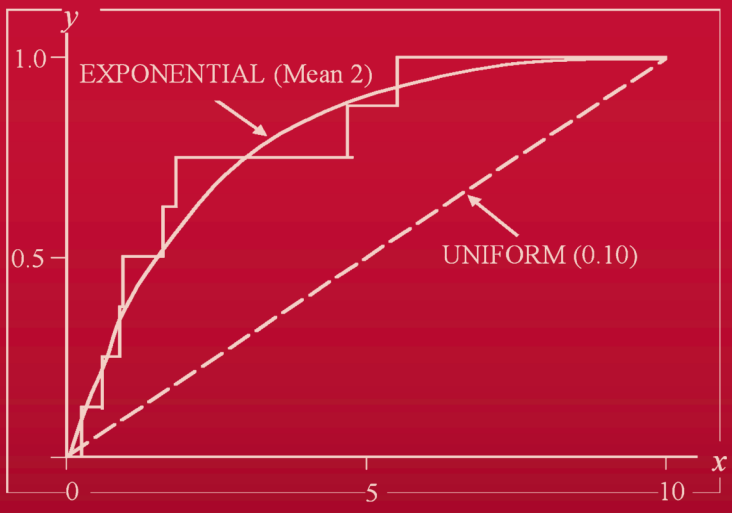
Texts in Statistical Science

Applied Nonparametric Statistical Methods

Fourth Edition



Peter Sprent and Nigel C. Smeeton

Chapman & Hall/CRC
Taylor & Francis Group

Texts in Statistical Science

Applied Nonparametric Statistical Methods

Fourth Edition

Texts in Statistical Science

Applied Nonