is a weighted mean of the  $b_{ij}$  with weights proportional to  $(x_j - x_i)^2$ . Jaeckel (1972) recommended taking as a point estimator of  $\beta$  the median of the weighted  $b_{ij}$  with weights  $w_{ij} = (x_j - x_i)/[\sum_{i < j}(x_j - x_i)]$ . The procedure has some optimum properties when there are no outliers but these and other weighting schemes that have been proposed may be less satisfactory than, for instance, the Theil–Kendall method if there are rogue observations. In simulation studies Hussain and Sprent (1983) found that the Theil–Kendall method performed almost as well as least squares when relevant assumptions held for the latter and that it showed a marked improvement in efficiency for long-tail error distributions, whereas in the latter situation weighted medians performed no better than, or sometimes less well than, Theil–Kendall.

We indicated in [Section 8.1.3](#) that a problem arises with the Theil–Kendall method when there are tied values of  $x_i$ . The  $b_{ij}$  corresponding to such a pair becomes infinite. For only a small proportion of ties there will be little loss of efficiency if such points are replaced

by one data entry with the  $y$  value set equal to the mean of the  $y$  values for all the points with that tied  $x$  value. This is usually preferable to artificial tie-breaking devices such as splitting the ties by making arbitrary small changes to tied  $x$ -values to separate them since that process may lead to large or even bizarre  $b_{ij}$  associated with such splits because of the small denominators. An alternative with greater appeal is to exclude all comparisons between points with a common  $x_i$  value but to consider all joins of each of these points to points with other values  $x_j$ , i.e. where  $x_j \neq x_i$ . An extreme example is when there are only two distinct  $x$  values  $x_1$  and  $x_2$ , and repeated  $y$  at each. The problem is then equivalent to the two-sample problem for means or medians of the  $y$ . If we consider the slopes of all pairwise joins between  $y$  values associated with the two  $x$  it is easy to see that these are similar to the differences computed for the Hodges–Lehmann estimator in [Section 5.2.4](#), indeed they are simply these differences divided by the constant  $x_2 - x_1$ , so the procedure is exactly equivalent to the WMW test.

### 8.1.5 Joint estimation of slope and intercept

Estimating slope is often the main aim of a regression analysis but when we want to use the fitted line for forecasting or prediction we must also estimate  $\alpha$  so that we are in a position to predict using the estimate of  $E(Y|x) = \alpha + \beta x$  for some new specified  $x$ . The intercept corresponds to the case  $x = 0$ . Whereas estimation of  $\beta$  is unchanged if we replace  $x$  by  $x' = x - h$  where  $h$  is any constant, this is not the case for  $\alpha$ .

Even in classic parametric least squares estimates of  $\alpha$  and  $\beta$  are in general correlated. Only in the special case when the mean of the  $x_i$  is zero and the estimator of  $\alpha$  reduces to  $\hat{\alpha} = \bar{y}$  is this estimate uncorrelated with that for  $\beta$  and then the customary normality assumption implies independence, so in that case we can make inferences about  $\alpha$  without having to worry about what is now a ‘nuisance’ parameter  $\beta$ . Unfortunately, when we move to distribution-free methods this simplification does not hold because even when the mean of the  $x_i$  is zero, the statistics used to estimate  $\alpha$  and  $\beta$  are correlated.

Many of the implications are discussed in detail by Maritz (1995, Section 5.3) and more informally by Sprent (1988, Section 10.6). A broad class of statistics relevant to joint estimation of  $\alpha$  and  $\beta$  were discussed by Adichie (1967) and these lead to inferences which in practice call for iterative methods of estimation starting with an estimate of  $\beta$  that is used to obtain an estimate of  $\alpha$ , then using this

estimate of  $\alpha$  to get a revised estimate of  $\beta$  and so on, the process usually converging after a few cycles. Some of the statistics used have links with and similarities to those used for estimation of slopes in Sections 8.1.2 and 8.1.3.

Here we give two variants of a method used in practice that often gives satisfactory estimates of  $\alpha$  providing we start with what we accept as a good estimate of  $\beta$ . For example, for the data in Example 8.1 the estimate  $\hat{\beta} = 0.567$  obtained by the methods used in Sections 8.1.2 or 8.1.3 is clearly a reasonable slope estimate for a line through all points other than (6, 11.1) if that is what is of interest. The assumptions in Section 8.1.2 imply that all  $D_i = y_i - \beta x_i$  (as defined in Section 8.1) have identical distributions with median  $\alpha$ . We do not know  $\beta$ , but if  $\hat{\beta}$  is a reasonably good estimate of  $\beta$  it is not unreasonable to assume that  $d_i = y_i - \hat{\beta} x_i$  will have a distribution with a median close to  $\alpha$ . An obvious procedure is to compute all  $n$  values of  $d_i$  and to take the median of these  $d_i$  as an estimate of  $\alpha$ . This is equivalent to the optimum estimator based on the sign test. If one makes the further assumption that the  $D_i$  are symmetrically distributed about  $\alpha$  the Hodges–Lehmann estimator based on the Walsh averages of the  $d_i$  is arguably more appropriate. We stress however that these results are only approximate and will clearly be influenced by the choice of  $\hat{\beta}$ .

### *Example 8.5*

*The problem.* Given the data in Example 8.1, i.e.

$x$	0	1	2	3	4	5	6
$y$	2.5	3.1	3.4	4.0	4.6	5.1	11.1

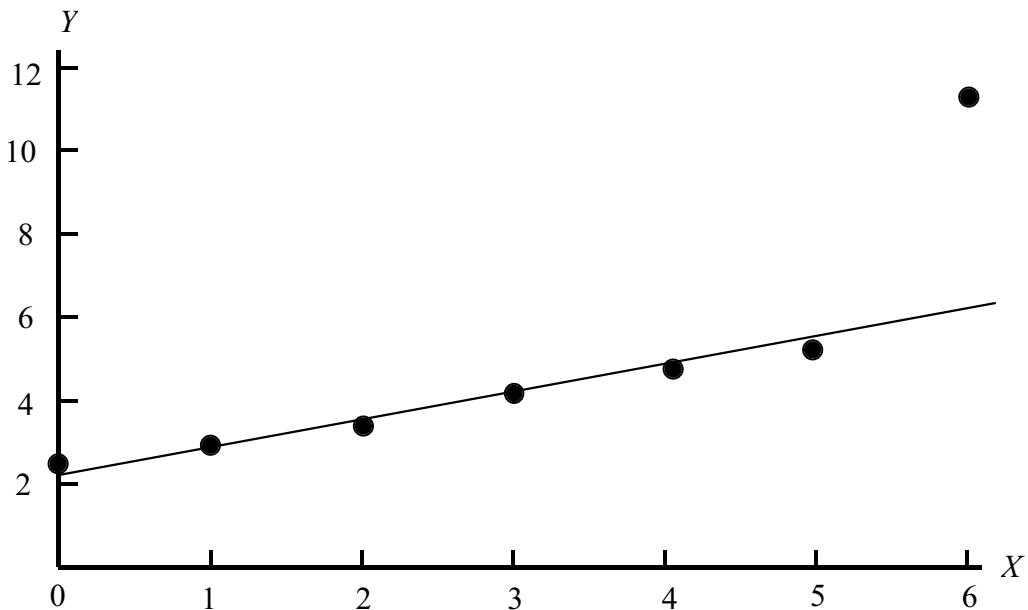
and assuming a reasonable estimate of  $\beta$  is  $\beta = 0.567$  obtain an estimate of the intercept  $\alpha$ .

*Formulation and assumptions.* We compute all  $d_i = y_i - 0.567x_i$  and take as our estimate either (i) the median of the  $d_i$  or (ii) the Hodges–Lehmann estimator which is the median of the Walsh averages.

*Procedure.* The  $d_i$  are respectively 2.500, 2.533, 2.266, 2.299, 2.332, 2.265, 7.698. The median of these 7 values is 2.332, there being three smaller and three larger values. The Hodges–Lehmann estimator is easily obtained from any program that computes the Walsh averages or a program such as the dedicated Hodges–Lehmann estimation program in StatXact. For these data the estimate (the median of the Walsh averages) is 2.399.

*Conclusion.* Two possible estimates of  $\alpha$  are 2.332 or 2.399.

*Comments.* 1. One may argue that the median of the  $d_i$  should be the preferred estimator because the extreme value  $d_7 = 7.698$  suggests that the symmetry assumption needed to justify Hodges–Lehmann estimation may not



**Figure 8.2** The Theil–Kendall regression line fitted to the data in Example 8.5.

hold. The line of best fit using the Theil–Kendall estimator of slope and the median of the  $d_i$  as intercept estimator is shown in Figure 8.2.

2. It is tempting to apply either the sign-test method or Hodges–Lehmann procedure to obtain a confidence interval for  $\alpha$ . However, this may not be appropriate because we use only an estimate of  $\beta$  since we do not know its true value. It is easily shown that alternative choice of the estimate of  $\beta$  may profoundly influence both the pattern of the  $d_i$  as well as the estimate based upon them. For instance, in Example 8.1 we established that if the last point were ignored the least squares estimate of  $\beta$  would be  $\beta = 0.517$  and for this the values of the  $d_i$  are 2.500, 2.583, 2.366, 2.449, 2.532, 2.515, 7.998 with median 2.515 reflecting a general upward shift in the  $d_i$ . The pattern changes even more markedly if we assume  $\beta = 0.5$  when the  $d_i$  apart from the last are even more concentrated, being 2.5, 2.6, 2.4, 2.5, 2.6, 2.6 and 8.1. Indeed to obtain reasonable approximate confidence intervals for  $\alpha$  one must use a more sophisticated approach. As indicated above, some that are used require an iterative approach for the joint estimation of  $\alpha$  and  $\beta$ . One possible approach is given by Sprent (1998, Examples 10.7 and 10.8) where it is in fact shown that in this example an approximate interval based on the estimate  $\hat{\beta} = 0.567$  may not be unreasonable as iteration does not improve the joint estimates of slope and intercept.

Some alternative proposals for estimation of  $\alpha$  and the more general problem of estimating  $\text{med}(Y|x)$  are discussed in Sprent (1998, Section 10.6) and by Maritz (1995, Section 5.3).

### 8.1.6 Comparison of slopes for several lines

Suppose we have  $k$  mutually independent sets of paired observations of variables  $(x, y)$  where the  $i$ th set contains  $n_i$  observations ( $i = 1, 2, \dots, k$ ) and  $k \geq 2$ . Assume also that the  $j$ th observation ( $j = 1, 2, \dots, n_i$ ) in the  $i$ th set [i.e.  $(x_{ij}, y_{ij})$ ] satisfies the relationship

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

where the  $\varepsilon_{ij}$  are realized values of a  $N(0, \sigma^2)$  random variable and the  $\alpha_i$ ,  $\beta_i$  and  $\sigma^2$  are all unknown. This model specifies  $k$  regression lines all with the same error distribution. The model has been widely studied in parametric analysis in particular in the context of testing the hypothesis that all the  $\beta_i$  are equal against an alternative of some inequality. Equality of the  $\beta_i$  implies all  $k$  lines are parallel. An extended hypothesis of equality of all  $\alpha_i$  in addition to equality of all  $\beta_i$  implies the  $k$  regression lines are identical.

If we drop the normality assumptions on the distribution of the  $\varepsilon_{ij}$  and assume only that they all have median zero and are otherwise identically distributed appropriate methods of estimation of  $\beta$  within each of the  $k$  data sets include those given in Sections 8.1.2 to 8.1.4. The importance of the ‘pairwise’ slopes denoted in the methods in those sections by  $b_{ij}$  (not to be confused with the  $i, j$  subscripts as in  $x_{ij}$  in this section) here suggests that a comparison of these for the different data sets might produce a useful base for tests for parallelism. Unfortunately this intuitively reasonable approach is complicated by the lack of independence between pairwise slopes **within** any one set. This complication does not arise, however, in the case of the abbreviated Theil method. If this is applied to each data set the pairwise slopes involved are independent within that set and the independence of observations in different sets implies they are independent of the pairwise slopes obtained by the same method in the other sets. If, in addition, the  $x$  are equally spaced all pairwise slopes have the same distribution within a particular set. If the slopes of all sets are the same they will have identical distributions in all sets. If the  $\beta_i$  are not all equal (i.e. if the lines are not all parallel) the distributions of the pairwise slopes for each set will differ only in their medians. Thus in this very restricted case an appropriate test for parallelism against the alternative of at least one different in slopes is the Kruskal–Wallis test (or the WMW test when  $k = 2$ ).

We do not know the Pitman efficiency of the above procedure but it may not be high because the abbreviated Theil procedure does not make full use of all relevant information in the data.

More sophisticated tests for parallelism are discussed by Sen (1969) and a very general test that is asymptotically distribution-free under broad assumptions developing some of Sen's ideas in combination with proposals due to Adichie (1984) is described with an example by Hollander and Wolfe (1999, Section 9.5).

### 8.1.7 Efficiency power and sample size

It is well known in classic least squares straight line regression that the standard error of the estimator  $\hat{\beta}$  of  $\beta$  is highly dependent upon the positioning of the  $x_i$ . If one has a choice of the  $x_i$  one aim should be to select these to make this standard error as small as possible because then the confidence interval for  $\beta$  at any chosen level will be the shortest available. It is also well known that for least squares the shortest confidence interval is obtained when  $n$ , the sample size, is even by setting half the  $x_i$  at the least possible value and the other half at the greatest possible value. In practice this ideal design for an experiment to estimate  $\beta$  may be impossible to attain because of experimental constraints. Also, other factors may have to be taken into consideration. One of these may be the need to test whether a straight line is adequate to describe the systematic component of the model. To test this one or more  $x$  values intermediate between the two extreme values must be included. In many practical situations there are major restrictions on the available  $x$  values. For example if the study is designed to determine whether a straight line provides a reasonable relationship between weight and age of chickens from shortly after hatching to maturity one may have to select a sample from available chickens. There may be doubt about whether a straight-line relationship is adequate, so one would be unwise to choose only day old chickens and fully mature birds. However, it would be helpful if one could reduce the standard error of the estimate of  $\beta$  by choosing, if possible, relatively large numbers of young and of mature birds and just a few intermediate in size. There may well be practical limitations on the extent to which this can be done – for example, there may be only a few mature birds available. For valid inferences if more than one bird is available at a given  $x$  and only some of these are included, those included should be chosen at random.

The often quite profound influence of design (in the sense of choosing  $x$ ) on the precision of estimates of  $\beta$  carries over to nonparametric situations. Suppose, for example that the data in Example 8.1 had been generated by adding 'errors' to some  $Y$

generated by the systematic relationship  $Y = x + 2$ . Thus, for instance, for the first point  $(0, 2.5)$  an ‘error’ 0.5 has been added to the systematic  $Y = 0 + 2 = 2$ . It is easily shown for the remaining points (see [Exercise 8.10](#)) that the ‘errors’ are 0.1, -0.6, -1, -1.4, -1.9, 3.1. Suppose now that we retain  $Y = x + 2$  as the systematic part of our model but replace the values  $x = 4, 5, 6$  in Example 8.1 by values  $x = 99, 100, 101$  and impose upon the true  $Y$  given by  $Y = x + 2$  the same errors as those just given. For example we replace the point  $(5, 5.1)$  by the point  $(100, 100.1)$  since when  $x = 100$ ,  $Y = 100 + 2 = 102$  and the associated error is -1.9, whence  $y = 102 - 1.9 = 100.1$ . Proceeding this way our new data set is

$x$	0	1	2	3	99	100	101
$y$	2.5	3.1	3.4	4.0	99.6	100.1	101.1

In [Exercise 8.10](#) we ask you to verify that the point estimate of  $\beta$  given by the Theil–Kendall method is now 0.985 and that a nominal 95 per cent confidence interval for  $\beta$  given by that method is  $(0.50, 1.037)$ . Although the lower limit is the same as that obtained in [Example 8.3](#) the upper limit is appreciably lower and the point estimator is now near the top end of the interval. The reduction in width of the interval implies an increase in power attributable to the design change of altering dramatically three of the  $x$ -values. We strongly urge readers to spend a few minutes considering the implications of these findings. In particular, one should consider whether the model with systematic part  $Y = x + 2$  and the given errors seems a reasonable model in relation to the amended data values, and why the point estimator changes so dramatically between that in [Example 8.3](#) and that obtained here.

We drew attention in earlier sections to a relationship between some linear regression models and the procedures used to analyse them and methods such as WMW and the sign test. Where relevant, results for Pitman efficiency under various error structures carry over to regression. However, power studies become very complicated not only because, as in the simpler situations, they are often highly dependent upon what distributional assumptions are made, but as we have just indicated, the power is strongly influenced by design factors in choosing  $x$ , as well upon the size of the sample. Further, changes in sample size may well introduce new  $x$ -values that may or may not in themselves enhance the power. In view of the complexity we do not pursue the matter further here. Other useful comments on efficiency are given by Hollander and Wolfe (1999, Section 9.8).

## 8.2 MULTIPLE REGRESSION

In parametric multiple linear regression in place of  $n$  paired observations  $(x_i, y_i)$  we have  $n$  observations of  $p + 1$  variables  $(x_1, x_2, \dots, x_p, Y)$  where the  $j$ th set,  $j = 1, 2, \dots, n$ , is  $(x_{1j}, x_{2j}, \dots, x_{pj}, y_j)$ . The straight line regression model which assumes that  $y_i$  is an observed value of a random variable  $Y_i$  with conditional mean given by (8.1) and for all  $x_i$   $\text{Var}(Y_i) = \sigma^2$ , usually with an additional assumption of normality, is generalized to one in which

$$E(Y_j|x_{1j}, x_{2j}, \dots, x_{pj}) = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj}$$

with  $\text{Var}(Y_i) = \sigma^2$ , usually one again with an assumption of normality. The model may also be written

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj} + \varepsilon_j$$

where the  $\varepsilon_j$  are independently distributed  $N(0, \sigma^2)$ .

As for straight-line regression the analogous nonparametric multiple regression analysis replaces the assumption of normality by one where the  $\varepsilon_j$  are assumed to be identically distributed with median zero.

Readers familiar with parametric multiple regression will know that this is a particular case of what is known as a general linear model which is also fundamental to analysis of variance. As Maritz (1995, Section 6.1) points out it is theoretically possible to develop a distribution-free approach with similarities to classic methods but adds that it soon becomes clear that exact inferential statements can only be made about rather uninteresting questions, unless one considers generalizations with certain special qualities which can perhaps be labelled properties of orthogonality.

Both the theoretical explanations of possible procedures and their application require a deeper treatment than other topics covered in this book, so we omit details. The flavour of how some topics are treated is indicated by Maritz (Chapter 6). Sprent (1998, Section 11.2) discusses, with an example, the extension of some of the distribution-free methods for straight-line regression discussed in this chapter to the case of two explanatory or independent variables  $x_1, x_2$ , mainly with the aim of highlighting practical difficulties arising with exact theory for all but trivial data sets. Hollander and Wolfe (1998, Section 9.6) demonstrate both the theoretical and practical problems that must be addressed when using an asymptotically distribution-free inference procedure specifically for investigating assumptions

that some of the  $\beta_i$  may be zero and they apply the method to a substantial data set.

Because they are useful in practice some more productive approaches to nonparametric regression analysis that have no parametric analogues are discussed briefly in the next section.

### 8.3 NONPARAMETRIC REGRESSION MODELS

The multiple regression model

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj} + \varepsilon_j \quad (8.13)$$

has a systematic or deterministic component  $\alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj}$  and a random component  $\varepsilon_j$ . The key feature of the methods developed in Sections 8.1.1 to 8.1.4 where  $p = 1$  was a relaxation of assumptions about the random component when making inferences about  $\beta_1$  while the form of the systematic component remained unaltered. A well-known parametric generalization of (8.13) is to replace the deterministic component  $\alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj}$  by a general function of some specified form  $f(x_{1j}, x_{2j}, \dots, x_{pj}, \boldsymbol{\theta})$  where  $\boldsymbol{\theta}$  is a vector of unknown parameters. The classic parametric regression model then becomes

$$y_j = f(x_{1j}, x_{2j}, \dots, x_{pj}, \boldsymbol{\theta}) + \varepsilon_j \quad (8.14)$$

where the usual assumption is that the  $\varepsilon_j$  are independently  $N(0, \sigma^2)$ . As in the case of the model (8.13) it is possible in theory at any rate to replace the normality assumption for the  $\varepsilon_i$  by one of identical distributions with median zero. That approach has been developed for many particular families  $f(x_{1j}, x_{2j}, \dots, x_{pj}, \boldsymbol{\theta})$  but this is a specialized topic we do not pursue here.

In this section we consider for the case of one explanatory variable,  $x$ , completely nonparametric models in the sense that no specific form is suggested for either the systematic or the random portions of the model. All we specify are some properties that each must obey. Some but not all of these methods extend to the situation in (8.14) with  $p$  explanatory variables. Many methods used with these completely nonparametric models are theoretically sophisticated and also need advanced software facilities for their implementation as well as expert guidance for their interpretation. For these reasons we only indicate the flavour of some possible approaches and give an indication when each may be useful and where more information about them may be found.

For our first example we consider a bivariate model that stipulates only that the systematic part of the model is a function  $f(x, \theta)$  which is a monotonic increasing (or decreasing) function of  $x$ , while the random component is identically distributed for any  $x$  with median zero, or equivalently, that  $\text{med}(Y|x) = f(x, \theta)$ .

The concept is useful when a regression of  $Y$  on  $x$  is clearly not a straight line but exhibits the property that  $\text{med}(Y|x)$  or that  $E(Y|x)$  increases as  $x$  increases in which case the regression is said to be monotonic increasing. Similarly if  $\text{med}(Y|x)$  decreases as  $x$  increases we say the regression is monotonically decreasing.

Conover (1999, Section 5.6) indicates one approach to fitting a monotonic regression, giving an example involving the time to fermentation ( $Y$ ) of grape juice when various amounts of sugar ( $x$ ) are added. It is reasonable to suppose in the light of experience and a knowledge of the chemistry of fermentation that  $E(Y|x)$  should decrease as  $x$  increases. It is also clear from a plot of the data that the relationship between  $E(Y|x)$  cannot be represented by a straight line. Because of a random component the observed  $(x_i, y_i)$  do not exhibit strict monotonicity due to variations from expectation for individual observations at the given  $x$ -values.

The theory behind the approach adopted by Conover is that if the regression or systematic part of the relationship between two variables is monotonic then there will be a linear relationship between their ranks. This in turn implies that if a monotonic regression model is appropriate for  $n$  paired observations  $(x_i, y_i)$  the near-monotonic relationship between observations will transform to a near-linear relationship between ranks. The essence of the method proposed by Conover is to transform the  $(x_i, y_i)$  to ranks  $(r_i, s_i)$  using mid-ranks for ties. A least squares straight-line regression is fitted to these ranks. A back-transforming procedure, involving linear interpolation if needed, is then used in a way described by Conover to obtain estimates  $\hat{y}_i$  of the mean or median of  $Y$  corresponding to each  $x_i$ . The estimated regression curve consists of piecewise linear segments joining points  $(x_i, \hat{y}_i)$  for adjacent values of  $x_i$ . In summary, the monotonicity property is encapsulated in the fitted curve by a series of straight-line segments each of which is hopefully a reasonable approximation to the unspecified true monotonic curve in that neighbourhood. Readers wishing to apply this method should refer to Conover for details and also for references to alternative methods that have been proposed for monotonic regression.

The idea of approximating to a true but unknown function representing a regression relationship by a sequence of joined-up

straight-line segments each of which might be expected to be a locally good approximation to some true unspecified function extends to situations where monotonicity is no longer assumed, but then rank-based methods can no longer be used. A number of approaches to this and more sophisticated modelling are discussed by Hollander and Wolfe (1999, Section 9.7) and by Ryan (1997, Chapter 10).

The processes described are called **smoothing processes** and are especially relevant to large data sets, often with  $n$  considerably greater than 50, and subject to what is conveniently described as appreciable fuzziness.

We assume a model of the form

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n$$

where the  $\varepsilon_i$  have median zero for all  $i$ . The form of  $f(x)$  is completely unspecified except that either  $\text{med}(Y_i|x_i)$  or  $E(Y_i|x_i) = f(x_i)$ . The aim is to estimate  $f(x_i)$  at each of the  $x_i$ . The resulting points should lie on a smoother curve than that for the original data.

A simple approach due to Cleveland (1979) uses what is known as a **running line** smoother. Suppose  $n = 60$  and that the  $(x_i, y_i)$  are arranged in ascending order of the  $x_i$ . We might take the first seven points and fit a straight line to them by least squares (or one of the other methods described in [Section 8.1](#) if we believed that this were more appropriate). We use this fitted line to give the usual linear regression estimate  $\hat{y}_4$  of  $f(x)$  at the mid-value  $x_4$  of the selected first seven  $x_i$ . The process is repeated replacing the first point  $(x_1, y_1)$  by  $(x_8, y_8)$  to give an estimate  $\hat{y}_5$  of  $f(x)$  at  $x_5$ . Continuing this way, using finally the last seven points we get estimates of  $f(x)$  at each  $x_i$  from  $x_4$  to  $x_{57}$ . There are two difficulties. First, why choose groups of seven successive observations? Second, we have no estimates corresponding to the end values  $x_1$  to  $x_3$  or  $x_{58}$  to  $x_{60}$ .

The number of points in each group, usually referred to as the window size, is arbitrary. Clearly a very small window size (say 3) will not give much smoothing and too large a window size (say 55) will force something like a straight-line relationship on the data even if this is not valid and will only give values at a handful of points in the middle of the range. Hastie and Tibshirani (1987) suggest that a window size including between 10–15 per cent of the total number of points is often reasonable. The method is unsatisfactory for dealing with values near the end of the data range.

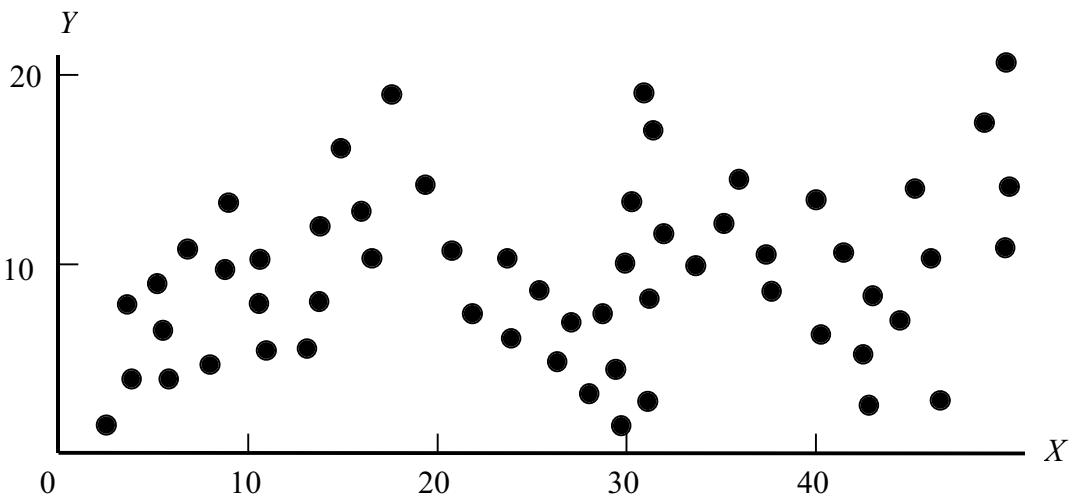
Kernel regression smoothing is a more sophisticated approach that also uses windows spanning groups of neighbouring points but gives

more weight to points near the centre of each window. Thus if, as in the situation discussed in the previous paragraph, we use a window spanning the values  $x_1$  to  $x_7$  to estimate  $f(x)$  at  $x_4$  more weight is now given to points at or near  $x_4$  than to the remote points  $(x_1, y_1)$  and  $(x_7, y_7)$ . The precise allocation of weights is determined by what is called a **kernel function**. There are a number of modifications to either running line or kernel smoothing.

One such alternative makes use of a curve fitting technique based upon the concept of splines. Splines are constructed by piecing together a number of polynomials that differ between adjacent intervals. The simplest are linear splines and indeed the straight line segments we described in the discussion of monotonic regression are linear splines. The points where the slopes change (in that example the  $x_i$ ) are called knots. More sophisticated splines may consist of, say, cubic polynomials determined and joined in such a way that there is continuity at the knots both in the  $y$  values and also in the first derivatives or slopes.

Another sophisticated approach to nonparametric regression uses roughness penalties. These are especially appropriate to data like that in [Figure 8.3](#). Visual examination suggests some sort of ‘wavy’ curve might be appropriate to indicate a relationship between the mean or median of  $Y$  for each  $x$ , but if there is little empirical or theoretical knowledge to suggest a specific form of  $f(x)$  a nonparametric fit seems appropriate. One might use a running line or kernel estimator smoother but the results may prove highly dependent upon the choice of window size. Instead, we may introduce what are called penalty functions. We give only a brief account. A full treatment of the theory and practical implications of the technique is given by Green and Silverman (1994).

The function  $f(x)$  in the systematic part of the model may take any form. In general, the more parameters that are needed to specify the model, the more flexible that model becomes. At one extreme if a model with sufficient flexibility is chosen the resulting estimation of parameters leads to a curve passing through all points. For example if there are  $n$  distinct points  $(x_i, y_i)$  we can always find a polynomial of degree not greater than  $n - 1$  to pass through all the points. This is a generalization of the obvious and well-known fact that one can always find a straight line to pass through two points and a quadratic to pass through three points (which reduces to a straight line if the points are collinear). One would usually not be happy to use the resulting polynomial for prediction or interpolation purposes. This is a clear case of over-smoothing.



**Figure 8.3** A noisy data set

The key idea behind curve fitting with a smoothness penalty (sometimes called penalized least squares regression) is that a term is added to the sum of squares  $\sum_i [y_i - f(x_i)]^2$  that is minimized in least squares. This additional term is given a form that reduces the effect of roughness while avoiding the difficulties of over-smoothing. For a reasonably smooth curve a measure of such roughness in a broad sense is the amount of change in the second derivative over the interval  $(a, b)$  where typically  $a = x_1$  and  $b = x_n$  for a data set with the  $x_i$  in ascending order. This is intuitively obvious if  $f(x)$  happens to be a polynomial because for a straight line the second derivative is zero so the roughness is always zero. There is no smoother curve than a straight line! For a quadratic the second derivative is constant. For a polynomial of higher degree,  $n$ , the second derivative is a polynomial of degree  $n - 2$  at most and in general this means the measure of roughness acts as a penalty because it tends to increase as the degree of the polynomial increases.

In a more general context where  $f(x)$  need not be a polynomial improved fit still tends to produce more roughness so the idea of a roughness penalty is that there is a pay-off between the closeness of fit at individual points and the smoothness of the fitted curve overall. The function to be minimized takes the form

$$U(f) = \sum_i [y_i - f(x_i)]^2 + \alpha \int_a^b [f''(x)]^2 dx$$

where  $\alpha$  is a positive constant and  $f''$  represents the second derivative of  $f$ . We denote by  $f^*$  the function that minimizes  $U(f)$ . The value of  $\alpha$  has to be chosen and clearly the greater the value of  $\alpha$  the greater the influence of the roughness penalty  $\int_a^b [f''(x)]^2 dx$  on the

solution. When  $\alpha \rightarrow 0$  the solution approaches least squares but when  $\alpha \rightarrow \infty$  for minimization of  $U(f)$  the second derivative  $f''(x)$  is forced towards zero, implying that the fitted curve does not change in slope and thus approaches a straight-line regression no matter what  $f$  is chosen.

A remarkable property of the method is that once  $\alpha$  is chosen we do not need to specify the family  $f$  to estimate  $f^*$ . This is because the latter can always be expressed in terms of cubic splines. A description of these or the fitting procedures is beyond the scope of this book. A detailed account of both the theory and application of nonparametric regression using penalty functions is given by Green and Silverman (1994), their Chapters 2 and 3 being devoted to practical aspects of the use of cubic splines in fitting.

Another approach to smooth curve fitting is described by Efron and Tibshirani (1993, Section 7.3) using what they call the **loess** method and a further robust method of fitting uses what are called  $M$ -estimators. Although we do not describe their application to regression a brief introduction to  $M$ -estimators is given in Section 11.4.

Pettitt (1983, 1985) discusses various aspects of regression with censored data.

## 8.4 OTHER MULTIVARIATE DATA PROBLEMS

Although they are not regression problems it is convenient to mention briefly here several other aspects of nonparametric analysis of multivariate data.

The univariate concept of the histogram is often associated with ‘smoothing techniques’ to iron out inherent data discontinuities. For multivariate data more sophisticated techniques are used. One is known as kernel density estimation and it has links with splines and other concepts used in sophisticated regression problems. A number of applications may be found in the literature. Kasser and Bruce (1969) measured a number of variables on patients with coronary heart disease and on normal patients. These included heartbeat rates at rest and after exercise. A three-dimensional plot of a non-parametrically estimated joint density function of these two variables for diseased and normal patients is given by Izenman (1991) and illustrates dramatically the differences in pattern for the two groups. Izenman also reviews the subject, giving over 200 references that include applications in fields as diverse as medicine and the interpretation of satellite pictures of the earth.

Ingenious extensions of nonparametric centrality tests to bivariate data were proposed by Brown (1983). He applied angular tests analogous to the sign test for ‘location’ shifts in a specific direction involving what he termed the **spatial median**. This is the centrality measure in two dimensions that minimizes absolute distances to observations. Brown et al. (1992) review various aspects of bivariate sign tests. Multivariate sign tests are discussed by Randles (1989) and Peters and Randles (1990) develop a multivariate signed-rank test for one-sample centrality problems.

## 8.5 FIELDS OF APPLICATION

Regression is a technique used (sometimes overused) in nearly all fields of application, so in this section we use a different format from that in other chapters, making only general comments on the use of the technique. In bivariate regression where least squares is widely used, the Theil–Kendall method provides a robust alternative especially useful when one is prepared to, or wishes to, minimize the influence of observations that do not follow a regression pattern that suggests a straight-line relationship for the bulk of the data. This is appropriate if one wishes to use the fitted line for forecasting or prediction in situations where the dominant pattern suggested by the bulk of the data is of major interest.

Before fitting any regression curve to bivariate data it is wise to plot the points. This will indicate whether a straight line is adequate for fitting or prediction purposes and whether we are likely to need a robust method of fitting due to presence of outliers, or whether a more sophisticated curve (perhaps a polynomial or a hyperbola) might be more appropriate, or whether a monotonic regression, for example, may be desirable.

In multiple regression or in situations where no simple model for the systematic part of a bivariate relationship appears obvious some of the approaches briefly indicated in [Section 8.3](#) should be considered.

## 8.6 SUMMARY

**Least squares regression** may be appropriate providing errors are independent and are symmetrically distributed all with the same variance even without a normality assumption. However, although

inferences may be based on the relevant permutation distribution there is a lack of robustness against certain departures from homogeneity of error distributions.

When the  $x_i$  are equally spaced, rank based methods computing **Spearman's rho** ([Section 8.1.2](#)) between the  $y_i$  and  $d_i = y_i - bx_i$  may be more robust than classic least squares.

The **Theil–Kendall method** ([Section 8.1.3](#)) often gives similar results to methods based on Spearman's rho, is easier to apply and is not restricted to equally spaced  $x_i$ . The point estimator is the median of the slopes of all pairwise joins of data points with different  $x$  coordinates. An **abbreviated Theil procedure** ([Section 8.1.4](#)) uses only a small subset of slopes that join independent pairs of points. Confidence intervals based on the Theil–Kendall method depend on the distribution of Kendall's tau for the complete procedure and upon the sign test for the abbreviated procedure.

Nonparametric analogues of some bivariate regression procedures extend in theory to **multiple regression** but software for their implementation is not readily available so one may have to resort to asymptotic results more frequently than one would ideally wish.

A more sophisticated development is completely **nonparametric regression** ([Section 8.3](#)) where no specific form is given a priori to the systematic part of the model and only very general assumptions are made about the error structure. The aim is often smoothing of rather diffuse data sets primarily to discern trends and if appropriate then to make predictions or forecasts. Models range from those for **monotonic regression** to sophisticated models using **penalty functions** designed to avoid over-smoothing.

## EXERCISES

- 8.1 Show that if a sample of  $n_1$  observations has mean  $m_0$  and another sample of  $n_2 = n - n_1$  has mean  $m_1$  and the first sample values are ‘indexed’ by  $x = 0$  and the second sample values by  $x = 1$  then  $\beta$  given by [\(8.3\)](#) equals  $m_1 - m_0$ .
- 8.2 For the data in [Example 8.1](#) verify that the sum of residuals  $e_i$  and the sum of the products of residuals with the corresponding  $x_i$  is zero. Explain why this is an inevitable consequence of the least squares estimation procedure.
- 8.3 The numbers of rotten oranges ( $y$ ) in 10 randomly selected boxes from a large consignment are counted after they have been kept in storage for a stated number of days ( $x$ ). Use the Theil–Kendall method to compute the

slope of a straight line fitted to these data and obtain an appropriate estimate of the intercept.

$x$	3	5	8	11	15	18	20	25	27	30
$y$	2	4	7	10	17	23	29	45	59	73

- 
- Plot these data and the fitted line. Does the fit seem reasonable?
- 8.4 Using the asymptotic result relevant to the Theil–Kendall method given in [Section 8.1.3](#) obtain an approximate 95 per cent confidence interval for the data used in [Example 8.3](#). Compare this with the exact permutation-based interval obtained in that example.
- 8.5 Verify the confidence limits quoted in [Example 8.4](#) using the abbreviated Theil method.
- 8.6 Comment critically on and explain the meaning of the following statement: Given  $n$  observations  $(x_i, y_i)$  where the  $x_i$  are equally spaced, the Cox–Stuart test for trend ([Section 3.2.3](#)) applied to the  $y_i$  is essentially equivalent to testing whether the abbreviated Theil estimator is or is not consistent with zero slope.
- 8.7 Mattingley (1987) gives the following data based on the US census of agriculture which gives at approximately 10 year intervals from 1920 to 1980 the percentages of US farms with tractors and farms with horses. Explain why it would be pointless, or wrong, to fit a linear regression for tractor percentage on horse percentage to these data. Suggest what alternative type(s) of regression might be more appropriate.

<i>Per cent tractors</i>	9.2	30.9	51.8	72.7	89.9	88.7	90.2
<i>Per cent horses</i>	91.8	88.0	80.6	43.6	16.7	14.4	10.5

- 
- 8.8 Gat and Nissenbaum (1976) give ammonia concentration ( $y \text{ mg l}^{-1}$ ) at various depths ( $x \text{ m}$ ) in the Dead Sea. Fit a linear regression for concentration on depth using the Kendall–Theil method and obtain an approximate 95 per cent confidence interval for  $\beta$ .

$x$	25	50	100	150	155	187	200	237	287	290	300
$y$	6.13	5.51	6.18	6.70	7.22	7.28	7.22	7.48	7.38	7.38	7.64

- 
- 8.9 Katti (1965) gave data for weight of food eaten ( $x$ ) and weight gain ( $y$ ) for 10 pigs fed on one type of food ( $A$ ) and for 10 fed on a second type ( $B$ ). Use the abbreviated Theil method to fit linear regressions to each and test whether the hypothesis that the true slopes  $\beta_1, \beta_2$  are equal is supported.

<i>A</i>	$x$	575	585	628	632	637	638	661	674	694	713
	$y$	130	146	156	164	158	151	159	165	167	170
<i>B</i>	$x$	625	646	651	678	710	722	728	754	763	831
	$y$	147	164	149	160	184	173	193	189	200	201

8.10 In [Section 8.1.7](#) (p. 298) we gave the data set

---

$x$	1	2	3	99	100	101
$y$	2.5	3.1	3.4	99.6	100.1	101.1

---

Verify that the  $y$  values are consistent with the explanation given in that section of the model used to obtain them. Also use the Theil–Kendall method to obtain a point estimate of  $\beta$  together with a nominal 95 per cent confidence interval.

# Categorical Data

## 9.1 CATEGORIES AND COUNTS

Data often consist of counts of the number of units (people, institutions, towns, countries, items, etc.) with given attributes. It is often convenient to present these in one-, two-, three- or higher-dimensional tables usually referred to as one-way, two-way, three-way, etc. contingency tables. Each *dimension* or *way* corresponds to a classification into categories representing attributes. These may be **explanatory** (e.g. dose levels of a drug; the names of several different drugs; gender; psychiatric diagnoses; ethnic groups; income levels). Alternatively they may be **responses** (e.g. side-effects of drugs classified as none, slight, moderate, severe; blood pressure levels after administration of a drug; examination grades). The attributes are often qualitative; if there is no natural ordering they are described as **nominal** (e.g. psychiatric diagnoses; ethnic groups). Attributes that may be arranged in a natural order are described as **ordinal** (e.g. reactions to a drug classified as slight, moderate, severe; grouping by age under 50 and age 50 and over).

This and the next chapter deal with problems that at first sight appear different from any previously considered, yet many are solved using procedures we have already developed. This is because we can re-express many problems met in earlier chapters in an equivalent contingency table format.

This chapter is mainly about two-way tables consisting of two or more rows and columns. Typically, each of the rows represents either a level of an explanatory attribute or a level of response, and each of the columns represents a level of response. [Table 9.1](#) is an example with two rows and five columns. A question often asked is whether there is evidence that the incidence rate of side-effects differs between drugs. The null hypothesis of independence is often expressed as one of **no association** between row and column categories. Generally this means that the population distribution of the possible column outcomes is the same for each row category.

**Table 9.1** Numbers of patients showing side-effects for two drugs.

	Side-effect level				
	None	Slight	Moderate	Severe	Fatal
Drug A	23	8	9	3	2
Drug B	42	8	4	0	0

In [Table 9.1](#) the row categories are explanatory and nominal. However, had each row referred to different dose levels of the same drug they would have been ordinal explanatory categories. The column categories refer to ordinal responses that reflect seriousness.

We have already met situations where we formed contingency tables deduced from measurement data by counting the numbers of measurements falling into defined categories. For instance, Tables 5.2 and 6.4 involve counts from two and six samples, respectively, of numbers in each of two categories (above and below an overall sample median  $M$ ). The column categories were ordinal responses (' $<M$ ' and ' $>M$ '). The row categories (samples) were explanatory. In that context whether the latter are ordered depends upon the nature of the samples. If successive samples corresponded to increasing dose levels of a drug they would be ordinal; if they represented psychiatric diagnoses or each represented people of a different ethnic group they would be nominal.

Although the advantage of doing so is not immediately obvious we may also present the data in [Example 5.1](#) as a  $2 \times 21$  contingency table. The rows correspond to Group A and Group B and each column corresponds to one of the times taken to carry out the calculations, these times being arranged in ascending order. The numbers in the cells of this table are either 0 or 1. An entry 1 is placed in the cell in the first row of a time column if someone from Group A took that time to complete the calculations and similarly a 1 is entered in the appropriate cell in the second row for those in Group B. All other cell entries are zero. This leads to [Table 9.2](#). We show why this is useful in [Sections 9.3.1](#) and [10.2.6](#).

In [Example 5.4](#) a contingency table format was used for what were for practical purposes 'tied' data and there a contingency table was a natural way to present the data. As we have already indicated this correspondence between tables of measurements and 'counts' of units or subjects exhibiting each measurement provides a useful tool for showing relationships between methods.

**Table 9.2** The data in Example 5.1 presented in a contingency table.

Time	17	18	19	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	39	41
Group A	1	1	1	0	1	1	0	1	1	0	0	1	0	1	0	1	0	0	0	0	0
Group B	0	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	1	1	1	1	1

Many tests for lack of association applicable to nominal categories in both rows and columns may still be applied when either or both of row and column categories are ordered. Generally, however, more powerful procedures exist that take account of ordering.

Sometimes both rows and columns represent response categories. For instance, a particular antibiotic can be used to treat both bronchitis and otitis media (infection of the middle ear) in the same patient (Hornbrook, Hurtado and Johnson, 1985). **Table 9.3** shows a data set where the antibiotic is prescribed to 60 patients who have both health problems. Its effect is noted on changes (worse, the same, better) in bronchitis (rows) and otitis media (columns). Such a variation in outcome might occur in practice, as a substantial minority of the patients may not take the prescribed antibiotic leading to a worsening in their condition. Since, for each illness the three outcomes are mutually exclusive, all subjects can be classified into one and only one combination of row and column categories.

This may not be the case in a different study; in asthma the outcomes of interest might be no symptoms, cough and wheeze. It is possible for individuals to have both cough and wheeze so a fourth category of 'both' would be needed to take this into account.

We have shown row and column totals and the total number of patients, 60, in the bottom right of Table 9.3. This is a common practice for contingency tables because totals play a prominent part in analyses. In Table 9.3 both row and column categories are ordinal;

**Table 9.3** Outcome for patients who receive a prescription for a particular antibiotic.

Bronchitis	Otitis media			Total
	Worse	The same	Better	
Worse	13	5	6	24
The same	1	19	4	24
Better	4	0	8	12
Total	18	24	18	60

in the study of cough and wheeze in asthma outlined above there is no natural ordering and these categories would be nominal.

### 9.1.1 Some basic concepts

Before giving a more formal treatment of the analysis of contingency tables we illustrate some basic concepts using the antibiotic treatment data above. The appropriate null hypothesis for [Table 9.3](#) asserts that there is no association between the changes in the severity of bronchitis and changes in the severity of otitis media. So, in the population of such patients who receive this antibiotic, the proportions of patients in the categories for outcome in otitis media are the same whatever the change (if any) in the severity of the bronchitis. Under this null hypothesis, the expected number of patients in a given cell can be calculated using the row and column totals. For instance, if outcomes in bronchitis and otitis media are indeed unrelated we would expect  $(18 \times 24)/60$  or 7.2 patients to deteriorate in both conditions. A table of expected values can be constructed for the cells, as shown in [Table 9.4](#).

If the sample reflected the null hypothesis exactly, the observed and expected values in each cell would be the same apart from the 'rounding' needed to produce integer values. Is it likely that the absolute differences observed - ranging in value from 0.4 to 9.4 - could have come about purely by sampling fluctuation if the null hypothesis is true for the population?

The classic way to address this problem is to calculate the chi-squared statistic (dealt with in detail in [Section 9.2.2](#)). This summary measure assesses the differences between observed and expected values, producing a number that is large if there is strong evidence

**Table 9.4** Expected numbers of patients in [Table 9.3](#) under the null hypothesis of independent outcomes for bronchitis and otitis media. Observed values from Table 9.3 are given in italics in brackets.

Bronchitis	Otitis media			Total
	Worse	The same	Better	
Worse	7.2 (13)	9.6 (8)	7.2 (6)	24
The same	7.2 (1)	9.6 (19)	7.2 (4)	24
Better	3.6 (4)	4.8 (0)	3.6 (8)	12
Total	<u>18</u>	<u>24</u>	<u>18</u>	<u>60</u>

against the null hypothesis but close to zero if there is little evidence. This statistic takes into account the sample size and the numbers of rows and columns (large tables tend to produce larger values of the statistic). Most statistical packages calculate this statistic along with a  $P$ -value that is used to assess the strength of the evidence against the null hypothesis. For [Table 9.4](#) association appears likely. Indeed, the calculated chi-squared value is large with corresponding  $P < 0.0001$ . Changes in condition in otitis media and bronchitis seem to be positively associated. Since row and column categories are ordered it is possible to get a more powerful test than the above if this ordering is taken into account.

### 9.1.2 Some models for counts in $r \times c$ tables

Books on categorical data analysis in both parametric and nonparametric contexts include Bishop, Fienberg and Holland (1975), Breslow and Day (1980), Plackett (1981), Agresti (1984, 1990, 1996), Christensen (1990), Everitt (1992) and Lloyd (1999).

In general  $r$  and  $c$  are used to denote the numbers of rows and columns in a table referred to as an  $r \times c$  (in speech ' $r$  by  $c$ ') table. For [Table 9.1](#),  $r = 2$  and  $c = 5$ . It is convenient to represent a general  $r \times c$  table in the form of [Table 9.5](#). The entry  $n_{ij}$  at the intersection of the  $i$ th row and  $j$ th column, referred to as cell  $(i, j)$ , represents the number of items having that combination of row and column attributes. The total count for row  $i$  is denoted by  $n_{i+}$ , i.e.  $n_{i+} = \sum_j n_{ij}$ . Similarly,  $n_{+j}$  is the total count for column  $j$ , and  $N = \sum_i n_{i+} = \sum_j n_{+j} = \sum_{i,j} n_{ij} = n_{++}$  is the total count over all cells.

The methods we develop are nonparametric in the sense that, although several parametric models may give rise to an observed distribution of counts in a contingency table, the inference procedures we use often do not depend upon which of these models we choose or upon the precise values of certain parameters, since these are not specified in our hypotheses. Technically, we avoid the problem of specifying parameter values by a device known as conditional inference. This makes use of properties of sufficient and ancillary statistics, a topic beyond the scope of this introductory text. Readers familiar with the basic theory of maximum likelihood will find an excellent account of inference theory for categorical data in Agresti (1990, Chapter 3) while Agresti (1996) deals at a slightly less technical level with both classic and modern methods of categorical data analysis.

**Table 9.5** A general  $r \times c$  contingency table.

				Column categories			<i>c</i>	Total
	1	2	3	.	.	.	$n_{1c}$	$n_{1+}$
	$n_{11}$	$n_{12}$	$n_{13}$	.	.	.	$n_{2c}$	$n_{2+}$
Row	1	$n_{21}$	$n_{22}$	$n_{23}$	.	.	.	.
categories	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
<i>r</i>	$n_{r1}$	$n_{r2}$	$n_{r3}$	.	.	.	$n_{rc}$	$n_{r+}$
Total	$\overline{n_{+1}}$	$\overline{n_{+2}}$	$\overline{n_{+3}}$	.	.	.	$\overline{n_{+c}}$	$\overline{N}$

When making inferences we nearly always regard row and column totals as fixed; this is the *conditional* feature. It is justified by the concept of conditional inference, but is **not** in general a consequence of the experimental set-up. Sometimes neither row nor column totals are fixed a priori by the nature of the experiment, in other cases one of these sets is fixed but not the other and only exceptionally are both fixed. For example, if we are comparing two drugs, A and B, that will result in either an improvement (a success, denoted by S) or no improvement (a failure, denoted by F) we may have 19 patients and allocate 9 patients at random to receive drug A, the remaining 10 to receive drug B, and observe the following results:

	S	F	Total	
Drug A	7	2	9	
Drug B	5	5	10	(9.1)
Total	12	7	19	

Here row totals are fixed by our choice of numbers to be allocated to each drug. If we regard the 19 patients as the entire population of interest then under the null hypothesis of *independence* or *no association* (*i.e.* that the drugs are equally effective) we are justified in regarding the total number of successes, 12, as fixed. This is irrespective of how we allocate patients to the two groups of 9 and 10 corresponding to the two drugs. Just as in the situation in Example 1.4, it is extreme outcomes under randomization like none

(or 9) of the 12 successes being among the 9 patients allocated to drug A that make us suspect the null hypothesis is not plausible. As in other permutation tests, we may compute the probability of each possible outcome under  $H_0$  when the marginal totals are fixed. The situation is different when we sample our patients as two groups of 9 and 10 from a larger population. Even under  $H_0$ , if we took another pair of samples of 9 and 10 (the fixed row totals representing our designated sample sizes) it is unlikely that the combined number of successes for these different patients would again be 12. We might get an outcome like that in (9.2).

	<i>S</i>	<i>F</i>	<i>Total</i>	
<i>Drug A</i>	6	3	9	
<i>Drug B</i>	3	7	10	(9.2)
<i>Total</i>	<u>9</u>	<u>10</u>	<u>19</u>	

However, the hypothesis of independence does not involve the actual probability of success with each drug, but stipulates only that the probability of success,  $\pi$ , say, is the same for each drug. In practice, we usually do not know the value of  $\pi$ . However, if we assume there is such a common value for both drugs, we can obtain an estimate based on the total number of successes, 12 from 19, in our sample of 19 in our observed situation, i.e. that in (9.1). We can use this information to compute the expected number of successes in sub-samples of 9 and 10 from the total of 19 under the assumption that  $H_0$  is true. Again, if we observe big discrepancies between the observed and expected numbers of successes computed in this way for the two drugs, this is an indication of association, i.e. that the probability of success for drug B is not the same as the probability for drug A. In effect, what we are doing is looking at a permutation distribution once more, but only in possible samples which have fixed column totals (in our case 12, 7 respectively) as well as fixed row totals. Such conditional inference is perfectly valid, and indeed it is the entries in the body of any such table with fixed marginal totals that provide the evidence for or against independence. As indicated above the column marginal totals in this example tell us something about the likely value of  $\pi$ , the probability of success if indeed  $H_0$  is true, but they contain little information as to whether there is really a common probability,  $\pi$ , of success for each drug. It is the individual cell values that contain virtually all the information

on this latter point. Similar arguments apply for regarding all marginal totals in an  $r \times c$  table as fixed for nearly all the models we consider in this and the next chapter. Methods for inference that do not require all marginal totals to be fixed have been proposed and are used. However, there is near-unanimity among statisticians that conditional inference with fixed marginal totals is the most appropriate way to test for independence (although independence is not the only thing we may want to test in contingency tables). Conditional inference is convenient, valid and overcomes technical problems with hypothesis tests in this context. For enlightening discussions about this approach and approaches that do not assume marginal totals are fixed see Yates (1984), Upton (1992) and a wide-ranging paper by Agresti (1992), also Freidlin and Gastwirth (1999). We show in this and the next chapter how many permutation distribution tests already met in this book can be reformulated as tests involving contingency tables with fixed row and column totals.

Experimentally, we might arrive at the outcome in (9.1) without pre-fixing even the row totals. For example, the data might have come from what is called a retrospective study, by simply going through hospital records of all patients treated with the two drugs during the last 12 months and noting the responses. Our method of testing for no association using conditional inference would be the same, i.e. conditional upon the observed row and column totals.

For elucidation, we describe briefly several models used to describe counts, indicating where each might be appropriate, and show that each leads to the same conditional inference procedures. We confine this introductory discussion largely to  $2 \times 2$  tables. The reader who is not familiar with the multinomial and Poisson distributions and the concept of conditional and marginal distributions may wish to omit the remainder of this section and move directly to [Section 9.2](#), where we describe applications.

A typical model for the situation exemplified by (9.1), where the row classifications are explanatory and the column classifications are responses, is to assume that each row has associated with it independent binomial distributions and that the cell entries are the numbers of occurrences of a response A and a response B for a  $B(n_{1+}, p_1)$  distribution in the first row and for a  $B(n_{2+}, p_2)$  distribution in the second row. In the case of (9.1) we have  $n_{1+} = 9$  and  $n_{2+} = 10$  so that under this model for each cell we obtain probabilities  $p_{ij}$  for the observed numbers  $n_{ij}$  as follows:

$$p_{11} = \Pr(n_{11} = 7) = \binom{9}{7} p_1^7 (1 - p_1)^2, \quad p_{12} = 1 - p_{11}, \\ p_{21} = \Pr(n_{21} = 5) = \binom{10}{5} p_2^5 (1 - p_2)^5, \quad p_{22} = 1 - p_{21}.$$

The hypothesis of no association implies  $H_0: p_1 = p_2$  but the associated common value  $\pi$ , say, is not specified. Independence of response from patient to patient and the additive property of binomial variables imply that under  $H_0$  the probability of observing a total of 12 favourable outcomes in 19 observations (the column total line) is

$$p_{+1} = \binom{19}{12} \pi^{12} (1 - \pi)^7.$$

Since the classifications of individuals are independent, it follows from standard definitions of conditional and independent probabilities that the probability of observing 7 successes ( $= n_{11}$ ) in 9  $n_{1+}$  trials, conditional upon having observed a total of 12 successes in 19 trials, is

$$P^* = p_{11} p_{21} / p_{+1}.$$

Thus, when  $\pi = p_1 = p_2$

$$\begin{aligned} P^* &= [(\binom{9}{7})\pi^7(1-\pi)^2 \times (\binom{10}{5})\pi^5(1-\pi)^5] / [\binom{19}{12}\pi^{12}(1-\pi)^7] \\ &= [(\binom{9}{7}) \times (\binom{10}{5})] / (\binom{19}{12}) \\ &= [(9!) \times (10!) \times (12!) \times (7!)] / [(19!) \times (7!) \times (2!) \times (5!) \times (5!)] = 0.180 \end{aligned}$$

As indicated in [Section 5.3.1](#), for a general  $2 \times 2$  table with cell entries

$$\begin{matrix} a & b \\ c & d \end{matrix}$$

this generalizes to

$$P^* = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!} \quad (9.3)$$

and this does not involve  $\pi$ .

Equation (9.3) represents the frequency or probability mass function for the **hypergeometric distribution**. We saw in [Section 5.3.1](#) that when we fix marginal totals,  $P^*$  depends essentially only on  $a$ , since, given  $a$ , then  $b, c$  and  $d$  are automatically fixed to give the correct marginal totals. The fact that fixing  $a$  fixes all cell entries is summarized by saying that a  $2 \times 2$  contingency table has one degree of freedom. The value of  $P^*$  given by (9.3) is relevant to testing for independence (no association). We have already

mentioned in Section 5.3 that the test based on the hypergeometric distribution is the Fisher exact test which we described with an application to a  $2 \times 2$  table in Example 5.6. By calculating  $P^*$  for all possible values of  $a$  we may set up critical regions in the tail of the distribution relevant to assessing the strength of evidence against  $H_0$  in either one- or two-tail tests, as we did in Example 5.6. We sometimes want estimates of the expected values of each  $n_{ij}$  under  $H_0$ . If we knew  $\pi$  we could obtain exact expected values, but when  $\pi$  is unknown we base our estimates on the marginal totals, using these to estimate  $\pi$ . Since  $n_{+1}$  has expected value  $N\pi$ , we estimate  $\pi$  by  $\hat{\pi} = (n_{+1})/N$  and the expected value of  $n_{11}$  is  $m_{11} = n_{1+}\hat{\pi} = (n_{1+}n_{+1})/N$ . Similarly, we establish the general result  $m_{ij} = (n_{i+}n_{+j})/N$ . These expected values sum to the marginal totals.

More generally, if there is association in a  $2 \times 2$  table, then clearly, in (9.1) for example, if the binomial probabilities were known to be  $p_1, p_2$  the expected values in that case would be

$$\begin{array}{ll} 9p_1 & 9(1-p_1) \\ 10p_2 & 10(1-p_2) \end{array} \quad (9.4)$$

The quotients,  $p_1/(1-p_1)$  and  $p_2/(1-p_2)$ , represent the odds on the outcome 'success' for each row. A quantity often used as an indicator of association is the odds ratio, defined as

$$\theta = \frac{p_1 / (1-p_1)}{p_2 / (1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

From (9.4) the odds ratio is also the ratio of cross products of the cell expected frequencies  $m_{ij}$ , i.e.  $\theta = (m_{11}m_{22} / m_{12}m_{21})$ . Clearly, if  $p_1 = p_2$  then  $\theta = 1$ . This is also true for the estimated expected values calculated under  $H_0$ , i.e.  $m_{ij} = (n_{i+}n_{+j})/N$ . If, as usual, we do not know  $p_1, p_2$  we may estimate  $\theta$  by the sample odds ratio, which is defined as  $\theta^* = (n_{11}n_{22} / n_{12}n_{21})$ . In Section 10.2 we give tests for association, including a test for whether  $\theta^*$  is consistent with  $H_0: \theta = 1$ .

If  $p_1 = 3/4$  and  $p_2 = 1/2$  it is intuitively obvious that we expect more successes with drug A than with drug B and from (9.4) we find  $m_{11} = 6.75$  and  $m_{21} = 5$ . These expected values no longer add to the column total as do the expected values under  $H_0$ ; this should cause no surprise as we are dealing with a situation where the probabilities of success are known and different for each drug. For these  $p_1, p_2$  values  $\theta = (3/4 \times 1/2) / (1/2 \times 1/4) = 3$ . If we interchange probabilities so that  $p_1 = 1/2$  and  $p_2 = 3/4$  we find  $\theta = 1/3$ . Both these cases reflect the same degree of association, but in opposite directions. This is

demonstrated more clearly if we use  $\varphi = \log \theta$  as a measure of association, for now

$$\log 3 = -\log (1/3)$$

and generally

$$\log \theta = -\log (1/\theta).$$

Possible values of  $\varphi$  range from  $-\infty$  to  $\infty$  and when there is no association  $\varphi = 0$ . While the above relationships are true for logarithms to any base, most formulae for testing and estimation are based on natural logarithms using the base 'e'. Such logarithms are sometimes denoted by 'In' rather than 'log', but we use the latter.

The binomial model given above is not appropriate if our data are for two response classifications, like those in the tableau (9.5) giving numbers of patients responding in specified ways to a drug.

		Blood cholesterol level		Total
		Increased	Unchanged	
Blood Pressure	Higher	14	7	21
	Lower	11	12	23
Total		<u>25</u>	<u>19</u>	<u>44</u>

Here all four cells are on a par in the sense that all we can say is that with an unknown probability  $p_{ij}$  any of the  $N = 44$  subjects may be allocated to cell  $(i, j)$ . The multinomial distribution is an extension of the binomial distribution with independent allocation of  $n$  items to  $k > 2$  categories with specified probabilities for each category. Here we have 4 categories, one corresponding to each cell. The probabilities are subject to the constraint  $p_{11} + p_{12} + p_{21} + p_{22} = 1$ , since each person is allocated to one and only one cell. We specify by  $p_{1+}, p_{2+}$  the marginal probabilities that items will fall in row 1 or row 2, and likewise we specify marginal probabilities  $p_{+1}, p_{+2}$  for columns. For independence between rows and columns, standard theory gives the relationship  $p_{ij} = p_{i+}p_{+j}$  for  $i, j = 1, 2$ .

Even if there is independence we still do not know the value of the  $p_{ij}$ , but this presents no problem if we use conditional inference. Assuming no association (i.e. independence), we base our estimates on marginal totals, estimating  $p_{it}$  by  $n_{it}/N$  and  $p_{+j}$  by  $n_{+j}/N$ . Since  $p_{ij} = p_{it}p_{+j}$  if there is independence we estimate  $p_{ij}$  by  $p_{ij} = n_{it}n_{+j}/N^2$ .

It follows from properties of the multinomial distribution that we may then obtain expected estimated counts as  $\hat{m}_{ij} = N \hat{p}_{ij} = n_{i+}n_{+j}/N$ , exactly as in the two-sample binomial model, and proceed to tests as in that model. It is easy to show that under independence the odds ratio, here defined as  $\theta = (p_{11}p_{22})/(p_{12}p_{21})$ , is unity (hint: put  $p_{11} = p_1p_{11}$ , etc.). If the  $p_{ij}$  are estimated under  $H_0$  in the way indicated above, using these estimates or the estimated expected frequencies  $\hat{m}_{ij}$  gives  $\theta = 1$ . The sample odds ratio is  $\theta^* = (n_{11}n_{22})/(n_{12}n_{21})$ .

Another model often relevant is one where counts in each cell have independent Poisson distributions with a mean (Poisson parameter)  $m_{ij}$  for cell  $(i, j)$ . If  $\theta = (m_{11}m_{22})/(m_{12}m_{21}) = 1$ , there is no association between rows and columns. In general, we do not know the  $m_{ij}$ , but it can be shown that, again under  $H_0$ , these may be estimated from marginal totals in the way we do for the row-wise binomial and the overall multinomial models.

Extension of the three models above to  $r \times c$  tables is reasonably straightforward. For  $c > 2$ , binomial sampling within rows generalizes to multinomial sampling within rows. For this, or the overall multinomial sample or the Poisson count situation, under the null hypothesis of no association the expected cell frequencies are estimated in the notation of [Table 9.5](#) by  $\hat{m}_{ij} = n_{i+}n_{+j}/N$ ,  $i=1, 2, \dots, r$  and  $j = 1, 2, \dots, c$ . An odds ratio may be defined for any two rows  $i, k$  and any two columns  $j, l$  as

$$\theta_{ijkl} = (\hat{m}_{ij}\hat{m}_{kl})/(\hat{m}_{il}\hat{m}_{kj})$$

and  $\theta_{ijkl} = 1$  for all  $i, j, k, l$  if there is no association. The sample odds ratios are  $\theta^*_{ijkl} = (n_{ij}n_{kl})/(n_{il}n_{kj})$ .

The Fisher exact test extends to  $r \times c$  tables in a way described in [Section 9.2.1](#).

## 9.2 NOMINAL ATTRIBUTE CATEGORIES

Tests developed in this section are appropriate when row and column categories are both nominal. They are often used for tests of no association when attributes are ordered, but tests described in [Section 9.3](#) or in [Chapter 10](#) are then often more appropriate.

We describe three commonly used test procedures. Confusion exists in the literature because each often (but not invariably) leads to essentially equivalent permutation distributions for  $2 \times 2$  tables, but to different asymptotic test statistics even in this case. For general  $r \times c$  tables both the exact permutation distribution of the

test statistics and the asymptotic test statistics differ. However, one should not exaggerate the importance of these differences, for unless the choice of a precise significance level is critical it is seldom that the three possible exact tests lead to markedly different conclusions. However, there may be considerable differences between exact and asymptotic tests in some circumstances. Numerical values of the three test statistics often differ considerably, even for moderately large samples, but each statistic has asymptotically a chi-squared distribution with  $(r - 1)(c - 1)$  degrees of freedom under the hypothesis of independence (no association) between row and column classifications.

The three tests are

- The **Fisher exact test** (also known as the Fisher-Irwin or as the Fisher-Freeman-Halton test).
- The **Pearson chi-squared test**.
- The **Likelihood ratio test**.

Practice has often been to use the Fisher exact test when asymptotic theory is clearly inappropriate and either of the others when asymptotic results are thought to be acceptable. Despite differences in numerical value of the relevant test statistics and the fact that exact permutation distribution theory differs in each case, there seems no compelling reason to depart from this practice.

Tests based on exact permutation theory, except in the case of  $2 \times 2$  tables (where the Fisher exact test is commonly used) are not practical without relevant software. StatXact allows exact permutation tests based on all three methods and some facilities are also included in Testimate and in more general software, so there is now less reason to use asymptotic theory when it is of doubtful validity. This is particularly relevant for 'sparse' contingency tables in which there are many low cell counts even though  $N$  may be large. That situation is not uncommon in medical applications where responses may be recorded on a limited number of diseased patients because resources are in short supply or responses in certain categories are few in number (see e.g. Exercise 9.23).

### 9.2.1 The Fisher exact test

The Fisher exact test for  $2 \times 2$  tables based on the hypergeometric distribution specified in (9.3) was described by Fisher in various editions of his famous book *Statistical Methods for Research Workers* dating from the 1930s and was extended to  $r \times c$  tables by Freeman and Halton (1951). In the notation of [Table 9.5](#), the

generalized hypergeometric distribution relevant to an  $r \times c$  table gives the probability of an observed configuration under  $H_0$ , conditional upon fixed marginal totals as

$$P^* = \frac{\prod_i (n_{i+} !) \prod_j (n_{+j} !)}{n! \prod_{i,j} (n_{ij} !)} \quad (9.6)$$

The  $P$ -value for the test of no association is the sum of all probabilities for configurations with the given marginal totals that have the same or lower probability.

### Example 9.1

*The problem.* Scarlet fever is a childhood infection that among other symptoms gives rise to severe irritation of the nose, throat and ears. In a developing country in which scarlet fever is still a problem, six districts A to F were chosen. In each district, patients were located and the parents were asked to state the site at which they thought their child's irritation was the worst. The numbers in each response category are:

	<i>District</i>						
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>Total</i>
<i>Nose</i>	1	1	0	1	8	0	11
<i>Throat</i>	0	1	1	1	0	1	4
<i>Ears</i>	1	0	0	0	7	1	9
<i>Total</i>	2	2	1	2	15	2	24

Is there evidence of association between districts and sites of greatest perceived irritation?

*Formulation and assumptions.* We use the Fisher exact test for independence and the program in StatXact.

*Procedure.* StatXact gives the probability of observing this configuration under  $H_0$ : *no association* to be 0.0001 and the probability of this or a less probable configuration is  $P = 0.0261$ . It also gives a probability of 0.0039 of obtaining a configuration with the same probability as that observed, indicating that more than one configuration has this probability.

*Conclusion.* Since  $P = 0.0261$  there is fairly strong evidence against  $H_0$ .

*Comments.* 1. It is evident that the main cause of association is the relatively high number of responses of greatest irritation at the nose or ears (but not throat) in district E. Clearly, this district has responses out of line with the general trend, whereas the other districts produce reasonably consistent results for all three

sites of irritation. It may be that in district E more importance is placed on visual appearance compared with the other five districts.

2. The finding of dependence is not surprising, so it is relevant to note that asymptotic theory here does not indicate significance at a conventional 5 per cent level. We discuss the more commonly used asymptotic statistics in Sections 9.2.2 and 9.2.3, but see also the remarks below under *Computational aspects*.

3. While it is impractical to compute manually all relevant permutation distribution probabilities, it is not difficult to verify that for the given configuration (9.6) shows  $P^* = 0.000058$ , as indicated (after rounding) by StatXact (see Exercise 9.1).

4. We are not testing whether incidence levels differ between districts, but whether the proportions at irritation sites differ between districts (or equivalently, whether the proportions for districts differ between irritation sites). Had there, in addition, been 8 cases of greatest irritation in the throat in district E we would have found little evidence against the hypothesis of *no association*, although clearly the numbers of cases at all three irritation sites would be much higher for district E.

5. The test of association is essentially two-tail, the situation being broadly analogous in this aspect to that in the Kruskal-Wallis test where we test for differences in centrality which may be in any direction, or in different directions in populations corresponding to different samples.

*Computational aspects.* 1. A program like that in StatXact is invaluable in situations where asymptotic results may be misleading. Although seldom used in practice, a statistic is given in StatXact which is a monotonic function of  $P^*$  given by (9.6). Asymptotically this statistic has a chi-squared distribution with  $(r - 1)(c - 1)$  degrees of freedom. This statistic, denoted by  $F^2$ , takes the value 14.30 for this example, and  $\Pr(\chi^2 > 14.30) = 0.1597$ . This is in sharp contrast to the result from the exact test, for we do not establish significance at any conventional level with the asymptotic test. In practice, the asymptotic tests given in [Sections 9.2.2](#) and [9.2.3](#) are more widely used.

2. The Fisher exact test for  $r \times c$  tables is available in Stata. If the table has many cells and the sample size is large the computation of the P-value may take several minutes. Many general packages include the Fisher exact test for  $2 \times 2$  tables only.

Example 9.1 is not atypical; the StatXact manual gives very similar data for location of oral lesions in a survey of three regions in India. The asymptotic test has  $(r - 1)(c - 1)$  degrees of freedom in an  $r \times c$  table because, if we allocate arbitrarily cell values in, say, the first  $r - 1$  rows and  $c - 1$  columns the last row and column values are fixed to ensure consistency with the row and column totals.

## 9.2.2 The Pearson chi-squared test

An alternative statistic for testing independence of row and column categories is the Pearson chi-squared statistic. Its appeal is

- Ease of computation.
- Fairly rapid convergence to the chi-squared distribution as count size increases.
- Intuitive reasonableness.

An exact test is based on the permutation distribution of this statistic over all sample configurations having the fixed marginal totals. Computation of P-values is no easier than it is for the Fisher exact test and only for  $2 \times 2$  tables are the exact tests often equivalent.

Denoting as before the expected number in cell  $(i, j)$  by  $m_{ij} = n_i \cdot n_{+j}/N$ , then  $(m_{ij} - n_{ij})^2$  is a measure of departure from expectation under  $H_0$ . It is reasonable to weight this measure inversely by the expected count, since a departure of 5 from an expected count of 100 is less surprising than a departure of 5 from an expected count of 6 for any of the models given in Section 9.1.2. Pearson (1900) proposed the statistic

$$X^2 = \sum_{i,j} [(m_{ij} - n_{ij})^2/m_{ij}] \quad (9.7)$$

as a test for evidence of association.

For computational purposes it is easier to use the equivalent form

$$X^2 = \sum_{i,j} (n_{ij}^2/m_{ij}) - N \quad (9.8)$$

### Example 9.2

*The problem.* For the data in Example 9.1 calculate  $X^2$ . Test for significance using the asymptotic result that  $X^2$  has a chi-squared distribution.

*Formulation and assumptions.* We calculate each  $m_{ij} = n_i \cdot n_{+j}/N$ , then use formula (9.8) for  $X^2$ .

*Procedure.* Using the row and column totals for the data in Example 9.1 we obtain the following table of expected frequencies (an explanation of how to get these was given in [Section 9.1.1](#)):

	A	B	C	D	E	F	Total
Nose	11/12	11/12	11/24	11/12	55/8	11/12	11
Throat	1/3	1/3	1/6	1/3	5/2	1/3	4
Ears	3/4	3/4	3/8	3/4	45/8	3/4	9
Total	—2	—2	—1	—2	—15	—2	—24

whence, from (9.8)

$$X^2 = [1^2/(11/12) + 1^2/(11/12) + \dots + 7^2/(45/8) + 1^2/(3/4)] - 24 = 14.96$$

With 10 degrees of freedom chi-squared tables **show that significance** at the 5 per cent level requires  $X^2 \geq 18.31$ . In fact  $\Pr(X^2 \geq 14.96) = 0.1335$ .

*Conclusion.* There is little evidence against  $H_0$ .

*Comment.* An asymptotic result may be unsatisfactory with many sparse cells, but here that using  $X^2$  behaves marginally better than that based on F (Example 9.1). See also Exercise 9.23.

*Computational aspects.* Nearly all standard statistical packages include a program for tests based on  $X^2$  using the asymptotic result. StatXact also computes  $P = 0.1094$  using the exact permutation distribution for  $X^2$ . This not only shows that the Fisher exact test and the exact version of the Pearson chi-squared tests are not equivalent, but suggests that the tests may differ in power.

For  $2 \times 2$  tables the Fisher exact test and the exact tests based on  $X^2$  are often equivalent, though the asymptotic test statistics generally differ. It is customary to use the latter in asymptotic tests, and when doing so a better approximation is obtained by using a continuity correction commonly known in this context as Yates' correction. With this correction we compute

$$X^2 = \sum_{i,j} [(|n_{ij} - m_{ij}| - \frac{1}{2})^2 / m_{ij}],$$

i.e. we subtract  $\frac{1}{2}$  from the magnitude of each difference in the numerator before squaring.

An alternative form useful for computation is

$$X^2 = n[(|n_{11}n_{22} - n_{12}n_{21}|) - \frac{1}{2}N]^2 / (n_{+1}n_{+1}n_{+2}n_{+2}). \quad (9.9)$$

Omitting the term  $\frac{1}{2}N$  gives the  $X^2$  value without continuity correction. The correction is used only for  $2 \times 2$  tables.

### Example 9.3

*The problem.* A new drug (A) is only available in a pilot experiment for 6 patients who are chosen at random from a total of 50 patients; the remaining 44 receive the standard drug (B). Of those receiving drug A only one shows a side-effect, while 38 of those receiving drug B do. Test the hypothesis of no association between side-effects and drug administered.

*Formulation.* From the information above we form the following  $2 \times 2$  table:

	Side-effect	None	Total
Drug A	1	5	6
Drug B	38	6	44
Total	39	11	50

We use (9.9) to compute  $X^2$  and test using chi-squared with 1 degree of freedom.

*Procedure.* Substituting in (9.9) gives

$$X^2 = 50[ |(1 \times 6 - 5 \times 38)| - 25 ]^2 / (6 \times 39 \times 44 \times 11) = 11.16$$

Tables or software for the chi-squared distribution indicate  $P = \Pr(\chi^2 \geq 11.16) = 0.0008$ .

*Conclusion.* There is strong evidence against  $H_0$ : no association.

*Comment.* The permutation  $X^2$  test indicates  $P = 0.0012$ . Without Yates' correction  $X^2 = 14.95$  which over-estimates the evidence against  $H_0$ .

*Computational aspects.* Most general statistical packages include a program for the Pearson chi-squared test for  $r \times c$  tables, often using Yates' correction (or providing it as an option) for  $2 \times 2$  tables.

### 9.2.3 The likelihood ratio test

An alternative statistic for a test of association is the likelihood ratio

$$G^2 = 2 \sum_{i,j} n_{ij} \log(n_{ij}/m_{ij}) \quad (9.10)$$

where 'log' is the natural logarithm, i.e. to the base 'e'. Asymptotically  $G^2$  has a chi-squared distribution with  $(r - 1)(c - 1)$  degrees of freedom if there is no association. For the data in Example 9.1,  $G^2 = 17.33$ . Although this does not indicate significance at a conventional 5 per cent level, it is a better approximation to the exact  $P$  than that given by  $X^2$ . The test has some optimal properties for the overall multinomial sampling model and convergence to the chi-squared distribution under  $H_0$  is often faster than that of  $X^2$ .

An exact permutation theory test may be based on  $G^2$  and, except for  $2 \times 2$  tables, the results will generally differ from those for exact tests based on  $F^2$  or  $X^2$ . For the data in Example 9.1 the exact test based on  $G^2$  gives  $P = 0.048$ .

In each of the three tests for the data in Example 9.1 the asymptotic test is an unsatisfactory approximation. Although the three exact permutation distribution tests are not equivalent, two give strong indications of evidence that there is association and all give lower  $P$ -values than those for asymptotic theory. The sample has a preponderance of sparse cell entries and although such situations do arise in practice, in many contingency tables very few cells have low expected values under  $H_0$ . There is a variety of conflicting advice in the literature about when it is safe to use asymptotic results with  $G^2$  or  $X^2$ . Generally speaking, the larger  $r$  and  $c$  are, the less we need worry about a small proportion of sparse cell counts. Programs such as StatXact or Testimate that provide both exact and asymptotic  $P$ -values, so that these may be compared, are invaluable for those who frequently meet contingency tables with sparse data.

Since the tests in [Section 9.2](#) apply to nominal categories, reordering rows or columns does not affect the value of the test statistic. As already indicated, for ordered categories there are more appropriate tests that take account of any natural ordering in rows and columns.

## 9.3 ORDERED CATEGORICAL DATA

We consider first nominal row explanatory categories and ordered column response categories; then ordered row explanatory and ordered column response classifications or both ordered row and ordered column response categories.

### 9.3.1 Nominal row and ordered column categories

**Table 9.6** repeats [Table 9.1](#) with the addition of row and column totals. The explanatory row categories, type of drug, are nominal but column responses are ordinal. Here the Wilcoxon-Mann-Whitney (WMW) test is appropriate, but the data are heavily tied, so, e.g. the 65 responses 'tied' as *none* are allocated the mid-rank 33, the 16 responses tied as slight are given the mid-rank of ranks 66 to 81, i.e. 73.5, and so on. Use of the Mann-Whitney formulation obviates the need to specify mid-ranks, for we simply count the numbers of recipients of drug B showing the same or more severe side effects than that for each recipient of drug A, scoring an equality as  $\frac{1}{2}$  and an excess as 1. Thus the 23 responses *none* for drug A each give rise to 42 ties, scoring 21 and  $8 + 4$  higher ranks, contributing a total score  $23(\frac{1}{2} \times 42 + 8 + 4) = 759$  to the Mann-Whitney  $U_m$ .

**Table 9.6** Numbers of patients showing side-effects for two drugs.

	Side-effect level					
	None	Slight	Moderate	Severe	Fatal	Total
Drug A	23	8	9	3	2	45
Drug B	42	8	4	0	0	54
Total	65	16	13	3	2	99

### *Example 9.4*

*The problem.* For the data in [Table 9.6](#) compute the Mann-Whitney statistic and test for evidence of a difference in levels of side-effects.

*Formulation and assumptions.* We compute either  $U_m$  or  $U_n$  in the way described above, and which was also illustrated in Example 5.4. Because the sample is large we might use an asymptotic result but if a suitable computer program is available, with so many ties an exact test may be preferred.

*Procedure.* We calculate  $U_m = 23(\frac{1}{2} \times 42 + 8 + 4) + 8(\frac{1}{2} \times 8 + 4) + 9(\frac{1}{2} \times 4) = 841$ . Using (5.7) (i.e. ignoring ties) gives  $Z = 2.63$ , whereas the appropriate modification of (5.6) gives  $Z = 3.118$ . Using StatXact, the exact permutation test gives a two-tail  $P = 0.0019$ , equal to that using  $Z = 3.118$ .

*Conclusion.* With sample sizes 45 and 54, having allowed for ties, an asymptotic result should be reasonable and this indicates strong evidence against  $H_0$  (because  $P < 0.01$ ). This is confirmed by the exact  $P = 0.0019$ .

*Comment.* If we ignore ordering and use the Fisher exact test it gives an exact  $P = 0.0127$  while for the asymptotic Pearson chi-squared test  $P = 0.0193$ . These results are in line with our general remark that we gain power by taking account of ordering. Nevertheless, for some configurations we sometimes observe marginally smaller  $P$ -values when we ignore ordering.

*Computational aspects.* Since asymptotic results appear reasonable here, any statistical package for the WMW test that uses mid-ranks and gives the significance level based on asymptotic theory that takes account of ties should be satisfactory. A program such as the one in StatXact does, however, give additional confidence in that it compares not only exact and asymptotic results for the WMW test, but also allows us to perform the Fisher exact test or a Pearson chi-squared test (exact or asymptotic) and compare results.

In [Section 5.3.1](#) we recommended the Fisher exact test when using a median test to compare treatments. Since we had ordered categories, why did we not recommend the WMW test if we have a program for exact permutation distribution tests? It turns out that it does not matter which we use for  $2 \times 2$  tables, for the exact tests are then equivalent, although the asymptotic forms of the two are not quite the same. The position is more complicated for the median test in [Section 6.2.6](#), where we have more than two samples. We discuss this in [Section 9.3.2](#).

Nearly all tests for centrality in earlier chapters can be formulated as tests involving contingency tables with fixed marginal totals. This is a key to efficient computation of exact permutation distributions and is how StatXact handles data for these tests. We gave an example in [Table 9.2](#) for a data set encountered earlier. In Example 1.4 we developed a permutation test for what we now recognize as a specific example of the WMW test. In that example

we had sets of 4 and 5 patients ranked 1 to 9. In contingency table format we regard the drugs as the row classifications and the ordered ranks 1 to 9 as column (response) classifications. We count the number of items allocated to each row and column classification; e.g. this count will be 0 in row 1, column 4 if no patient given drug A was ranked 4. It will be 1 in row 2 column 5 if a person given drug B is ranked 5. There were no ties in Example 1.4, so if we observed the ranks 1, 2, 3, 6 for the four patients given drug A (implying those receiving drug B were ranked 4, 5, 7, 8, 9) we may construct the contingency table:

	Rank								
	1	2	3	4	5	6	7	8	9
Drug A	1	1	1	0	0	1	0	0	0
Drug B	0	0	0	1	1	0	1	1	1

In this table the row totals are 4, 5 (the sample sizes) and in this no-tie situation each column total is 1. In the Wilcoxon formulation of the WMW test the test statistic is  $S_i = \sum_j n_{ij}$ , where either  $i = 1$  or  $i = 2$ . In this context  $j$  is a column score which here is the rank. The permutation distribution in Example 1.4 is obtained by permuting all possible cell entries in the above table consistent with the fixed row and column totals and calculating  $S_i$  for each permutation. Rank is only one possible column score. Replacing the score  $j$  in column  $j$  by other appropriate ordered scores  $s_j$  we obtain in the form  $S_i = \sum_j s_j n_{ij}$  where  $i = 1$  or  $i = 2$ , various other statistics discussed in [Chapter 5](#) for testing centrality. For example, the  $s_j$  may be the van der Waerden scores or the ordered raw data, the latter giving rise to the raw scores Pitman permutation test. Choices of column scores in  $2 \times c$  tables for many well-known tests are discussed by Graubard and Kom (1987). In the case of ties in the WMW test, cell entries correspond to the number of ties in the relevant sample (row) at any given mid-rank and the column scores are the appropriate mid-rank values. Each column total is the total number of ties assigned that mid-rank score. Permutation is over all configurations giving the fixed row and column totals.

In a 'no-tie' situation, under  $H_0$  we expect a good mix of ones and zeros over the cells of the  $2 \times c$  table. A concentration of ones at opposite corners of the table indicates association and possible

rejection of  $H_0$  e.g. a configuration with the marginal totals above that indicates rejection is

	Rank								
	1	2	3	4	S	6	7	8	9
Drug A	0	0	0	0	0	1	1	1	1
Drug B	1	1	1	1	1	0	0	0	0

We look more fully at the approach used here in [Section 10.2.6](#).

When there are more than two nominal explanatory categories, each corresponding to a row and the column responses are ordinal the obvious extension from the WMW test is to the Kruskal-Wallis test (with ties if appropriate). Unless some samples are very small or there is a concentration of many entries in one or two columns (heavy tying) asymptotic results are usually satisfactory. Except for fairly small samples, even programs like StatXact cannot cope with the exact permutation distribution for the Kruskal-Wallis test, but the Monte Carlo approximation available in StatXact works well in practice.

### *Example 9.5*

The problem. The following information on times to failure of car tyres due to various faults is abstracted from a more detailed and larger data set kindly made available to us by Timothy P. Davis. The classified causes of failure are:

- A Open joint on inner lining
- B Rubber chunking on shoulder
- C Loose casing low on sidewall
- D Cracking on the sidewall
- E Other causes

One cause of failure, cracking of the tread wall, is omitted from this illustrative example, for it clearly gave a very different pattern of times to failure from that for other faults. Depending on one's viewpoint, the row categories A, B, C, D, E might be regarded as nominal explanatory or as nominal response categories. In the sense that they are different types of failure they represent different responses; on the other hand, each tyre failing for a particular reason might be regarded as one of a sample of tyres doomed to fail in that way and if we are interested in times to failure at which different faults occur we might regard the different faults as explanatory variables. For illustrative purposes in this example we take the latter stance, although the analysis would still be appropriate if we regarded these as distinct response categories. The column classification is time to failure recorded to the nearest hour, grouped into intervals <100, 100-199, 200-299, 300+ hours.

Cause	Times to failure (hrs)			
	<100	100-199	200-299	300+
A	2	6	9	0
B	0	1	7	0
C	4	10	6	0
D	1	6	5	0
E	0	8	10	2

*Formulation.* We perform a Kruskal-Wallis test with 5 samples. Mid-ranks are calculated regarding all units in the same column as ties. Thus, the seven ties in the column labelled <100 are all given the mid-rank 4, and so on. The Kruskal-Wallis statistic is calculated as in Section 6.2.3 and the statistic has asymptotically a chi-squared distribution with  $r - 1 = 4$  degrees of freedom. Computation is tedious without a computer program.

*Procedure.* The computation outlined above gives the value 10.44 for the Kruskal-Wallis statistic, and tables or software indicate  $P = \Pr(\chi^2 \geq 10.44) = 0.0336$ .

*Conclusion.* There is reasonably strong evidence of association between rows and columns, implying that different types of fault do not all have the same median times to occurrence.

*Comment.* Despite heavy tying, in the light of our finding in earlier chapters, one hopes the asymptotic result is reasonable here. That this is so was confirmed by the StatXact Monte Carlo procedure for estimating the exact  $P$ . This gave an estimated  $P = 0.0293$ , with the 99 per cent confidence interval for the true  $P$  being (0.0256, 0.0330). This suggests the asymptotic result is conservative, falling just above the upper confidence limit for the permutation distribution result.

*Computational aspects.* Many packages compute the Kruskal-Wallis statistic, so the asymptotic test is readily available. StatXact will compute exact  $P$ -values with small numbers of rows and columns but computation may fail on some PCs or prove very time consuming for examples with a moderate number of ties. The Monte Carlo method is relatively fast. The estimate above of  $P = 0.0293$ , was based on 14 000 simulations. Increasing the number of simulations shortens the confidence interval for the true  $P$ .

In a no-tie situation with  $r$  samples and a total of  $N$  observations the Kruskal-Wallis test may be formulated in contingency table terms as an  $r \times N$  table with rows corresponding to samples and all cell entries 1 or 0. The columns are ordered by ranks and the observation of rank  $j$  is recorded as 1 in the cell in the  $j$ th column in the row corresponding to the sample in which it occurs. All other entries in that column are zero in the no-tie case. The row totals are

the sample sizes  $n_i$  and each column total is 1. Ranks are the scores used to generate the exact permutation distribution and the statistic is a function of the rank sums for each row. Permutation is over all cell entries in the  $r \times N$  table consistent with row and column totals.

### 9.3.2 Ordered row and column categories

We consider first the case where row categories are ordinal explanatory categories; columns are responses. One appropriate test of no association is then the Jonckheere-Terpstra test with each row corresponding to one of the ordered samples. Asymptotic results may be reasonable if allowance is made for ties. We illustrate the need to allow for ties in a situation not uncommon in drug testing, where side-effects may not occur very often, but may be serious if they do.

#### *Example 9.6*

The problem. Side-effects (if any) experienced by patients at increasing dose levels of a drug are classified as none, mild, moderate or severe. Do the following data indicate that side-effects increase with dose level?

Dose	Side-effects			
	None	Mild	Moderate	Severe
100 mg	50	0	1	0
200 mg	60	1	0	0
300 mg	40	1	1	0
400 mg	30	1	1	2

*Formulation.* The Jonckheere-Terpstra statistic is relatively easy to compute as an extension of the procedure used for the Mann-Whitney formulation of the WMW test in Example 9.4, scoring the relevant ties in any column as  $\frac{1}{2}$ . Those in the first column make a substantial contribution to the total.

*Procedure.* Without a computer program it is tedious but not difficult to compute the relevant statistic, denoted by  $U$  in Example 6.3. It is

$$U = 50[\frac{1}{2}(60 + 40 + 30) + 7] + 60[\frac{1}{2}(40 + 30) + 6] + 40(\frac{1}{2} \times 30 + 4) + 5 + 3.5 + 3 + 2.5 = 6834.$$

The asymptotic test (6.3), with no correction for ties, gives  $Pr(U > 6834) = 0.2288$ . The two-tail probability is  $P = 0.4576$  so there is no evidence of association and the hypothesis  $H_0$ : *side-effect levels are independent of dose* is acceptable.

*Conclusion.* We would not reject  $H_0$  even in a one-tail test, which might be justified on the general grounds that increasing dose levels of a drug are more likely to produce side-effects if indeed there are any.

*Comments.* 1. A clinician is unlikely to be satisfied with this finding. Side-effects are not common, but there is to a clinician (and even to the layman) an indication that these are more likely and more severe at the highest dose level. Either the evidence is too slender to reach a firm conclusion or our test is inappropriate; in particular, can we justify an asymptotic result that does not allow for ties when there is so high an incidence of these?

2. The maximum safe dose of a drug is often determined by giving the drug to independent groups of healthy volunteers in increasing doses, the first group receiving the smallest dose. Once a serious side-effect has been observed the study is stopped; to continue the investigation with healthy volunteers would be unethical. Higher doses of the drug may have to be used with genuine patients in order to achieve an improvement. The potential gains would then have to be balanced against the likely side-effects.

*Computational aspects.* Some general statistical packages include the asymptotic Jonckheere-Terpstra test without allowance for ties. We show below that correcting for ties may have a dramatic effect. StatXact includes an asymptotic test that allows for ties.

### Example 9.7

*The problem.* Use the exact Jonckheere-Terpstra test on the data in Example 9.6 to determine if there is acceptable evidence that side-effects become more common as dose increases.

*Formulation.* The permutation test requires computation of probabilities for the Jonckheere-Terpstra statistic for all permutations of the contingency table in Example 9.6 subject to the same marginal totals and with the same or a lesser probability than that observed.

*Procedure.* StatXact provides a program for these computations which indicates  $\Pr(U \geq 6834) = 0.0168$ .

*Conclusion.* There is clear evidence against  $H_0$  with a one-tail  $P = 0.0168$ .

*Comment.* This is an illustration of a situation where an exact permutation test leads to a markedly different conclusion to a commonly used asymptotic test despite a large sample size.

We mentioned in [Section 6.2.5](#) an amended formula for  $\text{Var}(U)$  which allows for ties. The formula is adapted from one given in Kendall and Gibbons (1990, p. 66) and quoted by Lehmann (1975, p. 235). The notation here matches that in [Table 9.5](#).

We calculate

$$\text{Var}(U) = U_1/d_1 + U_2/d_2 + U_3/d_3 \quad (9.11)$$

where

$$\begin{aligned}
U_1 &= N(N-1)(2N+5) - \sum_i n_{i+}(n_{i+}-1)(2n_{i+}+5) - \sum_j n_{j+}(n_{j+}-1)(2n_{j+}+5) \\
U_2 &= [\sum_i n_{i+}(n_{i+}-1)(n_{i+}-2)][\sum_j n_{j+}(n_{j+}-1)(n_{j+}-2)] \\
U_3 &= [\sum_i n_{i+}(n_{i+}-1)][\sum_j n_{j+}(n_{j+}-1)] \\
d_1 &= 72, d_2 = 36N(N-1)(N-2), d_3 = 8N(N-1).
\end{aligned}$$

When there are no ties, all  $n_{ij} = 1$  and  $U_2 = U_3 = 0$ , and  $\text{Var}(U)$  reduces after algebraic manipulation to that in (6.3). In Example 9.6 the relevant values are  $N = 188$ ,  $n_{1+} = 51$ ,  $n_{2+} = 61$ ,  $n_{3+} = 42$ ,  $n_{4+} = 34$ ,  $n_{+1} = 180$ ,  $n_{+2} = n_{+3} = 3$ ,  $n_{+4} = 2$ . Substituting in (9.11) reduces the denominator in (6.3),  $\sqrt{[\text{Var}(U)]}$ , from 415.36 (ignoring ties) to 145.18 (Exercise 9.2). The corresponding  $Z = 2.125$ , and standard normal distribution tables give  $\Pr(Z \geq 2.125) = 0.0168$ , agreeing with the exact permutation result obtained in Example 9.7.

In clinical trials situations like that in Examples 9.6 and 9.7 are common. They arise not only for relatively rare side-effects of drugs but also in other situations where some responses are rare, e.g. incidence of lung cancer in populations exposed to adverse environmental factors. We may observe several thousand people in each explanatory category, yet only a few develop lung cancer.

In Section 6.2.6 we recommended the Fisher exact test for the median test for  $r > 2$  samples. This ignores the fact that the responses  $< M$  and  $\geq M$  are ordered. If the sample classifications are nominal an alternative is the Kruskal-Wallis test. If the samples are ordered so that the natural alternative hypothesis is some systematic increase or decrease in the median, we may do better using the Jonckheere-Terpstra procedure. With only two columns (ordered response categories) there will often be little difference between results using the Fisher test or Kruskal-Wallis or (if appropriate) the Jonckheere-Terpstra procedure. If only asymptotic results are used in this last test a correction for ties is important. In the special case of only two ordered response categories we show below that the Jonckheere-Terpstra test is equivalent to the WMW test.

### *Example 9.8*

*The problem.* In Exercise 6.21 we gave data supplied by Chris Theobald for blood platelet counts and spleen sizes for 40 patients. The investigators were interested in whether the proportion of patients with an abnormally low platelet count increased with increasing spleen size. If a count below 120 is abnormal the data in Exercise 6.21 give the contingency table below. Does the Jonckheere-Terpstra test indicate an association between spleen size and platelet count?

*Formulation and assumptions.* Exercise 6.21 indicated that platelet counts tend to decrease as spleen size increases: there ties were minimal as the actual counts were given. Here we must take ties into account in an asymptotic test.

Spleen size	Blood platelet count	
	Abnormal	Normal
0	0	13
1	2	10
2	5	6
3	3	1

**Procedure.** Proceeding as in Examples 9.6 and 9.7, we find  $U = 183$  and  $E(U) = 287.5$ . Using (9.11) we find the standard deviation of  $U$  is  $\sqrt{[Var(U)]} = 30.67$ , whence  $Z = -3.407$ . From tables of the standard normal distribution we find  $\Pr(Z \leq -3.407) = \Pr(Z \geq 3.407) = 0.0003$ .

**Conclusion.** There is strong evidence that the proportion of individuals with abnormal counts increases as spleen size increases.

**Comments.** 1. Computing  $Z$  by the formula on p. 205 ignoring ties gives  $Z = -2.551$ , indicating a one-tail  $P = 0.0054$  (compared with  $P = 0.0003$  obtained above). An exact test using StatXact with these data gave  $P = 0.0002$  for a one-tail test, in close agreement with the asymptotic result taking ties into account.

2. In general, ties have a greater effect on  $Var(U)$  when they occur between samples (in columns) than if they occur largely within one sample (in a row).

3. Regarding counts of less than 120 as 'abnormal' is arbitrary. Using an extended data set, Bassendine et al (1985) considered counts below 100 as 'abnormal'. With such arbitrary decisions about categories, one choice may lead to a result significant at a formal level such as  $a = 0.05$  or  $a = 0.01$  while another does not. In [Section 9.4.2](#) we discuss a similar problem arising with grouped continuous data.

4. Since we have only two columns, they remain 'ordered' if we interchange them. We might look upon 'abnormal' and 'normal' as designating samples from two populations and spleen sizes as rankings for those populations, and apply the WMW test as in Example 9.4. Doing so, we get exactly the same results as we do with Jonckheere-Terpstra.

The expression for  $Var(U)$  in (9.11) is unaltered if we interchange rows and columns in an  $r \times c$  table. However, the statistic  $U$  and the mean  $E(U)$  are affected by this change. It is not difficult to show that if we write  $U_1$  for the Jonckheere-Terpstra statistic for the original  $r \times c$  table and  $U_2$  for the corresponding statistic when rows and columns are interchanged then  $U_2 = U_1 + \frac{1}{4}(\sum_i n_{i+} - \sum_j n_{+j})$  and  $E(U_2) = E(U_1) + \frac{1}{4}(\sum_i n_{i+} - \sum_j n_{+j})$ . Clearly, the asymptotic tests using  $U_1$  and  $U_2$  are identical. Although  $U_1$  and  $U_2$  differ by the constant value given above, their permutation distributions are otherwise identical, i.e., for all  $k$ ,  $\Pr(U_1=k)=\Pr[U_2=k+\frac{1}{4}(\sum_i n_{i+} - \sum_j n_{+j})]$

This equivalence may seem surprising as the Jonckheere-Terpstra test was introduced as a test for a monotone trend between samples.

However, if there are no ties in  $p$  samples with  $n_1, n_2, \dots, n_p$  observations respectively the problem may be stated in a contingency table format with  $p$  rows and  $N = \sum_i n_i$  columns. In each column all cells except one have zero entries and the remaining cell has an entry unity. The entries in row  $i$  are constrained so that the sum of the cell entries that are unity (all others being zero) is  $n_i$ . If we interchange rows and columns the equivalent test problem is one in which we have  $N$  samples each of one observation and these take only  $p$  different values of which there are  $n_1$  tied at the lowest value,  $n_2$  tied at the next lowest, and so on. An extreme case is that in which we have  $N$  samples, each of one observation and none of these is tied. We may represent this situation by an  $N \times N$  table with all column totals and all row totals equal to 1. This implies that in any row or column there is one cell value of unity and all others are zero. In this case clearly  $U_2 = U_1$ . As pointed out in [Section 7.1.4](#) the Jonckheere-Terpstra test procedure is here equivalent to that for Kendall's tau;  $U_1$  or  $U_2$  are the numbers of concordances in this 'no-tie' situation, and this is sufficient to determine Kendall's  $t_k$ .

Any situation with ties, and where the concept of linear rank association implicit in Kendall's tau is relevant, may be looked upon as generating an  $r \times c$  matrix which 'shrinks' an  $N \times N$  matrix of zeros and ones by super-imposing rows with identical 'x' values and superimposing columns with identical 'y'. In particular, if the  $i$ th sample has  $n_i$  observations we shrink  $n_i$  adjacent rows to a single row by adding all column entries in those  $n_i$  adjacent rows. The x variable values in this situation are effectively labels that distinguish the (ordered) samples. We have pointed out that in the 'no-tie' case the Jonckheere-Terpstra statistic is equivalent to the number of concordances in Kendall's  $t_k$ . With ties, the two statistics use slightly different counting systems. The Kendall statistic counts concordances as unity, ties as zero and discordances as -1; for the Jonckheere-Terpstra statistic, concordances count as 1, ties as  $\frac{1}{2}$  and discordances as zero. With appropriate adjustments to certain formulae the two counting systems are interchangeable and, used correctly, lead to the same conclusions.

If rows and columns are both response categories we may be interested in whether there is a patterned association (akin to a correlation) between row and column classifications in that high responses in one classification tend to be associated with high responses in the other, or high responses in one are associated with low responses in the other. Again the Jonckheere-Terpstra test is relevant.

Several statistics are used as measures of association but none is completely satisfactory in all circumstances. For some of those suggested, an exact Jonckheere-Terpstra test is appropriate to test for significant association in the form of a monotonic trend. However, for non-monotonic associations such tests may fail to detect patterns of association. For example, in an ordered-categories contingency table with cell entries

$$\begin{matrix} 0 & 17 & 0 \\ 5 & 0 & 5 \\ 3 & 0 & 3 \end{matrix} \quad (9.12)$$

there is a clear - but obviously not monotonic - association. Here it is easily verified (Exercise 9.4) that the asymptotic Jonckheere-Terpstra statistic  $Z$  takes the value 0, giving  $P = 0.50$  for a one-tail test! Apart from a small discontinuity effect the exact test confirms this probability.

A desirable property of a measure of monotonic association, akin to that of a correlation coefficient, is that it takes the value 1 when there is complete positive association, the value zero if there is no rank association, and the value -1 if there is inverse association, e.g. high category row classifications are associated with low category column classifications and vice-versa. An appropriate test for linear rank association using such statistics is often the Jonckheere-Terpstra test, but to measure degree of association there is a case for having a statistic that behaves like a correlation coefficient. We discuss one such coefficient here. This and other measures of association are described in considerable detail by Kendall and Gibbons (1990, Chapter 3) and by Siegel and Castellan (1988, Chapter 9). These and other authors give guidance on relevant asymptotic tests which take account of the tied nature of the data.

The **Goodman-Kruskal gamma statistic** is usually denoted by  $G$ . We count the number of concordances and discordances between row and column classifications as for Kendall's  $t_k$ , but make no allowance for ties. When both row and column classifications are ordered, for each count in cell  $(i, j)$  there is concordance between that count and the count in any cell below and to the right. Thus, if  $N_{ij}^+$  denotes the sum of all counts below and to the right of cell  $(i, j)$ , which itself has count  $n_{ij}$  the total number of concordances is

$$C = \sum_{ij} n_{ij} N_{ij}^+, \quad 1 \leq i \leq r-1, \quad 1 \leq j \leq c-1.$$

This resembles the count for the Jonckheere-Terpstra test, except that for counts in cell  $(i, j)$  we ignore ties in column  $j$  (each counted as  $\frac{1}{2}$  in Jonckheere-Terpstra). Similarly, denoting the sum of all

elements to the left of and below the cell  $(i, j)$  by  $N_{ij}^-$ , the total number of discordances is

$$D = \sum_{ij} n_{ij} N_{ij}^-, \quad 1 \leq i \leq r - 1, \quad 2 \leq j \leq c.$$

The Goodman-Kruskal statistic is  $G = (C - D)/(C + D)$ . Clearly, if  $D=0$ ,  $G=1$ . If  $r \geq c$ ,  $G=1$  if all entries in any column except the last are zero unless  $i=j$ ; in the last column they must be zero if  $i < c$ . If  $c \geq r$ ,  $G=1$  if all entries in any row except the last are zero unless  $i=j$ ; in the last row they must be zero if  $j < r$ . If  $r=c$ ,  $G=1$  if all entries for which  $i \neq j$  are zero. In Exercise 9.5 we ask for confirmation that  $G=1$  for each of the tables

				7	0	0
3	0	0	0	0	1	0
0	4	2	7	0	0	6
				0	0	7
				0	0	5

It is easy (Exercise 9.6) to obtain analogous conditions for  $G = -1$ . When there is a good scatter of counts over all cells,  $C$  and  $D$  will take similar values and  $G$  will be close to zero. Near-zero values of  $G$  are also possible if association is not monotonic. It is easily verified (Exercise 9.7) that  $C = D$ , hence  $G = 0$ , in the highly patterned contingency table (9.12).

Like Kendall's tau (as distinct from Kendall's tau-b),  $G$  takes no account of ties in both  $x$  and  $y$ , yet unlike Kendall's tau it may equal 1 even when there is heavy tying because the denominator is  $C + D$ . Clearly this falls well short of the denominator in Kendall's tau for the no-tie case. In [Section 7.1.4](#) we introduced Kendall's  $t_b$  in preference to  $t_k$  when there are ties. This is a compromise between ignoring them (as in  $t_k$ ) and deleting their effect if in the same row or column as we do in  $G$ . Calculation of  $t_b$  for contingency tables is straightforward and follows the method outlined in [Section 7.1.4](#). Other measures of association and some appropriate tests are given by Siegel and Castellan (1988, Chapter 9); we concur with their comment that all of them should be useful when appropriately applied'. Other approaches to exploring association in contingency tables are given in [Chapter 10](#).

## 9.4 GOODNESS-OF-FIT TESTS FOR DISCRETE DATA

The Pearson chi-squared test and the likelihood ratio test ([Sections 9.2.2](#) and [9.2.3](#)) may be looked upon as tests of goodness-of-fit of

data in a contingency table to a model of independence (no association) between row and column classifications. The asymptotic chi-squared test is also widely used as a goodness-of-fit test of data to any discrete distribution. This topic is covered in most standard courses in statistics so we give only a brief resume here drawing attention to a few points that are sometimes not covered specifically in introductory treatments.

Tests of goodness-of-fit to any discrete distribution (often appropriate for counts) are usually based on the Pearson chi-squared statistic,  $X^2$  in (9.7). Expected values are calculated for a hypothesized distribution which may be binomial, negative binomial, Poisson, uniform or some other discrete distribution. Sometimes one or more parameters must be estimated from the data. If we have  $r$  counts or cells the test will have  $r - 1$  degrees of freedom if no parameters are estimated; if a parameter (e.g.  $p$  for the binomial or  $\lambda$  for the Poisson distribution) is estimated from data, one further degree of freedom is lost for each parameter estimated. While tests of the Kolmogorov-Smirnov type are sometimes also applied for discrete distributions, generally the test criteria are no longer exact and the tests are often inefficient.

#### *Example 9.9*

*The problem.* It is often suggested that recorded ages at death are influenced by two factors. The first is the psychological wish to achieve ages recorded by 'decade', e.g. 70, 80, 90, and that a person nearing the end of a previous age decade who knows his or her days may be numbered will by sheer willpower or perhaps even by a change in lifestyle (e.g. by stopping smoking or abstaining from alcohol) strive to attain such an age before dying. If so, ages at death with final digits 0, 1 should be more frequent than those with higher final digits. A second factor is misstatement of age, elderly people may be imprecise about giving their age, tending to round it; e.g. if they are in the mid-seventies they tend to say they are 75 if anywhere between about 73 and 77. Similarly a stated age of 80 may correspond to a true age a year or two above or below. If these factors operate, final digits in recorded ages at death would not be uniformly distributed. [Table 9.7](#) gives final digits at age of death for the 117 males recorded at the Badenscallie burial ground (see Appendix). Is the hypothesis  $H_0$ : *any digit is equally likely* acceptable?

*Formulation and assumptions.* For 117 deaths the expected number of occurrences of each of the 10 digits, if all are equally likely, is 11.7. Since the denominator in  $X^2$  is always the expected number,  $m = 11.7$ , we may calculate  $X^2$  by taking the differences between the observed numbers  $n$ ; and 11.7, squaring these, adding and finally dividing the total by 11.7.

**Table 9.7** Recorded last digit of age at death, Badenscallie males.

Digit	0	1	2	3	4	5	6	7	8	9
Freq	7	11	17	19	9	13	9	11	13	8

*Procedure.* The differences  $n_i - m$  are respectively -4.7, -0.7, 5.3, 7.3, -2.7, 1.3, -2.7, -0.7, 1.3, -3.7. These differences sum to zero - a well-known property of deviations from the mean. The sum of squares of differences is 136.1. Division by the expected number  $m = 11.7$  gives  $X^2 = 136.1/11.7 = 11.63$ . This is well below the value 16.92 for  $\chi^2$  that corresponds to  $P = 0.05$  when there are  $10 - 1 = 9$  degrees of freedom; indeed it corresponds to  $P = 0.235$ . StatXact gives the exact permutation  $P = 0.239$ .

*Conclusion.* There is no real evidence against  $H_0$  the digits may be random.

*Comments.* 1. In posing the problem we suggested a possible build-up of digits 0, 1 and perhaps also 5. The observed variation in digit frequency is not in line with such a pattern. Any build-up is around 2 and 3. As the overall result is not significant we must not read too much into this.

2. In this data set ages at death ranged from less than 1 to 96 years. In particular, there is some evidence of an increased risk of death for very young children. In all, 41 of the 117 recorded ages at death were less than 60 years; the influence of elderly people is therefore diluted.

3. Survival until a notable family event (e.g. the birth of a grandchild) may be a stronger incentive to an elderly person than the completion of an age decade.

4. Misstatement of age occurs across a wide range of ages for a variety of reasons, discussed in detail by Cox (1978).

#### 9.4.1 Goodness-of-fit with an estimated parameter

We often meet data that might belong to a particular distribution such as the Poisson distribution with unknown mean. We may use the statistic  $X^2$  to test this, although for the Poisson and some other distributions alternative parametric tests may be more powerful.

For the asymptotic test using  $X^2$  a difficulty arises if some expected numbers are small. Traditional advice is to group such cells to give a group expected number close to 5 (but this is only a guide). There is a corresponding reduction in degrees of freedom.

##### Example 9.10

*The problem.* A factory employs 220 people. The numbers experiencing 0, 1, 2, 3, . . . accidents in a given year are recorded

Number of accidents	0	1	2	3	4	5	$\geq 6$
Number of people	181	9	4	10	7	4	5

Are these data consistent with a Poisson distribution?

*Formulation and assumptions.* The maximum likelihood estimate,  $\hat{\lambda}$ , of the Poisson parameter  $\lambda$ , is the mean number of accidents per person. The expected numbers having  $r$  accidents are  $E(X = r) = 220 \hat{\lambda}^r \exp(-\hat{\lambda})/r!$ ; this follows because for a Poisson distribution with parameter  $\lambda$ ,  $\Pr(X = r) = \lambda^r \exp(-\lambda)/r!$ .

*Procedure.* To obtain the mean number of accidents per person we multiply each number of accidents by the number of people having that number of accidents, add these products and divide the sum by 220, giving

$$\hat{\lambda} = [(0 \times 181) + (1 \times 9) + (2 \times 4) + (3 \times 10) + (4 \times 7) + (5 \times 4) + (6 \times 5)]/220 = 125/220 = 0.568$$

This is approximate because we treat 'six or more' as 'exactly 6'. In practice, this has little influence on our result, but it is a limitation we should recognize. To get expected numbers an iterative algorithm may be used, noting that

$$E(X = r + 1) = E(X = r) \times \hat{\lambda}/(r + 1).$$

The expected numbers are

Number of accidents	0	1	2	3	4	5	6 or more
Expected numbers	124.7	70.8	20.1	3.8	0.5	0.1	0

Grouping results for 3 or more accidents gives an associated total expected number 4.4. We calculate  $X^2$  using (9.8) as

$$X^2 = (181^2/124.7) + (9^2/70.8) + (4^2/20.1) + (26^2/4.4) - 220 = 198.3.$$

There are 2 degrees of freedom since we have four cells in the final test and we have estimated one parameter,  $\lambda$ .

*Conclusion.* It is clear even if using only conventional chi-squared tables that there is very strong evidence that the data are not consistent with a Poisson distribution since with 2 degrees of freedom  $\Pr(\chi^2 > 13.81) = 0.001$ .

*Comment.* Accident data seldom follow a Poisson distribution; they would if accidents occurred entirely at random and each person had the same probability of experiencing an accident, and having one accident did not alter the probability of that person having another. In practice, a better model is one that allows people a differing degree of accident proneness. The negative binomial distribution assumes that individual proneness has a gamma distribution within the population; this is very flexible and has been shown to give a good fit to numbers of episodes of mental illness experienced by adults in a fixed period of several years (Smeeton, 1986). If an individual experiences an accident, the probability of a further accident may be reduced if he or she becomes more careful, or increased if concentration or confidence is lowered, making that person more accident-prone. These factors may act differently from person to person, or for individuals exposed to different risks. Multiple accidents, in which a number of people are involved in one incident, also affect the distribution of numbers of accidents per person in some accident data.

*Computational aspects.* Nearly all standard statistical packages include a program for chi-squared goodness-of-fit tests.

#### 9.4.2 Goodness-of-fit for grouped data

The chi-squared test is sometimes applied to test goodness-of-fit of grouped data from a specified continuous distribution, commonly a normal distribution with specified mean and variance, or with these estimated from the data. This is not recommended unless the grouped data are all that are available and grouping is on some natural basis. If the grouping is arbitrary, one arbitrary grouping may result in rejection of a hypothesis while some other grouping may not.

A situation where a test for normality based on Pearson's  $X^2$  might be used is that for sales of clothing of various sizes. A large retailer might note numbers of sales of ready-made trousers with nominal leg lengths (cm) 76, 78, 80, etc. The implication is that customers requiring some 'ideal' leg length between 75 and 77 cm will purchase trousers with leg length 76, those requiring leg lengths between 77 and 79 will purchase trousers of length 78, and so on. The sizes are the centre value for each group. To test whether sales at nominal lengths are consistent with a normal distribution with unspecified mean and variance, these parameters must be estimated from the grouped data and 2 degrees of freedom deducted to allow for this.

Given complete sample data from a continuous distribution it is better to use goodness-of-fit tests of the Kolmogorov/Lilliefors type as appropriate or other tests relevant to particular continuous distributions. If grouping is used, and tests are to be based on  $X^2$ , Kimber (1987) points out that when both grouped and ungrouped data are available anomalies may arise unless parameter estimates are based on the grouped data.

### 9.5 EXTENSION OF McNEMAR'S TEST

There is an asymptotic chi-squared approximation equivalent to the normal approximation to the binomial for McNemar's test ([Section 4.2](#)). For the data in Table 4.3, for example, we tested in effect whether 9 successes and 14 failures (or vice versa) are consistent with a binomial distribution with  $p = \frac{1}{2}$  and  $n = 9 + 14 = 23$ . We may perform a goodness-of-fit test using  $X^2$ . The expected numbers of successes or failures is 11.5 and the test statistic is  $X^2 = (2.5)^2/11.5 + (2.5)^2/11.5 = 1.09$ . This is compared with the relevant critical value for chi-squared with 1 degree of freedom. Recalling that the relevant cells in a  $2 \times 2$  table for the McNemar test are the

off-diagonal cells  $n_{12}$ ,  $n_{21}$ , it is easily verified that with these numbers of successes or failures  $n_{12} = 9$  and  $n_{21} = 14$  in Table 4.3) then

$$X^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21}).$$

In Section 6.4.4 we introduced Cochran's test for comparing several treatments with binary (0,1) responses. When  $t = 2$  it can be shown (Exercise 9.17) that Cochran's Q is equivalent to the form given above for  $X^2$ . In this sense the Cochran test is a generalization of McNemar's test.

Another generalization of McNemar's test is from a  $2 \times 2$  contingency table to an  $n \times n$  table in which we test for off-diagonal symmetry. Bowker (1948) proposed a relevant test. The test might be used if there is a new formulation of a drug which it is hoped will reduce side-effects. If we have a number of patients who have been treated with the old formulation and records are available of any side-effects we might, if there are not ethical reasons to preclude this course, now treat each of these patients with the new formulation and note incidence of side-effects. Table 9.8 shows a possible outcome for such an experiment.

In the table we see that each off-diagonal count below the diagonal exceeds that in the symmetrically placed cell above the diagonal, e.g.  $n_{31} (= 4)$  is greater than  $n_{13} (= 3)$ , etc. This gives the impression that the trend is towards less severe side-effects with the new formulation, although a few who suffered no side-effects with the old formulation do show some with the new formulation.

Bowker proposed a test statistic to determine whether at least one pair of probabilities associated with the symmetrically placed-off diagonal cells differ in an  $n \times n$  table. It is a generalization of the  $X^2$  statistic used in McNemar's test and takes the form

$$X^2 = \sum [(n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})] \quad (9.13)$$

**Table 9.8** Side-effects with old and new formulations of a drug

		Side-effect levels - new formulation			Total
		None	Slight	Severe	
Side-effect levels - old formulation	None	83	4	3	90
	Slight	17	22	5	44
	Severe	4	9	11	24
	Total	104	35	19	158

where the summation is over all  $i$  from 1 to  $n - 1$  and  $j > i$ . Under the null hypothesis of symmetry the asymptotic distribution of  $X^2$  in (9.13) is chi-squared with  $\frac{1}{2}n(n - 1)$  degrees of freedom.

### *Example 9.11*

*The problem.* Do the data in Table 9.8 provide evidence that side-effects are less severe with the new formulation of the drug?

*Formulation and assumptions.* The Bowker statistic given by (9.13) is appropriate. Here  $n_{12} = 4$ ,  $n_{13} = 3$ ,  $n_{21} = 17$ , etc. The null hypothesis is that for all off-diagonal counts in the table the associated probabilities are such that all  $p_{ij} = p_{ji}$ . The alternative is that for at least one such pair  $p_{ij} \neq p_{ji}$ .

*Procedure.* Substitution of the relevant values in (9.13) gives

$$X^2 = (4 - 17)^2/21 + (3 - 4)^2/7 + (5 - 9)^2/14 = 9.33$$

Under the null hypothesis,  $X^2$  has a chi-squared distribution with 3 degrees of freedom, and appropriate tables or software establishes that with 3 degrees of freedom  $P = Pr(\chi^2 \geq 9.33) = 0.026$ .

*Conclusion.* There is evidence of a differing incidence rate for side-effects under the two formulations. From Table 9.8 it is clear that this difference is towards less severe side-effects under the new formulation.

*Comment.* Marascuilo and McSweeney (1977, Section 7.6) discuss an alternative test based on marginal totals due to Stuart (1955; 1957). Agresti (1990, Chapter 10) discusses other generalizations of McNemar's test.

## 9.6 FIELDS OF APPLICATION

We indicate a few situations where tests for independence for categorical data may be appropriate.

### **Rail transport**

A railway company may be interested in whether its image among standard class passengers is different from that among first class passengers. It may ask samples of each to grade service received as excellent, good, fair or poor, setting up a  $2 \times 4$  table of ordered response numbers.

### **Television viewing**

A public service broadcasting channel competes with a commercial service. Samples of men and women are asked which they prefer. The results may be expressed in a  $2 \times 2$  table, any differences between the sexes in preference ratings being of interest.

## **Drug addiction**

A team of doctors compares 3 treatments allocated at random to subjects for curing drug addiction. For each subject, withdrawal symptoms are classed as severe, moderate, mild or negligible. The resulting  $3 \times 4$  table of counts is tested to see if there is evidence of an association between treatments and severity of withdrawal symptoms. If the treatments can be ordered (e.g. by degree of physiological changes induced) a test for association of the Jonckheere-Terpstra type might be used. In the absence of such ordering a Kruskal-Wallis test would be appropriate.

## **Sociology**

A sociologist may be interested in variations in the perception of stigma related to asthma within white, Afro-Caribbean and Asian ethnic groups. A  $2 \times 3$  table could be used in a test for lack of association between ethnic group and presence or absence of perceived stigma.

## **Public health**

After a contaminated food episode on a jumbo jet some passengers show mild cholera symptoms. The airline wants to know if those previously inoculated show a higher degree of immunity. They find out which passengers have and have not been inoculated; records also show how many in each category exhibit symptoms, so they can test for association.

## **Rain making**

In a low-rainfall area rain-bearing clouds are 'seeded' to induce precipitation. Randomly selected clouds are either seeded or not seeded on a sequence of occasions and we observe after each whether or not local rainfall occurs within the next hour, the results being expressed in a  $2 \times 2$  table, a test being made for evidence of association.

## **Educational research**

Children are shown a video of a roller-coaster ride. This may reassure a child who is frightened of such rides but arouse fear in one who was initially unconcerned. McNemar's test may be used to indicate whether the video does more harm than good.

## **Medicine**

To test whether a drug is effective it and a placebo are given to patients in random order; it is noted which, if either, they claim to be effective. Some claim both are effective, some neither, some one but not the other. McNemar's test could be used to see if results favour the placebo, the new drug, or neither.

Here are examples of goodness-of-fit tests.

## **Genetics**

Genetic theory may specify the proportions of plants in a cross that are expected to produce blue or white flowers, round or crinkled seeds, etc. Given a sample in which we know the numbers in each colour/seed-shape combination, we may use a chi-squared goodness-of-fit test to see if these are consistent with theoretical proportions.

## **Sport**

It is often claimed that the starting positions in horse racing, athletics or rowing events may influence the probability of winning. If we know the starting positions and winners for a series of rowing events in each of which there are six starters we might test the hypothesis that the numbers of wins from each starting position is consistent with a uniform distribution.

## **Horticulture**

The positions at which leaf buds or flowers form on a plant stem are called nodes. Some theories suggest a negative binomial distribution for the node number (counted from the bottom of the stem) at which the first flower forms. Given data, a chi-squared goodness-of-fit test is appropriate.

## **Commerce**

A car salesman may doubt the value of advertising. Each week he advertises in either 0, 1, 2, 3 or 4 newspapers. After a long period he might compare weekly sales with the number of advertisements. He would expect a uniform distribution for sales if advertising were worthless. However, in this situation it might be necessary to adjust for trends in sales over time; the effect of this might be minimized if weeks were chosen at random for each number of advertisements.

## Queues

Numbers of people entering a bank, post office or supermarket during one-minute intervals are recorded over a long period. If the process were completely random with constant mean, the numbers of such intervals in which 0, 1, 2, 3, . . . people enter should follow a Poisson distribution. Would you be surprised if a test produced little evidence against this hypothesis? If so, why?

### 9.7 SUMMARY

Tests for independence between row and column classifications in contingency tables are usually based on conditional inference, regarding the marginal totals as fixed ([Section 9.1](#)).

For  $r \times c$  tables with nominal categories exact tests for association are commonly based on the **Fisher exact test** ([Section 9.2.1](#)), while asymptotic tests are usually based on the **Pearson chi-squared statistic**  $X^2$  ([Section 9.2.2](#)) or the **likelihood ratio statistic**  $G^2$  ([Section 9.2.3](#)). Exact permutation tests may also be based on  $X^2$  and  $G^2$ ; while the permutation distributions differ from that for the Fisher exact test (except in  $2 \times 2$  tables) conclusions for all three approaches are often broadly in line.

For categorical tables with ordered rows and columns, tests equivalent to the **Jonckheere-Terpstra test** may be used ([Section 9.3.2](#)). Adjustments for ties are important if asymptotic theory is used. If only columns are ordered the **Kruskal-Wallis test** ([Section 9.3.1](#)) may be relevant. The **Goodman-Kruskal gamma** statistic,  $G$ , ([Section 9.3.2](#)) is closely related to Kendall's tau as a measure of association.

The **chi-squared goodness-of-fit test** ([Section 9.4](#)) is appropriate for testing goodness-of-fit to discrete distributions. It uses the Pearson  $X^2$  statistic, the degrees of freedom depending on the number of parameters that are estimated from the data.

**McNemar's test** for off-diagonal symmetry in  $2 \times 2$  tables may be carried out using a statistic ([Section 9.5](#)) which has an asymptotic chi-squared distribution. This may be extended to **Bowker's test** for off-diagonal symmetry in  $n \times n$  tables.

## EXERCISES

- 9.1 Use (9.6) to verify that the probability of the configuration in Example 9.1 given by the Fisher exact test is  $P = 0.000058$ .
- 9.2 Using the modified form of  $\text{Var}(U)$  for the Jonckheere-Terpstra test given in (9.11) show, for the data used in Examples 9.6 and 9.7, that the standard deviation of  $U$  is 145.18, whereas it is 415.36 if ties are ignored.
- 9.3 Confirm the results quoted in Example 9.8 for association between blood platelet count and spleen size.
- 9.4 Verify for the contingency [table \(9.12\)](#) that the asymptotic Jonckheere Terpstra statistic  $Z$  takes the value zero. [It is not necessary to calculate  $\text{Var}(U)$  to confirm this.]
- 9.5 Confirm that the Goodman-Kruskal statistic  $G$  takes the value  $G = 1$  for each of the contingency tables on p.339.
- 9.6 Determine conditions similar to those given for the case  $G = 1$  to ensure that the Goodman-Kruskal statistic takes the value  $G = -1$ .
- 9.7 Verify that  $C = D$  and  $G = 0$  for the Goodman-Kruskal statistic for the [table \(9.12\)](#) on p. 338.
- 9.8 In a psychological test for pilot applicants, each candidate is classed as extrovert or introvert and is subjected to a test for flying aptitude that he or she may pass or fail. Do the results suggest an association between aptitude and personality type?

	<i>Introvert</i>	<i>Extrovert</i>
<i>Pass</i>	14	34
<i>Fail</i>	31	41

- 9.9 A manufacturer of washing machines issues instructions for their use in English for the UK and US markets, French for the French market, German for the German market and Portuguese for the Portuguese and Brazilian markets. The manufacturer conducts a survey of randomly selected customers in each of these markets and asks them to classify the instructions (in the language appropriate to that country) as excellent, reasonable, or poor. Do the responses set out below indicate the instructions are more acceptable in some countries than in others?

	<i>Excellent</i>	<i>Reasonable</i>	<i>Poor</i>
<i>UK</i>	42	30	28
<i>USA</i>	20	41	19
<i>France</i>	19	29	12
<i>Germany</i>	26	22	12
<i>Portugal</i>	18	31	21
<i>Brazil</i>	31	42	7

Note: It is sensible to consider UK/USA separately - also Portugal/Brazil - because of differences in idiom that may affect understanding of the instructions: e.g. British houses have 'taps', American houses have 'faucets'.

- 9.10 In Palpiteria all those who visit a doctor must pay. A political party claims that this inhibits poorer people from seeking medical aid. Data in the table below are for a random sample of wage earners, and incomes are stated in palpiliras (P), the country's unit of currency. Do they substantiate the claim that the poor make proportionately less use of the services?

Time since last visit to doctor	Income		
	Over 10 000P	5000-10 000P	Under 5000P
Under 6 months	17	24	42
6-12 months	15	32	45
Over 12 months	27	142	271
Never been	1	12	127

Your test should take into account the fact that the row and column categories are both ordered.

- 9.11 Would your conclusions in Exercise 9.10 have been different if the data had been in only two income groupings, under 5000P and 5000P or more?
- 9.12 Prior to an England v Scotland football match 80 English, 75 Scottish and 45 Welsh supporters are asked who they think will win. Do the numbers responding each way indicate that the proportions expecting each side to win are influenced by nationality?

	English	Scottish	Welsh
English win	55	38	26
Scottish win	25	37	19

- 9.13 A machine part is regarded as satisfactory if it operates for 90 days without failure. If it fails in less than 90 days it is unsatisfactory and this results in a costly replacement operation. A supplier claims that for each part supplied the probability of a satisfactory life is 0.95. Each machine requires 4 of these parts and all must be functional for satisfactory operation. To test the supplier's claim, a buyer runs each of 100 machines initially fitted with 4 new parts for a 90-day test period. The numbers of original parts (0-4) still functioning after 90 days are recorded for each machine as follows:

No. surviving	0	1	2	3	4
No. of machines	2	2	3	24	69

Do these results substantiate the supplier's claim?

- 9.14 It is claimed that a typesetter makes random errors at an average rate of 3 per 1000 words set, giving rise to a Poisson process. 100 randomly chosen sets of 1000 words from his output are examined and the mistakes in each counted. Are the results below consistent with the above claim?

<i>No. of errors</i>	0	1	2	3	4	5	6	7
<i>No. of samples</i>	6	11	26	33	12	6	4	2

- 9.15 Responses to emotive questions may be influenced by factors such as the age, race, sex and social background of the questioner. A random sample of 500 women aged between 30 and 40 are further divided into 5 groups of 100, and each group is allocated to one of the following interviewers:

- A: A 25-year-old white female with secretarial qualifications
- B: A middle-aged clergyman
- C: A retired army colonel
- D: A 30-year-old Pakistani man
- E: A non-white female university student

Each interviewer asks each of the 100 people allocated to him or her: 'Do you consider marriages between couples of different ethnic groups socially desirable?' The numbers answering 'yes' in each group are given below. Assess the evidence that the type of person conducting the interview may influence response.

<i>Interviewer</i>	A	B	C	D	E
<i>No. of 'yes'</i>	32	41	18	57	36

- 9.16 To measure abrasive resistance of cloth, 100 samples of a fabric are each subjected to a 10-minute test under a series of 5 scourers, each of which may or may not produce a hole. The number of holes (0 to 5) is recorded for each sample. Are the data consistent with a binomial distribution with  $n = 5$  and  $p$  estimated from the data? (Hint: Determine the mean number of holes per sample. If this is  $x$  then an appropriate estimate of  $p$  is  $x/5$ .)

<i>No. of holes</i>	0	1	2	3	4	5
<i>No. of samples</i>	42	36	14	3	4	1

- 9.17 The McNemar test data in [Table 4.3](#) on climbs can be reformulated in a way that makes the Cochran test given in [Section 6.4.4](#) appropriate with  $t = 2$ . Denoting a success by 1 and a failure by 0, we may classify each of the 108 climbers' outcomes for the first and second climb as either 0 or 1 in a  $2 \times 108$  table. Show that the Cochran Q statistic is in this case identical with the McNemar  $X^2$  statistic given in [Section 9.5](#).

- 9.18 Aitchison and Heal (1987) give numbers of OECD countries using significant amounts of only 1, 2, 3 or 4 fuels in each of the years 1960,

1973, 1983. Are the proportions in the different categories of use changing significantly with time?

Think carefully about an appropriate test and the interpretation of your findings. The data are in no sense a random sample from any population of countries. The categories are ordered, but do you consider that the monotonic trends in association implicit in the conditions to validate, say, a Jonckheere-Terstra test are likely to be relevant for these data?

No. of offuels	Year		
	1960	1973	1983
1	7	10	1
2	13	11	13
3	5	4	9
4	0	0	2

- 9.19 Marascuilo and Serlin (1979) report a survey in which a number of women were asked whether they considered the statement 'The most important qualities of a husband are determination and ambition' to be true or false. The respondents were asked the same question at a later date. Numbers making the possible responses were as follows:

First response	Second response	Numbers
<i>True</i>	<i>True</i>	523
<i>True</i>	<i>False</i>	345
<i>False</i>	<i>True</i>	230
<i>False</i>	<i>False</i>	554

Is there evidence that experience significantly alters attitudes of women towards the truth of the statement?

- 9.20 Jarrett (1979) gives the following data for numbers of coal mine disasters involving 10 or more deaths between 1851 and 1962.

Day of week Number	Sun	Mon	Tue	Wed	Thu	Fri	Sat
	5	19	34	33	36	35	29
<hr/>							
Month Number	Jan	Feb	Mar	Apr	May	Jun	Jul Aug
	14	20	20	13	14	10	18 15
<hr/>							
Month Number	Sep	Oct	Nov	Dec			
	11	16	16	24			

Test whether accidents appear to be uniformly spread over days of the week and over months of the year. What are the implications? Do they surprise you?

- 9.21 Dansie (1986) gives data for a survey in which 800 people were asked to rank 4 makes of car A, B, C, D in order of preference. The number of times each rank combination was specified is given below in brackets after each order. Do the data indicate that preference may be entirely random? Is there a significant preference for any car as first choice?

ABCD(41), ABDC(44), ACBD(37), ACDB(36), ADBC(49), ADCB(41),  
 BACD(38), BADC(38), BCAD(25), BCDA(22), BDAC(33), BDCA(25),  
 CABD(31), CADB(26), CBAD(40), CBDA(33), CDAB(33), CDBA(35),  
 DABC(23), DACB(39), DBAC(30), DBCA(21), DCAB(26), DCBA(34).

- 9.22 Noether (1987b) asked students to select by a mental process what they regarded as random pairs from the digits 1, 2, 3, repeating that process four times. Noether recorded frequency of occurrences of the last digit pair written down by each of 450 students. The results were:

		First digit		
		1	2	3
Second digit	1	31	72	60
	2	57	27	63
	3	53	58	29

What would be the expected numbers in each cell of the above table if pairs were truly random? Test whether one should reject the hypothesis that the students are choosing digits at random. How do you interpret your finding?

- 9.23 For the following  $2 \times 12$  table calculate the Pearson  $X^2$  and the likelihood ratio  $G^2$  statistics. Explain the basic cause of any difference between their values. If suitable computer software is available, determine the exact tail probabilities corresponding to each test statistic.

0	0	0	0	0	0	0	0	0	0	1	1
3	4	17	2	5	1	8	6	4	11	3	0

# Association in categorical data

## 10.1 THE ANALYSIS OF ASSOCIATION

When we reject the hypothesis of independence in contingency tables we often want to analyse the nature of the association. In [Section 9.3.2](#) we considered one aspect of association in tables with ordered row and column categories, recommending the Jonckheere-Terpstra test for monotonic association. We consider further aspects of association in this chapter; the subject is a large one and our treatment is indicative rather than comprehensive. Approaches used to assess association range from parametric modelling to distribution-free methods. Parametric approaches are usually formulated to reflect knowledge about the mechanism generating the counts. One such model is the logistic regression model for binary responses such as success and failure where the proportions responding depend upon levels of one or more explanatory variables. General descriptions of that model are given by Agresti (1990, Chapter 4 and 1996, Chapter 5) and a detailed treatment covering many applications is presented in Cox and Snell (1989).

In [Chapter 9](#) we considered two-way  $r \times c$  tables. Three or more way tables arise in practice and may require detailed and sometimes subtle analysis to elucidate patterns of association. The techniques used range from extensions of some of the nonparametric models outlined for a few special cases in this chapter to analyses based on generalized linear models. A detailed account of the latter is given in McCullagh and Nelder (1989); Dobson (1990) gives a more elementary treatment illustrated by many practical applications.

A common way of presenting data from a three-way classification is by means of two-way cross-sectional tables. [Table 10.1](#) provides an illustration where the counts cover 197 patients in a study of coronary heart disease where the first categorization is into presence (CHD) or absence (no CHD) of the disorder. The second classification is at three blood cholesterol levels (A, B, C) and the third is into five blood pressure groups (I to V).

**Table 10.1** Categorization of 197 patients in a study of coronary heart disease.

		Blood pressure group					<i>Total</i>
		<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>	
<i>CHD</i>	<i>Cholesterol</i>						
	<i>A</i>	2	3	1	0	4	10
	<i>B</i>	2	1	5	3	0	11
	<i>C</i>	4	7	8	6	2	27
	<i>Total (1)</i>	<u>8</u>	<u>11</u>	<u>14</u>	<u>9</u>	<u>6</u>	<u>48</u>
<i>No CHD</i>	<i>A</i>	16	14	11	8	6	55
	<i>B</i>	22	18	5	3	2	50
	<i>C</i>	15	13	10	5	1	44
	<i>Total (2)</i>	<u>53</u>	<u>45</u>	<u>26</u>	<u>16</u>	<u>9</u>	<u>149</u>
	<i>Total (1+2)</i>	61	56	40	25	15	197

Table 10.1 shows two cross-sections of the second and third classifications determined or 'stratified' by each of the two levels of the first classification (disease present or absent). Marginal totals relevant to each cross-section and a set of totals for each blood pressure group across the two strata are also given. The data may be presented in other ways (Exercise 10.1).

Tests for complete independence extend from two- to multiway tables, and may also be applied to two-way cross-sectional tables, but these analyses are usually insufficient (and often inefficient) because we are more interested in exploring the nature of associations.

A particular type of three-way classification that has received considerable attention is the  $k \times 2 \times 2$  table; the method of analysis depends on what questions are of interest and the nature of the categories in each classification. In particular it is often productive to look upon such tables as representing  $k$  strata, each stratum consisting of a  $2 \times 2$  table, a situation we consider in Sections 10.2.2 to 10.2.5.

For  $2 \times 2$  tables an appropriate measure of association is the odds ratio or, often better, the logarithm of the odds ratio which we introduced in Section 9.1.2.

## 10.2 SOME MODELS FOR CONTINGENCY TABLES

### 10.2.1 The loglinear model

In any  $r \times c$  table if there is independence the expected frequency in cell  $(i, j)$  is  $m_{ij} = n_{i+}n_{+j}/N$ , where  $n_{i+}$ ,  $n_{+j}$  are the  $i$ th row and  $j$ th column totals and  $N$  is the total of all counts in the table ([Section 9.1.1](#)). Taking logarithms

$$\log m_{ij} = \log n_{i+} + \log n_{+j} - \log N$$

This means that under independence, for each cell the logarithm of the expected number is a linear function of the logarithms of the row, column and grand totals.

Readers familiar with the normal theory linear model for a randomized block design, basic to many analyses of variance, will recognise (10.1) as an analogue of the additive model expressing a response (e.g. an expected yield) as the sum of an overall experimental mean plus a block effect plus a treatment effect. With that model the observed yield for any unit differs from the expected yield by an added amount usually regarded as a random error or departure from expectation. By analogy, even when there is no association (i.e. independence) between row and column classifications in an  $r \times c$  contingency table in general  $m_{ij}$  will not equal the observed count  $n_{ij}$  for cell  $(i, j)$ . For the model based on (10.1) we may look upon the difference  $\log n_{ij} - \log m_{ij} = \log(n_{ij}/m_{ij})$  as an additive 'error' or 'departure' term when the model for no association is accepted as adequate. This quantity appears in the likelihood ratio statistic  $G^2$  given in (9.10).

Association in contingency tables has analogies with interaction in factorial treatment structures in the analysis of variance and these analogies are exhibited in extensions of (10.1) that allow for association. A detailed treatment is beyond the scope of this book but we illustrate the basic ideas for  $2 \times 2$  and  $2 \times 2 \times 2$  contingency tables and briefly indicate some extensions to simple cases for  $r \times c$  tables. Readers not familiar with the analysis of variance for designed experiments may find the rest of this section difficult. If so it may be wise at first reading to skim through it briefly to grasp basic ideas only rather than to try and master detail.

We digress to comment on fundamentals concerning additive effects and interactions in a simple  $2 \times 2$  factorial treatment structure where observations are measured variables assumed to have some continuous distribution. The  $2 \times 2$  factorial treatment structure has two factors each at two levels. For instance, we may allow a

chemical reaction to proceed for two different times - 2 hours and 3 hours. These are the two levels of a first factor, *time*. We may also carry out the experiment at two different temperatures - 75°C and 80°C - the two levels of a second factor, *temperature*. The responses or yields are the amounts of some chemical produced by a given amount of input material for each of the factor combinations

- 2 hours at 75°C;
- 2 hours at 80°C;
- 3 hours at 75°C;
- 3 hours at 80°C.

Such an experiment is repeated (or replicated) a specified number of times and either the average or the total output for the same number of replicates for each factor combination is recorded. Apart from random variation in the form of an additive error, which in the analysis of variance model is assumed to be distributed  $N(0, \sigma^2)$ , the output  $x_{ij}$  for the first factor at level  $i$  and the second at level  $j$  ( $i, j = 1, 2$ ) is specified by a linear model. If each factor has purely additive effects we speak of a **no-interaction** model.

To clarify these concepts we assume for the moment an unrealistic state of perfection where there is no random variation to create error. Suppose we find operating for 2 hours at 75°C gives an output  $X$  and that this increases to  $X + 3$  if we leave time unaltered but increase temperature to 80°C. In addition, suppose that output increases to  $X + 8$  if we leave temperature at 75°C but increase time to 3 hours. We say we have no interaction if the result of increasing both time from 2 to 3 hours and temperature from 75°C to 80°C is to increase yield from  $X$  to  $X + 3 + 8 = X + 11$ . These results for expected yield are summarized in [Table 10.2](#).

In general, for a combination of level  $i$  of the row factor with level  $j$  of the column factor we denote the true yield (including the additive error) by  $y_{ij}$  and the corresponding **expected** yield (output apart from error) by  $x_{ij}$ . Thus in [Table 10.2](#)  $x_{11} = X$ ,  $x_{12} = X + 3$ ,  $x_{21} = X + 8$ ,  $x_{22} = X + 3 + 8 = X + 11$ . Thus  $x_{11} + x_{21} = 2X + 11 = x_{12} + x_{22}$ , i.e. the diagonal or 'cross' sums in [Table 10.2](#) are equal. This is the fundamental characteristic of an additive (or no interaction) model.

In many real life situations we may find the effect of increasing both time and temperature is either to boost or diminish the effect of changing just one of these factors only. This would be the situation if, for example,  $x_{22}$  in [Table 10.2](#) took the value  $X + 17$  or  $X + 2$  instead of  $X + 11$ , the other  $x_{ij}$  being unaltered. The cross sums are then no longer equal. For the  $2 \times 2$  factorial model with no interaction the key requirement is

**Table 10.2** Expected yields in a no-interaction model.

Process time (hr)	Temperature (°C)	
	75	80
2	$X$	$X + 3$
3	$X + 8$	$X + 3 + 8$

$$x_{11} + x_{22} - x_{12} - x_{21} = 0. \quad (10.2)$$

This equality will not hold for the observed mean yields  $y_{ij}$ , or the corresponding total yields, because of random variation. The standard analysis of variance test for no interaction in this model is essentially one of testing whether the observed  $y_{11} + y_{22} - y_{12} - y_{21}$  departs from zero by an amount which is too great to be attributed to random variation; described in classic statistical jargon as being 'significantly different from zero'. More formally, asserting there is an interaction implies preference for an hypothesis

$$H_1: x_{11} + x_{22} - x_{12} - x_{21} = I, \text{ where } I \neq 0 \quad (10.3)$$

over  $H_0: I=0$ .

In Section 9.1.2 we established that for a  $2 \times 2$  contingency table the condition for no association (independence) between row and column categories was that the odds ratio  $\theta = (m_{11}m_{22})/(m_{12}m_{21}) = 1$  or equivalently that  $m_{11}m_{22} = m_{12}m_{21}$ . Taking logarithms we get the analogue of (10.2) for the loglinear model, i.e.

$$\log m_{11} + \log m_{22} - \log m_{12} - \log m_{21} = 0 \quad (10.4)$$

as the criterion for no association.

For the observed counts  $n_{ij}$  the empirical odds ratio  $(n_{11}n_{22})/(n_{12}n_{21})$  does not usually equal 1. A test of association (or interaction) in a contingency table becomes one of determining whether this ratio differs sufficiently from 1 to provide strong evidence against  $H_0: \theta = 1$ , or equivalently in classic terms whether  $\log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}$  differs significantly from zero, i.e. whether we prefer

$$H_1: \log m_{11} + \log m_{22} - \log m_{12} - \log m_{21} = I, \text{ where } I \neq 0 \quad (10.5)$$

to  $H_0: I = 0$ . Here (10.5) is the analogue of (10.3).

The model (10.3) is called a **first-order** or **two-factor** interaction model, and it is the only possible kind of interaction with a  $2 \times 2$  factorial treatment structure. The models extend to  $2 \times 2 \times 2$

factorial experiments and to  $2 \times 2 \times 2$  contingency tables. The linear model applies now to 3 factors each at two levels. We now represent expected yields at level  $i$  of the first factor, level  $j$  of the second factor and level  $k$  of the third factor by  $x_{ijk}$  where  $i, j, k$  may each take values 1 or 2. Again the observed yields  $y_{ijk}$  differ from these only by additive random errors or departures. If we consider the first two factors at level  $k = 1$  of the third factor we have a first-order interaction between factors 1 and 2 at this level of factor 3 if  $x_{111} + x_{221} - x_{121} - x_{211} = I$ , providing  $I \neq 0$ . Also, we have a first-order interaction between factors 1 and 2 at level 2 of factor 3 if  $x_{112} + x_{222} - x_{122} - x_{212} = J$ , providing  $J \neq 0$ . If  $I \neq J$  we say there is a **second-order** or **three-factor** interaction. If  $I = J$  and they are not both zero, there is only a **first-order** interaction and if  $I = J = 0$  there is no interaction. If the  $x_{ijk}$  are replaced by the observed  $y_{ijk}$  in general the equalities  $I = J$  or  $I = J = 0$  will not be satisfied even if these are relevant. Appropriate tests are then used to decide whether departures from such equalities provide sufficient evidence to imply that the equalities for expected values do not hold. In the context of a log-linear model we replace the  $x_{ijk}$  in the above conditions by  $\log m_{ijk}$  where the  $m_{ijk}$  are the expected counts based on marginal totals in a way described below in Example 10.1. For  $2 \times 2 \times 2$  tables the no association (no interaction) model corresponds to that of independence between classifications and in terms of odds ratios may be written

$$(m_{111}m_{221})/(m_{121}m_{211}) = (m_{112}m_{222})/(m_{122}m_{212}) = 1$$

Dependence or association may be first or second order (first- or second-order interaction in the loglinear sense). For a first-order interaction

$$(m_{111}m_{221})/(m_{121}m_{211}) = (m_{112}m_{222})/(m_{122}m_{212}) = k, k \neq 1 \quad (10.6)$$

and for a second-order interaction

$$(m_{111}m_{221})/(m_{121}m_{211}) \neq (m_{112}m_{222})/(m_{122}m_{212}).$$

By taking logarithms the analogy with the factorial treatment structure model outlined above becomes obvious. The reader should consult a specialist text such as Bishop, Fienberg and Holland (1975), Fienberg (1980), Plackett (1981), Agresti (1984, 1990, 1996) or Everitt (1992) for a detailed discussion of loglinear models and their use in measuring association.

We now consider association when we have a set of  $k$  cross-sectional tables each  $2 \times 2$  with  $k \geq 2$  which may be formed from or

combined as a  $k \times 2 \times 2$  table. How to answer some typical questions that may be asked will be illustrated using two specific data sets, the first given in Tables 10.3 and 10.4 and the second in Table 10.5. That in Tables 10.3 and 10.4 is a classic set first discussed by Bartlett (1935). The data are numbers of surviving (alive) and non-surviving (dead) plum rootstock cuttings in each of four batches of 240 cuttings. The four batches were subjected to treatments having a  $2 \times 2$  factorial structure. The first factor was time of planting - either early or late - and the second was length of cutting - either long or short. There are two possible responses at each level of each factor - *alive* or *dead*. Tables 10.3 and 10.4 represent two useful cross-sectional tables for comparing survival rates. Both contain the same information, but it is given in different order. In Table 10.3 we say the results are stratified by time of planting since each of the component  $2 \times 2$  tables (left and right) is for a different planting time. In Table 10.4 stratification is by length of cutting.

In Table 10.5 we break down a  $4 \times 2 \times 2$  table into four  $2 \times 2$  tables where the  $k = 4$  strata correspond to different age groups. In the  $2 \times 2$  tables within each age group the explanatory variables are two drugs designated as A and B and the responses to each are the column categories *side-effect* or *no side-effect*. The  $k$  age groups forming the strata are also explanatory variables, often referred to in the statistical literature as **covariates**.

**Table 10.3** Survival numbers for early and late planted long and short cuttings.

Length of cutting	Planted early		Planted late	
	Alive	Dead	Alive	Dead
Long	156	84	84	156
Short	107	133	31	209

**Table 10.4** Survival numbers for long and short cuttings planted early and late.

Planting time	Long cuttings		Short cuttings	
	Alive	Dead	Alive	Dead
Early	156	84	107	133
Late	84	156	31	209

**Table 10.5** Side-effects of two drugs grouped by ages.

Age group (yr)	Drug	Side-effect status	
		No side-effect	Some side-effect
20-29	A	8	1
	B	11	1
30-39	A	14	2
	B	18	0
40-49	A	25	3
	B	42	2
over 50	A	39	3
	B	22	6

For the data in [Tables 10.3](#) and [10.4](#) the Fisher exact test, the Pearson chi-squared test or the likelihood ratio test all indicate association in each cross-sectional table. Tables 10.3 and 10.4 both convey the same information though we look at different cross-sections (stratification by planting time in Table 10.3 but by cutting length in Table 10.4). A question of interest about association in the basic  $2 \times 2 \times 2$  table is whether there is evidence of only first-order association or whether there might be second-order association. Remember that for the  $2 \times 2$  cross-sectional tables independence implies that the expected odds ratios must each be unity while first order association implies they should be equal but the common value is not unity, while second-order association indicates they should not be equal. For the data in [Table 10.5](#) the possibilities are more varied. However, the ones likely to be of most interest are whether or not there is evidence of association and if there is any association whether or not it appears to be similar in nature, i.e. imply a common odds ratio, over all strata. Further questions of interest arise if the odds ratio appears to change between strata. For instance, in Table 10.5 do the odds ratios of the cross-sectional  $2 \times 2$  tables relating drugs and side-effects tend to increase with age (the stratification covariate)? Sometimes, with small numbers in individual strata a test for independence applied separately in each strata may give no evidence to indicate association; however it is then wise to consider whether, collectively, there may be some evidence of association if we use appropriate tests. We consider the case of only two strata in [Section 10.2.2](#) and the more general case of  $k \geq 2$  strata in Sections 10.2.3 to 10.2.5.

### 10.2.2 Analysis of two cross-sectional $2 \times 2$ tables

Although some more general methods given in the next section apply to the case  $k = 2$ , once it has been established that there is evidence of association the main point of interest is often whether a first-order model adequately describes the association. When  $k = 2$  this implies that (10.6) holds. We do not of course know the values of the expectations  $m_{ijk}$  in (10.6) but we may obtain from the data estimates  $\hat{m}_{ijk}$  of the expected counts where these are chosen to satisfy the condition

$$(\hat{m}_{111} \hat{m}_{221})/(\hat{m}_{121} \hat{m}_{211}) = (\hat{m}_{112} \hat{m}_{222})/(\hat{m}_{122} \hat{m}_{212}) \quad (10.7)$$

The  $\hat{m}_{ijk}$  no longer have the simple values that they have for the separate cross-sectional tables under the hypothesis of independence. However, it is easily verified that if we retain the condition that all marginal totals in the complete  $2 \times 2 \times 2$  table are fixed, there is only one degree of freedom and once we determine one of the  $\hat{m}_{ijk}$  in (10.7) all the others are fixed. Once these expected values are determined an associated  $P$ -value may be found using either a likelihood ratio test or a Pearson chi-squared test. Exact permutation tests are possible using these statistics but the asymptotic chi-squared distribution results may be used for large samples. In practice the likelihood ratio statistic  $G^2$  has some advantages when we generalize from  $k = 2$  to  $k > 2$ , a situation where there is more than one degree of freedom. We show in [Section 10.3.2](#) that this latter statistic may be partitioned into additive single degree of freedom components in a sense that is not in general true for the Pearson  $X^2$  statistic. The procedure is best illustrated by an example.

#### Example 10.1

*The problem.* Determine whether the data in [Table 10.3](#) are consistent with a model that specifies only first order association.

*Formulation and assumptions.* We postulate a first-order association model (10.6) after first establishing that a 'no association' or independence model is not appropriate. If (10.7) holds the condition that the marginal totals are fixed means that relationship can be used to calculate a value  $x$  for  $\hat{m}_{111}$  under  $H_0$ : *association is first-order only*. The  $G^2$  statistic may then be calculated using (9.10) and summing over all 8 cells of the two  $2 \times 2$  subtables in Table 10.3.

*Procedure.* For convenience we reproduce below the data in Table 10.3. The Fisher exact test, a Pearson chi-squared test or a likelihood ratio test applied to the two  $2 \times 2$  constituent tables show clear evidence of association with all  $P < 0.0001$ .

Length of cutting	Planted early		Planted late	
	Alive	Dead	Alive	Dead
Long	156	84	84	156
Short	107	133	31	209

If we denote  $\hat{m}_{111}$  by  $x$  it is easily shown that the fixed marginal totals condition implies  $\hat{m}_{121} = 240 - x$ ,  $\hat{m}_{211} = 263 - x$ ,  $\hat{m}_{221} = x - 23$ ,  $\hat{m}_{112} = 240 - x$ ,  $\hat{m}_{122} = x$ ,  $\hat{m}_{212} = x - 125$ ,  $\hat{m}_{222} = 365 - x$  and thus (10.7) is satisfied if  $x$  is chosen so that

$$\frac{x(x-23)}{(240-x)(263-x)} = \frac{(240-x)(365-x)}{x(x-125)}$$

This cubic equation in  $x$  may be solved numerically with an appropriate computer algorithm. The relevant solution is one that gives expected numbers that are all positive and turns out to be  $x = \hat{m}_{111} = 161.1$ . The remaining expectations may be calculated using the relationships given above and are summarized as

Length of cutting	Planted early		Planted late	
	Alive	Dead	Alive	Dead
Long	161.1	78.9	78.9	161.1
Short	101.9	138.1	36.1	203.9

Calculating the likelihood ratio statistic  $G^2$  using (9.10) where summation is over all 8 cells gives  $G^2 = 2.30$  (Exercise 10.2) which, with 1 degree of freedom gives  $P = 0.1294$ . The Pearson chi-squared statistic is  $X^2 = 2.27$  giving  $P = 0.1319$ .

*Conclusion.* There is insufficient evidence to reject the hypothesis that there is only a first-order interaction.

*Comments.* 1. Accepting that there is only a first-order interaction is equivalent to saying that the odds ratios for the cross-sectional tables do not differ significantly.

2. We would reach exactly the same conclusion had we worked with the cross-sectional tables in [Table 10.4](#) (Exercise 10.3). This is because each table contains essentially the same information.

3. We give an alternative analysis of these data in Example 10.2 and other methods developed in Section 10.2.3 to 10.2.5 for the case  $k \geq 2$  may also be applied to this example where they are relevant.

*Computational aspects.* General statistical packages, e.g. SPSS, are tending to increase coverage of loglinear models giving at least asymptotic results. StatXact includes exact permutation tests for several of the methods described in the next section.

We pointed out in [Section 9.1.2](#) that the odds ratio or its logarithm contains all the information on association in a  $2 \times 2$  table. Example 10.1 showed one way of using that information. We also pointed out in [Section 9.1.2](#) that for asymptotic tests normality is approached more rapidly if we use  $\varphi = \log \theta$  rather than  $\theta$  itself. If we use natural logarithms and denote the empirical log odds ratio by  $\varphi^*$  i.e.  $\varphi^* = \log [(n_{11}n_{22})/(n_{12}n_{21})]$  then (see e.g. Agresti 1990, Section 3.4.1)  $\varphi^*$  is asymptotically normally distributed under the hypothesis of independence with mean zero and

$$\text{Var}(\varphi^*) = 1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}. \quad (10.8)$$

If any  $n_{ij} = 0$  both  $\varphi^*$  and  $\text{Var } \varphi^*$  become infinite. This difficulty is avoided if we replace all  $n_{ij}$  ( $i, j = 1, 2$ ) by  $n_{ij} + 0.5$  and differences in inference based on  $\varphi^*$  and

$$\varphi^+ = \log \{[(n_{11} + 0.5)(n_{22} + 0.5)]/[(n_{11} + 0.5)(n_{22} + 0.5)]\}$$

are small when most  $n_{ij}$  are moderate or large. Association implies  $\varphi = \varphi_0 \neq 0$  and that the empirical odds ratio  $\varphi^*$  is asymptotically distributed with mean  $\varphi_0$  and variance given by (10.8).

If we have two  $2 \times 2$  tables with expected odds ratios  $\theta_1$  and  $\theta_2$  then  $H_0: \theta_1 = \theta_2$  implies  $\theta_2/\theta_1 = 1$  or  $\varphi_2 - \varphi_1 = 0$ . Thus, under  $H_0$  it follows that asymptotically

$$Z = \frac{\varphi_2^* - \varphi_1^*}{\sqrt{\text{Var}(\varphi_1^*) + \text{Var}(\varphi_2^*)}} \quad (10.9)$$

has a standard normal distribution.

### *Example 10.2*

*The problem.* For the data in [Table 10.3](#) show that the hypothesis  $H_0: \theta_1 = \theta_2$  implying equal association (or independence if both equal 1) is plausible.

*Formulation and assumptions.* We evaluate  $\varphi_k^* = \log \theta_k^*$  for  $k = 1, 2$  and base the test on (10.8) and (10.9).

*Procedure.* For Table 10.3 we find

$$\varphi_1^* = \log[156 \times 133/(107 \times 84)] = 0.8365,$$

$$\varphi_2^* = \log[84 \times 209/(31 \times 156)] = 1.2893$$

$$\text{Var}(\varphi_1^*) + \text{Var}(\varphi_2^*) = (2/156 + 2/84 + 1/107 + 1/133 + 1/31 + 1/209) = 0.0905$$

whence

$$Z = (1.2893 - 0.8365)/\sqrt{0.0905} = 1.51.$$

*Conclusion.* Since  $Z = 1.51$  corresponds to a two-tail  $P = 0.1310$  we have no strong evidence against  $H_0$ .

*Comments* 1. This result is in line with that in Example 10.1 where we found  $G^2 = 2.30$ . Under  $H_0$  the distribution of  $Z^2$  is chi-squared with 1 degree of freedom. The value  $Z^2 = 1.51^2 = 2.27$  is close to that of  $G^2$ .

2. In view of the large counts in each cell, having accepted the hypothesis that the odds ratios do not differ it is reasonable to use the normal approximation to test for each table the hypothesis  $\phi_i = 0$ ,  $i = 1, 2$  implying independence. The standard test statistic  $Z = \phi_i^*/\sqrt{\text{Var}(\phi_i^*)}$  shows strong evidence that there is association (Exercise 10.4).

3. The method used here avoids the need to estimate the expected values  $\hat{m}_{ijk}$  needed in Example 10.1

4. An asymptotic confidence interval for the true difference  $\phi_2 - \phi_1$  may be obtained in the usual way. For example, the approximate 95 per cent interval is of the form

$$\phi_2^* - \phi_1^* \pm 1.96\sqrt{[\text{Var}(\phi_1^*) + \text{Var}(\phi_2^*)]}$$

For the above data the interval is (-0.1369, 1.0423). Because zero is in the interval a hypothesis of equal odds ratios is acceptable.

### 10.2.3 Analysis of $2 \times 2$ tables for several strata

We saw in [Section 9.1.2](#) that binomial sampling with two levels of an explanatory variable is one model often associated with a  $2 \times 2$  table. A typical situation where this happens is one where presence or absence of side-effects (responses) is noted for each of two medicines (explanatory categories). Another example arises in a production process where two types of machine (explanatory categories) are being investigated in manufacturing runs from each of which the numbers of satisfactory and of flawed items (responses) are noted. We may introduce further explanatory variables or covariates leading to situations like that exemplified in [Table 10.5](#) where each of the  $k = 4$  covariates is an age group. These groupings are usually referred to as strata. In that table each stratum consists of a  $2 \times 2$  table of counts classified by drugs and side-effects. An informal inspection of Table 10.5 suggests that side-effect incidence tends to increase slightly with age for drug B, but there is no such tendency for drug A. Apart from this there looks to be little difference in side-effect incidence between the drugs.

Two obvious questions that may be of interest are:

- Is there an association between drugs and side-effects?
- If there is an association does it change with age, i.e. does the odds ratio change between strata?

An intuitive approach is to test for evidence of association within each stratum and if there is no evidence to accept the hypothesis of independence. A disadvantage is that if the counts in some cells are small the tests may have low power against alternatives that specify

a slight association. Intuitively one feels that if the information could be combined in some way for all strata the larger sample size may detect such associations if these are similar for all strata. Care is needed in such analyses, for they may give misleading results especially if the degree of association or type of association differs between strata, a matter we explore further in [Section 10.3.1](#).

In Examples 10.1 and 10.2 we considered the special case of only two strata (i.e.  $k = 2$ ) and the large numbers in each cell enabled us to detect strong evidence of association within each stratum.

The situation is different for the data in [Table 10.5](#), where exact tests applied to each of the four strata specified by age show no evidence of association. However, as we shall see in Example 10.4 the suspicion aroused by a visual inspection of the data that there are more side-effects for drug B among older patients is not without justification, although the evidence for this is not strong.

We proceed by first assessing whether there is evidence that the odds ratios are equal for all strata. If there is strong evidence that these ratios are not all equal, a test for independence across all strata (i.e. that the within-strata odds ratios  $\theta_1, \theta_2, \dots, \theta_k$  have a common value unity) is then not appropriate since we have already decided they are probably not all equal.

Thus a sensible procedure is to test first whether the hypothesis of equality of odds ratios holds and only if there is little evidence against this should we test whether the common value may be unity.

The methods used in Examples 10.1 and 10.2 for the case  $k = 2$  do not generalize easily to the more general  $k > 2$ . Further, we gave only asymptotic results and these are not appropriate if there are many cells with small counts. They are unlikely to be satisfactory for the data in Table 10.5, for example. The approaches using asymptotic and exact results are somewhat different so we consider them separately.

#### 10.2.4 Asymptotic tests about odds ratios in $k$ $2 \times 2$ tables

For any  $k \geq 2$  three common questions are

- Is there evidence of association (i.e.  $\theta_s \neq 1, s = 1, 2, \dots, k$ )?
- If there is association are all  $\theta_s$  equal?
- If all  $\theta_s$  are equal how do we estimate the common value?

Several asymptotic tests help answer these questions and when these are not adequate exact permutation tests are possible using packages like StatXact. We outline some basic ideas behind these procedures, quoting relevant formulae without derivation. Some

asymptotic tests are provided in general statistical packages and may even be conducted manually in simple cases. Many inference procedures in this and the next section rely heavily on the fact that once we have the observed or expected count for cell (1, 1) in any 2 × 2 table this fixes the observed or expected odds ratio if marginal totals are fixed.

We pointed out above that it is desirable to test whether a hypothesis of equality of odds ratios is tenable before testing whether that common ratio may be unity. If we reject the first of these possibilities the second is automatically rejected.

Data often suggest that if there is any association then it may be similar within all strata, e.g. all the empirical odds ratios  $\theta_s = (n_{11s}n_{22s})/(n_{12s}n_{21s})$  might take values slightly greater than 1.

Care is needed in choosing an estimate of a common odds ratio appropriate for testing the hypothesis that there is indeed such a common ratio. With different stratum sizes as in [Table 10.5](#), and perhaps different allocations of numbers to rows or columns within strata, a simple average of all stratum odds ratios is too naive. Mantel and Haenszel (1959) proposed the estimator

$$\theta^* = \frac{\sum_s (n_{11s}n_{22s} / N_s)}{\sum_s (n_{12s}n_{21s} / N_s)} \quad (10.10)$$

This is a weighted mean of the strata odds ratios  $\theta_s$  with weights  $n_{12s}n_{21s}/N_s$  where  $N_s$  is the total count for the cells in stratum  $s$ . The motivation for choosing these weights is that when all  $\theta$  are close to unity they are nearly the reciprocals of the variances of the  $\theta_s$ .

Breslow and Day (1980) proposed a test for whether data were consistent with an estimated common odds ratio given by (10.10). We first compute the expected count  $m_{11s}$  in cell (1, 1,  $s$ ) for each of the  $k$  stratum conditional upon the common odds ratio being  $\theta^*$  given by (10.10). This amounts to choosing for each stratum  $s$  an  $x = m_{11s}$  which is the positive root not exceeding  $n_{1+s}$  or  $n_{+1s}$  of the equation

$$\frac{x(N_s - n_{1+s} - n_{+1s} + x)}{(n_{1+s} - x)(n_{+1s} - x)} = \theta^*. \quad (10.11)$$

If odds ratios are equal across all strata the differences  $|m_{11s} - n_{11s}|$ ,  $s = 1, 2, \dots, k$  should all be small and Breslow and Day (1980, Chapters 4 and 5) discuss appropriate tests for homogeneity of odds ratios in detail. They proposed a statistic

$$M_{BD}^2 = \frac{\sum_s (m_{11s} - n_{11s})^2}{Var(n_{11s})} \quad (10.12)$$

where

$$\text{Var}(n_{11s}) = [(1/m_{11s}) + (1/m_{12s}) + (1/m_{21s}) + (1/m_{22s})]^{-1} \quad (10.13)$$

Asymptotically  $M_{BD}^2$  has a chi-squared distribution with  $k - 1$  degrees of freedom under  $H_0$ : all  $\theta_s$  are equal. The expected frequencies in the expression for  $\text{var}(n_{11s})$  are subject to the fixed marginal totals for stratum s given  $m_{11s}$ .

If we accept  $H_0$  we may obtain a confidence interval for the common value  $\theta$  which we estimate by  $\theta^*$ . Asymptotically the distribution of  $\varphi^* = \log \theta^*$  approaches normality more rapidly than that of  $\theta^*$ .  $E(\varphi^*) = \varphi = \log \theta$  and several estimates have been proposed for  $\text{Var}(\varphi^*)$ . That recommended in StatXact was proposed by Robins, Breslow and Greenland (1986) and has been shown to work well for small  $k$  and large counts or for large  $k$  even if some of the  $2 \times 2$  tables are fairly sparse. The expression for this estimate is complicated but we give it for completeness. For each stratum s we define

$$a_s = (n_{11s} + n_{22s})/N_s, \quad b_s = (n_{12s} + n_{21s})/N_s, \quad c_s = n_{11s}n_{22s}/N_s,$$

$$d_s = n_{12s}n_{21s}/N_s$$

then write

$$c_+ = \sum_s c_s, \quad d_+ = \sum_s d_s$$

and

$$P = [\sum_s (a_s c_s)]/(2c_+^2), \quad Q = [\sum_s (a_s d_s + b_s c_s)]/(2c_+ d_+), \quad R = [\sum_s (b_s d_s)]/(2d_+^2)$$

then

$$\text{Var}(\varphi^*) = P + Q + R.$$

Approximate 95 per cent asymptotic confidence limits for  $\varphi$  are

$$\varphi^* \pm 1.96 \sqrt{\text{Var}(\varphi^*)}. \quad (10.14)$$

For all but small data sets (where asymptotic procedures are not appropriate) a computer program is highly desirable for testing and estimation, but we illustrate salient features of the computation in Example 10.3. Approximate confidence limits for  $\theta$  can be obtained by back-transformation. Everitt (1992, Section 2.8.1) develops an alternative formula basing the interval on  $\theta^*$  directly.

**Table 10.6** Production counts for two machines with three raw material sources.

Material source	Machine	Output status	
		Satisfactory	Faulty
A	Type I	42	2
	Type II	33	9
B	Type I	23	2
	Type II	18	7
C	Type I	41	4
	Type II	29	12

*Example 10.3*

*The problem.* For the data in **Table 10.6** explore whether the hypothesis of equality of odds ratios is acceptable. If it is, obtain an approximate 95 per cent confidence interval for the common odds ratio  $\theta$ .

*Formulation and assumptions.* The material sources form three strata ( $k = 3$ ) and the two machine types are explanatory categories and the output categories represent two responses. The solution requires the following steps: (1) estimation of a common odds ratio using (10.10); (2) calculation of  $\chi^2$  from (10.12) for use in an asymptotic test of  $H_0$ : all  $\theta_i$  are equal against  $H_1$ :  $\theta_i \neq \theta_j$  for at least one  $i \neq j$ ; (3) if  $H_0$  is accepted, a confidence interval for the common value  $\theta$  is obtained by back-transformation from the interval for  $\phi$  given by (10.14).

*Procedure.* Although we recommend using widely available software for this test we outline the main steps in the computation. The reader should verify all quoted numerical values (Exercise 10.5). Using (10.10) gives

$$\theta^* = \frac{(42 \times 9) / 86 + (23 \times 7) / 50 + (41 \times 12) / 86}{(33 \times 2) / 86 + (18 \times 2) / 50 + (29 \times 4) / 86} = 4.702$$

The expected value  $m_{111}$  is the relevant root of (10.11), which here is

$$\frac{x(86 - 75 - 44 + x)}{(75 - x)(44 - x)} = 4.702$$

This gives  $x = m_{111} = 41.68$  (the other root  $x = 100.54$  exceeds both row and column marginal totals). Similarly, we find  $m_{112} = 23.07$  and  $m_{113} = 41.25$ . Then (10.13) gives

$$\text{Var}(n_{111}) = (1/41.68 + 1/2.32 + 1/33.32 + 1/8.68)^{-1} - 1.6660$$

and in like manner  $\text{Var}(n_{112}) = 1.3181$ ,  $\text{Var}(n_{113}) = 2.4550$ , whence (10.12) gives

$$M_{BD}^2 = 0.32^2/1.6660 + 0.07^2/1.3181 + 0.25^2/2.4550 = 0.09.$$

Assuming an asymptotic chi-squared distribution with  $k - 1 = 2$  degrees of freedom there is clearly no evidence against  $H_0$ ; indeed, the fit is remarkably good ( $P = 0.96$ ). To obtain the relevant confidence interval we note that  $\varphi^* = \log \theta^* = 1.5480$ ; tedious but otherwise straightforward computation gives  $\text{Var}(\varphi^*) = 0.1846$  whence (10.14) gives the 95 per cent confidence limits for  $\varphi$  as

$$1.5480 \pm 1.96\sqrt{0.1846}$$

giving an interval  $(0.7059, 2.3901)$ . Taking natural antilogarithms gives an interval  $(2.0256, 10.9145)$  for  $\theta$ .

*Conclusion.* We accept the hypothesis of a common odds ratio and estimate it (after sensible rounding) to be  $\theta = 4.70$  with a 95 per cent confidence interval  $(2.03, 10.92)$ .

*Comments.* 1. The confidence interval appears rather wide and is not symmetric about the point estimate. This reflects the markedly skew distribution of  $\theta$  referred to in [Section 9.1.2](#).

2. Since independence or lack of association implies  $\theta = 1$  there is strong evidence of association;  $\theta = 1$  is clearly outside the above confidence interval.

*Computational aspects.* 1. The above results are confirmed by the program in StatXact but the estimate and confidence limits given there are the reciprocals of those given here. This is because StatXact defines the odds ratios as reciprocals of those we give. This leads to similar inferences because, as we pointed out, ratios  $\theta$  or  $1/\theta$  represent the same degree of association, only in opposite directions.

2. Some software may give a different confidence interval to that obtained here, because, as indicated above, there are alternative estimates of variance to that given by the Robins, Breslow and Greenland formula.

The method in Example 10.3 tests whether a common odds ratio across all strata is an acceptable hypothesis, and if it is we may then decide with the aid of a confidence interval whether that common value may be  $\theta = 1$  implying independence. There is another widely used test for independence known usually as the Mantel-Haenszel test as it was proposed by Mantel and Haenszel (1959) although Cochran (1954) proposed an effectively identical test. The latter is described by Everitt (1992, Section 2.7.2). We give the test in the Mantel-Haenszel formulation as we find this intuitively appealing because of similarities to the Breslow-Day test given above.

Basically the test requires the calculation in each stratum of the cell  $(1, 1)$  expectations under the hypothesis of independence in that stratum, i.e.  $m_{11s} = n_{1+s}n_{+1s}/N_s$ . The statistic is

$$M^2 = \frac{(m_{11+} - n_{11+})^2}{\sum_s \text{Var}(n_{11s})} \quad (10.15)$$

where  $m_{11+} = \sum_s m_{11s}$ ,  $n_{11+} = \sum_s n_{11s}$  and

$$\text{Var}(n_{11s}) = n_{1+s}n_{+1s}n_{2+s}n_{+2s}/[N_s^2(N_s - 1)]$$

Asymptotically (10.15) has a chi-squared distribution with 1 degree of freedom under  $H_0$ : all  $\theta_s = 1$ . If the expected and observed values in the cell of each table differ appreciably and these differences are all in the same direction this suggests odds ratios are consistently above or consistently below 1 and large values of  $M^2$  are likely. Circumstances may arise however where differences between expected and observed values in cell (1, 1) for each table are in opposite directions in some tables to that in others. In this case we may still get small values of  $M^2$  even when there is clear evidence of association (albeit in opposite directions) within some strata. For this reason the Mantel-Haenszel test is not recommended without first using a test such as the Breslow-Day test for equality of odds ratios. However, if the asymptotic theory is accepted for the Breslow-Day test we may, as shown above, compute approximate confidence intervals for a common odds ratio and accept or reject a hypothesis of independence on the basis of these, making the Mantel-Haenszel test somewhat superfluous. Since it is widely used we do however illustrate how it works.

#### *Example 10.4*

*The problem.* For the data in [Table 10.5](#) compute  $M^2$ , and assuming all  $k = 4$  strata have the same expected odds ratio  $\theta$  is it reasonable to accept the hypothesis  $H_0: \theta = 1$ ?

*Formulation and assumptions.* Assuming all  $\theta_s$  are equal we compute  $M^2$  using (10.15) and obtain a  $P$ -value assuming a chi-squared distribution with 1 degree of freedom.

*Procedure.* We find  $m_{111} = 9 \times 19/21 = 8.14$ ,  $m_{112} = 15.06$ ,  $m_{113} = 26.06$ ,  $m_{114} = 36.60$  whence  $m_{11+} = 8.14 + 15.06 + 26.06 + 36.60 = 85.86$  and  $n_{11+} = 8 + 14 + 25 + 39 = 86$ . Also  $\text{Var}(n_{111}) = (19 \times 9 \times 12 \times 2)/(21^2 \times 20) = 0.4653$ ,  $\text{Var}(n_{112}) = 0.4832$ ,  $\text{Var}(n_{113}) = 1.1213$ ,  $\text{Var}(n_{114}) = 1.9096$  whence  $\sum_s \text{Var}(n_{11s}) = 0.4653 + 0.4832 + 1.1213 + 1.9096 = 3.9794$  so that

$$M^2 = (0.14)^2/3.9794 = 0.0049.$$

*Conclusion.* The extremely small value 0.0049 means there is no evidence against the hypothesis of independence.

*Comments.* 1. Some writers recommend a continuity correction subtracting  $\frac{1}{2}$  from  $|m_{11+} - n_{11+}|$  before squaring. This may avoid over-weighing of evidence against  $H_0$  in borderline cases.

2. The small differences between corresponding  $m_{11s}$  and  $n_{11s}$  point towards independence. However, as we indicated above situations arise where these differences may be large and of opposite sign in different strata but the numerator in (10.15) remains small. This is why we prefer the Breslow-Day test

with confidence intervals for a common  $\theta$  if that hypothesis is accepted. For the data in this example it can be shown (Exercise 10.6) that  $M_{BD}^2 = 6.497$  with an associated  $P = 0.0898$  so there is no substantial evidence against equality of odds ratios, although room for a little doubt. The point estimator of the odds ratio is  $\theta^* = 1.0334$  with a 95 per cent confidence interval (0.4036, 2.6462). In the light of these findings it is not surprising that the Mantel-Haenszel test strongly supports independence. The Breslow-Day test, however, does raise a small doubt about the hypothesis of equality of odds ratios.

### 10.2.5 Exact tests for equality of odds ratios in $k$ $2 \times 2$ tables

Zelen (1971) gives an exact permutation test for a common odds ratio and Gart (1970) gives a method for obtaining confidence intervals if a hypothesis of a common odds ratio is accepted. Stafact includes programs for both procedures. Unlike many other nonparametric tests the asymptotic procedures described in the previous section are not based on the asymptotic properties of the statistics used in the exact test. As computing facilities such as those provided by StatXact are needed to apply the exact methods, we only sketch their nature. The rationale for both procedures is illustrated for small data sets by Sprent (1998, Section 12.5) and is explained more fully in the StatXact manual.

Zelen's test for homogeneity of the ratios is based on a property of the hypergeometric probabilities arising in the Fisher exact test for  $2 \times 2$  tables. Subject to the condition that all  $n_{11s}$  sum over  $s$  to the observed marginal total  $n_{11+}$ , the product of the hypergeometric probabilities over all possible  $k$   $2 \times 2$  tables satisfying this condition is a maximum when the observations strongly support a common odds ratio  $\theta$  whatever the value of  $\theta$ , and this product decreases steadily as that support weakens. Under the hypothesis that all odds ratios are equal, the actual probability  $P^*$  of observing any particular set of within-stratum outcomes in which the cell (1, 1) entries sum to the observed  $n_{11+}$  is given by the product of the  $k$  hypergeometric probabilities for those outcomes divided by the sum of all such products over all possible cell (1, 1) values consistent with marginal totals for the relevant  $2 \times 2$  tables and which sum to  $n_{11+}$ . If we denote the probability for our observed tables by  $P_0^*$ , the corresponding  $P$ -value is the sum of the probabilities  $P^*$  associated with all outcomes for which  $P^* \leq P_0^*$ . If the test is applied to the data in Table 10.5 StatXact gives the probability associated with the observed outcome as  $P_0^* = 0.008213$  and the relevant  $P = 0.0689$  compared with the asymptotic Breslow-Day value  $P = 0.0898$  given in Comment 2 on Example 10.4.

The procedure proposed by Gart for estimating the common odds ratio if the Zelen test supports that hypothesis uses a different approach. We noted in the asymptotic Breslow-Day procedure that if we estimate expected values under an assumption that the common odds ratio takes a particular value  $\theta^*$  say, that in general if  $m_{11s}$  is the expected value in cell (1, 1) of table s then  $\sum_s m_{11s} \neq n_{11+}$ . Large discrepancies indicate that such a value of  $\theta^*$  is an unsatisfactory value for the common odds ratio and of course there may be no value of  $\theta^*$  for which this discrepancy is not large if the hypothesis that there is a common odds ratio is not supported. Gart's procedure for obtaining a confidence interval for a common odds ratio  $\theta$  is in essence to compute a statistic  $T$  which is the sum of the *observed*  $n_{11s}$ . One then considers its position in the distribution of all possible values that this statistic might take if each observed  $n_{11s}$  were replaced by any possible value consistent with the marginal totals for the sth  $2 \times 2$  table; i.e. the values are no longer constrained to sum to  $n_{11+}$ . We may then find values  $\theta^-$ ,  $\theta^+$  that give values of  $T$  which would just be accepted as hypothetical estimates of the common odds ratio in a test at the  $100(\frac{1}{2}\alpha)$  per cent significance level and these provide  $100(1 - \alpha)$  per cent confidence limits for  $\theta$ . Some modifications are needed to these arguments if the data indicate acceptance of either a zero or infinite common odds ratio. Details of many subtle aspects of the Gart procedure are given in the StatXact 4 manual, Section 15.3.2.

For the data in [Table 10.6](#) the exact 95 per cent confidence interval given by Gart's method is (1.9149, 12.306), slightly longer than the asymptotic interval (2.0256, 10.9145) obtained in Example 10.3. The latter depends on the validity of certain variance estimates, involves a back-transformation from natural logarithms and assumes a reasonable rate of convergence to normality. From the practical viewpoint the difference between the asymptotic result and that for exact permutation theory should not cause alarm bearing in mind that there are also questions of discontinuity in the distribution of Gart's  $T$  statistic so the interval may not be an exact 95 per cent interval.

## 10.2.6 The linear-by-linear association model

The linear-by-linear association model is a relatively simple yet versatile loglinear model with only one additive interaction term to represent association. For an  $r \times c$  table a score  $u_i$  is allocated to row  $i$ ,  $i = 1, 2, \dots, r$  subject to constraints  $u_1 \leq u_2 \leq \dots \leq u_r$  and a score

$v_j$  to column  $j$ ,  $j = 1, 2, \dots, c$  subject to constraints  $v_1 \leq v_2 \leq \dots \leq v_c$ . If  $m_{ij}$  is the expected count in cell  $(i, j)$  the model may be written

$$\log m_{ij} = \text{independence terms} + \beta u_i v_j \quad (10.16)$$

where  $\beta$  is a constant. The model reduces to that for independence (no association) between row and column categories when  $\beta = 0$ . It is sometimes convenient to use the modified form

$$\log m_{ij} = \text{independence terms} + \beta(u_i - m_u)(v_j - m_v) \quad (10.17)$$

where  $m_u, m_v$  are the means of the  $u_i$  and of the  $v_j$  respectively. Clearly (10.17) is equivalent to (10.16) if in the latter we replace  $u_i, v_j$  by  $u_i^* = u_i - m_u$  and  $v_j^* = v_j - m_v$ .

A number of standard nonparametric test statistics have exact permutation distributions based on this model. It is relevant to tables of counts where both row and column categories are ordered and when, if the classifications are not independent, they show a monotonic trend across row and column categories. When the scores are ranks they are often referred to as unit interval scores because the difference between the scores associated with any pair of neighbouring rows or of neighbouring columns is in each case unity.

The parameter  $\beta$  may be estimated to reflect the extent of any relevant departure from independence. We do not discuss estimation of  $\beta$  but indicate what we mean by relevant departures in Example 10.5. Cells for which  $|\beta(u_i - m_u)(v_j - m_v)|$  is large show the greatest departure from independence.

#### Example 10.5

*The problem.* For the case  $r = 5, c = 7$  with rank scores, i.e.  $u_i = i, v_j = j$ , determine for each cell in the  $r \times c$  table the contribution to  $\log m_{ij}$  of the interaction term  $\beta(u_i - m_u)(v_j - m_v)$  in (10.17).

**Table 10.7** Interaction terms in  $\log m_{ij}$  for rank scores in a  $5 \times 7$  contingency table.

Column score Row score	-3	-2	-1	0	1	2	3
-2	$6\beta$	$4\beta$	$2\beta$	0	$-2\beta$	$-4\beta$	$-6\beta$
-1	$3\beta$	$2\beta$	$\beta$	0	$-\beta$	$-2\beta$	$-3\beta$
0	0	0	0	0	0	0	0
1	$-3\beta$	$-2\beta$	$-\beta$	0	$\beta$	$2\beta$	$3\beta$
2	$-6\beta$	$-4\beta$	$-2\beta$	0	$2\beta$	$4\beta$	$6\beta$

*Formulation and assumptions.* In this example  $m_u = 3$  and  $m_v = 4$  so the interaction term in cell  $(i,j)$  in (10.17) is  $\beta(i-3)(j-4)$ .

*Procedure.* For any  $\beta$  the interaction term in each cell is given in [Table 10.7](#).

*Conclusion.* If  $\beta > 0$ ,  $\log m_{ij}$  is greater than its value under independence for cells at the top left and bottom right of Table 10.7 and less for cells at the top right or bottom left. If  $\beta < 0$  the situation is reversed. If  $\beta = 0$  there is no association.

The name linear-by-linear association reflects the characteristic that within any row, say row  $i$ , for any chosen  $u_i, v_j$  the interaction contribution across columns is a linear function of the  $v_j$  with slope  $\beta(u_i - m_u)$  and that within any column  $j$  the interaction contribution across rows is a linear function of the  $u_i$  with slope  $\beta(v_j - m_v)$ .

The relevance of this model to permutation theory is that many permutation tests depend upon a statistic,  $S$ , that may be written

$$S = \sum_{ij} u_i v_j n_{ij} \quad (10.18)$$

with appropriate choices of  $u_i, v_j$ . Independence ( $\beta = 0$ ) corresponds to the null hypothesis; we reject  $H_0$  if our observed outcome expressed in the appropriate contingency table format indicates a general pattern of the form exemplified for a particular case in Table 10.7 when  $\beta \neq 0$ . This corresponds to extreme values of  $S$  given by (10.18) whereas intermediate values indicate lack of association.

### Example 10.6

*The problem.* In Example 5.1 we applied the WMW test to a data set we presented in a contingency table format in Table 9.2. Show that if, in that format, we use row scores 0, 1 corresponding to Group A and Group B respectively and column scores corresponding to the ranks of the observations in the combined groups, then  $S$  is a statistic for the Wilcoxon rank sum test.

*Formulation.* In [Table 9.2](#) the columns were ordered by increasing times taken to complete the task. We now allocate to the columns in that table the rank scores  $1, 2, 3, \dots, 21$  and compute the Statistic  $S$  in (10.18) using these scores.

**Table 10.8** The information in Table 9.2 with times replaced by corresponding ranks for column scores and 0, 1 assigned as row scores.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
		<i>Group &amp; score</i>																				
A	0	1	1	1	0	1	1	0	1	1	0	0	1	0	1	0	1	0	0	0	0	
B	1	0	0	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	

*Procedure.* With the above formulation by referring to Table 9.2 it is easy to obtain the relevant Table 10.8 with the scores shown. It follows immediately that  $S$  given by (10.18) is the sum of the ranks for Group B and here  $S = 155$ .

*Conclusion.* The Wilcoxon rank sum test is one for independence in a linear-by-linear association model if the above row and column scores are used.

*Comments.* 1. If we interchange Group A and Group B, i.e. swap rows and retain the same row scores, the statistic  $S$  becomes the sum of the Group A ranks, the alternative Wilcoxon rank sum statistic used in Example 5.1. Hence in this two-sample situation the ordering of the groups does not matter; there is an 'opposite' association with the two different orderings.

2. In the context of the WMW test independence or lack of association corresponds to a situation when there is no difference between group medians and is associated with a broad scatter of zeros and ones in the cells of a table like Table 10.8. There is strong linear-by-linear association if the ones (corresponding to an observation) tend to concentrate towards the left or right in the first row and towards the opposite extreme in the second row, a situation we saw in a simpler case in Section 9.3.1.

The Spearman rank correlation coefficient provides a more informative illustration. If there are  $n$  paired observations without ties the rank outcome pattern can be presented as an  $n \times n$  contingency table with rows and columns representing  $x$  ranks and  $y$  ranks respectively. If an  $x$ -rank  $i$  is paired with a  $y$ -rank  $j$  we enter 1 in cell  $(i, j)$ . All other entries in row  $i$  or in column  $j$  are 0. The row and column totals are all 1. We illustrate the pattern for three cases when  $n = 5$ .

<i>Case A</i>	<i>x-ranks</i>	1	2	3	4	5
	<i>y-ranks</i>	4	2	5	3	1
<i>Case B</i>	<i>x-ranks</i>	1	2	3	4	5
	<i>y-ranks</i>	2	1	3	5	4
<i>Case C</i>	<i>x-ranks</i>	1	2	3	4	5
	<i>y-ranks</i>	1	2	3	4	5

These are represented by the  $5 \times 5$  tables (10.19), (10.20), (10.21).

<i>Case A</i>	<i>y-ranks</i>	1	2	3	4	$S$
	<i>x-ranks</i>	1	2	3	4	
	1	0	0	0	1	0
	2	0	1	0	0	0
	3	0	0	0	0	1
	4	0	0	1	0	0
	5	1	0	0	0	0

(10.19)

Case B	y-ranks	1	2	3	4	5	(10.20)
	x-ranks	1	0	1	0	0	
	1	0	1	0	0	0	
	2	1	0	0	0	0	
	3	0	0	1	0	0	
	4	0	0	0	0	1	
	5	0	0	0	1	0	

Case C	y-ranks	1	2	3	4	5	(10.21)
	x-ranks	1	1	0	0	0	
	1	1	0	0	0	0	
	2	0	1	0	0	0	
	3	0	0	1	0	0	
	4	0	0	0	1	0	
	5	0	0	0	0	1	

In case A, the 1s are well scattered over rows and columns. In case B there is a discernable tendency for the 1 s to drift towards the top left and bottom right, hinting at some association between  $x$ - and  $y$ -ranks. Case C clearly represents the closest possible association and corresponds to  $r_s = 1$ . The resemblance of these tables to scatter diagrams (with conventional axes rotated clockwise through  $90^\circ$ ) is apparent if we imagine the Os to be deleted and the 1s to be replaced by dots.

It is implicit from our remarks in [Section 7.1.1](#) about permutation tests for the Pearson coefficient that the sum  $\sum_i r_s i$  may be used here as a test statistic for the Spearman coefficient in an exact permutation test. It immediately follows that if we take the ordered  $x$ -ranks and  $y$ -ranks as scores for a linear-by-linear association test in contingency tables of the form (10.19) to (10.21) that S given by (10.18) equals  $\sum_i r_s i$ . The relevant permutation distribution is over all such possible  $5 \times 5$  tables with row and column sums of unity. It is easily seen that there are  $5!$  such tables in each case.

Similar arguments show that for  $n$  paired observations  $(x_i, y_i)$  and no ties in either variable the statistic  $\sum_i x_i y_i$  appropriate for an exact permutation test for the Pearson product moment correlation coefficient ([Section 7.1](#)) is equivalent to (10.18) with scores  $u_i = x_{(i)}$ ,  $v_j = y_{(j)}$ . The relevant permutation is over all  $n \times n$  tables with row and column sums all unity.

The models extend easily to tied situations. For example, using mid-ranks for the Spearman coefficient with ties consider the case:

<i>x-ranks</i>	1.5	1.5	3	4	5
<i>y-ranks</i>	1	4	4	2	4

The relevant contingency table with mid-rank scores is now:

<i>y-ranks</i>	1	2	4	Total
<i>x-ranks</i>				
1.5	1	0	1	2
3	0	0	1	1
4	0	1	0	1
5	0	0	1	1
Total	1	1	3	5

To generate the relevant permutation distribution we calculate S using mid-rank scores for all permutations of this table with the given marginal totals. We may alternatively, if we wish, use the centralized scores  $u_i - m_u, v_j - m_v$ .

StatXact has a program to compute relevant *P*-values for any linear-by-linear association model based on permutation of such tables subject to fixed marginal totals for arbitrarily assigned scores. This is a powerful tool but it is important to choose relevant and sensible scores. We consider such scores in a re-analysis of the data in Example 9.6. The tests are essentially applications of the permutation test for a Pearson product moment correlation coefficient with many ties and scores (*x, y* values) chosen to reflect what may be regarded as appropriate measures of 'distance' between categories.

#### *Example 10.7*

*The problem.* Given the data on side-effect of dose levels of a drug in Example 9.6, i.e.:

Dose	None	Side-effects		
		Mild	Moderate	Severe
100 mg	50	0	1	0
200 mg	60	1	0	0
300 mg	40	1	1	0
400 mg	30	1	1	2

allocate appropriate scores and perform a linear-by-linear association test for evidence of association between dose level and side-effects.

*Formulation and assumptions.* One reasonable choice of scores would be row and column ranks. For rows this corresponds to dose levels in 100 mg units. For columns it represents an ordering of side-effects consistent with limited information. Another possibility would be to allocate column scores in a way a clinician might interpret the data. He or she might accept 1 and 2 as reasonable scores for no and slight side-effects but regard moderate side-effects as 10 times as serious as slight ones and severe side-effects as 100 times as serious as slight ones, giving logical column scores  $v_1 = 1$ ,  $v_2 = 2$ ,  $v_3 = 20$ ,  $v_4 = 200$ . The StatXact linear-by-linear association program will perform the relevant tests using either of these (or any other chosen) scoring system. If no suitable program is available we give below an asymptotic result but this should be used with some reservations when, as here, there are many cells with small counts.

*Procedure.* Using the StatXact linear-by-linear association program with row and column rank scores gives a one-tail  $P = 0.0112$  and doubling this gives a two-tail  $P = 0.0224$ . With the alternative column scores 1, 2, 20, 200 the corresponding tail probabilities are  $P = 0.0071$  and  $P = 0.0142$ .

*Conclusion.* A one-tail  $P$ -value is relevant as it is logical to expect side-effects to increase rather than decrease as dose increases. Whichever scoring system we use there is strong evidence that side-effects increase with dose.

*Comments.* 1. In Example 9.6 the Jonckheere-Terpstra test indicated strong evidence of association. As we pointed out in [Section 9.3.2](#) that test is equivalent to one using Kendall's tau. Using ranks for row and column scores is more like using Spearman's rho without taking ties into account, while other row and column scores give a test equivalent to that for a Pearson coefficient with those scores playing the roles of  $x$ ,  $y$  values.

2. One might have performed a test exactly equivalent to that for Spearman's rho by allocating mid-rank scores for ties. Because there are 51 patients 'tied' at the 100 mg dose the score for row 1 would then be 26, that for row 2 would be  $\frac{1}{2}(52 + 112) = 82$ , and so on. A valid criticism of the use of these mid-ranks in this highly tied situation is that the mid-ranks depend heavily on the number of subjects allotted to each treatment (here dose levels). There must be some unease about such a scoring system, especially if the numbers allocated to each treatment differ substantially. There is a better case for basing scores on the nature of treatments or responses rather than upon the number of patients allocated to each treatment. Possible unsatisfactory influences of mid-rank scores with heavy tying and unbalanced row and column totals have been pointed out by a number of writers including Graubard and Kom (1987).

3. In a group of tests on the same data each using different but plausible scores  $P$ -values often differ more substantially than they did in this example. An important caveat about any scoring system is that it should be chosen for its relevance to the problem at hand and not on a hint inspired by an inspection of the data that suggests that a particular scoring system may enhance the prospects of getting a small  $P$ -value. The speed of modern computers tempts one to try many analyses for the same data. The temptation is best avoided for linear-by-

linear association tests by choosing a scoring system before data are obtained. Once again we refer the reader to the penultimate paragraph in [Section 1.5](#).

We do not prove it, but it can be shown that for  $S$  given by (10.8)

$$E(S) = [(\sum_i u_i n_{i+})(\sum_j v_j n_{+j})]/N$$

and

$$\text{Var}(S) = \frac{[\sum_i u_i^2 n_{i+} - (\sum_i u_i n_{i+})^2][\sum_j v_j^2 n_{+j} - (\sum_j v_j n_{+j})^2]}{N^2(N-1)}$$

and asymptotically  $Z = [S - E(S)]/\sqrt{\text{Var}(S)}$  has a standard normal distribution.

### *Example 10.8*

*The problem.* Apply an asymptotic test to the data in Example 10.7 using rank scores for rows and columns.

*Formulation and assumptions.*  $E(S)$  and  $\text{Var}(S)$  are computed by the formulae given above and the value of  $Z$  is calculated.

*Procedure.* The ranks scores are  $u_i = i$ ,  $v_j = j$  and we easily verify that  $n_{1+} = 51$ ,  $n_{2+} = 61$ ,  $n_{3+} = 42$ ,  $n_{4+} = 34$ ,  $n_{+1} = 180$ ,  $n_{+2} = 3$ ,  $n_{+3} = 3$ ,  $n_{+4} = 2$ ,  $N = 188$ .

For the given data simple calculations give  $S = 484$ ,  $E(S) = 469.7$  and  $\text{Var}(S) = 35.797$  whence  $Z = 2.39$ , and  $P = \Pr(Z > 2.39) = 0.0084$ .

*Conclusion.* There is strong evidence against  $H_0$ .

*Comment.* This result is in line with that for the exact test. In Exercise 10.7 we establish with the second choice of scores that the asymptotic  $P = 0.0101$ . Despite many ties and many small cell counts in this example the asymptotic results are not misleading.

Many other tests already encountered may be formulated as linear-by-linear association tests. If the rank column scores in Example 10.6 are replaced by raw data scores this leads to a Pitman test. Other alternatives might use the van der Waerden scores or, if appropriate, log-rank scores.

A test for trends in  $r \times 2$  contingency tables known as the Cochran-Armitage test may also be formulated as a linear-by-linear association test. The test applies basically to a situation where the rows represent explanatory variables and the columns two mutually exclusive and exhaustive outcomes. Typically the rows are ordered explanatory variables such as age groups or increasing doses of a drug. It is assumed that columns represent binomial responses, the first for an event A and the second for an event B. For the  $i$ th row,  $i = 1, 2, \dots, r$  for each of the  $n_{i+}$  units in the counts to that row

the probability of the event A is  $p_i$  and of the opposite event B is  $q_i = 1 - p_i$ . The Cochran-Armitage test proposed independently by Cochran (1954) and Armitage (1955) is used to test the hypothesis  $H_0 : p_1 = p_2 = \dots = p_r$  against  $H_1 : p_1 \leq p_2 \leq \dots \leq p_r$ , where at least one inequality is a strict inequality (or a similar inequality with all signs reversed). Effectively Cochran and Armitage proposed a test which is a linear-by-linear association test with column scores 0, 1 and row scores 1, 2, 3, ...,  $r$ . In Example 10.9 we apply it to a set of data given by Graubard and Korn (1987); these data are also used to illustrate the procedure in the StatXact Manual.

*Example 10.9.*

*The problem.* Table 10.9 indicates whether or not congenital sex organ malformation was found among children born to mothers whose stated alcohol consumption (number of drinks per day) during pregnancy fell within various ranges. There is a slight indication that the probability that malformation is absent may decrease with increasing maternal alcohol consumption since the maximum likelihood estimates of  $p_i$ , for each row are  $p_1 = 17066/(17066 + 48) = 0.9972$  and similarly  $p_2 = 0.9974$ ,  $p_3 = 0.9937$ ,  $p_4 = 0.9921$ ,  $p_5 = 0.9737$ .

*Formulation and assumptions.* Our alternative to equality of the  $p_i$  is one of a decreasing trend and we perform an exact Cochran-Armitage test as a linear-by-linear association test with row scores 1, 2, 3, 4, 5 and column scores 0, 1.

*Procedure.* Using StatXact the exact one-tail P = 0.1046.

*Conclusion.* There is not strong evidence against  $H_0$ : all  $p_i$  are equal.

*Comments.* 1. A one-tail test is appropriate on the grounds that there is no medical or other evidence to suggest a beneficial effect of increased alcohol uptake in this context. The situation may be different for other health issues. There is some evidence, for example, that moderate consumption of red wine may protect against the onset of heart disease.

**Table 10.9** Numbers of children with or without malformations in relation to mother's stated alcohol consumption (Graubard and Korn, 1987).

Mother's alcohol consumption (drinks per day)	Malformation	
	Absent	Present
0	17066	48
<1	14464	38
1–2	788	5
3–5	126	1
≥6	37	1

2. The choice of row scores is arbitrary. While there may be a temptation to use mid-rank scores for rows as an intuitively reasonable scoring system the disadvantage noted in Comment 2 on Example 10.7 is more marked in this example. It is easily verified that the mid-rank scores for rows are 8557.5, 24365.5, 32013, 32473 and 32555.5. The final three rows (where most of the changes in malformation probability, if any, would seem to take place) get almost identical scores. For this scoring system  $P = 0.2861$ .

3. Temptation to experiment with many scoring systems should be avoided. When no other system is obviously appropriate the default scoring system 1, 2, 3, 4 ... has much to commend it. However, if there are clinical grounds for adopting some other scoring system it is sensible to do this. For this example Graubard and Korn suggest using the mean alcohol intake for row scores, i.e. 0, 0.5, 1.5, 4 and 7 where the choice of 7 is arbitrary but not unreasonable. This results in an exact one-tail  $P = 0.0168$ , suggesting that there is indeed some evidence against  $H_0$ . Sprent (1998, Section 13.5) suggests yet another scoring system that might well get clinical support. It gives a  $P$ -value not very different from that using the mean intake scores and this is comforting because it suggests that the analysis is not too sensitive to differences between scoring systems that may appear realistic under what most people would regard as rational assumptions.

Computational aspects. Availability of programs like that in StatXact for exact linear-by-linear association tests with arbitrary scores adds flexibility to the Cochran-Armitage test as originally proposed with the restricted row scores 1,2,3,4,...

### 10.2.7 Capture-recapture techniques

Capture-recapture analysis was originally proposed to estimate the number of animals in a population when it was impracticable or even impossible to do this using a complete count or census. For example, to estimate the total number of squirrels in a wood one might use a benign form of trapping to capture 25 animals which are then marked for identification and released to the wild. At a later date a further 40 animals might be trapped and these might include 5 marked, i.e. recaptured, animals, the remaining 35 being unmarked. If the unknown number of animals in the wood is  $N$  the first sample marking procedure may be looked upon as dividing the population into two sub-populations - the first consisting of the 25 marked animals and the second of the  $N - 25$  unmarked animals. If all animals have the same probability of being included in the first and second trapping then the ratio of the number of marked animals to unmarked animals in the second sample is an intuitively reasonable estimate of the proportions in the population. Equating these ratios, i.e. setting  $25/(N - 25) = 5/35$  and solving gives  $N = 200$ . It is easily verified that this is equivalent to using the ratio of marked to total

populations, i.e.  $251N = 5/40$ . More generally if the first or capture sample size is  $n_1$  and the second or recapture sample size is  $n_2$  and in this second sample there are  $b$  marked animals then we estimate  $N$  by equating the ratios  $n_1/N = b/n_2$  giving the estimate  $N = n_1n_2/b$ .

The assumptions that all animals have an equal probability of being included in each sample and that the population size  $N$  remains constant between the capture and the recapture sample are crucial but may not hold in practice. Some animals may be more inquisitive than others and thus more likely to be trapped; sometimes the probability of a marked animal being retrapped may be decreased by some aspect of the marking process such as being frightened by the initial trapping or marking procedure or increased because the bait used in the trap was a desirable food item. These factors negate the equal probability assumption. Further, the population size may vary between samples due to births, deaths or migration. We show below that when the model leading to the simple estimate given above is valid approximate confidence limits for the true population size may be obtained by formulating the problem using an incomplete  $2 \times 2$  contingency table. When the basic assumptions needed for validity of the simple estimate break down it is well known to users of capture-recapture methods that some of these can be dealt with by taking more than two samples. This leads to a multiple capture-recapture process where in each succeeding sample one identifies which animals have been captured previously and at which sample or samples this happened. Analysis of such data is often based on loglinear models to describe interactions or associations. A detailed treatment is beyond the scope of this elementary text but we outline the basic ideas behind a contingency table approach. We see in Example 10.10 and the subsequent discussion that the basic ideas of capture-recapture analysis have been taken beyond the estimation of animal populations and are now employed in wider contexts. The use of loglinear models for capture-recapture analysis is described by Bishop, Fienberg and Holland (1975) while Seber (1982) provides a comprehensive review of basic capture-recapture procedures. Cormack (1989) gives a detailed description of practical uses of log-linear models in this context.

For the two sample procedure outlined at the start of this section we may express results in a  $2 \times 2$  table like [Table 10.10](#) where rows represent numbers from the total population of unknown size  $N$  that are included or are not included in the first sample, and columns represent similar information for the second sample. It is easily seen

**Table 10.10** A contingency table for a simple capture-recapture model. Interrogation marks indicate an entry to be estimated.

		<i>Present in second sample</i>		<i>Total</i>
		<i>Yes</i>	<i>No</i>	
<i>Present in first sample</i>	<i>Yes</i>	<i>b</i>	$n_1 - b$	$n_1$
	<i>No</i>	$n_2 - b$	?	?
<i>Total</i>		$n_2$	?	?

that in the standard notation  $n_{ij}$  for the count in cell  $(i, j)$  relevant numerical entries in Table 10.10 are  $n_{11} = b$ ,  $n_{12} = n_1 - b$ ,  $n_{21} = n_2 - b$  and the first row and column totals are respectively  $n_1$ ,  $n_2$ . In any particular application these all have known numerical values. However, we do not know  $n_{22}$ , which in our example is the number of squirrels that are not observed in either sample. This means we do not know the second row and column marginal totals nor do we know the grand total,  $N$ . The unobservable number  $n_{22}$  is often called a structural zero. If we denote the unknown population number by  $N$  an intuitively reasonable estimate of  $N$  is obtained by choosing it so that the observed number in each cell  $n_{ij}$  equals the expected number  $m_{ij} = n_{i+}n_{+j}/N$  under the assumption of independence.

Remember that in a  $2 \times 2$  table knowing  $m_{11}$  fixes all other  $m_{ij}$  to ensure correct marginal totals. We know  $n_{11} = b$  and from Table 10.10 it follows for any unknown total  $N$  that  $m_{11} = n_1 n_2 / N$ . Setting  $m_{11} = n_{11} = b$  implies an estimate  $\hat{N} = n_1 n_2 / b$ , the same as we got at the start of this section. This may look like using a sledgehammer to crack a nut, but the approach opens up useful ways to obtain a confidence interval for  $N$  and also has implications for developing loglinear models allowing for interactions when applied to more than two samples. Sekar and Deming (1949) derived an expression for the asymptotic variance of  $\hat{N}$ , namely

$$\text{Var}(\hat{N}) = \frac{n_{1+}n_{+1}n_{12}n_{21}}{n_{11}^3} \quad (10.22)$$

and asymptotic 95 per cent confidence limits for  $N$  are given by

$$\hat{N} \pm 1.96\sqrt{\text{Var}(\hat{N})}$$

For populations not much bigger than the combined sample sizes these asymptotic limits are sometimes unsatisfactory, giving a lower limit less than the total count of distinct units in the two samples.

Regal and Hook (1984) suggest an alternative approach that uses the likelihood ratio statistic  $G^2$ . If the capture-recapture estimate of the total number of cases is used in the  $2 \times 2$  table,  $G^2$  is zero, since the method effectively equates observed to expected numbers under independence. They suggested forming a 95 per cent confidence interval that included all  $N$  for which the fit of the independence model is adequate in a significance test at the 5 per cent level. As  $G^2$  is associated with a chi-squared distribution with one degree of freedom, values of  $N$  giving a  $G^2 < 3.84$  form the 95 per cent confidence interval. Unlike the interval calculated from the standard error, this gives a confidence interval that is asymmetric around  $\hat{N}$  with the lower limit being closer to  $\hat{N}$  (and necessarily no less than the total observed distinct number of population members).

Capture-recapture methods have a valuable role to play in the needs assessment of the most marginalized members of society such as the homeless (Fisher et al., 1994) and drug abusers (Hay and McKeganey, 1996). In addition they can, as the following example shows, be used to correct for underestimation of the prevalence of a medical problem in human populations when several incomplete registers (each corresponding to a capture or a recapture sample of patients) exist (Hook and Regal, 1982; Smeeton et al., 1999).

### *Example 10.10*

*The problem.* Guillain-Barré syndrome is a paralyzing neurological condition due to inflammation of the peripheral nerves. In the UK patients are listed on various registers, none of which is complete. Rees et al. (1998) estimated the number of such cases in the south east of England. Two lists were obtained, one from the British Neurological Surveillance Unit (BNSU) and another from hospital activity analysis (HAA); 23 patients were on the BNSU list, 68 were on the HAA list and 17 were on both. Find an estimate of the total number of cases in the south east of England, along with a 95 per cent confidence interval.

*Formulation and assumptions.* The relevant data can be presented in a  $2 \times 2$  table in the format of [Table 10.10](#):

		<i>Present in HAA list</i>		<i>Total</i>
		<i>Yes</i>	<i>No</i>	
<i>Present in BNSU list</i>	<i>Yes</i>	17	6	23
	<i>No</i>	51	?	?
	<i>Total</i>	68	?	?

As there are only two registers, interactions cannot be investigated; the two lists are assumed to be independent. We have  $n_{+1} = 68$ ,  $n_{1+} = 23$  and  $n_{11} = 17$ . The appropriate estimate  $\hat{N}$  of  $N$  is  $\hat{N} = 68 \times 23/17 = 92$ .

The asymptotic variance of this estimate given by (10.22) is

$$\text{Var}(\hat{N}) = (23 \times 68 \times 51 \times 6)/17^3 = 97.41.$$

This leads to 95 per cent confidence limits for  $N$  of  $92 \pm 1.96\sqrt{97.41}$  which lead after integer rounding towards the estimate  $\hat{N}$ , to the interval (73, 111). This interval is clearly unacceptable since we know from the lists there are at least  $69 + 23 - 17 = 74$  cases. Using the likelihood ratio statistic  $G^2$  gives a lower limit for  $N$  of 79 and an upper limit of 121. These limits do not include the total number of known cases and the corresponding interval is therefore more appropriate than the earlier one.

*Conclusion.* The estimated total number of cases is 92. On the basis of the approximate 95 per cent confidence interval the likely number of 'hidden' cases of Guillain-Barré syndrome in this area of England is between  $79 - 74 = 5$  and  $121 - 74 = 47$ .

*Comments* 1. With small samples, caution is needed when using the asymptotic standard error method to obtain a 95 per cent confidence interval for  $N$ . The lower limit of the interval should always be checked against the number of known cases to see if it makes sense.

2. Confidence intervals for small samples will almost certainly be wide and in this example may not shed much light on the number of cases missed by both lists. The upper limit of the confidence interval may exceed any reasonable estimate expected by individuals working in the field of study.

3. Additional registers are invaluable for improving the estimate of  $N$ . Using these extra 'samples', log-linear models allowing interaction terms can be investigated. Rees et al. had access to two further registers - a research database containing 22 cases and 5 death certificates giving this syndrome as a cause of death. This brought the total number of known cases to 79 with an estimate for  $N$  of 98. The new likelihood ratio 95 per cent confidence interval for  $N$  was (86, 120). As one would expect, further registers increased the number of known cases. Note that the 95 per cent confidence interval is less wide; the upper limit for  $N$  is slightly smaller. An interaction term was required for an acceptable model; this casts doubt on the unavoidable assumption of independence in our example.

4. A problem arises if the count  $b$  of numbers in both samples is zero, for the estimate of  $N$  using the above approach is infinite. Adjustments are available in these circumstances but it is usually better, when possible, to overcome the problem by taking larger samples.

*Computational aspects.* Nearly all programs that provide the chi-squared test for  $2 \times 2$  tables quickly give the likelihood ratio statistic  $G^2$  and these at least provide a basis for a trial and error process for forming a confidence interval based on this statistic. An improved interval might be based on that of the Fisher exact test.

The contingency table approach generalizes to three or more samples in a multiple capture-recapture program. For three samples

the outcomes may be presented in an incomplete  $2 \times 2 \times 2$  table where the three dimensions refer to the respective samples and the categories in each are presence or absence. We write the observed cell count in cell  $(i, j, k)$  as  $n_{ijk}$  where a subscript value 1 implies a population unit is observed in the corresponding sample and a subscript 2 that it is not observed. Thus, for example,  $n_{121}$  is the number of animals, patients, or whatever observed in samples 1 and 3 but not in sample 2 while  $n_{111}$  is the number observed in all three samples. Because we do not know  $N$  there is one unobservable or structural zero cell count corresponding to the unknown  $n_{222}$ , the number in the population that are never seen.

The expected numbers  $m_{ijk}$  in any cell other than the one with a structural zero can be modelled by a log-linear model that allows for possible interactions due to factors such as varying probabilities of capture for individual animals within or between samples, population changes due to births, deaths, immigration or emigration, etc. The detail is beyond the scope of this book and the interested reader should refer to Cormack (1989). In the rest of this section we make some general remarks about the method that indicate the versatility of the approach.

The modelling process differs from that considered in Section 10.2.1 due to the lack of a specific count in cell  $(2, 2, 2)$ . This means that in the three-sample case we can no longer make inferences about or test for a second-order interaction. We can however test hypotheses about first-order interactions. If samples are taken on  $k$  occasions the results can be presented in a  $2^k$  contingency table with one structural zero cell and inferences may be made about interactions up to order  $k - 2$ .

The loglinear models appropriate to such analyses are a special case of an important class of models known as generalized linear models. How well a model fits the data can be examined using the likelihood ratio test statistic  $G^2$ , which is often referred to as the deviance. For the independence model with  $k$  samples,  $G^2$  is associated with a chi-squared distribution having  $2^k - k - 2$  degrees of freedom. If the  $P$ -value from this goodness-of-fit test is low, the model has an inadequate fit to the data so a more complex model is sought. If an extra interaction term is added to the model, the number of degrees of freedom is reduced by one. To create models that are as simple as possible, it is usual to accept the extra term only if the decrease in the deviance is statistically significant ( $P < 0.05$ ); in other words the deviance has to decrease by at least 3.84. If a

model provides a very poor fit this is invariably the case. As extra terms are added the *P*-value rises (the fit is better).

The estimate for  $N$  should be given with an appropriate confidence interval. Asymptotic formulae for the calculation of confidence intervals in the case of multiple samples have been developed (Bishop, Fienberg and Holland, 1975; Selvin, 1995). However, as with the basic two-samples-only situation, these may lead to confidence intervals that contain values less than the number of known cases. To avoid this problem, Regal and Hook (1984) suggest the calculation of  $G^2_{\min}$ , the minimum value of the likelihood ratio statistic for the loglinear model under consideration. This is obtained when using the capture-recapture estimate for  $N$ . Using similar reasoning to the two-sample case, the appropriate confidence interval for that particular model contains all those values of  $N$  for which  $G^2$  is no more than  $G^2_{\min} + 3.84$ .

In applying capture-recapture methods to real data, it is sometimes necessary to introduce a subjective element in the model selection process. A more complex model is sometimes preferred if this allows the *P*-value of the overall fit to rise above 0.05, even if the decrease in the deviance is modest. One should bear in mind, however, that additional terms in a model may lead to wider confidence intervals for  $N$ . A more complex model may need to be rejected even if there is a statistically significant decrease in the deviance if the confidence interval contains implausible values from an ecological or clinical point of view (Hay, 1997).

Smeeton et al. (1999) used capture-recapture methods in order to estimate the prevalence of congenital malformation of the heart in the first year of life. All such infants are severely ill and doctors would in the light of experience question a predicted rate of more than 20 cases per 1000. In a practical study, a model with three interaction terms gave an estimate of 14.5 cases per 1000 and a 95 per cent confidence interval for the rate from 5.7 to 79.9 cases per 1000. Such findings would not be taken seriously by the medical profession. The removal of one of the interaction terms reduced the estimate to 6.8 cases per 1000 and the upper limit of the confidence interval to a much more realistic rate of 13.3 cases per 1000.

A further discussion of interval estimation of population size using overlapping lists or records is given by Regal and Hook (1999).

Capture-recapture analysis of constant populations can be performed with SPSS which allows definition of the missing cell value referred to as a structural zero. The estimate for  $N$  is calculated from the coefficients of the loglinear model. If  $N$  varies and birth,

death and migration need to be taken into account the package GLIM can be used in the way described in Cormack (1989).

## 10.3 COMBINING AND PARTITIONING OF TABLES

Intuitively one may feel that if several tables refer to the same row and column categories then combining these into one table to get larger counts in cells may be helpful, for example, in increasing the power of tests for detecting association. In other circumstances one may feel that some breaking down or partitioning of tables, perhaps by combining certain rows and columns or breaking down into strata may reveal information not easily discernable in the original table. Combining or partitioning should be done with caution. Detailed discussions are given by Agresti (1990, 1996). The treatment here covers only certain aspects of these topics to give some indication of when they can be useful and points out some of the pitfalls that may go with incautious use of the methods.

### 10.3.1 Simpson's paradox and combining tables

Suppose we are given the information in [Table 10.11](#) for numbers of individuals in urban and rural areas responding to a standard cough medicine and a new medicine where for each person it is recorded whether there is no effect or it cures the cough.

Visual inspection of Table 10.11 shows that the proportion of cures in both urban and rural areas is higher with the new medicine than it is with the standard. This is confirmed if we carry out the Fisher exact test on each table separately giving  $P < 0.0001$  for urban areas and  $P = 0.0017$  for rural areas, both providing strong evidence against a hypothesis that the medicines are equally effective. If we now ignore the split into urban and rural areas and combine the data for each giving the  $2 \times 2$  table

	No effect	Cure
Standard medicine	837	444
New medicine	1167	557

and apply the Fisher exact test to this table we find  $P = 0.0946$ . So much for a hope that combining the information would enhance the evidence of association. Indeed if we look at this table we see that the cure rates do not look very different for the two medicines, roughly a third of all those treated being cured, irrespective of which

**Table 10.11** Responses to two cough medicines in urban and rural areas.

	Urban		Rural	
	No effect	Cure	No effect	Cure
Standard medicine	498	103	339	341
New medicine	1042	367	125	190

medicine they are given. If anything, the standard medicine appears to do slightly better! This situation where the individual tables indicate association (and here association in the same direction) but the combined table does not indicate association (or even suggests that any association might be in the opposite direction) illustrates Simpson's Paradox, first described formally by Simpson (1951). This paradox may arise when we combine data for two strata where different models hold, especially when, as here, the sample sizes in each stratum are different. In the above example the models for urban and rural areas are clearly different because the proportion of cures for each medicine in these areas are markedly different; indeed the percentages in each calculated from [Table 10.11](#) are

	Urban	Rural
Standard medicine	17.14	50.15
New medicine	26.05	60.32

Agresti (1990, Section 5.2.2) gives an excellent real data example of this paradox. The moral is that one should only combine tables if one is confident the relevant identical distributional model is valid for each table. It is probably only worth doing this if counts are small in corresponding cells in all tables that are to be combined.

### 10.3.2 Partitioning of tables

We saw in [Sections 9.2.2](#) and [9.2.3](#) that both the Pearson  $X^2$  statistic and the likelihood ratio  $G^2$  statistic have asymptotically a chi-squared distribution with  $(r - 1)(c - 1)$  degrees of freedom. It is well known that a variate having a chi-squared distribution with  $v$  degrees of freedom may be expressed as a sum of  $v$  independent variates each with 1 degree of freedom. The  $G^2$  statistic may be partitioned similarly into components, each of which has asymptotically a chi-squared distribution with 1 degree of freedom. Rules for partitioning are complicated in the general  $r \times c$  case and

there is usually more than one possible partitioning. Necessary conditions for the single degree of freedom components to be additive are given by Agresti (1990, Sections 3.3.5-7) and are:

- Subtable degrees of freedom must sum to the degrees of freedom of the original table.
- Each cell count in the original table must appear in one and only one subtable.
- Each marginal total of the original table must be a marginal total for one and only one subtable.

Lancaster (1949) suggested partitioning an  $r \times c$  table into a total of  $(r - 1)(c - 1)$  tables each  $2 \times 2$  with 1 degree of freedom having the form

$$\begin{array}{cc} \sum_{s < i} \sum_{t < j} n_{st} & \sum_{s < i} n_{sj} \\ \sum_{t < j} n_{it} & n_{ij} \end{array}$$

For example, for a  $3 \times 3$  table

$$\begin{array}{ccc} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ n_{31} & n_{32} & n_{33} \end{array}$$

this partitioning gives the four tables

$$\begin{array}{lll} \text{(i)} & n_{11} & n_{12} \\ & n_{21} & n_{22} \end{array} \quad \begin{array}{lll} \text{(ii)} & n_{11} + n_{21} & n_{12} + n_{22} \\ & n_{31} & n_{32} \end{array} \quad \begin{array}{lll} \text{(iii)} & n_{11} + n_{12} & n_{13} \\ & n_{21} + n_{22} & n_{23} \end{array}$$
  

$$\begin{array}{lll} \text{(iv)} & n_{11} + n_{12} + n_{21} + n_{22} & n_{13} + n_{23} \\ & n_{31} + n_{32} & n_{33} \end{array}$$

It is easy to verify (Exercise 10.14) that this partitioning satisfies the above necessary conditions. The partitioning is not unique. For example, if rows and columns of the original table are permuted the chi-squared statistic for the whole table is unchanged but clearly the partitioning above will be altered numerically. Although under the null hypothesis of independence asymptotically the chi-squared statistics for subtables are independent and additive the Pearson  $X^2$  statistics computed for each of the 4 tables (i) - (iv) above do not in general add to the corresponding statistic for the original table. However, if the likelihood ratio statistic  $G^2$  is used the components are additive.

Partial partitioning where not all components correspond to single degrees of freedom is also possible. In practice one should choose a

partitioning that helps to answer relevant questions about association.

### Example 10.11

*The problem.* Two drugs used in chemotherapy are tested on 100 patients, 60 receiving drug A and 40 receiving drug B. Numbers treated with each drug who exhibit the presence or absence of two specific side-effects, hair loss (HL) and visual impairment (VI) are as follows:

		Side-effect status			
		HL	VI	HL + VI	None
Drug A	9	4	16	31	
	3	16	2	19	

Partition  $G^2$  to examine association between single side-effects, between single and double side-effects and between overall side-effect status for the two drugs.

*Formulation and assumptions.* We may show that the necessary conditions for additivity given above are satisfied if we compare column 1 with column 2, then columns 1 + 2 with column 3, then columns 1 + 2 + 3 with column 4. These components represent appropriate tests for association (i) between single side-effects for each drug, (ii) between a single side-effect (either HL or VI) and a double side-effect (both HL and VI) and (iii) at least one side-effect and no side-effect. Tests for each may be carried out using the  $G^2$  statistic in either an exact or asymptotic test if the latter is felt appropriate.

*Procedure.* The relevant partitioned  $2 \times 2$  tables are

(i)	HL	VI	(ii)	HL or VI	HL + VI
Drug A	9	4	Drug A	13	16
Drug B	3	16	Drug B	19	2
			Side-effect	None	
			Drug A	29	31
			Drug B	21	19

It is easily verified (Exercise 10.15) that for the complete table  $G^2 = 22.13$  and for the partitioned  $2 \times 2$  tables that for (i)  $G^2 = 9.72$ , for (ii)  $G^2 = 12.24$  and for (iii)  $G^2 = 0.17$ . The sum of these components is  $9.72 + 12.24 + 0.17 = 22.13$ , the  $G^2$  value for the complete  $2 \times 4$  table. StatXact may be used to compute both exact and asymptotic  $P$ -values for the original table and each of the  $2 \times 2$  components and gives the following values:

$2 \times 4$ table	(i)	(ii)	(iii)
Exact $P$	0.0001	0.0035	0.0010
Asymptotic $P$	0.0001	0.0018	0.0005

*Conclusion.* The extremely low  $P$ -value for the full table indicates strong evidence of association between drugs and side-effect status. The low  $P$  for subtable (i) reflects the fact that if there is only one side-effect then it is more likely to be hair loss for drug A and visual impairment for drug B. The low  $P$  for subtable (ii) reflects the fact that occurrence of both side-effects is more common with Drug A, while the high  $P$  for subtable (iii) shows there is no evidence of association between drugs and the incidence of some side-effect, i.e. that both drugs seem equally likely to give rise to some side-effects.

*Comments.* 1. The action a clinician might take in the light of these findings depends upon the importance of each side-effect. If visual impairment were slight and temporary it might be rated less serious than hair loss and this would indicate a preference for drug B. If visual impairment were severe and permanent it would be more serious than hair loss, but again if both are moderate two side-effects may be regarded as more serious than just one.

2. The above partitioning is not unique; we might reorder the columns, e.g. place column 4 first and then combine columns in the order used above. The order we have used leads to more logical interpretation of the components of  $G^2$ .

3. We have used a particular case of a rule for partitioning  $G^2$  into  $c - 1$  additive components each with 1 degree of freedom when we have a  $2 \times c$  table. The rule is to first form the subtable with columns 1 and 2, next the subtable with column  $1 + 2$  as the first column and column 3 as the second column, next the subtable with the sum of columns 1, 2, 3 and column 4, and so on until the final table uses the sum of columns 1, 2, 3, ...,  $c - 1$  and column  $c$ .

4. Although the Pearson  $X^2$  statistic and  $G^2$  have the same asymptotic distributions under  $H_0$ , if  $X^2$  is partitioned as above the component values of the Pearson statistic will not in general sum to  $X^2$ .

Best (1994) used a partitioning of chi-squared into components analogous to linear and quadratic components in a regression type analysis of variance. He was specifically interested in looking at centrality and dispersion differences for data obtained in tasting trials, but appropriate use of this partitioning extends to other types of data and it can also be generalized to look at other features such as skewness. Best's approach uses a partitioning proposed by Lancaster (1953) and an account together with a comparison with some alternatives is given by Sprent (1998, Section 13.2).

### 10.3.3 Partitioning in a different kind of contingency table

To compare preferences for competing brands of a product like tomato soup, or for similar types of wine made from the same grape in different countries, or for different varieties of peas or beans, a widely used procedure is to ask a number of potential consumers,  $N$ , to rank each of  $k$  brands of the product from 1 to  $k$  in order of preference. In practice the ranking is usually in descending order of preference, the

**Table 10.12** Numbers of potential purchasers giving various rankings to each of three models of car.

Model	Rank			Total
	1	2	3	
I	27	47	16	90
II	35	31	24	90
III	28	12	50	90
Total	90	90	90	270

brand ranked 1 being the first choice. The results might be expressed in a  $k \times N$  table like Table 7.7. If there are no tied rankings the results may also be expressed in a  $k \times k$  table like Table 10.12 with each row corresponding to one of the  $k$  brands and each column to a rank, the  $i$ th column corresponding to rank  $i$ . Here the rows have no natural ordering but the columns are ordered. The entry in cell  $(i, j)$  is the number among the  $N$  tasters who gave rank  $j$  to brand  $i$ . A classic paper on this topic by Anderson (1959) gave data for a study in which 123 consumers were each asked to rank three varieties of snap beans in order of preference. Table 10.12 is a similar table for rankings by 90 potential purchasers for three models of broadly similar cars.

The basic difference between this contingency table and those considered earlier in this chapter is that the cell entries are not sampled independently from 270 units. Instead, each of the 90 would-be purchasers independent of one another allocates the ranks 1, 2, 3 to each of the three models. Since each purchaser uses each rank he or she is represented once in each column. This means one may validly carry out a goodness-of-fit test to see, for example, if the hypothesis that there is the same probability of each variety being ranked 1 is supported by considering only the data in column 1, because what happens in the allocations in the remaining columns is irrelevant to that question. The appropriate test is one for a uniform distribution using the methods of Section 9.4 leading to an  $X^2$  statistic that has asymptotically a chi-squared distribution with 2 degrees of freedom. If there is no preferred first choice the expected numbers in column 1 are each  $90/3 = 30$ . A similar argument could be applied to column 2 or column 3 but clearly these tests would not be independent of one another for if they were we could add the components and get a chi-squared with 6 degrees of freedom for the whole table, whereas with fixed marginal totals there are only 4 degrees of freedom.

associated with [Table 10.12](#). A moment's reflection will show that if we performed a conventional Pearson  $X^2$  test with these fixed marginal totals the value of  $X^2$  (with 4 degrees of freedom) would be exactly the same as that obtained by adding the goodness-of-fit statistics for the no-preference goodness-of-fit test applied to each row with 6 associated degrees of freedom. Anderson conjectured that if we reduced the usual  $X^2$  statistic by a factor  $2/3$  (i.e. the ratio 4 to 6 of the degrees of freedom) we might get a statistic that had asymptotically a chi-squared distribution with 4 degrees of freedom under a hypothesis of independence. He proved this was so and that the result extended to the case of  $k$  items to be ranked where we have a  $k \times k$  table where our conventional Pearson  $X^2$  statistic is reduced by a factor  $(k-1)/k$ . The modified statistic is  $X'^2 = (k-1)X^2/k$ . For the data in Table 10.12 we find  $X^2 = 42.8$ . Since  $k = 3$  the chi-squared statistic for the overall test is  $X'^2 = 2 \times 42.8/3 = 28.53$  with 4 degrees of freedom. It is easily verified that  $P < 0.0001$  indicating, not surprisingly in view of the data pattern, strong evidence of association between models and purchaser preferences.

Anderson discusses partitioning  $X'^2$  into single additive degrees of freedom but as in the previous section these are not unique and the important thing is to choose a partitioning that has a sensible interpretation. For a  $3 \times 3$  table one partitioning into contrasts proposed by Anderson and the interpretation of each contrast is:

$C_1: n_{21} - n_{23}$	Linear ranking effect for model II.
$C_2: n_{11} - n_{13} - (n_{31} - n_{33})$	Difference between linear rankings for models I and III.
$C_3: n_{21} - 2n_{22} + n_{23}$	Quadratic ranking effect for model II.
$C_4: n_{32} - n_{12}$	Difference between quadratic rankings for models I and III.

Interpretation of  $C_4$  is clearer if we write it in an alternative form:

$$[(n_{11} - 2n_{12} + n_{13}) - (n_{31} - 2n_{32} + n_{33})]/3.$$

Equivalence is established by noting that  $n_{i1} + n_{i2} + n_{i3} = N$ , ( $i = 1, 3$ ).

Anderson also establishes that each single degree of freedom component of  $X'^2$  takes the form  $C_i^2/\text{Var}(C_i)$ . In determining  $\text{Var}(C_i)$  one must take into account correlations between the various  $n_{ij}$  arising because the same observers are involved in all counts. The calculations of the  $\text{Var}(C_i)$  are explained in Anderson's original paper and outlined in Sprent (1998, Section 13.3) and we only quote the result, viz.  $\text{Var}(C_1) = \text{Var}(C_4) = 2N/3$  and  $\text{Var}(C_2) = \text{Var}(C_3) = 2N$ .

An interesting feature of these contrasts is that if we add the components  $X^2$  for the first two contrasts their sum is equal to the Friedman statistic T given by (6.5) based on the rank sums for each variety given by all 90 purchasers, equivalent to Kendall's coefficient of concordance. We illustrate these points in an example.

### *Example 10.12*

*The problem.* For the data in Table 10.12 partition the statistic  $X^2$  into components corresponding to the contrasts  $C_1$  to  $C_4$  described above and verify the assertion that addition of the first two components gives the Friedman statistic T in this case.

*Formulation and assumptions.* The necessary computations are based on the description of components given above.

*Procedure.* We see immediately from Table 10.12 that  $C_1 = 35 - 24 = 11$  and since  $\text{var}(C_1) = 2N/3$  and  $N = 90$  it follows that the corresponding component of  $X^2$  is  $11^2/[2 \times 90/3] = 2.02$  and in like manner the remaining components are found to be 6.05, 0.05 and 20.42 summing to  $X^2 = 28.54$ , the value quoted above apart from rounding error. To calculate the Friedman statistic for these data we note that the sum of the ranks given by all 90 purchasers for model I is  $27 + 2 \times 47 + 3 \times 16 = 169$  and similarly the sums for models II and III are 169 and 202 whence the Friedman statistic is easily found to be  $T = 8.07$ , equal to the sum  $2.02 + 6.05 = 8.07$  of the first two components above of  $X^2$ .

*Conclusion.* Clearly if an asymptotic chi-squared test is applied to each component there is strong evidence of association between models and consumer preference for components  $C_2$  and  $C_4$ , but not for  $C_1$  or  $C_3$ .

*Comments.* 1. Components  $C_3$  and  $C_4$  are associated with departures from linearity of ranking. Linearity broadly speaking represents average rankings and departures therefrom represent differences in dispersion. For the data in Table 10.12 it is clear that there is a monotonic fall-off for rank preferences for model II, but that the pattern is different for models I and III. Model I is most frequently ranked 2, and least frequently ranked 3. On the other hand model III is seldom ranked 2; a large number rank it 3, but it gets one more first ranking than does model I. Thus model III shows a greater dispersion of ranks than do the other models.

2. Anderson (1959) suggests that in many practical situations the experimenter may be more interested in model or varietal contrasts at the more favourable ranks, 1 and 2, in which case a more useful set of contrasts would be

$$D_1: n_{31} - n_{11}, \quad D_2: n_{31} - 2n_{21} + n_{11}, \quad D_3: n_{12} - n_{13} - (n_{32} - n_{33}), \quad D_4: n_{23} - n_{22}.$$

The components  $D_1$  and  $D_2$  are linear and quadratic comparisons for the first rankings (of main interest in this approach) while the remaining two examine differences in the second and third rankings.

Problems similar to those briefly discussed in this section have been considered by Scholz and Stephens (1987), Best (1994), Best and Rayner (1996) and others.

## 10.4 POWER

Increasing attention has been given in recent years to power and sample size relationship for tests involving contingency tables. Exact power computations in practice are limited to fairly restricted and clearly defined models such as the comparison of binomial populations in  $2 \times k$  (or  $k \times 2$ ) tables, particularly when there is ordering in the  $k$  categories. The subject is a complex one. An account of the practical difficulties involved together with examples of applications is given by Mehta and Patel (1999, Chapter 25).

We showed in [Section 5.7.1](#) where we considered power calculations for the median test that exact power calculations were possible if we could determine the binomial probabilities relevant under  $H_0$  and a specific alternative of interest. The latter binomial probability was in that context dependent upon what assumptions we made about the population distributions giving rise to our samples. However, exact power computations are feasible once that binomial probability,  $p_1$ , is ascertained, the procedure then being one for determining power of the Fisher exact test of  $H_0: p = \frac{1}{2}$  against  $H_1: p = p_1$ . The procedure generalizes easily to any  $p_0$  specified in  $H_0$ .

StatXact (version 4.0) includes a facility for exact power calculations for  $2 \times k$  tables where the columns correspond to  $k$  binomial samples or the rows correspond to two multinomial samples each with  $k$  possible outcomes. These cover many potential situations of practical importance providing we can specify meaningful binomial or multinomial probabilities for the alternative hypothesis of interest. When we can the programs may be applied to tests such as the Cochran-Armitage test and the extensions to it with arbitrary scores given in [Section 10.2.6](#) as well as to two-independent sample permutation tests including WMW, normal scores, logrank scores or even the Pitman raw data test. The reader should refer to the StatXact manual for technical details and to see examples of what can be achieved.

An example of how relevant binomial probabilities may be assigned for power calculations for a Cochran-Armitage test in a clinical trial context using a logistic regression model is given in Mehta, Patel and Senchaudhuri (1998).

## 10.5 FIELDS OF APPLICATION

Tests of association may be relevant in any situation where independence may not be an acceptable hypothesis. In three-way or

higher dimensional tables especially, the nature of association is often of special interest.

## **Drug testing**

Situations like that in Example 10.7 where side-effects of drugs are of interest are common in clinical trials. In such cases the nature of association between dose level and side-effects is often the main interest in the study. An appropriate linear-by-linear association model with suitable choice of row and column scores adequately describes many such associations.

## **Medicine**

In medical research it is often felt that physiological abnormalities may produce undesirable responses and that the seriousness of these may increase with the severity of the disorder. An example of such a trend occurred with the spleen size/blood platelet count data considered in Example 9.8. Similarly, responses to environmental factors such as different levels of a known carcinogen nearly always show an ordered response and again linear-by-linear association models may be appropriate.

## **Administration of justice**

There has been considerable interest in recent years in differences in criminal trends and in the way the courts treat offenders in different ethnic groups. Agresti (1984, p. 32) considers the analysis of counts of death penalty verdicts based on ethnic grouping of victims and defendants using data given by Radelet (1981). In civil proceedings relating to redundancy there have often been claims that company policy on laying off staff differs between age groups or between ethnic groups, or perhaps gender, even when staff are performing similar functions. Many examples illustrating the use of nonparametric methods in a legal context are given by Gastwirth (1988).

## **Sociology**

Many studies have been made of association between socio-economic background and educational or career achievement and between that background and attitudes towards social problems. Loglinear models are often relevant to studies of such associations.

## Ornithology

In order to obtain information on the number of birds of a particular species in a defined area, samples of birds can be caught, ringed and released on several occasions. A study of the overlap between the samples can be used to obtain capture-recapture estimates of the population size.

### Consumer preferences

While the Kendall coefficient of concordance is adequate for a study of many linear (averaging) aspects of consumer preferences more sophisticated aspects such as dispersion often require an approach based on partitioning into contrasts that give rise to additive components of the  $X^2$  or  $G^2$  statistics. For example, one might be interested in ascertaining whether the degree of sweetness that consumers of soft drinks prefer shows different patterns in different countries, or whether men and women show different preferences.

## 10.6 SUMMARY

The loglinear model ([Section 10.2.1](#)) is used for the study of many aspects of association. It has close parallels to the linear model for treatment comparisons such as those arising in the analysis of factorial treatment structures in the analysis of variance. A no-interaction model corresponds to independence and a model with interactions to patterns of association. These can be applied to capture-recapture analysis ([Section 10.2.7](#)) in which one of the cell values is missing (the unobserved cases).

Sets of  $k$  tables each  $2 \times 2$  (Sections 10.2.2 to 10.2.5) are often associated with comparisons of binomial responses to each of two treatments at each of  $k$  levels of an explanatory variable (e.g. different age groups or different sources of a raw material). This explanatory variable is often called a covariate. Both asymptotic and exact test procedures are available to test whether the odds ratios may be supposed equal for all  $2 \times 2$  tables and if they are whether or not they all take the common value unity. The latter implies independence. Equality at a value other than unity implies a first-order interaction.

The linear-by-linear association model ([Section 10.2.6](#)) is a log-linear model of special importance in nonparametric methods as many well-known tests may be formulated as special cases and may also be extended by modification of the scoring system used in the original test. The Cochran-Armitage test for monotonic trends provides a good example of the possibility for such modifications in binomial probability models with the parameter  $p$  changing monotonically with covariate values.

Combining and partitioning of contingency tables ([Section 10.3](#)) is often a useful tool. The former must be used with caution as is illustrated by Simpson's paradox ([Section 10.3.1](#)). Partitioning into subtables each corresponding to a single degree of freedom in a statistic having an asymptotic chi-squared distribution is often useful in elucidating the structure of associations but most partitionings are not unique and certain necessary conditions must be met to ensure additivity of components ([Section 10.3.2](#)).

In tasting tests where panels of  $N$  tasters are each required to rank in order of preference a range of similar products such as  $k$  different brands of tomato soup or varieties of apple or plum the numbers allocating each rank to each product may be presented in a  $k \times k$  contingency table. Cell entries are the numbers of tasters allocating each designated rank to each product. The appropriate analysis for association (consistency between tasters) or no particular preferences (independence) requires a modified analysis that takes account of the fact that although there are only  $N$  tasters the total number of counts (entries in all  $k^2$  cells) is  $Nk$ . Partitioning of a chi-squared statistic associated with such tables is often informative about the nature of any association ([Section 10.3.3](#)).

## EXERCISES

- 10.1 Prepare cross-sectional tables for the data in [Table 10.1](#) to show
  - (i) separately for each cholesterol level, presence or absence of CHD at each blood pressure level;
  - (ii) separately for each blood pressure level, presence or absence of CHD for each cholesterol level.
- 10.2 Verify the value  $G^2 = 2.30$  obtained in Example 10.1.
- 10.3 Verify that an analysis similar to that in Example 10.1 applied to [Table 10.4](#) in place of [Table 10.3](#) leads to identical conclusions about the nature of any association.

- 10.4 Use an asymptotic test based on  $\varphi = \log\theta$ , to show that for each of the  $2 \times 2$  subtables in [Table 10.3](#) there is strong evidence against the hypothesis of no association.
- 10.5 Verify the numerical results stated in Example 10.3.
- 10.6 Perform the computations requested in Comment 2 on Example 10.4.
- 10.7 Confirm the result stated in the comments on Example 10.8 that with column scores 1, 2, 20, 200 the linear-by-linear test gives an asymptotic  $P = 0.0101$ .
- 10.8 O'Muircheartaigh and Sheil (1983) gave the following data for the numbers of players with scores (i) par or better  $P_1$ , (ii) over par  $P_2$ , for low and high handicap golfers under two wind conditions,  $W_1$  and  $W_2$ . Does a first-order interaction model adequately describe the data?

	<i>Low handicap</i> $P_1$	<i>High handicap</i> $P_1$	<i>Low handicap</i> $P_2$	<i>High handicap</i> $P_2$
$W_1$	9	35	1	49
$W_2$	37	51	12	115

- Use at least two different methods of analysis and compare your results. Also obtain a 95 per cent confidence interval for the common odds ratio if you accept the hypothesis that they are the same for each  $2 \times 2$  table.
- 10.9 Howarth and Curthoys (1987) give numbers of males and female students in English and Scottish universities in the years 1900-01 and 1910-11. Are proportions in the sexes independent between countries for each year? Is a first-order interaction model (i) necessary and (ii) sufficient to explain the observations?

	<i>1900-01</i>		<i>1910-11</i>	
	<i>M</i>	<i>F</i>	<i>M</i>	<i>F</i>
<i>England</i>	11755	2080	16038	3579
<i>Scotland</i>	4432	719	5137	1599

- 10.10 Considering only the data for coronary heart disease (CHD) in [Table 10.1](#) and assuming that blood pressure levels and cholesterol levels are both in increasing orders, use an appropriate test to decide if there is evidence of an association between blood pressure and cholesterol levels. If you were told also that the blood pressure levels were I = normal or below normal, II = between 1 and 10 per cent above normal, III = 11 to 20 per cent above normal, IV = 21-50 per cent above normal and V = more than 50 per cent above normal and that the cholesterol levels are A = normal or less, B = up to 50 per cent above normal, C = over 50 per cent above normal and including some patients as much as 200 per cent above normal, how might you take this information into account in your analysis?
- 10.11 In an English parliamentary electoral constituency a random sample of 400 voters are classified by age and political affiliation as follows:

Political affiliation	Age group			
	30 or under	31-40	41-55	56 or over
Conservative	31	32	39	34
Liberal Democrat	16	19	25	31
Labour	36	27	58	52

Is there evidence of an association between political affiliation and age? It is generally (though not universally) accepted that the Conservative, Liberal Democrat and Labour parties represent an ordering of right, middle and left in the political spectrum.

- 10.12 Agresti (1984) quotes the following data on cross-classification of attitudes towards abortion and amounts of schooling based on the US General Social Survey, 1972. Test these data for evidence of association between attitudes and educational background.

Schooling	Attitude towards abortion		
	Disapprove	Neutral	Approve
Less than high school	209	101	237
High school	151	126	426
More than high school	16	21	138

- 10.13 Use a Pearson chi-squared test or a likelihood ratio test to determine whether either of the  $2 \times 2$  subtables in Table 10.11 indicates association, and if so whether it is reasonable to suppose a first-order association model is adequate.
- 10.14 Verify that the partitioning shown on p. 391 for a  $3 \times 3$  table into four  $2 \times 2$  tables satisfies the necessary conditions quoted on that page.
- 10.15 Verify the value of  $G^2$  and each component thereof in Example 10.11.
- 10.16 Thirty adults are each asked to indicate their preferences for each of three brands of competing detergents by ranking them 1, 2, 3. The results are presented in the following tables of numbers giving each rank:

Brand	Preference ranking		
	1	2	3
A	12	2	16
B	9	12	9
C	9	16	5

Partition the Anderson statistic into components  $C_1$  to  $C_4$  similar to those given in Section 10.3.3. (Note that if we had four type of detergent we would get a  $4 \times 4$  table with an adjusted chi-squared statistic with 9

degrees of freedom which could be split into 9 contrasts. Choice of appropriate contrasts and calculation of the variance of each is appreciably more complicated than it is for a  $3 \times 3$  table.)

- 10.17 The Scottish Office Statistical Bulletin, Nov. 1995 published by the UK Government Statistical Service gave for the academic years indicated the number of first-year undergraduate student awards in Scotland for students aged under 21 and those aged 21 or over. Do these data indicate an increasing trend in the proportion of adult students who receive grants?

<i>Year</i>	<i>1990–1</i>	<i>1991–2</i>	<i>1992–3</i>	<i>1993–4</i>	<i>1994–5</i>
<i>Under 21</i>	18349	21379	23412	25325	26606
<i>21 &amp; over</i>	5166	8079	10269	12535	13355

# 11

## Robust estimation

### 11.1 WHEN ASSUMPTIONS BREAK DOWN

In Section 1.1.2 and elsewhere we have emphasized that distribution-free does not mean assumption free, pointing out that some procedures need stronger assumptions than others for validity. Sometimes we assumed all samples were from populations whose cumulative distribution functions differed if at all only in their centrality measures, e.g. in means or in medians. In other cases we assumed only that the population distributions were symmetric, in yet others that samples came from populations where any differences were in some sense ordered. Independence both between observations within a sample and between different samples was often of major importance. Our most 'relaxed' assumption was that associated with the sign test applied to a single sample where the null hypothesis required only that each observation came from unspecified distributions (not necessarily all the same) provided only that each had the same median and that all were independent. In many examples we either showed, or mentioned in comments, that some procedures were more sensitive than others to breakdowns in assumptions. The 'breakdown' often took the form of a few sample values being inconsistent with the pattern associated with the bulk of the observations. Such observations are often referred to as **outliers** and they may arise in several ways. They may be:

- values that are incorrectly measured or recorded;
- measurements made on units that are atypical of the bulk of those in the population under study;
- measurements made with less accuracy or precision than the bulk of the observations.

There are many causes of incorrect data. A measurement of 3.6 cm may be recorded as 33.6 cm or as 36 cm. Repeating a digit is a common error with keyboard data entry, as is omission of a decimal point. A temperature measurement might be incorrectly reported in

degrees Fahrenheit when it should have been in degrees Centigrade. An observer who forgets to take a measurement at a critical time and fears a reprimand might cheat and insert a faked or guessed value. Such incorrectly reported data are not always easy to detect. For temperature measurements in a hot climate a reading given in degrees Fahrenheit when it should have been given in degrees Centigrade is readily picked up. This is not the case in polar climates, however, where the readings on either scale can be plausible.

It is not always easy to find the cause of incorrectly recorded but atypical data. One machine among many may produce poor quality goods because it is operating at an abnormal temperature, a factor that may not be observed, or thought to be relevant, by the person collecting the data. The milk production of one cow in a herd may be exceptionally low because it suffers from a yet-to-be-diagnosed illness that affects yield.

A frequent complication when data from several sources are combined is that those from some sources may be less reliable than those from other sources. For example, if the calcium content of milk samples is measured in several laboratories, most may make very precise and accurate determinations while one or two may make less precise measurements (increasing dispersion about the correct values). Another possibility is that a laboratory may consistently return values that are inaccurate, nearly all being too high (or too low). Higher dispersion around a similar mean could indicate the use of an incorrect scale, for instance at a polar research establishment where temperature measurements are recorded using the Centigrade scale at all sites apart from one where the Fahrenheit scale is used. The Fahrenheit readings will have a greater spread of values, a factor that might draw attention to the different scales in use.

How one proceeds in such circumstances depends to some extent on one's objectives. Clearly one should eliminate or correct measurement or recording errors as far as possible, but some such errors nearly always sneak through especially in large routine data collections. Although one might also try and reduce the other sources of outliers such as atypical experimental units or poor laboratory performance one often has to live at least in part with these, but it is often important to draw attention to them and to state clearly what action, if any, one takes regarding them.

Recognition of these problems has led to several major statistical developments especially since the 1970s. Generally referred to as

robust methods these aim to minimize the influence of outliers or other anomalous observations when these lead to a breakdown in basic assumptions while at the same time performing almost as well as the optimum methods when relevant assumptions hold. Alternatively robust methods may be highly efficient while requiring only minimal assumptions.

We give a brief introduction to two approaches that have proved valuable in the context of possible outliers and refer briefly to some others in passing.

The first of these - known as the **bootstrap** - has the added benefit that it is also valuable in situations where existing theory is intractable or difficult to apply. It has both a parametric and a nonparametric form, but we consider only the latter here. An interesting feature of the nonparametric version is that it often comes close to being assumption free. The one key assumption always needed is that the sample cumulative distribution function, which we met in [Section 3.3.1](#), is a reasonably good approximation to the population cumulative distribution function. When compared to some longer-established nonparametric methods there is a price to be paid for this relaxation of assumptions, but that price is often less than any that may arise from making unjustifiable assumptions.

The second main approach we consider is known as  **$M$ -estimation** with the key property that the resulting estimators behave like conventional maximum likelihood estimators when the latter are appropriate while protecting against certain types of outliers in the sense that they reduce the influence of these in determining the outcome of the analysis.

The basic ideas behind both the bootstrap and  $M$ -estimation are straightforward, but their practical application often requires care if some not-always-obvious pitfalls are to be avoided. We only outline a few basic principles here, referring the reader to more detailed texts for practical details. As a preliminary, we give a brief description of a feature called **influence** which helps to clarify the relationship between outliers and estimation and we also say something about detecting outliers.

## 11.2 OUTLIERS AND INFLUENCE

### 11.2.1 Nature and detection of outliers

This topic is discussed fully by Barnett and Lewis (1994). In broad terms an outlier is an observation so remote from other observations

as to cause surprise. Whether an observation surprises us is subjective and may depend upon what we know about the source of the data. For example, given the data set 0, 5, 9, 8, 3, 0, 125, 9, 17 without further information most people would say the observation 125 was sufficiently remote from the others to cause surprise. However, an entomologist who knew these were counts of numbers of aphids on each of 9 plants of the same rose species may not consider 125 to be odd, for in many situations involving insect populations heavy infestation on just one or two among a large number of plants is common. Barnett and Lewis (p. 16) quote even more extreme data from Fisher, Corbet and Williams (1943) for numbers of moths caught in a light trap, viz:

11 54 5 7 4 15 560 18 120 24 3 51 3 12 84

Given the observations 2.7, 3.3, 3.5, 2.8, 4.1, 4.3 and no other information none may seem surprising, but if we are told that these are weights in kg for a growing animal recorded at fortnightly intervals, a zoologist would have doubts about the validity of the fourth observation 2.8. A growing animal may suffer weight loss at some growth stage but the decrease is unlikely to be as large as this. If a loss of this magnitude were recorded the animal would probably be dead a fortnight later; it would certainly be unlikely to return to a weight consistent with a normal growth pattern within a further two weeks. This observation is not an outlier in the sense of being extreme but it is probably a contaminated observation, the contamination being a measurement error, one possibility being that a true weight of 3.8 was recorded as 2.8. This might happen if, with the balance used for weighing, the 1 kg and 2 kg weights placed in the scale pan were similar in size and design (differing mainly in density), or if the weights were read from a digital display it could be a careless misreading.

Clearly outliers or contaminated observations are a nuisance that must or should be taken into account when making inferences. Criteria have been evolved that give a basis for excluding outliers from an analysis. The aim here is often to make the remaining data more consistent with some parametric inference model. Barnett and Lewis (1994, Section 6.3) list 48 tests for outliers in normal distributions alone, most of these being optimal only for fairly specific alternatives to the null hypothesis that all data belong to the same normal distribution where neither or one or both parameters may be known. Difficulties are often caused by a masking effect whereby the power of a test for one outlier is reduced by the

presence of others. Sprent (1998, Section 3.3) illustrates some of these points using just four of the tests described by Barnett and Lewis in appropriate examples.

If an outlier can be shown to be an error clearly this should be corrected if possible; if it cannot be corrected it should be rejected. When there is no clear indication that an outlier is a measuring or recording error the appropriate course of action is less clear. It depends upon the population of interest and on what questions are being asked about that population. For example, in an experiment to test a drug for reducing blood pressure it may be known that it is ineffective if the recipient drinks alcohol and all recipients might be instructed not to drink alcohol while taking the drug. If, in a group of 25 patients receiving the drug the decreases in systolic blood pressure in mm Hg are

-3	0	2	5	6	21	23	23	27	30	32	35	37
39	41	43	47	47	49	52	54	57	59	60	64	

a clinician might suspect that the first five readings were for patients who had almost certainly ignored the alcohol ban. How firmly the clinician held that view would be a matter of experience; if it were known, for instance, that the drug is ineffective for about 1 person in 50 whether or not they drink alcohol, it would be reasonable to suppose that one or two of the first five readings might be for such cases, but rather unlikely that all five would. If, on the other hand, there were strong grounds for believing it was almost certain that a substantial reduction in blood pressure would take place if alcohol was not consumed, and what was of interest was the mean or median reduction in blood pressure in such cases it makes sense to omit the five lowest readings. To assume these five had all ignored the alcohol ban might be unfair. It would be a matter of clinical judgement whether to ask those, or perhaps all, participants if they had taken alcohol and if so how much (a dose-response relationship may mean that a little alcohol does not have the same impact as a larger intake). There may of course be doubts about the truth of the answers given to such a question!

If there were indications that between 5 and 10 per cent of the population might not respond positively to the drug (perhaps for some genetic reason) the above results are consistent with that hypothesis. In that case if one were interested in the whole population one should not reject any observations when making inferences. If one were interested only in the population other than those who clearly did not respond positively, the five observations

should be rejected. The purpose of this hypothetical example is to show that there are no easy answers to dealing with outliers. It is of course important, and ethically sound, to indicate clearly what has been done - and why - if outliers are present or suspected.

### 11.2.2 A test for outliers

Many tests for outliers lack robustness. Some are notoriously bad at detecting more than one outlier in the same tail; others tend to miss a pair of outliers in opposite tails. A simple and reasonably robust test is to classify any observation  $x^*$  as an outlier if

$$|x^* - \text{med}(x_i)| / \{\text{med}[|x_i - \text{med}(x_i)|]\} > 5.$$

Here  $\text{med}(x_i)$  is the median of all observations in the sample and the denominator is a measure of spread called the **median absolute deviation**, often abbreviated to **MAD**. The choice of 5 as a critical value is motivated by the reasoning that if the observations other than outliers have an approximately normal distribution, it picks up as an outlier any observations more than about three standard deviations from the mean.

*Example 11.1*

*The problem.* Use the above test to detect any outliers in the data set

8.9    6.2    7.2    5.4    3.7    2.8    22.2    12.7    6.9    3.1    29.8

*Formulation and assumptions.* It is easiest to determine the median and MAD after ordering the observations. We first test the observation furthest from the median then stop if this is not an outlier. If it is an outlier we test the next most extreme in either tail proceeding until we find an observation that is not an outlier.

*Procedure.* The ordered observations are

2.8    3.1    3.7    5.4    6.2    6.9    7.2    8.9    12.7    22.2    29.8

The median is 6.9 and the absolute deviation of the observed 2.8 from this median is  $|2.8 - 6.9| = 4.1$ . Similarly the remaining absolute deviations are 3.8, 3.2, 1.5, 0.7, 0.0, 0.3, 2.0, 5.8, 15.3 and 22.9. Ordering these we easily find that their median, the MAD, is 3.2. Setting  $x^* = 29.8$ , the left-hand side of our statistic is  $22.9/3.2 = 7.16$  so we class 29.8 as an outlier. Setting  $x^* = 22.2$  we find  $15.3/3.2 = 4.78$  so we do not class this or any other observation as an outlier.

*Conclusion.* The data set contains one outlier,  $x^* = 29.8$ .

*Comments.* 1. Having decided 29.8 is an outlier we still have to decide what to do about it. An obvious line to follow is to check (a) whether it may be an

error and if it is whether it can be corrected; if it is not obviously an error then (b) is there anything peculiar about the experimental unit giving rise to that value?

2. The data look likely to have come from a skew rather than a normal distribution. The test we have given is not a test for normality, however. A test such as Lilliefors' test should be used to assess normality if that is relevant.

### 11.2.3 Influence and robustness

In robustness studies a key role is played by influence functions. We consider only briefly two simple influence functions both of which are discussed more fully in Sprent (1998, Section 3.4). A detailed treatment of influence functions including applications in various fields is given in Barnett and Lewis (1994, Section 3.1.3). The idea is due to Hampel who describes their role in robust estimation in Hampel (1974).

We consider first the effect of a single outlier in relation to the mean. Suppose we have  $n$  uncontaminated observations  $x_1, x_2, \dots, x_n$  which we call 'good' observations and one contaminated observation  $z$  which is an outlier in the sense that it takes a value greater than any good observation. Let  $\bar{x} = (\sum x_i)/n$  be the mean of the good observations and  $\bar{x}_a$  the mean of the augmented set that also includes  $z$ . It is easily seen that

$$\bar{x}_a - \bar{x} = \frac{n\bar{x} + z}{n+1} - \bar{x} = \frac{z - \bar{x}}{n+1}$$

whence it follows that the effect of  $z$  on the sample mean  $\bar{x}$  for the good observations is a linear function of  $z$  and that it tends to infinity as  $z \rightarrow \infty$ . In other words the effect of just one contaminated observation on the sample mean may be infinite! While (11.1) is a measure of the influence of one contaminated observation the influence function  $I_n(z)$  for a sample of  $n$  is obtained from (11.1) by multiplying by  $n+1$ , i.e.

$$I_n(z) = (n+1)(\bar{x}_a - \bar{x}) = z - \bar{x}.$$

As  $n \rightarrow \infty$ ,  $\bar{x} \rightarrow \mu$ , the population mean for the good observations, and the limiting function

$$I(z) = z - \mu$$

is the asymptotic influence function, but it is often referred to simply as the *influence function* as it is the form of greatest interest.

Broadly similar and easily modified arguments apply if  $z$  is an outlier in the lower tail, i.e. has a value markedly less than all good observations.

In contrast to the situation with the mean where one outlier may have an appreciable, indeed unbounded, influence the effect of a single outlier upon the median is usually small and indeed bounded providing  $n \geq 2$ . The situation is slightly but not essentially different depending upon whether the number of good data is odd or even. Consider first the case where we have  $n = 2m$  good data. If the observations are arranged in ascending order and we denote by  $x_{(i)}$  the  $i$ th largest sample value then the median of the good observations is  $\frac{1}{2}(x_{(m)} + x_{(m+1)})$ . If an outlier  $z$  has a value greater than  $x_{(2m)}$  the effect is to shift the median of the combined sample only to  $x_{(m+1)}$ . Similarly an outlier below  $x_{(1)}$  only shifts the median to  $x_{(m)}$ . If the number of good observations is odd and  $n = 2m + 1$  giving a median  $x_{(m+1)}$  it is easily seen that an outlier  $z > x_{(2m+1)}$  only shifts the median to  $\frac{1}{2}(x_{(m+1)} + x_{(m+2)})$  while any  $z < x_{(1)}$  shifts the median to  $\frac{1}{2}(x_{(m)} + x_{(m+1)})$ .

Because the underlying population distribution influences the values of the order statistics a different approach to that used for the mean is needed to determine the asymptotic influence function. Details are given in Barnett and Lewis (1994, Chapter 4) who show that an appropriate expression is

$$I(z) = \frac{\operatorname{sgn}(z - m)}{2f(m)} \quad (11.3)$$

where  $m$  is the median for the good observations that are distributed with a frequency function  $f(x)$  and  $\operatorname{sgn}(z - m) = +1, 0$  or  $-1$  depending upon whether  $z$  is greater than, equal to, or less than  $m$ . Clearly  $I(z)$  given by (11.3) is bounded for a continuous distribution and the supremum or greatest value of  $|I(z)|$  is  $1/2f(m)$ . This supremum is called the gross error sensitivity.

It is clear from the above that if we have a large sample from a  $N(\mu, \sigma^2)$  distribution and just one contaminated observation is added that the possible effect upon the sample mean is unbounded while the maximum effect upon the sample median cannot exceed  $1/2f(\mu) = \sigma\sqrt{(2\pi)/2} = 1.253\sigma$ , since the median of the normal distribution of the good values is  $\mu$ .

It is not difficult to see that if there are  $n = 2m + 1$  good observations and two contaminated observations are added, then whatever their magnitude the greatest effect upon the adulterated sample median is to shift it from  $x_{(m+1)}$  to either  $x_{(m)}$  or  $x_{(m+2)}$ . The argument

easily extends to 4 contaminants with bounds at most  $x_{(m-1)}$  or  $x_{(m+3)}$  and proceeding in this way, when  $2m$  contaminated observations are added the bounds are  $x_{(1)}$  and  $x_{(2m+1)}$ . Only when there are at least  $2m + 1$  contaminants is it possible for the median to include some of the contaminated observations and thus become unbounded if some of the contaminated values are unbounded. Thus, at least 50 per cent of the observations must be contaminated in a way that makes them outliers before the median becomes unbounded. A broadly similar argument may be used for an even number of good observations. The number of contaminated observations needed to make a mean or median unbounded determines what is called the breakdown point. This is sometimes expressed as a percentage, i.e. 50 per cent for the median, but more usefully as a fraction, i.e.  $\frac{1}{2}$  for the median. As we have seen, the breakdown point is  $1/(n+1)$  for the mean, since only one contaminated observation need be added to a sample of  $n$  to make the sample mean potentially unbounded. The breakdown point is an important measure of robustness because in many practical situations there is a strong suspicion, or even direct evidence, that an appreciable proportion of observations may be contaminated. It should now be intuitively clear why in earlier chapters we often found that estimators based on medians are more robust against outliers than are those based on means.

## 11.3 THE BOOTSTRAP

### 11.3.1 Motivation

We indicated that the motivation for the bootstrap is that the sample cumulative distribution function  $S(x)$  introduced in [Section 3.3.1](#) should for all but small samples reflect many characteristics of the population cumulative distribution function,  $F(x)$ , for the population from which the random sample was obtained. This was implicit in the Kolmogorov test introduced in that section where we used  $S(x)$  in deriving a statistic to determine whether a sample was consistent with some given  $F(x)$ . For the bootstrap we are often less concerned with hypotheses about any particular  $F(x)$  but more often with making inferences simply on the assumption that  $S(x)$  is a good approximator to some  $F(x)$  without being specific about what that  $F(x)$  is. The more vague our assumptions about a population distribution the more useful the bootstrap becomes.

The method is based on repeated resampling of data to tell us more about characteristics of the population from which the data are

a random sample. Resampling, as we saw as early as Example 1.4, and in numerous other examples of permutation test procedures throughout this book, is at the heart of permutation or randomization test theory. Permutation tests are based on resampling without replacement and where appropriate lead to exact tests and estimation procedures. Bootstrapping uses sampling with replacement and leads to approximate tests and estimation procedures; however, it is also applicable in many situations where no permutation test is available or appropriate. Although the method can be used without specification a priori of any distributional model the procedure itself can only be justified by fairly complicated mathematics. The main practical requirement for its application is suitable computing facilities. Although bootstrapping results are by their nature usually only approximate they are often more reliable and informative than those obtained by fitting a wrong model, e.g. assuming normality when that is not valid or is not justified by some asymptotic result like the central limit theorem. There is seldom a unique or best solution to a bootstrap problem. In practice the method is used mainly either because there is no tractable analytic solution, when a permutation test or the facility to carry it out is not available, or when there is doubt about whether conditions needed for a particular analytic solution or permutation test actually hold.

The approach is intuitively reasonable in many applications and confidence in the method is enhanced because when it is correctly used inferences are usually similar to those given by analytic or permutation solutions when these exist or are relevant. Important applications include those to complex data structures or ones that involve inferences about concepts such as correlation or ratios of variables where analytic results are not readily available except under very restrictive distributional assumptions.

Use of the bootstrap stems largely from work by Efron (1979). A straightforward introduction to it and related techniques is given by Efron and Gong (1983). A more detailed account of the elements of the bootstrap is given by Sprent (1998, Chapter 2) and full treatments at the elementary and intermediate level with many examples are given by Efron and Tibshirani (1993), Davison and Hinkley (1997) and Chernick (1999).

### 11.3.2 Bootstrap samples

Given a random sample of  $n$  observations  $x_1, x_2, \dots, x_n$  from some population a bootstrap sample is a random sample of size  $n$

obtained from these data by sampling with replacement. Thus some of the  $x_i$  may occur more than once and others not at all in a bootstrap sample. It is notionally possible to determine the distribution of all possible bootstrap samples and the distribution for many associated statistics such as the means or medians of the bootstrap samples. Such distributions are called true bootstrap distributions to distinguish them from estimates based on a random subset of all possible bootstrap samples. The latter are important in practical applications because determining true bootstrap distributions analytically is a formidable task for all but small  $n$  except in the case of a few statistics where some general analytic results hold (and in this latter case inferences can usually be made easily without bootstrapping). Sprent (1998, Section 2.3) investigates the complete bootstrap distribution for the mean and median for a sample of 3 observations. This is useful for illustration, but trivial in practice because so small a sample is not very informative about the distribution associated with a large population.

In practice useful bootstrap inferences are nearly always based on Monte Carlo sampling to generate a predetermined fixed number,  $B$ , say, of bootstrap samples. We denote a typical bootstrap sample by

$$x_1^*, x_2^*, \dots, x_n^*$$

where each  $x_i^*$  is equal to one of the original observations. Because sampling is with replacement some of the original sample values (the  $x_j$ ) may not appear and others may occur more than once among the  $x_i^*$ . For example, if  $n = 9$  and the observations are 2.5, 3.1, 4.2, 5.1, 5.3, 5.9, 6.7, 7.2, and 10.5 a typical bootstrap sample might be  $x_1^* = 4.2$ ,  $x_2^* = 6.7$ ,  $x_3^* = 5.1$ ,  $x_4^* = 6.7$ ,  $x_5^* = 7.2$ ,  $x_6^* = 6.7$ ,  $x_7^* = 2.5$ ,  $x_8^* = 10.5$ ,  $x_9^* = 5.1$ . If  $B$  bootstrap samples are generated the  $b$ th may be written  $x_1^{*b}, x_2^{*b}, \dots, x_n^{*b}$ . Vector notation provides a convenient shorthand if we write  $\mathbf{x}^*$  for any bootstrap sample and  $\mathbf{x}^{*b}$  for the  $b$ th sample. In bivariate or multivariate situations such as those in correlation or regression each  $x_i$  may itself be a vector. For each bootstrap sample  $\mathbf{x}^*$  we often compute statistics such as the sample mean or median or sample variance and use these to estimate the corresponding population distribution characteristic. For illustrative purposes we denote the parameter or other population characteristic we are interested in by  $\theta$  and the statistic used to estimate it from the bootstrap sample  $\mathbf{x}^*$  by  $s(\mathbf{x}^*)$  which gives an estimate  $\theta^* = s(\mathbf{x}^*)$  of  $\theta$ . Because we are sampling with replacement the numerical value of  $s(\mathbf{x}^*)$  changes from sample to sample and so the statistic has a distribution. As the number of bootstrap samples,

$B$ , tends to infinity the mean of any statistic  $s(\mathbf{x}^*)$  will tend to the mean computed for the true bootstrap sampling distribution, although as pointed out above, this is only known in a few special cases or can only be worked out for general cases for small values of  $n$ . Practical experience however shows that in many (but not in all) situations even for small  $n$  the mean of  $B$  bootstrap samples converges rapidly to the limiting value that holds when  $B \rightarrow \infty$ .

If we generate  $B$  bootstrap samples and denote the mean of the  $s(\mathbf{x}^{*b})$  by  $s(\cdot^*)$ , i.e.  $s(\cdot^*) = \sum_b [s(\mathbf{x}^{*b})]/b$ , then the appropriate estimator of the true bootstrap standard error of  $\theta^*$  which we denote by  $\text{se}_B(s)$  is

$$\text{se}_B(s) = \{\sum_b [s(\mathbf{x}^{*b}) - s(\cdot^*)]^2/(B - 1)\}^{1/2} \quad (11.4)$$

which is the usual estimator of a population standard deviation based upon a random sample of  $B$  from some population and tends to the true bootstrap standard deviation for the statistic  $s$  as  $B \rightarrow \infty$ . In practice the approximation is often good for  $B$  as low as 20 providing  $n$  is not too small. Even for small  $n$  reasonable estimates may be obtained with  $B = 100$ , although again we caution against indiscriminate use of the bootstrap with very small samples because then the sample distribution may not truly reflect a population characteristic of interest.

Pitfalls to watch out for when using the bootstrap with small samples include difficulties with bias that carry over to moderate or large samples in some contexts and are discussed with numerical examples in Sprent (1998, Section 2.4) and much more generally for a range of applications by Efron and Tibshirani (1993). It is important to realise that in applications using only a finite number,  $B$ , of bootstrap samples there are two sources of error. The first is the usual sampling error applicable to all sampling based inferences about a population no matter what method of inference is used. For example, in parametric inference about the mean of a normal population based on a sample of  $n$  the sample mean  $\bar{x}$  will in general not equal the population mean  $\mu$  but the sampling 'error' in taking  $\bar{x}$  as an estimator of  $\mu$  is measured by the standard error. The second source of error specific to bootstrap sampling is that made when we approximate to the true bootstrap standard error using only a finite number  $B$  of bootstrap samples. Although we do not prove it here, it can be shown that the true bootstrap standard deviation is approximately equal to the estimated population standard deviation for the corresponding estimator. Example 11.2 sheds some light on the way sampling variation is reflected in bootstrap estimates.

### Example 11.2

*The problem.* For the following sample data explore the use of a bootstrap for estimating the population median.

0.04	0.06	0.27	0.32	0.33	0.40	0.50	0.63	0.69	0.92
1.09	1.10	1.35	1.61	1.66	1.69	1.71	1.80	1.98	2.65
2.83	3.50	3.72	3.75	3.99	5.16	5.49	6.31	7.05	16.05

*Formulation and assumptions.* The data are in fact a random sample of 30 from a known distribution, but they are not unlike some that may arise in practice. They might for example be the percentages of some pollutant in various water sources, a situation where it is not uncommon to find small levels of pollution in many of the sources but quite high and wide ranging levels in others. The data have been arranged in ascending order to indicate more clearly the skewness with a long upper tail including the value 16.05 that might suggest itself as an outlier (Exercise 11.2). The sample median is  $\frac{1}{2}(1.66 + 1.69) = 1.675$ . We explore the information available by generating  $B = 40$  bootstrap samples and obtaining the median of each for use in estimating the bootstrap standard error of this statistic.

*Procedure.* We used Minitab to generate the 40 bootstrap samples using the facility therein for sampling with replacement. The median was computed for each of these 40 samples and ranged from 0.805 to 3.17. The estimated bootstrap standard error given by (11.4) was 0.4116.

*Conclusions.* To help interpret these results we now disclose that the data are a computer generated random sample from an exponential distribution with mean  $\mu = 3$ . The theoretical median of that distribution is  $\theta = 2.079$ . A crude 'rule of thumb' indication used in bootstrapping is that we should accept a hypothetical value of a parameter  $\theta$  if the sample equivalent estimator (here the sample median) lies within two standard deviations of that hypothesized value. Taking the estimated bootstrap standard error as a reasonable estimate of this standard deviation in this example the difference  $2.079 - 1.675 = 0.404$  is less than this estimated standard error, which here is 0.4116.

*Comments.* 1. There may be some unease as to whether 40 bootstrap samples give a reliable estimate of the true bootstrap standard error. That it does so is also an implicit requirement for validity of the crude rule of thumb used above as this is based on the assumption that the true bootstrap standard error is reasonably close to the best estimated standard error where this is known.

In this example where we sampled from a known distribution there is an analytic expression for the standard error of a median estimator from a sample of  $n$ , namely

$$se(\text{median}) = 1/(4nf^2)^{1/2} \quad (11.5)$$

where  $f$  is the ordinate of the probability density function at the median. For an exponential distribution with mean 3 it can be shown that  $f = 0.1667$  and when  $n = 30$  that  $se(\text{median}) = 0.5476$ . Here we see that our estimate of the bootstrap standard error, 0.4116, is an underestimate of the true standard error. This is a bias introduced by the fact that our chosen sample has a median  $m = 1.675$  which

is appreciably below the population median  $\theta = 2.079$ . In fact 19 of the 30 sample values are less than the population median. This is not an alarmingly departure from the expected number 15. The situation is not very different from that with many familiar tests. Recall that when sampling from a normal distribution the usual sample estimator  $s^2$  of  $\sigma^2$  may differ appreciably from the latter in many samples. This is why, when  $\sigma^2$  is unknown the *t*-test is invoked to allow for such uncertainty. It is interesting to note that if we assumed the population median was in fact equal to the sample median of 1.675, then the standard error of the median estimator assuming our sample was from an exponential distribution with this median would now be 0.4412, close to the bootstrap estimate of 0.4416 obtained above.

2. The bootstrap may also be used to obtain an estimated confidence interval for a parameter such as the median. There are a number of practical difficulties in obtaining such intervals but we briefly outline one approximate procedure in [Section 11.3.4](#).

3. In this example we considered a sample from a known distribution and so were able to invoke some theoretical results for the standard error of the median because  $f$  was known in (11.5). Had we not known  $f$  or we suspected the sample was from a mixture of distributions or that some observations may be outliers the bootstrap might be a serious competitor to, say, sign test procedures. We explore this point further in Example 11.5.

### 11.3.3 Bootstrapping versus permutation procedures

We have suggested that the bootstrap is most useful when no simple analytical procedure or no permutation procedure exists. However, the method may still provide an approximation to certain permutation procedures although its application in these circumstances is in general not as straightforward as that of exact procedures or even that for Monte Carlo approximations in permutation tests. A key difference between a permutation test and a bootstrap procedure is that the former is based on sampling without replacement while the latter is based on sampling with replacement. The consequences of this apparently small difference are far reaching and here we illustrate some of them for two-independent-sample problems.

Although we did not consider the Pitman permutation procedure for two independent samples from continuous distributions in much detail in [Chapter 5](#) because it lacks robustness, for hypothesis testing it is in theory straightforward and as in the case of most permutation tests there are many equivalent statistics. Given samples

$$\begin{aligned}x_1, x_2, \dots, x_m \\y_1, y_2, \dots, y_n\end{aligned}$$

it is well known and easily verified that equivalent permutation test statistics for the hypothesis  $H_0$ : *the population means are identical* against the usual one- or two-tail alternatives under the assumption that the samples are from populations with otherwise identical distributions include the usual  $t$ -statistic, the first sample sum  $\sum x_i$ , the second sample sum  $\sum y_j$ , the first sample mean  $\bar{x}$ , the second sample mean  $\bar{y}$  and the difference between sample means  $\bar{x} - \bar{y}$ . This equivalence follows from the fact that under permutation both the denominator in the usual  $t$ -statistic and the sum of the combined sample values remain constant so that all the above statistics are linear functions of one another and have the same ordering and the same  $P$ -value is associated with each corresponding value in that ordering. We must remember that if this permutation  $t$ -statistic is used it does not have the tabulated  $t$ -distribution relevant to the normal parametric test.

The above equivalences all assume sampling is without replacement and they do not carry over to the bootstrap where we sample with replacement. Then the sum of all  $m + n$  values will change between bootstrap samples depending upon which of the original sample values do not appear at all, appear only once, or appear more than once in a given bootstrap sample. Thus we will arrive at different conclusions if we base our bootstrap inferences on each of the above possible permutation distribution statistics. We have already pointed out that there is in general no unique and best bootstrap for a particular situation. What is important is to use a statistic that is intuitively reasonable in a particular case. If we believe our samples come from identical distributions that differ only in mean it makes sense to base bootstrap inferences on the sample mean difference  $\bar{x} - \bar{y}$ . How we might proceed is shown in Examples 11.3 and 11.4.

### *Example 11.3*

*The problem.* In Example 5.1 we considered two data sets giving times to perform some calculations. These were

<i>Group A</i>	23	18	17	25	22	19	31	26	29	33
<i>Group B</i>	21	28	32	30	41	24	35	34	27	39

Assuming that these are from populations with distributions that differ if at all only in their means  $\mu_1$  and  $\mu_2$  use a bootstrap analysis to assess the strength of evidence against the hypothesis  $H_0: \mu_1 = \mu_2$  when the alternative is  $H_1: \mu_1 \neq \mu_2$ .

*Formulation and assumptions.* The analysis in Example 5.1 was based on the Wilcoxon rank sum test. However here we base our analysis on the raw data so that the bootstrap may be regarded as an approximation to a Pitman permutation

test. We consider 1000 bootstrap samples and for each we compute  $\bar{x}^* - \bar{y}^*$  in the way described under *Procedure* below. We estimate the relevant P-value as the proportion of the  $\bar{x}^* - \bar{y}^*$  that exceed the observed  $\bar{x} - \bar{y}$  in magnitude.

*Procedure.* In a Pitman permutation test where sampling is without replacement each permutation sample of size  $m$  is obtained by drawing a sample of  $m$  observations from the combined sample of  $m + n$  observations. The  $n$  observations not selected automatically form the complementary permutation sample of size  $n$ . In this example  $m = 10$  and  $n = 11$ . To obtain an analogous bootstrap sample we first select a sample of  $m + n = 21$  with replacement from the combined Group A and Group B data. The first 10 values chosen for this sample are designated as a bootstrap sample of size  $m$ . The remaining  $n$  sample values constitute the bootstrap sample of size  $n$ . For each such bootstrapped samples we compute  $\bar{x}^* - \bar{y}^*$  and compare its magnitude with that of the observed  $\bar{x} - \bar{y} = -7.245$  for the given data.

Using this procedure, one sample of 21 we obtained by sampling with replacement from the given data was

17 22 22 27 31 25 41 34 26 17 30 23 28 21 33 21 27 34 23 33 19

leading to bootstrap samples of 10 and 11 values respectively

17 22 22 27 31 25 41 34 26 17  
30 23 28 21 33 21 27 34 23 33 19

whence it is easily verified that  $\bar{x}^* - \bar{y}^* = 26.2 - 26.545 = -0.345$ .

An estimate of the two-tail test  $P$ -value based on  $B$  bootstrap samples is then given by  $P^* = (\text{number of samples for which } |\bar{x}^* - \bar{y}^*| \geq 7.245)/B$ . For 1000 such samples we found  $|\bar{x}^* - \bar{y}^*| \geq 7.245$  in 15 cases implying  $P^* = 0.015$ . For a second sampling again with  $B = 1000$  we found  $P^* = 0.016$ .

*Conclusion.* There is strong evidence against the hypothesis that the means are equal.

*Comments.* 1. For these data a Pitman permutation test gives the exact  $P = 0.014$ , in close agreement with the bootstrap result obtained above. We pointed out in [Section 5.1.2](#) that the Pitman test usually gave similar results to the *Mest* even when assumptions relevant to the latter did not hold. For the above data few statisticians would have reservations about use of the parametric *t*-test and indeed for these data the two-tail  $P$ -value given by that test is  $P = 0.011$ , again in broad agreement. Also in Example 5.1 we showed that using the exact WMW test for these data gave  $P = 0.016$ , so there is good agreement between all tests considered here.

2. Using the bootstrap statistic  $\bar{x}^* - \bar{y}^*$  is intuitively reasonable under the assumption that the only possible population distributional differences involved the means. A different approach is appropriate if we drop the assumption of identical distributions under  $H_0$ . We do not pursue this further here but this problem is discussed briefly by Sprent (1998, Section 2.7) and more fully, together with some alternatives to the approach used in this example by Efron and Tibshirani (1993, Chapters 15 and 16). We remind readers that even in simple parametric situations choice of appropriate procedures depends even more strongly on relevant assumptions; for instance validity of the *t*-test depends

not only on an assumption of normality but also upon one of equal population variances.

*Computational aspects.* Nearly all major statistical packages have a facility for random sampling with replacement and even if no direct bootstrapping program is included for the simple example considered here it is usually easy to write a macro to form many bootstrap samples quickly. We used Minitab. For more sophisticated applications of the bootstrap the package S-PLUS is especially well suited.

Bootstrapping following the broad pattern in Example 11.3 may also be applied in situations where WMW methods are used. With one possible modification the situation is essentially the same as that for the raw data except that these are replaced by ranks. The one modification that might be considered is that if we proceed in this way we are not strictly using a WMW procedure in the bootstrap. Why this is so is illustrated by a trivial example. We consider the situation in Example 1.4 where patients were ranked 1 to 9 in response to two drugs, four of them having received one drug and the remaining 5 another. We learnt in Chapter 5 that essentially what we did in that example was a WMW test. If we follow the procedure in Example 11.3 given the combined sample raw data 1, 2, 3, 4, 5, 6, 7, 8, 9 we may obtain a combined bootstrap sample, say, 1, 2, 9, 2, 8, 1, 3, 3, 9. Continuing as we did in Example 11.3 we would split this into bootstrap samples 1, 2, 9, 2 and 8, 1, 3, 3, 9 of size 4 and 5. To do so is not unreasonable but proceeding in this way we are not performing a strict WMW bootstrap because we have not reranked the combined sample bootstrap data to allow for ties induced by the resampling. Doing so using appropriate mid-ranks would replace 1, 2, 9, 2, 8, 1, 3, 3, 9 by 1.5, 3.5, 8.5, 3.5, 7, 1.5, 5.5, 5.5, 8.5 giving bootstrap samples 1.5, 3.5, 8.5, 3.5 and 7, 1.5, 5.5, 5.5, 8.5. In practice the difference tends to be small for samples of reasonable size although it may be appreciable in small samples. We do not recommend the bootstrap for small samples for a reason already explained - namely that a small sample may not be a good representation of an underlying distribution (in itself a limitation in any form of inference) and in the case of the bootstrap this is compounded by the bootstrap sampling error.

#### *Example 11.4*

*The problem.* For the data in Example 11.3 explore the use of a bootstrap in the estimation of a  $P$ -value using ranked data.

*Formulation and assumptions.* Bootstrap sampling may be used to obtain an approximation to the exact  $P = 0.016$  for a two-tailed test obtained in Example

5.1 using the WMW procedure. The process used follows broadly the pattern in Example 11.3.

*Procedure.* When the combined samples are ranked we established in Example 5.1 that the rank sums were  $S_m = 76$  and  $S_n = 155$ . These statistics are inappropriate for bootstrapping since for bootstrap samples the corresponding  $S_m^*$  and  $S_n^*$  no longer satisfy the condition  $S_m^* + S_n^* = \frac{1}{2}(m+n)(m+n+1)$  in general. Analogous to the situation in Example 11.3 we used the difference between means of our bootstrap samples as the appropriate statistic proceeding for 1000 samples in the manner described in Example 11.3 using (a) the data obtained by resampling ranks directly and (b) that obtained by modification by reranking the resampled data to allow for ties introduced by resampling with replacement. The respective two-tail  $P$ -values obtained in the way described in Example 11.3 were  $P = 0.012$  for (a) and  $P = 0.008$  for (b). While the result using (a) is close to the exact result that for (b) is disappointing. We took a second sample of 1000 bootstrap estimates using (b) and this time we found  $P = 0.022$ . Combining the two samples gave an estimated  $P = 0.015$  in close agreement with the WMW result (*see* Comment 1 below).

*Conclusion.* There is strong evidence against  $H_0$ .

*Comments.* 1. The quite marked disagreement between the two estimates of  $P$  from two different samples of 1000 using method (b) suggests that our chosen statistic may be somewhat unstable in its properties. This phenomenon is familiar to serious users of the bootstrap especially with small samples. It may arise from two causes; we may simply be unlucky with the way the sampling scheme is working for the particular samples that were selected, but in this particular example we conjecture that the effect may in part be due to introducing the extra step of modifying the resampled values by reranking them to allow for ties. It is possible that some modification of the statistic we used (difference in means) might not then be the most appropriate statistic; however we have not explored this further.

2. Despite the discrepancy discussed in Comment 1 the results lead to essentially the same conclusion as that using an exact WMW test and in practice would be comparable to those obtainable using the same number of Monte Carlo samples for a permutation test.

#### 11.3.4 Bootstrap confidence intervals

Bootstrapping is often used to obtain approximate confidence intervals and much has been written about their interpretation and properties. Simplistic approximations suffer from two defects, the first being one of bias which we cover briefly in [Section 11.3.5](#) and the other is that they often tend to provide less than the claimed coverage, e.g. an interval computed in a manner that purports to give a 95 per cent coverage may only give about 91 or 92 per cent coverage. At a basic level Efron and Tibshirani (1993, Section 14.2) list 5 different ways to compute confidence intervals using the bootstrap. Each works tolerably well when appropriate but it is hard

to know which is appropriate if little is known about distributional properties of the population from which the data are obtained - a situation when we are particularly likely to want to use the bootstrap. Another approach to bootstrap confidence intervals due to Hall (1992) is described by Manly (1997, Chapter 3).

We start our discussion from the well-known result that given a sample of  $n$  from a normal distribution with variance  $\sigma^2$  the 95 per cent confidence limits for the population mean  $\mu$  are  $\bar{x} \pm 1.96\sigma/(\sqrt{n})$  where  $\bar{x}$  is the sample mean. If  $\sigma$  is unknown we replace the true standard error  $\sigma/(\sqrt{n})$  by its usual sample estimate  $s/(\sqrt{n})$  and 1.96 by the appropriate quantile of the  $t$ -distribution with  $n - 1$  degrees of freedom where  $s = \sqrt{\{[\sum_i(x_i - \bar{x})^2]/(n - 1)\}}$ . For values of  $n > 30$ , the  $t$ -distribution quantiles closely approach those of the standard normal distribution. Modifications for other confidence levels are straightforward. In practice the above limits are widely used even if there is little evidence that the sample is from a normal distribution, faith in the outcome relying on the central limit theorem holding, despite its asymptotic nature, even for moderate  $n$ .

This fundamental result for parametric inference stimulated two approaches to forming bootstrap confidence intervals. The first is that for reasonably large samples approximate 95 per cent confidence limits for a parameter  $\theta$  may be based on a sample estimator  $\hat{\theta}$  of  $\theta$  and are given by  $\hat{\theta} \pm 2se^*(\hat{\theta})$  where  $se^*(\hat{\theta})$  is an estimate of the true bootstrap standard error based on  $B$  bootstrap samples. For samples from a normal population this works for the mean because it can be shown that in that case the bootstrap standard error is a good approximation to the usual estimated standard error (although it shows a small bias). However, as we have emphasized the bootstrap is most useful when analytic theory is non-existent or is highly distribution dependent and therefore any parametric theory may not hold for a particular sample. Experience has shown that this bootstrap method does not translate well to such situations. This is partly due to potential bias in bootstrap standard errors but more importantly because confidence intervals based on symmetry about an estimator  $\hat{\theta}$  are inappropriate if that estimator has a skew distribution.

A more fruitful and relatively easy to apply approach to bootstrap confidence intervals is based on the distribution of appropriate bootstrap estimators,  $s(x^{*b})$  of each of  $B$  bootstrap samples. When  $B$  is large, then providing the  $s(x^{*b})$  are reasonably free from bias the distribution of  $s(x^{*b})$  is likely to approach the sampling distribution of the estimator of the parameter  $\theta$  we are interested in. An obvious

way to estimate a 95 per cent confidence interval is to take the 0.025 and 0.975 quantiles of the distribution of the  $B$  computed  $s(\mathbf{x}^{*b})$ . Thus if  $B = 1000$  for a 95 per cent confidence interval the limits are the 25th and 975th largest values of the  $s(\mathbf{x}^{*b})$ . More generally for a  $(1 - 2\alpha)$  100 per cent interval with  $B$  bootstrap samples the limits are the  $\alpha B$ th and the  $(1 - \alpha)B$ th largest sample value of  $s(\mathbf{x}^{*b})$ . For nonintegral values of such quantiles it usually suffices to round to the nearest integer. Confidence intervals formed in this way are called **quantile-based intervals**. The main defects of these intervals are that they may be misleading due to bias and they tend to give less than the nominal coverage. The effect of bias is that the true probabilities associated with each tail are not equal to each other, and which is the greater depends upon the direction of bias. Efron and Tibshirani (1993, Chapter 14) present two commonly used corrections to remedy these weaknesses, but we do not discuss these here-

### *Example 11.5*

*The problem.* For the data in Example 11.2 obtain 95 and 99 per cent quantile based confidence intervals for the population median using 2000 bootstrap samples. The data are

0.04	0.06	0.27	0.32	0.33	0.40	0.50	0.63	0.69	0.92
1.09	1.10	1.35	1.61	1.66	1.69	1.71	1.80	1.98	2.65
2.83	3.50	3.72	3.75	3.99	5.16	5.49	6.31	7.05	16.05

*Formulation and assumptions.* For each of 2000 bootstrap samples we obtain the median and these are arranged in order and the appropriate limits are obtained in the way described under Procedure.

*Procedure.* We obtained 2000 medians of bootstrap samples using Minitab and arranged these in ascending order. The 95 per cent limits were the 50th and 1950th largest. For our sample these turned out to be 0.92 and 2.83. Similarly 99 per cent limits are given by the 10th and 1990th largest and turned out to be 0.69 to 3.50.

*Conclusion.* Estimated quantile-based 95 and 99 per cent bootstrap confidence intervals for the population median are (0.92, 2.83) and (0.69, 3.50).

*Comments.* 1. It is interesting to compare these with confidence intervals based on the sign test procedure. Under that procedure the interval (0.92, 2.83) has an exact 95.72 per cent coverage and the interval (0.69, 3.50) has an exact 99.48 per cent coverage, so there is virtually no evidence here of the potential undercoverage of quantile-based intervals that we referred to above. This may be fortuitous for another sample of 2000 bootstrap medians would be unlikely to lead to the same intervals.

2. For our 2000 samples the estimated standard error of the bootstrap median turned out to be 0.4213, in close agreement with the estimate 0.4116 obtained in

Example 11.1 from only 40 samples. It is common experience with bootstrapping that while samples of 100 or less give good estimates of standard errors, samples of 1000 or more are needed for satisfactory quartile-based or related estimates of confidence intervals.

3. Because of the asymmetric sample [which we in fact know came from an asymmetric distribution (see Example 11.2)] approximate confidence intervals based on the estimated bootstrap standard error are not appropriate even if that standard error estimate is itself good. In this example using this approach approximate 95 per cent confidence limits are  $1.675 \pm 2 \times 0.4213$ , giving the unsatisfactory interval (0.83, 2.52).

*Computational aspects.* See remarks under this heading in Example 11.3.

When a statistic of interest is calculated for a large number of bootstrap samples it is often useful to examine the distribution of these bootstrap statistics informally, perhaps presenting them in a histogram or as a box and whisker plot - techniques familiar in what are often called exploratory data analyses.

Although the data set is too small to give bootstrap results likely to be relevant to a larger population we consider some bootstrap samples using the data in Example 8.1 to show what we mean.

#### *Example 11.6*

*The problem.* In Example 8.1 we gave the data set

Hours from start of thaw (x)	0	1	2	3	4	5	6
Flow in cubic metres/sec (y)	2.5	3.1	3.4	4.0	4.6	5.1	11.1

which was displayed in Figure 8.1. In that example and in Example 8.3 least squares and Theil-Kendall estimates of the regression line slope  $\beta$  were obtained. Use a bootstrap with 1000 samples to obtain approximate 95 per cent inter-quartile based confidence intervals for each of these estimates. Use histograms to illustrate the main characteristics of bootstrap estimates of  $\beta$  in each case and explain the main characteristics of and differences between the forms of these histograms.

*Formulation and assumptions.* One (but not the only) basis for bootstrapping in a bivariate regression problem with  $n$  observed data points  $(x_i, y_i)$  is to take samples of  $n$  of these points with replacement. Using this approach for least squares bootstrap estimates  $b^*$  of  $\beta$  we obtain the usual least squares estimate for each bootstrap sample. Similarly for the Theil-Kendall estimator our bootstrap estimator for each sample is the median of the pairwise  $b_{ij}$  (as defined in Section 8.1.3) for that sample.

*Procedure.* For each bootstrap sample of data points we may compute the bootstrap least-squares or Theil-Kendall estimator of  $\beta$  in the way described above. For 1000 samples the least squares estimators in one run gave a quantile-based 95 per cent confidence interval for  $\beta$  as (0.49, 2.01), in reasonable agreement with the classic least squares interval (0.23, 1.98). The Theil-Kendall

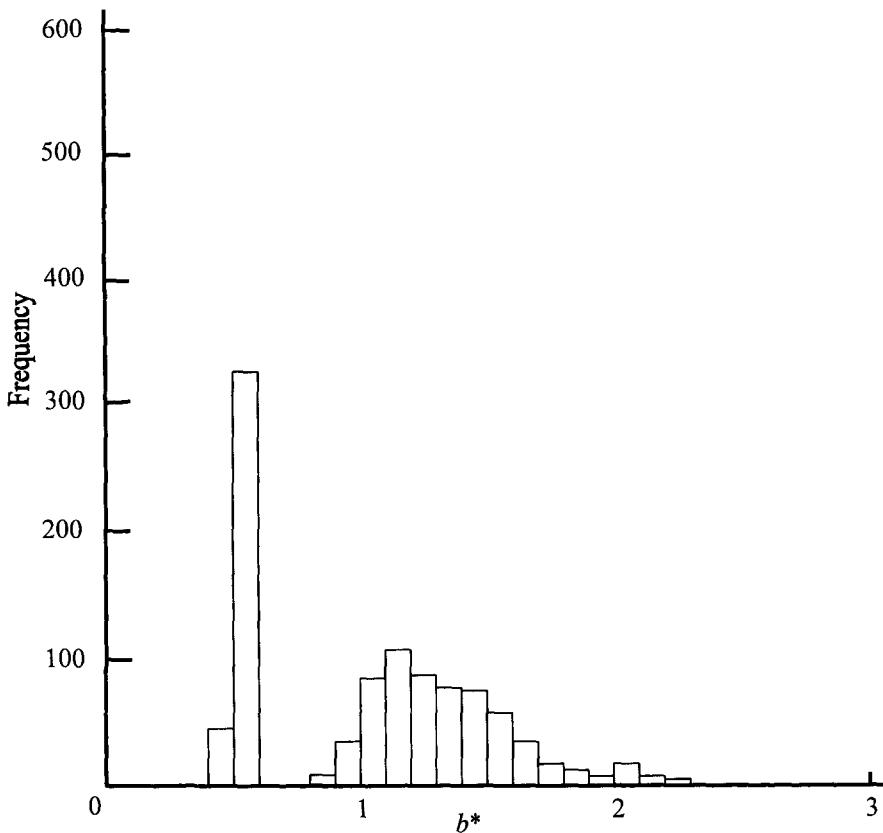
estimator for a different sample of 1000 gave an interval (0.50, 2.37), rather longer than the theory-based interval (0.50, 1.93) obtained in Example 8.3. Figures 11.1 and 11.2 show histograms of the 1000 bootstrap estimators of  $\beta$  obtained for each method using a class interval of 0.1. The class interval for 0.8 to 0.9, for example, contains all  $b^*$  such that  $0.8 < b^* \leq 0.9$ , with similar inequalities applying to other intervals.

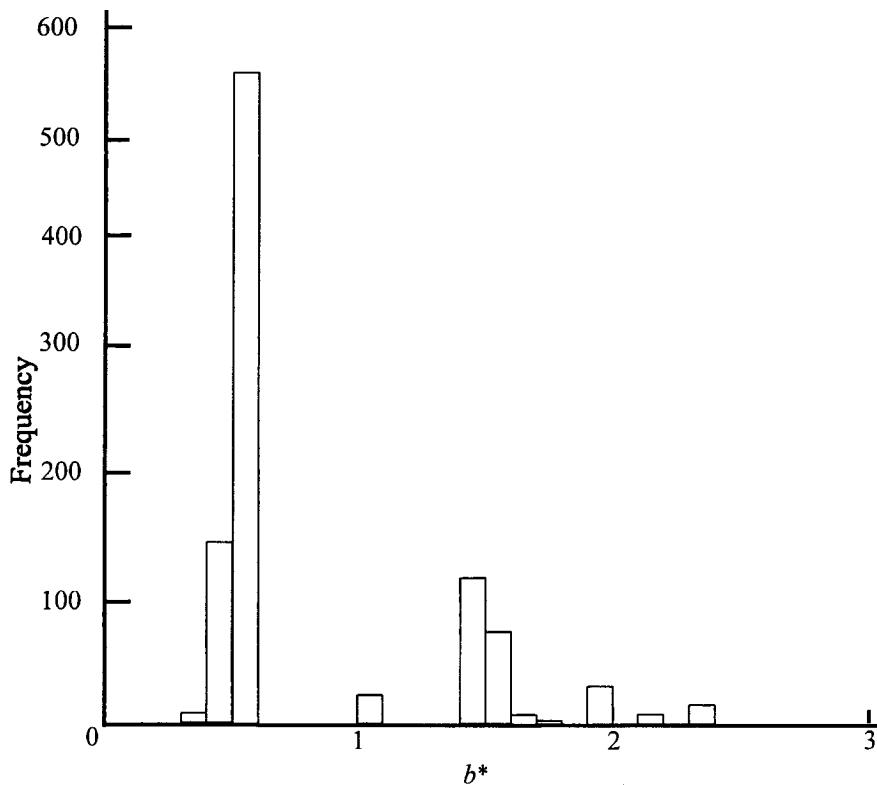
*Conclusions.* Owing to appreciable discontinuities in possible values of the  $b^*$  (especially in the case of the Theil-Kendall estimator) not too much reliance should be placed in confidence intervals for so small a value of  $n$ . The most interesting feature of the histograms is the evident bimodality (or in the case of the Theil-Kendall estimator perhaps multimodality). A moment's reflection would suggest that nearly all bootstrap samples that do not include the aberrant point (6, 11.1) give a  $b^* \leq 0.6$  and the breaks in the histograms between 0.6 and 0.8 in both Figures 11.1 and 11.2 suggest that values of  $b^* > 0.8$  arise from samples where the point (6, 11.1) occurs at least once. The fact that  $b^* \leq 0.6$  for only 36.2 per cent of the samples for least squares while  $b^* \leq 0.6$  for 71.5 per cent of all samples using Theil-Kendall reflects the more robust nature of the Theil-Kendall estimator. Most samples in which the point (6, 11.1) occurs only once are likely to have little influence upon the median of the  $b_{ij}$ , whereas only one occurrence of this point in a sample will influence the least squares estimator appreciably. In Section 8.1.4 we pointed out that the least squares estimator can be expressed as a weighted mean of the  $b_{ij}$  and the robustness properties noted here are broadly in line with our findings on influence in [Section 11.2.2](#).

*Comments.* 1. In the case of the bootstrap because of resampling the number of  $b_{ij}$  in many bootstrap samples will be less than the total of 21 for the original sample because of repeated values and also many of the pairwise estimates will be identical if there are several repeated values. These factors may also influence robustness in a rather complicated manner. With either estimator in a very extreme case the bootstrap sample may consist of 7 replicates of the same point in which case the estimate of  $\beta$  is undefined both for least squares and the Theil-Kendall method.

2. The results discussed here are based on one sample of 1000 for least squares estimators and on another sample of 1000 for Theil-Kendall estimators. It would of course be feasible and indeed interesting to compare both estimators for the same sample. We referred above to discontinuities in the Theil-Kendall estimators. [Table 11.1](#) indicates the extent of these for the particular sample we used. These arise because a variety of samples give rise to the same sampling median (which must be an observed  $b_{ij}$  or the mean of two observed  $b_{ij}$ ).

The simple illustrations of bootstrapping given here are meant to convey no more than the flavour of this approach and to indicate in an intuitive way some of its strengths and weaknesses. As we have pointed out the method is most useful in situations where there is no simple analytic method or we are dealing with data that are samples from essentially unknown distributions. Correlation problems and inferences based on ratios of two variables are situations often met





**Figure 11.2** Histogram for 1000 bootstrap Theil-Kendall slope estimators for the data in Example 11.6. Ten values of  $b^*$  exceeded 2.4.

**Table 11.1** Values of 1000 bootstrap slope estimates  $b^*$  and number of times ( $r$ ) that each occurs using the Theil-Kendall estimator for Example 11.6 data.

$b^*$	$r$	$b^*$	$r$	$b^*$	$r$	$b^*$	$r$
0.3	4	0.535	1	0.583	12	1.762	3
0.45	17	0.537	17	0.6	183	1.9	1
0.5	132	0.546	10	1.017	18	1.925	29
0.51	6	0.55	63	1.1	3	1.983	1
0.512	4	0.558	18	1.433	125	2.146	3
0.52	90	0.562	4	1.517	7	2.367	6
0.522	13	0.567	86	1.6	64	3.25	5
0.525	52	0.575	3	1.679	5	6.0	5

adjustments may be made to remove bias and bootstrap methods exist for estimation of bias but in many practical applications bias is more easily estimated by a related but older technique called the jackknife. The idea was introduced specifically for bias estimation by Quenouille (1949), but the name jackknife is due to Tukey (1958). The jackknife will not always work; in particular it will not work well when the median is used as an estimator. Many commonly used estimators are biased except perhaps when rather strong distributional assumptions are made. Indeed one of the few universally unbiased estimators in common use is the sample mean, which is unbiased when used as an estimator of the population mean. A well-known example of a biased estimator is the sample variance  $s^2 = \sum(x_i - \bar{x})^2/n$  as an estimator of a population mean  $\sigma^2$ . Here the bias is easily estimated because it is well known that  $E(s^2) = (n-1)\sigma^2/n = \sigma^2 - \sigma^2/n$  so the bias is  $-\sigma^2/n$ .

Given a sample of  $n$ , for any parameter  $\theta$  for which an estimator is the sample analogue of the parameter being estimated and is denoted by  $\hat{\theta}$  we form  $n$  further estimators by replacing the original sample by a set of samples identical with the original except that one observation is omitted from each in turn. These are called jackknife samples. Thus the  $i$ th jackknife sample is

$$x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n, \quad i = 1, 2, \dots, n.$$

We denote the estimator of  $\theta$  based on the sample with  $x_i$  omitted by  $\hat{\theta}_{(i)}$  and the mean of the  $\hat{\theta}_{(i)}$  by  $\hat{\theta}_{(.)}$ , i.e.  $\hat{\theta}_{(.)} = \sum_i \hat{\theta}_{(i)} / n$ . Finally, the jackknife estimator,  $\hat{\theta}^\dagger$ , of  $\theta$  is

$$\hat{\theta}^\dagger = n\hat{\theta} - (n-1)\hat{\theta}_{(.)} \quad (11.6)$$

and the jackknife estimator of bias is

$$\text{bias}(\hat{\theta}) = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}) \quad (11.7)$$

The motivation for (11.6) and (11.7) is that for both the sample mean and the sample variance as estimators of the corresponding population parameters equation (11.6) gives unbiased estimators of these parameters and (11.7) correctly estimates the bias of the sample mean as estimator of the population mean as zero and that of the sample variance as estimator of the population variance as  $-\sigma^2/n$ . These results are easily verified by calculating expectations. In general (11.7) does not reduce the bias of every estimator to zero nor

does (11.6) always provide an unbiased estimator of any parameter. However, the latter will usually reduce any bias in  $\hat{\theta}$ .

The jackknife and its use in association with the bootstrap to reduce bias and make other improvements to bootstrap estimation is discussed in detail both by Efron and Tibshirani (1993, especially Chapter 11) and by Davison and Hinkley (1997).

Another widely used technique that bears some resemblance to the bootstrap is cross-validation. In fields like regression and classification methods, parameter estimation is often a part of the model building process to produce a model that will be used to make predictions when new data become available. One source of concern is how good these predictions will be when applied to the new data. Ideally we should aim to use all available data to fit a model and then take further samples from the original population and see how well our fitted model acts as a predictor. This may be costly or impossible, in which case an alternative is cross-validation. For this the data are divided into two or more portions and a model is fitted in turn to all but one of these portions. The prediction error of the fitted model when it is fitted to the omitted portion is then calculated. Each portion is omitted in turn and a combined estimated prediction error is obtained. An extreme case especially suited to relatively small data sets is the leave-one-out cross-validation procedure where each of  $n$  observations is omitted in turn and a model is fitted to the remaining  $n - 1$  data, a predicted value is then obtained for the omitted observation using that model. For example, in a regression context if, when the  $i$ th observation is omitted the predicted value of  $y_i$  is  $y_{(i)}^*$  the cross-validation prediction error is defined as  $\sum_i (y_i - y_{(i)}^*)^2/n$ . The computational procedure is similar to that for the jackknife but the objective is different. A more detailed description of cross-validation and its relationship to the bootstrap is given by Efron and Tibshirani (1993, Chapter 17).

## 11.4 M-ESTIMATORS AND OTHER ROBUST ESTIMATORS

Throughout this book we have met estimators that show varying degrees of tolerance to breakdowns in assumptions especially those associated with a few observations being out of line with the general patterns. We indicated in [Section 11.2](#) and elsewhere that the median is generally more robust than the mean in the presence of a few rogue observations - behaviour explained by the influence

functions. The price to be paid for this robustness may be an increase in the standard error of our estimator or perhaps the introduction of bias in procedures such as the bootstrap. There are several compromise estimators that retain some of the desirable properties of the mean as an estimator but either down-weight or eliminate the more extreme observations. Two such estimators are the so-called **trimmed mean** in which we eliminate a proportion (typically the 10 per cent most extreme observations) and base inferences on the mean of the remaining observations and the **Windsorized mean** where extreme observations are shrunk to the value of the largest remaining observation. We do not discuss these methods in detail but these and related ones are fully covered in Barnett and Lewis (1994) in a range of contexts. The mean, median, trimmed mean and Windsorized mean are members of a class of estimators sometimes called **L-estimators**. Here  $L$  stands for linear, the name being given because if we write the  $i$ th ordered sample value  $x_{(i)}$  then any of these estimators take the form

$$\mu^* = \sum_i w_i x_{(i)}$$

where the  $w_i$  are weights. Putting  $w_i = 1/n$  for all  $i$  gives  $\mu^*$  as the sample mean. If  $n$  is odd say  $n = 2m + 1$  and we put  $w_{m+1} = 1$  and  $w_i = 0$  if  $i \neq m + 1$  then  $\mu^*$  is the sample median. For  $n = 2m$  the median is the mean of the middle pairs of values so the weights are  $w_m = w_{m+1} = \frac{1}{2}$  and  $w_i = 0$  otherwise. For a trimmed mean where the  $k$  largest and  $k$  smallest observations are trimmed the weights are  $w_i = 0$  if  $i \leq k$  or if  $i \geq n - k + 1$ , and  $w_i = 1/(n - 2k)$  otherwise. Other patterns of trimming are possible (see e.g. Exercise 11.6). For Windsorization if  $k$  values in each tail are shrunk to  $x_{(k+1)}$  and  $x_{(n-k)}$  respectively then  $w_i = 0$  if  $i \leq k$  or if  $i \geq n - k + 1$ , while  $w_{k+1} = w_{n-k} = (k + 1)/n$ , otherwise  $w_i = 1/n$ . Clearly since they include the mean L-estimators are not necessarily robust. They tend to gain robustness when extreme order statistics are downweighted. In passing we note that the median is a special case of the trimmed mean in which all but one observation is trimmed if  $n$  is odd and all but two are trimmed if  $n$  is even.

Inevitably there are often doubts about whether certain observations are or are not in some way distorted relative to the bulk of the observations. Even if there are reasonable models to cover such discrepancies there may still be either one or two gross departures from the general pattern or alternatively a large number of small perturbations, these often arising as a consequence of rounding or using measuring devices of limited accuracy. This has

led to the development of procedures that are almost as good as the best available procedures when there are no such disturbances and are little affected by just a few major or many minor disturbances.

The Huber-Hempel  $M$ -estimators are one such class. The ' $M$ ' indicates that they are like optimal maximum likelihood estimators when these are appropriate and are little disturbed by either a few grossly aberrant observations or by small perturbations in many observations. For all but trivial problems their use requires adequate computer programs. We demonstrate the basic ideas for estimation of the mean of a symmetric distribution where the maximum likelihood estimator is the sample mean.

In particular given a sample of  $n$  observations  $x_1, x_2, \dots, x_n$  from a normal distribution this is equivalent to least squares estimation where the estimator  $\hat{\mu}$  is the value of  $\mu$  that minimizes

$$U(x, \mu) = \sum_i (x_i - \mu)^2, \quad (11.8)$$

i.e. the sum of squares of deviations of the  $x_i$  from  $\mu$ . The influence function (11.2) shows that for any one observation  $\hat{\mu}$  is linearly dependent on the distance of that observation from the true mean. Thus an outlier  $z$  lying six standard deviations from the mean will have twice the influence of one lying three standard deviations from the mean.

Huber (1972) and others developed  $M$ -estimators to cope with the possible presence of a few outliers where one really wanted inferences applicable to the remaining data in situations where specific maximum likelihood estimators were available for the remaining data. The proposed estimators have a built-in mechanism for reducing the effect of any outliers but apart from this they are exactly or almost identical to maximum likelihood estimators.

The function  $(x - \mu)^2$  in (11.8) is an example of what is called a distance measure because it is a measure (the square) of the distances of the  $x_i$  from  $\mu$ . The complete function in (11.8) is called a distance function. Another example of a distance function is the function  $V(x, \mu) = \sum_i |x_i - \mu|$ , i.e. the sum of the absolute distances, which is minimized by setting  $\mu^* = \text{med}(x_i)$ . It is well known that in the case of  $U$  in (11.8) we find the minimum by differentiating with respect to  $\mu$  and equating that derivative to zero leading to the *normal* or *estimating* equation

$$\sum_i [-2(x_i - \mu)] = 0$$

with solution  $\hat{\mu} = (\sum_i x_i)/n = \bar{x}$ , the sample mean.

In general a distance measure is defined as a function  $d(t)$  such that for any  $t$

- (i)  $d(t) \geq 0$ ,
- (ii)  $d(t) = d(-t)$ ,
- (iii) the derivative  $\psi(t) = d'(t)$  is a nondecreasing function of  $t$  for all  $t$ .

However, condition (iii) is relaxed for some M-estimators.

For estimating the mean,  $\mu$ , of a symmetric distribution Huber (1972) proposed a function  $d(t)$  such that for some fixed  $k > 0$ ,

$$\begin{aligned} d(t) &= \frac{1}{2}t^2 && \text{if } |t| \leq k, \\ &= k|t| - \frac{1}{2}k^2 && \text{if } |t| > k. \end{aligned}$$

The derivative  $\psi(t) = d'(t)$  is

$$\begin{aligned} \psi(t) &= -k && \text{if } t < -k, \\ &= t && \text{if } |t| \leq k, \\ &= k && \text{if } t > k. \end{aligned}$$

We estimate  $\mu$  by the value  $\mu^*$  that satisfies the normal equation

$$\sum_i \psi(x_i - \mu) = 0. \quad (11.9)$$

If  $k \rightarrow \infty$ , then for all  $x$ ,  $\psi(x_i - \mu) = x_i - \mu$  and (11.9) gives the sample mean as the appropriate estimator, equivalent to the maximum likelihood estimator in the normal case. For finite  $k$  the form of  $d(t)$  implies that for all  $x_i$  for which  $|x_i - \mu| \leq k$  we minimize a function equivalent to that for least squares while for  $|x_i - \mu| > k$  we minimize a linear function of absolute differences. A suitable value for  $k$  needs to be chosen. In this simple problem practical experience has indicated that a useful choice is one such that the interval with end points given by  $\text{med}(x_i) \pm k$  contains between 70 and 90 per cent of all observations. In general (11.9) must be solved iteratively, and when  $k$  is chosen we proceed by first rewriting that equation in the form

$$\sum_i \frac{\psi(x_i - \mu)}{x_i - \mu} (x_i - \mu) = 0. \quad (11.10)$$

Setting  $w_i = \psi(x_i - \mu)/(x_i - \mu)$ , (11.10) becomes  $\sum_i w_i (x_i - \mu) = 0$  with solution

$$\mu^* = [\sum_i (w_i x_i)] / (\sum_i w_i). \quad (11.11)$$

The weights are functions of the unknown  $\mu^*$  so we need an estimate  $\mu_0$  of  $\mu^*$  to calculate initial weights  $w_0$  and use these in (11.11) to calculate a new estimate  $\mu_1$ , repeating this procedure until convergence, which is usually achieved in a few iterations. For even moderate sample sizes computation is tedious without a suitable computer program, but we illustrate the steps for small data sets.

*Example 11.7*

*The problem.* Obtain and compare Huber  $M$ -estimators of the mean for the three data sets

$$\begin{aligned} 0, & \quad 1.2, \quad 2.3, \quad 3.8, \quad 5.2, \quad 7.1, \quad 7.9, \quad 8.5 \\ 0, & \quad 1.2, \quad 2.3, \quad 3.8, \quad 5.2, \quad 7.1, \quad 7.9, \quad 19.7 \\ 0, & \quad 1.2, \quad 2.3, \quad 3.8, \quad 5.2, \quad 7.1, \quad 7.9, \quad 115.5 \end{aligned}$$

*Formulation and assumptions.* A suitable choice of  $k$  is required after which the iterative procedure outlined above is carried out for each data set.

*Procedure.* We illustrate the procedure for the second data set, leaving the reader to follow through the similar steps for the other sets (Exercise 11.7). The median of the second set is  $\frac{1}{2}(3.8 + 5.2) = 4.5$ . If we take this as our first estimate of  $\mu$  and choose  $k = 4$  the interval  $4.5 \pm 4$  includes 75 per cent of all observations and if we choose  $k = 5$  the interval  $4.5 \pm 5$  includes 87.5 per cent. Both are within the suggested range 70 to 90 per cent. We illustrate the procedure with  $k = 5$ . Only the last point falls outside the interval  $4.5 \pm 5$  (i.e. -0.5 to 9.5) and since we set  $\psi(x_i - \mu) = (x_i - \mu)$  for observations in the interval (-0.5, 9.5) it follows that  $w_i = 1$  for all but the observation  $x_8 = 19.7$ . For that observation  $\psi(x_i - \mu) = k = 5$ , so that  $w_8 = 5/(19.7 - 4.5) = 0.3289$ , whence our new estimator,  $\mu_1$ , is

$$\mu_1 = \frac{0 + 1.2 + 2.3 + 3.8 + 5.2 + 7.1 + 7.9 + 0.3289 \times 19.7}{1 + 1 + 1 + 1 + 1 + 1 + 0.3289} = 33.3733/7.3289 = 4.554$$

For the next iteration we retain the same  $k = 5$ . This gives full weight  $w = 1$  to all observations in the interval  $4.554 \pm 5$ , i.e. (-0.446, 9.554). The weight for  $x_8 = 19.7$  is now  $w_8 = 5/(19.7 - 4.554) = 0.3301$  whence a similar calculation to that above gives a new estimate  $\mu_2 = 4.639$ . For the next iteration it is again clear that only  $x_8$  has a weight less than 1 and this weight is easily seen to be  $w_8 = 5/(19.7 - 4.639) = 0.3320$  and with this weight our next estimate  $\mu_3 = 4.643$ . The new adjusted weight for  $x_8$  is  $w_8 = 5/(19.7 - 4.643) = 0.3321$ , leading to the new estimate  $\mu_4 = 4.643$ . To this degree of accuracy there is no improvement on the previous iteration so we conclude  $\mu^* = 4.643$ .

*Conclusion.* The  $M$ -estimator of the population mean based on the second data set is 4.643.

*Comments.* 1. With the same choice of sample median and  $k$  for the first data set it is immediately clear that there are no observations for which  $w_i \neq 1$  and that our estimator is thus the sample mean, i.e.  $\mu^* = 4.5$ . For the third data set

similar calculations to that for the second set with  $x_8$  there replaced by  $x_8 = 115.5$  lead after several iterations to the estimate  $\mu^* = 4.643$ .

2. As indicated under *Procedure* a choice  $k = 4$  may not be unreasonable. In Exercise 11.7 we ask that this choice be used.

3. For the three data sets here sample means are respectively 4.5, 5.9 and 17.875 indicating the strong influence of the outliers. The above M-estimation procedure leads to 4.5, 4.643, 4.643, as estimates of the population mean indicating a dramatic effect of downweighting the 'outliers'.

4. Other possible estimators in this case include trimmed means, Windsorization or the median. The sample median for each set is 4.5, while a trimmed mean excluding  $k = 1$  observation from each tail is easily shown to lead to estimators 4.58 in each case. Similarly, in each case the Windsorization estimates are 4.575 (Exercise 11.8).

Different choices of  $k$  or different amounts of trimming or Windsorization will give different estimates but in practice if estimators are robust against the outliers in the data there should be little difference between rational choices, i.e. choices that downweigh rogue observations substantially or even eliminate them from the computation while directly or indirectly giving strong support to good observations. We might at first thought conclude that the median does not obey these criteria, its computation taking into account at most two observations; however all observations do have a role in determining the median because the ordering of the observations determines which observations enter computationally into the calculation of the median. Only if the observations are sampled from a symmetric distribution, however, will the sample median be a sensible estimator of the population mean (and then only if the latter exists!). In many practical situations, even in the absence of any rogue observations the median estimator will, however, have a greater standard error than that for the mean.

Sprent (1998, Section 3.5) outlines several practical considerations that should influence the choice of  $k$  when using *M*-estimators. In particular if  $k$  is too small only a few observations get full weight and the estimator may be strongly influenced by rounding or grouping effects in those fully weighted observations. Also there may be unsatisfactory features relating to change of scale that can be overcome by using different weights to those used above. Numerous alternative distance functions have been proposed, but we omit details here. Many aspects of these procedures are discussed by Andrews et al. (1972).

Andrews (1974) introduced an *M*-estimator for robust linear multiple regression and Härdle and Gasser (1984) apply a similar type of estimator for fitting nonlinear curves.

Thomas (2000) describes the use of the bootstrap to determine the precision of robust estimates of centrality, illustrating the use of appropriate methods for ten data sets that are either highly skewed or contain outliers.

## 11.5 FIELDS OF APPLICATION

We have already referred to Efron and Tibshirani (1993) and Davison and Hinkley (1997) as sources for examples on many and varied applications of the bootstrap. We indicate one example here.

### Ratio estimates

Efron and Tibshirani (1993, Section 10.3) discuss use of the bootstrap in a problem involving a test criterion for bioequivalence used by the US Food and Drug Administration that is based on ratios. The procedure is useful here because there are virtually no analytic results for the properties of the ratio that are of interest.

All statisticians and experimenters handling more than a few data sets come face to face with data that well may be in some way contaminated in the sense that some of the observations do not appear to fit a model that is reasonable for the rest of the data. Robust methods may then have a valuable role in making inferences that are relevant to the uncontaminated data. Again we just indicate one situation where this may be the case.

### Laboratory determinations

Very often two or more laboratories are asked to carry out the same chemical determination on samples from the same population, e.g. to determine the percentage fat content in samples of milk or the amount of some contaminant, e.g. lead in 100 gm samples from zinc ingots. Precision may vary from laboratory to laboratory or some laboratories may make consistent errors (bias) in their measurements. Robust methods of estimation such as the use of  $M$ -estimators might be used to reduce the influence of extreme results from laboratories producing readings with low precision. Conventional parametric or nonparametric tests comparing results from different laboratories for a shift in mean or median may be useful to detect biases.

## 11.6 SUMMARY

Observations that cause surprise in the sense that they appear far removed from the bulk of the data or seem to be inconsistent with some specified model are generally called outliers. In any given data set what is classed as an outlier may be context dependent ([Section 11.2.1](#)). The appropriate action to be taken to deal with outliers depends both upon their nature and the objectives of the investigation in which they occur. Erroneous observations should be corrected if possible or else rejected. If it is believed outliers represent a small proportion of units present in the population the action to be taken may depend on whether one wants to make inferences relevant to the complete population or only to the population where the small proportion giving rise to outliers are to be ignored.

Influence studies ([Section 11.2](#)) indicate both the extent to which outliers may influence sample characteristics such as the mean or median and also explain why some estimators (e.g. the median) will be less affected by substantial numbers of outliers than others (e.g. the mean).

The bootstrap ([Section 11.3](#)) is a resampling method of inference that relies for its usefulness on the assumption that for not-too-small samples the sample cdf is a reasonable approximation to the population cdf. It differs from randomization or permutation distribution theory in that it involves sampling with replacement whereas the latter involves sampling without replacement. It is especially useful in situations where there are few analytic results for exact inference or where formulation of a precise model is difficult. Some of the problems with bias that are associated with certain bootstrap estimators may be eased by use of the jackknife ([Section 11.3.5](#)).

Two important categories of robust estimators ([Section 11.4](#)) are certain  $L$ -estimators such as trimmed or Windsorized means and the median and  $M$ -estimators which behave very like the relevant maximum likelihood estimators when applied to uncontaminated data but reduce the influence of outliers when these are present, making the method particularly suitable when the status of outliers is uncertain but we want to make inferences applicable to the population from which the bulk of the data was sampled.

## EXERCISES

- 11.1 Use the test given in [Section 11.2.2](#) to test for outliers in each of the following data sets:

*Set A*    3.2    4.9    1.3    9.7    12.9    -6.3    4.2    22.5    0.3    -7.1  
*Set B*    1    35    71    13    45    18    91    34    777    29    452

- 11.2 Use the test given in [Section 11.2.2](#) to test for outliers in the data set used in [Example 11.2](#).
- 11.3 Smith and Naylor (1987) give the following data for strengths of 15 cm lengths of glass fibre and suggest that the two smallest observations may be outliers. Does the test proposed in [Section 11.2.2](#) confirm this?

0.37 0.40 0.70 0.75 0.80 0.81 0.83 0.86 0.92 0.92 0.94 0.95  
0.98 1.03 1.06 1.06 1.08 1.09 1.10 1.10 1.13 1.14 1.15 1.17  
1.20 1.20 1.21 1.22 1.25 1.28 1.28 1.29 1.29 1.30 1.35 1.35  
1.37 1.37 1.38 1.38 1.40 1.40 1.42 1.43 1.51 1.53 1.61

- 11.4 A sample of paired observations of X, Y are

X	1	2	3	5	8
Y	3	7	9	7	12

Use a bootstrap based on at least 40 samples to obtain an estimate of the bootstrap standard error of the estimator  $\bar{x}/\bar{y}$  for the ratio of population means. In addition use 1000 bootstrap samples to obtain 95 per cent quantile-based confidence intervals for the population ratio. Appropriate computer software will be needed for the latter computation.

- 11.5 For the data in [Exercise 11.4](#) use the jackknife to estimate the bias in the estimator  $\bar{x}/\bar{y}$ .
- 11.6 Calculate the mean of the data in [Example 11.2](#) and also the trimmed means trimming the top and bottom 10 per cent (deciles) and 25 per cent (quartiles) of all observations. Explain any difference between these estimates of centrality. Which do you prefer and why?
- 11.7 For the first and third data sets in [Example 11.7](#) verify the values for the  $M$ -estimators quoted in that example when  $k = 5$ . Also obtain the corresponding estimates when  $k = 4$ .
- 11.8 Verify the values for the median, trimmed mean and Windsorized estimates given in [Comment 4](#) in [Example 11.7](#).

# Appendix

## BADENSCALLIE BURIAL DATA

Several examples in this book use data on the age at death of male members of four Scottish clans in the burial ground at Badenscallie in the Coigach district of Wester Ross, Scotland. The data were collected in June 1987. Clan names have been changed but the records are as complete as possible for four clans. There were a few missing values because names or dates were unreadable on a few headstones and several headstones appeared to be missing. Minor spelling variations, especially those of M', Mc and Mac, were ignored. Ages are given for complete years, e.g. 0 means before first birthday and 79 means on or after 79th but before 80th birthday, according to the information on the tombstone. Ages are given in ascending order within each clan.

### *McAlpha* (59 members)

0	0	1	2	3	9	14	22	23	29	33	41	41	42	44	52	56	57
58	58	60	62	63	64	65	69	72	72	73	74	74	75	75	75	77	77
78	78	79	79	80	81	81	81	82	82	83	84	84	85	85	86	87	87
88	90	92	93	95													

### *McBeta* (24 members)

0	19	22	30	31	37	55	56	66	66	67	67	68	71	73	75	75	78
79	82	83	83	88	96												

### *McGamma* (21 members)

13	13	22	26	33	33	59	72	72	72	77	78	78	80	81	82	85	85
85	86	88															

### *McDelta* (13 members)

1	11	13	13	16	34	65	68	74	77	83	83	87					
---	----	----	----	----	----	----	----	----	----	----	----	----	--	--	--	--	--

# Solutions to odd-numbered exercises

Many exercises in this book are open ended in that they require, as do most practical statistical problems, not simply numerical outcomes (a statistic, a  $P$ -value, confidence interval, etc.) but an interpretation of such concepts in the context of the real world problem giving rise to the data. The brief summaries below are not solutions in this sense, but rather a guide to indicate appropriate calculations. In many cases valid alternative tests or estimation procedures might have been used and these will often, but not invariably, lead to similar conclusions. Bearing such points in mind it is hoped this section will help overcome any difficulties the reader may have with the odd-numbered exercises.

## *Chapter 1*

**1.1** 2.55. **1.3** Open interval (156, 454). See [Example 2.10](#) for ways to refine the limits. **1.5**  $S = 13$ ;  $P = \Pr(S \leq 13) = 7/126 \approx 0.056$ , indicating slight evidence against  $H_0$  that might suggest a larger experiment desirable. **1.7** Open interval  $(-\infty, 454)$  approx. 98 per cent coverage. **1.9** If  $X$  is  $N(\mu, \sigma^2)$  then  $\Pr(X < \mu - \sigma) = \Pr(Z < -1) = 0.159$  because  $Z$  is  $N(0, 1)$ . Thus if  $\sigma > \mu$  there is high probability of at least one negative value in all but small samples. **1.11**  $P = 1/4845$ .

## *Chapter 2*

**2.1**  $P = 0.797$ . No evidence against  $H_0$ . **2.3** Student lifestyles not necessarily similar to those of others, not all ages and only one sex represented. **2.7** (i) Interval [1, 5] gives exact 96.3 per cent coverage using StatXact convention (see manual) for zero differences. (ii) Interval [3, 5] gives exact 91.5 per cent coverage using same convention but any extension of limits gives coverage exceeding 95 per cent. **2.9**  $P = 0.13$  (two tail). 95 per cent Wilcoxon interval is (200.5, 433). Normal theory interval (227.4, 405.9). **2.11** For  $H_0: \theta = 9$  see [Example 2.9](#). For  $H_0: \theta = 7.5$  one-tail  $P = 0.18$ . **2.13** Using method of [Example 2.12](#) one-tail  $P = 0.017$ . Strong evidence treatment is beneficial. **2.15** Two-tail exact  $P = 0.0093$  for Wilcoxon (asymptotic  $P = 0.0104$ ). For modified van der Waerden scores exact  $P = 0.0045$ , asymptotic  $P = 0.0052$ . These results are based on StatXact convention for dealing with tied ranks. **2.17**  $S_+ \leq 7$  (rounding up). **2.19** One-tail test appropriate. For sign test  $P = 0.090$ , for Wilcoxon test  $P = 0.011$ . Symmetry assumption not unreasonable so some evidence claim is not justified. **2.21** Sign test interval is (11, 36). Wilcoxon interval is (14, 37). Some reservations about assumption of symmetry due to observed time 145. A valid population might be parking times of cars in that area on observation day. If this were the Friday before Christmas parking times might have a very different distribution from that for times on a Friday in August.

## *Chapter 3*

**3.1** Denoting by  $r$  the minimum number of damaged fruit in a batch of  $n$  which implies rejection to meet producer's risk  $P = 0.10$  if  $p = 0.01$  and by  $pwr$  the corresponding power if  $p = 0.03$  one finds

$n$	100	200	300	400	500
$r$	3	5	6	8	9
$pwr$	0.58	0.72	0.89	0.92	0.96

Because of large discontinuities in exact  $P$ -values the exact power fluctuates markedly for small changes in sample size. A power of 0.95 is needed to meet the specified consumer's risk. Depending on available software the reader should experiment with several values of  $n$  between 400 and 500. For  $n = 440$ , for example, StatXact indicates that rejection if there are 8 or more damaged fruit gives an exact producer's risk  $P = 0.078$  and a power 0.954 when the alternative is  $p = 0.03$ , implying an exact consumer's risk of  $P^* = 1 - 0.954 = 0.046$ . Asymptotic results show reasonable agreement. **3.3** For smaller sample two-tail  $P = 0.24$ ; for larger sample two-tail  $P < 0.001$ . Power increased with sample size. The more informative 95 per cent confidence intervals for proportions approving are (0.17, 0.59) and (0.27, 0.40). **3.5** Cox–Stuart trend test gives two-tail  $P = 0.125$  corresponding to 6 plus and 1 minus. A slight suggestion of a trend but a larger sample would be needed to confirm this. **3.7** Both tests suggest strong evidence against normality.  $P \approx 0.011$  (Lilliefors').  $P = 0.0009$  (Shapiro–Wilk). **3.9**  $P = 0.165$ . No substantial evidence against  $H_0$ . **3.11** Test for runs above and below median gives a two-tail  $P = 0.0634$ , suggesting slight evidence against the null hypothesis of no effect. Runs above the median dominate for those tested later. Cox–Stuart test may be less powerful as monotonicity assumption is unlikely to be justified (two-tail  $P = 0.29$ ). **3.13** Hodges–Ajne  $P = 0.94$ . No evidence against hypothesis that all types of question are equally likely. **3.15** Sign test two-tail  $P = 0.109$ . Wilcoxon  $P = 0.049$ . Some evidence against  $H_0$ . No reservations.

## *Chapter 4*

**4.3** Two-tail  $P = 0.045$  indicating moderate evidence of median difference. Wilcoxon 95 and 99 per cent confidence intervals are (0.05, 0.85) and ( $-0.25$ , 1.05) and corresponding normal theory intervals are (0.03, 0.85) and ( $-0.15$ , 1.02). **4.5** Sign test appropriate; four M corresponds to 4 plus. Two-tail  $P = 0.049$  suggests moderately strong evidence that fathers show better understanding. **4.7** No strong evidence of change in attitudes, two-tail  $P = 0.126$ . **4.9** Exact two-tail  $P = 0.0029$  indicating strong evidence against  $H_0$ :  $p = 0.75$ . **4.11** For Wilcoxon test exact two-tail  $P = 0.125$  so no strong evidence against  $H_0$ .  $t$ -test  $P = 0.063$ . **4.13** Wilcoxon two-tail  $P = 0.031$  so fairly strong evidence against equal loss. **4.15** Wilcoxon two-tail  $P = 0.271$ . 95 per cent confidence interval ( $-2.95$ , 0.95). No evidence of one organization consistently returning higher percentages. **4.17** When  $\theta = 2$ ,  $p_1 = 0.8784$ . **4.19** For medians, 1, 2 we require  $\lambda = 0.6931$ , 0.3465 respectively, whence  $p_0 = 0.5$  and  $p_1 = 0.707$ . Sample size of 35 is needed.

## *Chapter 5*

**5.3**  $P^* = 0.0029$ . **5.5** WMW two-tail  $P = 0.0549$ ; not formally significant at 5 per cent level but a slight indication that hard specimens may be associated with lower temperature. Test valid if it is reasonable to suppose any difference is a median shift or dominance by one distribution. **5.7** Exact two-tail  $P = 0.208$  ( $S_m = 101$ ). **5.9** For WMW two-tail  $P = 0.139$ ; 95 per cent confidence interval  $(-0.2, 1.9)$ ; for  $t$ -test  $P = 0.097$ , interval is  $(-0.17, 1.83)$ . Little reason to doubt validity of normal theory test and result reflects slightly higher efficiency. **5.11** For two independent sample analysis two-tail  $P = 0.272$  using WMW test. Differences between individuals in each field swamps any systematic difference between fields for individuals. Analysis used here is inappropriate for these data since it ignores pairing. **5.15**  $P = 0.352$ ; hardly surprising in the light of findings in [Exercise 5.9](#). **5.17** WMW two-tail  $P = 0.099$ .  $t$ -test  $P = 0.18$ . Smaller  $P$  for WMW probably because WMW more robust against greater spread of Species B values. **5.19** For asymptotic WMW test allowing for ties two-tail  $P = 0.004$  suggesting strong evidence of shorter waiting times in 1983. **5.21** For asymptotic WMW test allowing for ties two-tail  $P = 0.280$  so no evidence of difference in average sentence length. Do the data suggest WMW test may not be appropriate? Might there not be distributional differences in features other than average sentence length? **5.23** Appropriate runs test one-tail  $P = 0.72$ . (This is probability of 11 or less runs under  $H_0$ .) No evidence of clustering.

## *Chapter 6*

**6.1**  $F = 6.78$  with 2, 15 degrees of freedom.  $P = 0.008$ . Similar to Kruskal–Wallis test. **6.3**  $Z = 2.85$ ,  $p = 0.002$ . **6.7**  $F = 2.93$  with 3, 5 degrees of freedom.  $P = 0.139$ . **6.9** Jonckheere–Terpstra  $U = 90$ . Exact one-tail  $P = 0.056$ . Rather slender evidence supporting Professor's claim. **6.11**  $T = 6.47$ ,  $P = 0.089$ . Little firm evidence of consistent preferences. **6.15** (i) StatXact Monte Carlo estimated  $P \approx 0.267$  (Asymptotic  $P = 0.262$ ). (ii) Exact  $P = 0.013$  (Asymptotic  $P = 0.020$ ). **6.17** StatXact Monte Carlo estimated  $P \approx 0.111$  (Asymptotic  $P = 0.109$ ). **6.19** Multiple comparison test on ranks indicates numbers significantly higher than controls for gibberellic acid, indole acetic acid and adenine sulphate. **6.21** Jonckheere–Terpstra  $U = 119$ . Exact or asymptotic  $P < 0.0001$ .

## *Chapter 7*

**7.5** No evidence of trend using Cox–Stuart test. Low power because information used is less than that for other coefficients. **7.7**  $r_s = 0$ ,  $t_k = 0.048$ . No evidence of association. **7.9**  $r_s = 0.90$ , exact two-tail  $P = 0.006$ .  $t_b = 0.78$ , exact two-tail  $P = 0.017$ . **7.11** (i)  $W = 0.367$ , Monte Carlo estimate  $P \approx 0.0012$  (Asymptotic  $P = 0.0025$ ). (ii)  $W = 0.549$ , Monte Carlo estimate  $P < 0.0005$  (Asymptotic  $P = 0.0005$ ). Examiner 5 consistently awards higher marks. **7.13** Asymptotic one-sided  $P = 0.0385$ .

## *Chapter 8*

**8.3**  $y = -13.2 + 2.375x$ . Plot indicates nonlinearity. **8.7** Plot indicates clear nonlinearity. One possibility would be to fit a monotonic regression. **8.9** WMW  $U = 5$ , exact one-tail  $P = 0.075$ . Hardly enough evidence to reject equality of slopes but samples are small for suggested test.

## *Chapter 9*

**9.9** Pearson chi-squared  $X^2 = 25.34$ . Asymptotic  $P = 0.0047$ . One Monte Carlo estimate of exact  $P$  using 10000 samples gave  $P = 0.0048$ . Likelihood ratio or Fisher–Freeman–Halton test also appropriate. **9.11** Any valid test (Pearson, Likelihood ratio, Fisher, Jonckheere–Terpstra) indicates overwhelming evidence ( $P < 0.001$ ) that poor make less use of the service. **9.13** Chi-squared goodness of fit exact  $P < 0.0001$ . Overwhelming evidence that manufacturer's claim not justified. (Hint: under  $H_0$  the probabilities that 0, 1, 2, 3, 4 parts survive have a binomial distribution with  $p = 0.95$ .) **9.15** Yes.  $P < 0.0001$  (from data expected number of positive responses under  $H_0$  is 36.8 per 100 interviewed). **9.19** McNemar  $X^2 = 23$ . Strong evidence of change in attitudes. Exact  $P < 0.0001$ . **9.21** If choices random expected number for each group is 33.33.  $X^2 = 38.11$ , asymptotic  $P = 0.043$ . Some evidence selection not random. First choices for A, B, C, D are respectively 248, 181, 198, 173. Expected numbers 200 for each.  $X^2 = 16.99$ . Asymptotic  $P = 0.0007$ . Strong evidence for preference for A and some dislike of D. **9.23**  $X^2 = 40.48$ , asymptotic  $P < 0.0001$ ;  $G^2 = 13.43$ , asymptotic  $P = 0.266$ . In both tests exact  $P = 0.0182$ . Asymptotic results completely unreliable with so many sparse cells giving rise to low expected frequencies. Different patterns in columns 11, 12 relative to rest of table dominate in determining  $X^2$ ,  $G^2$ ; the former is more susceptible to small expected frequencies in row 1.

## *Chapter 10*

**10.9** Breslow–Day statistic (10.12) is 52.86 and exact and asymptotic  $P < 0.0001$  providing virtually overwhelming evidence that there is not a common odds ratio and thus that a first-order interaction model does not suffice. **10.11** If we accept ordering of parties we might use a linear-by-linear association model with rank scores. This gives an asymptotic two-tail  $P = 0.20$  so no evidence of association. Other tests are possible. One might query whether any association need be monotonic with age. **10.13** For first table exact  $P < 0.0001$  using Pearson, likelihood ratio or Fisher test and for second table exact  $P = 0.0017$ . Breslow–Day test exact  $P = 0.58$ . First-order association model appears adequate. **10.17** Cochran–Armitage test with scores 1, 2, 3, 4, 5 shows overwhelming evidence of increasing trend. One-tail  $P < 0.0001$ .

## *Chapter 11*

**11.1** No outlier in set A. In set B 452 and 777 are outliers. **11.3** None detected. **11.5**  $-0.0088$ . **11.7** With  $k = 4$  estimators are (i) 4.57, (ii) 5.25, (iii) 5.25.

# References

- Adichie, J.N. (1967) Estimates of regression parameters based on rank tests. *Ann. Math. Statist.*, **38**, 894–904.
- Adichie, J.N. (1984) Rank tests in linear models. In *Handbook in Statistics, Vol. 4*, Ed. Krishnaiah, P.B. and Sen, P.K. pp. 229–257. Amsterdam: Elsevier.
- Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons.
- Agresti, A. (1990) *Categorical Data Analysis*. New York: John Wiley & Sons.
- Agresti, A. (1992) A survey of exact inferences for contingency tables. *Statistical Science*, **7**, 131–153.
- Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.
- Aitchison, J.W. and Heal, D.W. (1987) World patterns of fuel consumption; towards diversity and a low cost energy future. *Geography*, **72**, 235–239.
- Ajne, B. (1968) A simple test for uniformity of a circular distribution. *Biometrika*, **55**, 343–354.
- Albers, W. and Akritas, M.G. (1987) Combined rank tests for the two sample problem with randomly censored data. *J. Amer. Statist. Assoc.*, **82**, 645–655.
- Anderson, R.L. (1959) Use of contingency tables in the analysis of consumer preference studies. *Biometrics*, **15**, 582–590.
- Andrews, D.F. (1974) A robust method for multiple linear regression. *Technometrics*, **16**, 523–531.
- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.S., Rogers, W.H. and Tukey, J.W. (1972) *Robust Estimates of Location. Survey and Advances*. Princeton, NJ: Princeton University Press.
- Ansari, A.R. and Bradley, R.A. (1960) Rank sum tests for dispersion. *Ann. Math. Statist.*, **31**, 1174–1189.
- Arbuthnot, J. (1710) An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Phil. Trans. Roy. Soc.*, **27**, 186–190.
- Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.
- Arnold, H.J. (1965) Small sample power for the one sample Wilcoxon test for non-normal shift alternatives. *Ann. Math. Statist.*, **36**, 1767–1778.
- Atkinson, A.C. (1985) *Plots, Transformations and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Clarendon Press.
- Babu, G.J., Rao, C.R. and Rao, M.B. (1992) Nonparametric estimation of specific occurrence exposure rates in risks and survival analysis. *J. Amer. Statist. Assoc.*, **87**, 84–89.
- Bahadur, R.R. (1967) Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, **38**, 303–324.
- Bardsley, P. and Chambers, R.L. (1984) Multipurpose estimation from unbalanced samples. *Appl. Statist.*, **33**, 290–299.

- Barnett, V.D. and Lewis, T. (1994) *Outliers in Statistical Data*. 3rd edn. Chichester: John Wiley & Sons.
- Bartlett, M.S. (1935) Contingency table interactions. *J. Roy. Statist. Soc., Suppl.*, **2**, 248–252.
- Bassendine, M.F., Collins, J.D., Stephenson, J., Saunders, P. and James, O.W.F. (1985) Platelet associated immunoglobulins in primary biliary cirrhosis: a cause of Thrombocytopenia? *Gut*, **26**, 1074–1079.
- Berry, D.A. (1987) Logarithmic transformations in ANOVA. *Biometrics*, **43**, 439–456.
- Best, D.J. (1994) Nonparametric comparison of two histograms, *Biometrics*, **50**, 538–541.
- Best, D.J. and Rayner, J.C.W. (1996) Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics*, **52**, 1153–1156.
- Biggins, J.D., Loynes, R.M. and Walker, A.M. (1987) Combining examination results. *Brit. J. Math. and Statist. Psychol.*, **39**, 150–167.
- Bishop, Y.M.M., Fienberg, S.E. and Holland, P. (1975) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Blomqvist, N. (1950) On a measure of dependence between two random variables. *Ann. Math. Statist.*, **21**, 593–600.
- Blomqvist, N. (1951) Some tests based on dichotomization. *Ann. Math. Statist.*, **22**, 362–371.
- Bodmer, W.F. (1985) Understanding statistics. *J. Roy. Statist. Soc. A*, **148**, 69–81.
- Boos, D.D. (1986) Comparing  $K$  populations with linear rank statistics. *J. Amer. Statist. Assoc.*, **81**, 1018–1025.
- Bowker, A.H. (1948) A test for symmetry in contingency tables. *J. Amer. Statist. Assoc.*, **43**, 572–574.
- Bradley, J.V. (1968) *Distribution-free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.
- Breslow, N.E. and Day, N.E. (1980) *Statistical Methods in Cancer Research. I. The Analysis of Case-Control Studies*. Lyons: WHO/IARC Scientific Publications, No. 32.
- Brookmeyer, R. and Crowley, J. (1982) A confidence interval for the median survival time. *Biometrics*, **38**, 29–41.
- Brown, B.M. (1983) Statistical use of the spatial median. *J. Roy. Statist. Soc. B*, **45**, 25–30.
- Brown, B.M., Hettmansperger, T.P., Nyblom, J. and Oja, H. (1992) On certain bivariate sign tests and medians. *J. Amer. Statist. Assoc.*, **87**, 127–135.
- Büning, H. and Kössler, W. (1999) The asymptotic power of Jonckheere-type tests for ordered alternatives. *Australian and New Zealand J. Statist.*, **41**, 67–72.
- Cantor, A.B. (1996) Sample-size calculations for Cohen's kappa. *Psychol. Meth.*, **1**, 150–153.
- Capon, J. (1961) Asymptotic efficiency of certain locally most powerful rank tests. *Ann. Math. Statist.*, **32**, 88–100.
- Carter, E.M. and Hubert, J.J. (1985) Analysis of parallel line assays with multivariate responses. *Biometrics*, **41**, 703–710.

- Chandra, M., Singpurwalla, N.D. and Stephens, M.A. (1981) Kolmogorov statistics for test of fit for the extreme value and Weibull distributions. *J. Amer. Statist. Assoc.*, **76**, 729–731.
- Chernick, M.R. (1999) *Bootstrap Methods – A Practitioner’s Guide*. New York: John Wiley & Sons.
- Christensen, R.A. (1990) *Log Linear Models*. New York: Springer-Verlag.
- Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatter plots. *J. Amer. Statist. Assoc.*, **74**, 829–836.
- Cochran, W.G. (1950) The comparison of percentages in matched samples. *Biometrika*, **37**, 256–266.
- Cochran, W.G. (1954) Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, **10**, 417–454.
- Cohen, A. (1983) Seasonal daily effect on the number of births in Israel. *Appl. Statist.*, **32**, 228–235.
- Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.
- Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213–220.
- Conger, A.J. (1980) Integration and generalization of kappa for multiple raters. *Psychol. Bull.*, **88**, 322–328.
- Conover, W.J. (1980) *Practical Nonparametric Statistics*. 2nd edn. New York: John Wiley & Sons.
- Conover, W.J. (1999) *Practical Nonparametric Statistics*. 3rd edn. New York: John Wiley & Sons.
- Conover, W.J. and Iman, R.L. (1976) On some alternative procedures using ranks for the analysis of experimental designs. *Commun. Statist.*, **A5**, 1349–1368.
- Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. London: Chapman & Hall.
- Cormack, R.M. (1989) Log-linear models for capture–recapture. *Biometrics*, **45**, 395–413.
- Cox, D.R. and Snell, E.J. (1989) *Analysis of Binary Data*. 2nd edn. London: Chapman & Hall.
- Cox, D.R. and Stuart, A. (1955) Some quick tests for trend in location and dispersion. *Biometrika*, **42**, 80–95.
- Cox, P.R. (1978) *Demography*. 5th edn. Cambridge: Cambridge University Press.
- Cramér, H. (1928) On the composition of elementary errors. *Skand. Aktuarie-tidskrift*, **11**, 13–74, 141–180.
- Dansie, B.R. (1986) Normal order statistics as permutation probability models. *Appl. Statist.*, **35**, 269–275.
- Daniel, W.W. (1990) *Applied Nonparametric Statistics*. 2nd. edn. Boston: PWS-Kent Publishing Company.
- David, F.N. and Barton, D.E. (1958) A test for birth-order effects. *Ann. Hum. Eugenics*, **22**, 250–257.
- Davis, T.P. and Lawrence, A.J. (1989) The likelihood for competing risk survival analysis. *Scand. J. Statist.*, **16**, 23–28.
- Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

- Day, S. and Talbot, D.J. (2000) Editorial: Statistical guidelines for clinical trials. *J. Roy. Statist. Soc. A.*, **163**, 1–3.
- de Kroon, J. and van der Laan, P. (1981) Distribution-free test procedures in two-way layouts: a concept of rank interaction. *Statistica Neerlandica*, **35**, 189–213.
- Dietz, E.J. and Killen, T.J. (1981) A nonparametric multivariate test for trend with pharmaceutical applications. *J. Amer. Statist. Assoc.*, **76**, 169–174.
- Dinse, G.E. (1982) Nonparametric estimation for partially-complete time and type of failure data. *Biometrics*, **38**, 417–431.
- Dinse, G.E. (1986) Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *J. Amer. Statist. Assoc.*, **81**, 328–336.
- Dinse, G.E. and Lagakos, S.W. (1982) Nonparametric estimation of lifetime and disease onset distribution from incomplete observations. *Biometrics*, **38**, 921–932.
- Dobson, A.J. (1990) *An Introduction to Generalized Linear Models*. London: Chapman & Hall.
- Donahue, R.M.J. (1999) A note on information seldom reported via the *P* value. *Am. Statistician*, **53**, 303–306.
- Durbin, J. (1951) Incomplete blocks in ranking experiments. *Brit. J. Psychol. (Statist. Section)*, **4**, 85–90.
- Durbin, J. (1987) Statistics and statistical science. *J. Roy. Statist. Soc. A*, **150**, 177–191.
- Easterbrook, P.J., Berlin, J.A., Gopalan, R. and Matthews, D.R. (1991) Publication bias in clinical research. *Lancet*, **337**, 867–872.
- Edgington, E.S. (1995) *Randomization Tests*. 3rd edn. New York: Marcel Dekker, Inc.
- Efron, B. (1979) Bootstrap methods: another look at the Jackknife. *Ann. Statist.*, **7**, 1–26.
- Efron, B. (1981) Censored data and the bootstrap. *J. Amer. Statist. Assoc.*, **76**, 312–319.
- Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. *Am. Statistician*, **37**, 36–48.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Emerson, J.D. (1982) Nonparametric confidence intervals for the median in the presence of right censoring. *Biometrics*, **38**, 17–27.
- Eplett, W.J.R. (1982) The distribution of Smirnov type two-sample rank tests for discontinuous distribution functions. *J. Roy. Statist. Soc. B*, **44**, 361–369.
- Everitt, B.S. (1992) *The Analysis of Contingency Tables*. 2nd. edn. London: Chapman & Hall.
- Fienberg, S.E. (1980) *The Analysis of Cross Classified Categorical Data*. 2nd edn. Cambridge, MA: MIT Press.
- Fisher, N., Turner, S.W., Pugh, R. and Taylor, C. (1994) Estimating numbers of homeless and homeless mentally-ill people in north east Westminster by using capture-recapture analysis. *BMJ*, **308**, 27–30.
- Fisher, N.I. (1993) *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

- Fisher, R.A. (1929) Tests of significance in harmonic analysis. *Proc. Roy. Soc. London*, **A125**, 54–59.
- Fisher, R.A. (1935) *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Fisher, R.A. (1948) *Statistical Methods for Research Workers*. 10th edn. Edinburgh: Oliver & Boyd.
- Fisher, R.A. and Yates, F. (1957) *Statistical Tables for Biological, Agricultural and Medical Research*. 5th edn. Edinburgh: Oliver & Boyd.
- Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecol.*, **12**, 42–57.
- Fleiss, J.L. (1978) Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**, 378–382.
- Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*. 2nd edn. New York: John Wiley & Sons.
- Fleiss, J.L. and Cohen, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.*, **33**, 613–619.
- Fleiss, J.L., Lee, J.C.M. and Landis, J.R. (1979) The large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.*, **86**, 974–977.
- Fligner, M.A. and Rust, S.W. (1982) A modification of Mood's median test for the generalized Behrens–Fisher problem. *Biometrika*, **69**, 221–226.
- Freeman, G.H. and Halton, J.H. (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, **38**, 141–149.
- Freidlin, B. and Gastwirth, J.L. (1999) Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics*, **55**, 264–267.
- Freidlin, B. and Gastwirth, J.L. (2001) Should the median test be retired from general use? *Am. Statistician*, **55** (in press).
- Freidlin, B., Podger, M.J. and Gastwirth, J.L. (1999) Efficiency robust tests for survival or ordered categorical data. *Biometrics*, **55**, 883–886.
- Freund, J.E. and Ansari, A.R. (1957) *Two Way Rank Sum Tests for Variance*. Blacksburg VA: Virginia Polytechnic Institute Report to Ordnance Research and NSF, **34**.
- Friedman, M. (1937) The use of ranks to avoid the assumptions of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.*, **32**, 675–701.
- Galton, F. (1892) *Finger Prints*. London: Macmillan.
- Gart, J. (1970) Point and interval estimation of the common odds ratio in the combination of  $2 \times 2$  tables with fixed marginals. *Biometrika*, **57**, 471–475.
- Gastwirth, J.L. (1965) Percentile modifications of two sample rank sum tests. *J. Amer. Statist. Assoc.*, **60**, 1127–1141.
- Gastwirth, J.L. (1968) The first median test: a two-sided version of the control median test. *J. Amer. Statist. Assoc.*, **63**, 692–706.
- Gastwirth, J.L. (1985) The use of maximum efficiency robust tests in combining contingency tables and survival analysis. *J. Amer. Statist. Assoc.*, **80**, 380–384.
- Gastwirth, J.L. (1988) *Statistical Reasoning in Law and Public Policy*. Orlando: Academic Press.
- Gat, J.R. and Nissenbaum, A. (1976) Limnology and ecology of the Dead Sea. *Nat. Geog. Soc. Res. Reports*, **1976 Projects**, 413–418.

- Geffen, G., Bradshaw, J.L. and Nettleton, N.C. (1973) Attention and hemispheric differences in reaction time during simultaneous audio-visual tasks. *Quart. J. Expt. Psychol.*, **25**, 404–412.
- Gehan, E.A. (1965a) A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203–223.
- Gehan, E.A. (1965b) A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, **52**, 650–653.
- Gibbons, J.D. and Chakraborti, S. (1992) *Nonparametric Statistical Inference*. 3rd edn. New York: Marcel Dekker.
- Gideon, R.A. and Hollister, R.A. (1987) A rank correlation coefficient resistant to outliers. *J. Amer. Statist. Assoc.*, **82**, 656–666.
- Gill, A.N. and Mehta, G.P. (1989) Selection procedures for scale parameters using two-sample statistics. *Sankhya, Series B*, **51**, 149–157.
- Gill, A.N. and Mehta, G.P. (1991) Selection procedures for scalar parameters using two-sample *U*-statistics. *Austral. J. Statist.*, **33**, 347–362.
- Gillon, R. (1986) *Philosophical Medical Ethics*. Chichester: John Wiley & Sons.
- Good, P. (1994) *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer-Verlag.
- Graubard, B.I. and Korn, E.L. (1987) Choice of column scores for testing independence in ordered  $2 \times k$  contingency tables. *Biometrics*, **43**, 471–476.
- Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.
- Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969) Analysis of categorical data by linear models. *Biometrics*, **25**, 489–504.
- Groggel, D.J. and Skillings, J.H. (1986) Distribution-free tests for the main effects in multifactor designs. *American Statistician*, **40**, 99–102.
- Gumpertz M.L., Graham, J.M. and Ristiano, J.B. (1997) Autologistic model of spatial pattern of Phytophthora epidemic in Bell Pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 131–156.
- Hájek, J., Sidák, Z. and Sen, P.K. (1999) *Theory of Rank Tests*. 2nd edn. San Diego: Academic Press.
- Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- Hall, P. and Hart, J.D. (1990) Bootstrap tests for differences between means in nonparametric regression. *J. Amer. Statist. Assoc.*, **85**, 1039–1049.
- Hampel, F.R. (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.
- Hanley, J.A. and Parnes, M.N. (1983) Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics*, **39**, 129–139.
- Härdle, W. and Gasser, T. (1984) Robust nonparametric function fitting. *J. Roy. Statist. Soc. B*, **46**, 42–51.
- Hastie, T. and Tibshirani, R. (1987) Generalized additive models: some applications. *J. Amer. Statist. Assoc.*, **82**, 371–386.
- Hay, G. (1997) The selection from multiple data sources in epidemiological capture-recapture studies. *The Statistician*, **46**, 515–520.
- Hay, G. and McKeganey, N. (1996) Estimating the prevalence of drug misuse in Dundee, Scotland: an application of capture-recapture methods. *J. Epidemiol. Community Health*, **50**, 469–472.

- Hettmansperger, T.P. and McKean, J.W. (1998) *Robust Nonparametric Statistical Methods*. London: Arnold.
- Hill, N.S. and Padmanabhan, A.L. (1984) Robust comparison of two regression lines and biomedical applications. *Biometrics*, **40**, 985–94.
- Hinkley, D.V. (1989) Modified profile likelihood in transformed linear models. *Appl. Statist.*, **38**, 495–506.
- Hodges, J.L. Jr. (1955) A bivariate sign test. *Ann. Math. Statist.*, **26**, 523–527.
- Hodges, J.L. and Lehmann, E.L. (1963) Estimates of location based on rank tests. *Ann. Math. Statist.*, **34**, 598–611.
- Hollander, M. and Wolfe, D.A. (1999) *Nonparametric Statistical Methods*. 2nd edn. New York: John Wiley & Sons.
- Hook, E.B. and Regal, R.R. (1982) Validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimates in prevalence studies. *Am J. Epidemiol.*, **116**, 168–176.
- Hora, S.C. and Conover, W.J. (1984) The  $F$ -statistic in the two way layout with rank score transformed data. *J. Amer. Statist. Assoc.*, **79**, 668–673.
- Hornbrook, M.C., Hurtado, A.V. and Johnson, R.E. (1985) Health care episodes: definition, measurement and use. *Medical Care Review*, **42**, 163–218.
- House, D.E. (1986) A nonparametric version of Williams' test for a randomized block design. *Biometrics*, **42**, 187–190.
- Howarth, J. and Curthoys, M. (1987) The political economy of women's higher education in late nineteenth and early twentieth century Britain. *Hist. Res.*, **60**, 208–231.
- Huber, P.J. (1972) Robust statistics: a review. *Ann. Math. Statist.*, **43**, 1041–1067.
- Hussain, S.S. and Sprent, P. (1983) Nonparametric regression. *J. Roy. Statist. Soc. A*, **146**, 182–191.
- Hutton, J.L. (1995) Statistics is essential for professional ethics. *J. Appl. Philos.*, **12**, 253–261.
- Iman, R.L. (1974) A power study of rank transformation for the two-way classification model when interaction may be present. *Canadian J. Statist.*, **2**, 227–239.
- Iman, R.L. and Davenport, J.M. (1980) Approximations of the critical region of the Friedman statistic. *Communications in Statistics*, **A9**, 571–595.
- Iman, R.L., Hora, S.C. and Conover, W.J. (1984) Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *J. Amer. Statist. Assoc.*, **79**, 674–685.
- Izenman, A.J. (1991) Recent developments in nonparametric density estimation. *J. Amer. Statist. Assoc.*, **86**, 205–224.
- Jaeckel, L.A. (1972) Estimating regression equations by minimizing the dispersion of residuals. *Ann. Math. Statist.*, **43**, 144–158.
- Jarrett, R.G. (1979) A note on the intervals between coal mining disasters. *Biometrika*, **66**, 191–193.
- Jonckheere, A.R. (1954) A distribution free  $k$ -sample test against ordered alternatives. *Biometrika*, **41**, 133–145.
- Kasser, I.S. and Bruce, R.A. (1969) Comparative effects of aging and coronary heart disease on submaximal and maximal exercise. *Circulation*, **39**, 759–774.

- Katti, S.K. (1965) Multivariate covariance analysis. *Biometrics*, **21**, 957–974.
- Kendall, M.G. (1938) A new measure of rank correlation. *Biometrika*, **30**, 81–93.
- Kendall, M.G. and Gibbons, J.D. (1990) *Rank Correlation Methods*. 5th edn. London: Edward Arnold.
- Kerridge, D. (1975) The interpretation of rank correlations. *Appl. Statist.*, **24**, 257–258.
- Kimber, A.C. (1987) When is a  $\chi^2$  not a  $\chi^2$ ? *Teaching Statistics*, **9**, 74–77.
- Kimber, A.C. (1990) Exploratory data analysis for possibly censored data from skewed distributions. *Appl. Statist.*, **39**, 21–30.
- Kimura, D.K. and Chikuni, S. (1987) Mixtures of empirical distributions: an iterative application of the age-length key. *Biometrics*, **43**, 23–35.
- Klotz, J. (1962) Nonparametric tests for scale. *Ann. Math. Statist.*, **33**, 498–512.
- Klotz, J. (1963) Small sample power and efficiency for the one-sample Wilcoxon and normal scores tests. *Ann. Math. Statist.*, **34**, 624–632.
- Knapp, T.R. (1982) The birthday problem: some empirical data and some approximations. *Teaching Statistics*, **4**, 10–14.
- Koch, G.G. (1972) The use of nonparametric methods in the statistical analysis of the two-period change-over design. *Biometrics*, **28**, 577–594.
- Kolmogorov, A.N. (1933) Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari*, **4**, 83–91.
- Kolmogorov, A.N. (1941) Confidence limits for an unknown distribution function. *Ann. Math. Statist.*, **12**, 461–463.
- Kraemer, H.C. and Thiemann, S. (1987) *How many subjects? Statistical power analysis in research*. Newbury Park: Sage.
- Krantz, D.H. (1999) The null hypothesis testing controversy in psychology. *J. Amer. Statist. Assoc.*, **94**, 1372–1381.
- Kruskal, W.H. and Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.*, **47**, 583–621.
- Lancaster, H.O. (1949) The derivation and partitioning of  $\chi^2$  in certain discrete distributions. *Biometrika*, **36**, 117–129.
- Lancaster, H.O. (1953) A reconciliation of  $\chi^2$ , considering metrical and enumerative aspects. *Sankhya*, **13**, 1–9.
- Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.
- Leach, C. (1979) *Introduction to Statistics. A Nonparametric Approach for the Social Sciences*. Chichester: John Wiley & Sons.
- Lehmann, E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden Day, Inc.
- Light, R.J. (1971) Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol. Bull.*, **76**, 89–97.
- Lilliefors, H.W. (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Amer. Statist. Assoc.*, **62**, 399–402.
- Lindley, D.V. and Scott, W.F. (1995) *New Cambridge Elementary Statistical Tables*. 2nd edn. Cambridge: Cambridge University Press.
- Lindsey, J.C., Herzberg, A.M. and Watts, D.G. (1987) A method of cluster analysis based on projections and quantile–quantile plots. *Biometrics*, **43**, 327–341.

- Lloyd, C.J. (1999) *Statistical Analysis of Categorical Data*. New York: John Wiley & Sons.
- Lombard, F. (1987) Rank tests for change-point problems. *Biometrika*, **74**, 615–624.
- Lubischew, A.A. (1962) On the use of discriminant functions in taxonomy. *Biometrics*, **18**, 455–477.
- McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd edn. London: Chapman & Hall.
- Mack, G.A. and Skillings, J.H. (1980) A Friedman type rank test for main effects in a two-factor ANOVA. *J. Amer. Statist. Assoc.*, **75**, 947–951.
- Mack, G.A. and Wolfe D.A. (1981) *K*-sample rank test for umbrella alternatives. *J. Amer. Statist. Assoc.*, **76**, 175–181.
- McKean, J.W., Sheather, S.J. and Hettsmansperger, T.P. (1990) Regression diagnostics for rank based methods. *J. Amer. Statist. Assoc.*, **85**, 1018–1028.
- McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.
- Manly, B.F.J. (1997) *Randomization, Bootstrapping and Monte Carlo Methods in Biology*. 2nd edn. London: Chapman & Hall.
- Mann, H.B. and Whitney, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, **18**, 50–60.
- Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Res. Inst.*, **22**, 719–748.
- Marascuilo, L.A. and McSweeney, M. (1977) *Nonparametric and Distribution-free Methods for the Social Sciences*. Monterey, CA: Brooks/Cole Publishing Company.
- Marascuilo, L.A. and Serlin, R.C. (1979) Tests and contrasts for comparing change parameters for a multiple sample McNemar data model. *Brit. J. Math. and Statist. Psychol.*, **32**, 105–112.
- Mardia, K.V. (1972) *Statistics of Directional Data*. London: Academic Press.
- Mardia, K.V. and Jupp, P.E. (2000) *Directional Statistics*. Chichester: John Wiley & Sons.
- Maritz, J.S. (1995) *Distribution-free Statistical Methods*. 2nd edn. London: Chapman & Hall.
- Marriott, F.H.C. (ed.) (1990) *A Dictionary of Statistical Terms*. 5th. edn. Harlow: Longman Scientific & Technical.
- Mathisen, H.C. (1943) A method of testing the hypothesis that two samples are from the same population. *Ann. Math. Statist.*, **14**, 188–194.
- Mattingley, P.F. (1987) Pattern of horse devolution and tractor diffusion in Illinois, 1920–82. *Prof. Geographer*, **39**, 298–309.
- Mehta, C.R. and Patel, N.R. (1983) A network algorithm for performing Fisher's exact test in  $r \times c$  contingency tables. *J. Amer. Statist. Assoc.*, **78**, 427–434.
- Mehta, C.R. and Patel, N.R. (1986) A hybrid algorithm for Fisher's exact test on unordered  $r \times c$  contingency tables. *Communications in Statistics*, **15**, 387–403.
- Mehta, C.R. and Patel, N.R. (1999) *StatXact 4 for Windows – User Manual*. Cambridge, MA: Cytel Software Corporation.
- Mehta, C.R., Patel, N.R. and Gray, R. (1985) On computing an exact common odds ratio in several  $2 \times 2$  contingency tables. *J. Amer. Statist. Assoc.*, **80**, 969–973.

- Mehta, C.R., Patel, N.R. and Senchaudhuri, P. (1988) Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Statist. Assoc.*, **83**, 999–1005.
- Mehta, C.R., Patel, N.R. and Senchaudhuri, P. (1998) Exact power and sample size computations for the Cochran–Armitage trend test. *Biometrics*, **54**, 1615–1621.
- Mehta, C.R., Patel, N.R. and Tsaiatis, A.A. (1984) Exact significance tests to establish treatment equivalence for ordered categorical data. *Biometrics*, **40**, 819–825.
- Mood, A.M. (1940) The distribution theory of runs. *Ann. Math. Statist.*, **11**, 367–392.
- Mood, A.M. (1954) On the asymptotic efficiency of certain nonparametric two-sample tests. *Ann. Math. Statist.*, **25**, 514–522.
- Moses, L.E. (1952) Nonparametric statistics for psychological research. *Psychol. Bull.*, **49**, 122–143.
- Neave, H.R. (1981) *Elementary Statistical Tables*. London: George Allen & Unwin Ltd.
- Nelder, J.A. (1999) From statistics to statistical science (with comment). *The Statistician*, **48**, 257–270.
- Noether, G.E. (1984) Nonparametrics: the early years – impressions and recollections. *Am. Statistician*, **38**, 173–178.
- Noether, G.E. (1987a) Sample size determination for some common nonparametric tests. *J. Amer. Statist. Assoc.*, **82**, 645–647.
- Noether, G.E. (1987b) Mental random numbers: perceived and real randomness. *Teaching Statistics*, **9**, 68–70.
- Noether, G.E. (1991) *Introduction to Statistics: The Nonparametric Way*. New York: Springer-Verlag.
- Office for National Statistics (1998) *1996 Mortality Statistics: Injury and Poisoning*. (Series DH4, no. 21) England & Wales. London: The Stationery Office.
- O’Muircheartaigh, I.G. and Sheil, J. (1983) Fore or five? The indexing of a golf course. *Appl. Statist.*, **32**, 287–292.
- O’Quigley, J. and Prentice, R.L. (1991) Nonparametric tests of association between survival times and continuously measured covariates: the logit-rank and associated procedures. *Biometrics*, **47**, 117–127.
- Page, E.B. (1963) Ordered hypotheses for multiple treatments: a significance test for linear ranks. *J. Amer. Statist. Assoc.*, **58**, 216–230.
- Paul, S.R. (1979) Models and estimation procedures for the calibration of examiners. *Brit. J. Math. Statist. Psychol.*, **32**, 242–251.
- Pearce, S.C. (1965) *Biological Statistics – An Introduction*. New York: McGraw-Hill.
- Pearson, J.C.G. and Sprent, P. (1968) Trends in hearing loss associated with age or exposure to noise. *Appl. Statist.*, **17**, 205–215.
- Pearson, K. (1900) On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen in random sampling. *Phil. Mag.* (5), **50**, 157–175.
- Peters, D. and Randles, R.H. (1990) A multivariate signed rank test for the one-sample location problem. *J. Amer. Statist. Assoc.*, **85**, 552–557.

- Peto, R. and Peto, J. (1972) Asymptotically efficient rank-invariant test procedures. *J. Roy. Statist. Soc. A*, **135**, 185–206.
- Pettitt, A.N. (1979) A nonparametric approach to the change-point problem. *Appl. Statist.*, **28**, 126–135.
- Pettitt, A.N. (1981) Posterior probabilities for a change-point using ranks. *Biometrika*, **68**, 443–450.
- Pettitt, A.N. (1983) Approximate methods using ranks for regression with censored data. *Biometrika*, **70**, 121–132.
- Pettitt, A.N. (1985) Re-weighted least squares estimation with censored and grouped data: an application of the EM algorithm. *J. Roy. Statist. Soc. B*, **47**, 253–260.
- Pitman, E.J.G. (1937a) Significance tests that may be applied to samples from any population. *J. Roy. Statist. Soc., Suppl.*, **4**, 119–130.
- Pitman, E.J.G. (1937b) Significance tests that may be applied to samples from any population, II: The correlation coefficient test. *J. Roy. Statist. Soc., Suppl.*, **4**, 225–232.
- Pitman, E.J.G. (1938) Significance tests that may be applied to samples from any population. III. The analysis of variance test. *Biometrika*, **29**, 322–335.
- Pitman, E.J.G. (1948) Mimeographed lecture notes on nonparametric statistics. Columbia University.
- Plackett, R.L. (1981) *The Analysis of Categorical Data*. 2nd edn. London: Charles Griffin & Co.
- Posner, K.L., Sampson, P.D., Caplan, R.A., Ward, R.J. and Chenly, F.W. (1990) Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statist. Med.*, **9**, 1103–1116.
- Pothoff, R.F. (1963) Use of the Wilcoxon statistic for a generalized Behrens–Fisher problem. *Ann. Math. Statist.*, **34**, 1596–1599.
- Quade, D. (1979) Using weighted ranks in the analysis of complete blocks with additive block effects. *J. Amer. Statist. Assoc.*, **74**, 680–683.
- Quenouille, M.H. (1949) Approximate tests of correlation in time series. *J. Roy. Statist. Soc. B*, **11**, 18–84.
- Radelet, M. (1981) Racial characteristics and the imposition of the death penalty. *Amer. Sociol. Rev.*, **46**, 918–927.
- Rae, G. (1988) The equivalence of multirater kappa statistics and intraclass correlation coefficients. *Educ. and Psychol. Meas.*, **48**, 921–933.
- Randles, R.H. (1989) A distribution-free multivariate sign test based on interdirections. *J. Amer. Statist. Assoc.*, **84**, 1045–1050.
- Randles, R.H., Fligner, M.A., Policello, G.E. and Wolfe, D.A. (1980) An asymptotically distribution-free test for symmetry versus asymmetry. *J. Amer. Statist. Assoc.*, **75**, 168–172.
- Randles, R.H. and Wolfe, D.A. (1979) *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley & Sons.
- Rees, D.G. (1995) *Essential Statistics*. 3rd edn. London: Chapman & Hall.
- Rees, J.H., Thompson, R.D., Smeeton, N.C. and Hughes, R.A.C. (1998) Epidemiological study of Guillain–Barré syndrome in south east England. *J. Neurol. Neurosurg. and Psychiatry*, **64**, 74–77.
- Regal, R.R. and Hook, E.B. (1984) Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Statist. Med.*, **3**, 287–291.

- Regal, R.R. and Hook, E.B. (1999) An exact test for all-way interactions in a  $2^M$  contingency table: application to interval capture-recapture estimation of population size. *Biometrics*, **55**, 1241–1246.
- Robins, J., Breslow, N. and Greenland, S. (1986) Estimation of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**, 311–323.
- Robinson, J.A. (1983) Bootstrap confidence intervals in location-scale models with progressive censoring. *Technometrics*, **25**, 179–187.
- Rogerson, P.A. (1987) Changes in US national mobility levels. *Prof. Geographer*, **39**, 344–351.
- Rosenthal, I. and Ferguson, T.S. (1965) An asymptotic distribution-free multiple comparison method with application to the problem of  $n$  rankings of  $m$  objects. *Brit. J. Math. Statist. Psychol.*, **18**, 243–254.
- Ross, G.J.S. (1990) *Nonlinear Estimation*. New York: Springer-Verlag.
- Ryan, T.P. (1997) *Modern Regression Methods*. New York: John Wiley & Sons.
- Sackett, D.L., Haynes, R.B., Guyatt, G.H. and Tugwell, P. (1991) *Clinical Epidemiology: A Basic Science for Clinical Medicine*. 2nd edn. Boston: Little Brown.
- Sackrowitz, H. and Samuel-Cahn, E. (1999)  $P$ -values as random variables –expected  $P$  values. *Am. Statistician*, **53**, 326–331.
- Saunders, R. and Laud, P. (1980) The multidimensional Kolmogorov goodness of fit test. *Biometrika*, **67**, 237.
- Savage, I.R. (1956) Contributions to the theory of rank order statistics I. *Ann. Math. Statist.*, **27**, 590–615.
- Scheirer, C.J., Ray, W.S. and Hare, N. (1976) The analysis of ranked data derived from completely randomized factorial designs. *Biometrics*, **32**, 429–434.
- Scholz, F.W. and Stephens, M.A. (1987) K-sample Anderson–Darling tests. *J. Amer. Statist. Assoc.*, **82**, 918–924.
- Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977) The application of time series methods to the analysis of repeated surveys. *Internat. Statist. Rev.*, **45**, 13–28.
- Seber, G.A.F. (1982) *The Estimation of Animal Abundance and Related Parameters*. 2nd edn. London: Charles Griffin and Company.
- Sekar, C.C. and Deming, W.E. (1949) On a method of estimating birth and death rates and the extent of registration. *J. Amer. Statist. Assoc.*, **44**, 101–115.
- Selvin, S. (1995) *Practical Biostatistical Methods*. Belmont, CA: Wadsworth Publishing Company.
- Sen, P.K. (1968) Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.*, **63**, 1379–1389.
- Sen, P.K. (1969) On a class of rank order tests for the parallelism of several regression lines. *Ann. Math. Statist.*, **40**, 1668–1683.
- Senn, S.J. (1992) *Cross-over Trials in Clinical Research*. Chichester: John Wiley & Sons.
- Shapiro, S.S. and Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.
- Shapiro, S.S., Wilk, M.B. and Chen, H.J. (1968) A comparative study of various tests for normality. *J. Amer. Statist. Soc.*, **63**, 1343–1372.

- Shirley, E. (1977) A nonparametric equivalent of Williams' test for contrasting existing dose levels of a treatment. *Biometrics*, **33**, 386–389.
- Shirley, E.A.C. (1987) Applications of ranking methods to multiple comparison procedures and factorial experiments. *Appl. Statist.*, **36**, 205–213.
- Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioral Sciences*. 2nd edn. New York: McGraw-Hill.
- Siegel, S. and Tukey, J.W. (1960) A nonparametric sum of ranks procedure for relative spread in unpaired samples. *J. Amer. Statist. Assoc.*, **55**, 429–444.
- Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. B*, **13**, 238–241.
- Smeeton, N.C. (1986) Modelling episodes of mental illness: some results from the Second U.K. National Morbidity Survey. *The Statistician*, **35**, 55–63.
- Smeeton, N. and Wilkinson, G. (1988) The detection of annual clusters in individual patterns of parasuicide. *Journal of Applied Statistics*, **15**, 179–182.
- Smeeton, N.C., Rona, R.J., Sharland, G., Botting, B.J., Barnett, A. and Dundas, R. (1999) Estimating the prevalence of malformation of the heart in the first year of life using capture-recapture methods. *Am. J. Epidemiol.*, **150**, 778–785.
- Smirnov, N.V. (1939) On the estimation of discrepancy between empirical curves of distribution for two independent samples. (In Russian.) *Bull. Moscow Univ.*, **2**, 3–16.
- Smirnov, N.V. (1948) Tables for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, **19**, 279–281.
- Smith, R.L. and Naylor, J.C. (1987) A comparison of likelihood and Bayesian estimators of the three-parameter Weibull distribution. *Appl. Statist.*, **36**, 358–369.
- Snee, R.D. (1985) Graphical display of results of three treatment randomized block experiments. *Appl. Statist.*, **34**, 71–77.
- Spearman, C. (1904) The proof and measurement of association between two things. *Amer. J. Psychol.*, **15**, 72–101.
- Sprent, P. (1998) *Data Driven Statistical Methods*. London: Chapman & Hall.
- Stuart, A. (1955) A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412–416.
- Stuart, A. (1957) The comparison of frequencies in matched samples. *Brit. J. Statist. Psychol.*, **10**, 29–32.
- Sukhatme, B.V. (1957) On certain two sample nonparametric tests for variances. *Ann. Math. Statist.*, **28**, 188–194.
- Swed, F.S. and Eisenhart, C. (1943) Tables for testing randomness of grouping in a series of alternatives. *Ann. Math. Statist.*, **14**, 83–86.
- Sweeting, T.J. (1982) A Bayesian analysis of some pharmacological data using a random coefficient regression model. *Appl. Statist.*, **31**, 205–213.
- Terpstra, T.J. (1952) The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking. *Indag. Math.*, **14**, 327–333.
- Theil, H. (1950) A rank invariant method of linear and polynomial regression analysis, I, II, III. *Proc. Kon. Nederl. Akad. Wetensch. A*, **53**, 386–392, 521–525, 1397–1412.
- Thomas, G.E. (1989) A note on correcting for ties with Spearman's  $\tau$ . *J. Statist. Comp. Sim.*, **31**, 37–40.

- Thomas, G.E. (2000) Use of the bootstrap in robust estimation of location. *The Statistician*, **49**, 63–77.
- Thomas, G.E. and Kiwanga, G.E. (1993) Use of ranking and scoring methods in the analysis of ordered categorical data from factorial experiments. *The Statistician*, **42**, 55–67.
- Tukey, J.W. (1958) Bias and confidence in not quite large samples. Abstract. *Ann. Math. Statist.*, **29**, 614.
- Upton, G.J.G. (1992) Fisher's exact test. *J. Roy. Statist. Soc. A*, **155**, 395–402.
- van der Waerden, B.L. (1952) Order tests for the two-sample problem and their power, I. *Proc. Kon. Nederl. Akad. Wetensch. A*, **55**, 453–458. Correction, **56**, 80.
- van der Waerden, B.L. (1953) Order tests for the two-sample problem and their power; II, III. *Proc. Kon. Nederl. Akad. Wetensch. A*, **56**, 303–310, 311–316.
- von Mises, R. (1931) *Wrscheinlichkeitrechnung und ihre Anwendung in der Statistik und Theoretischen Physik*. Leipzig: F. Deuticke.
- Wald, A. and Wolfowitz, J. (1940) On a test whether two samples are from the same population. *Ann. Math. Statist.*, **11**, 147–162.
- Walsh, J.E. (1949a) Application of some significance tests for the median which are valid under very general conditions. *J. Amer. Statist. Assoc.*, **44**, 342–355.
- Walsh, J.E. (1949b) Some significance tests for the median which are valid under very general conditions. *Ann. Math. Statist.*, **20**, 64–81.
- Whitworth, W.A. (1886) *Choice and Chance*. Cambridge: Deighton Bell.
- Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.
- Willemain, T.R. (1980) Estimating the population median by nomination sampling. *J. Amer. Statist. Assoc.*, **75**, 908–911.
- Williams, D.A. (1986) A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. *Biometrics*, **42**, 183–186.
- Williamson, P., Hutton, J.L., Bliss J., Blunt J., Campbell, M.J. and Nicholson, R. (2000) Statistical review by research ethics committees. *J. Roy. Statist. Soc. A*, **163**, 5–13.
- Woolson, R.F. (1981) Rank tests and a one-sample log rank test for comparing observed survival data to a standard population. *Biometrics*, **37**, 687–696.
- Woolson, R.F. and Lachenbruch, P.A. (1980) Rank tests for censored matched pairs. *Biometrika*, **67**, 597–606.
- Woolson, R.F. and Lachenbruch, P.A. (1981) Rank tests for censored randomized block designs. *Biometrika*, **68**, 427–435.
- Woolson, R.F. and Lachenbruch, P.A. (1983) Rank analysis of covariance with right-censored data. *Biometrics*, **39**, 727–733.
- Yates, F. (1984) Tests of significance for  $2 \times 2$  contingency tables. *J. Roy. Statist. Soc. A*, **147**, 426–463.
- Zelen, M. (1971) The analysis of several  $2 \times 2$  contingency tables. *Biometrika*, **58**, 129–137.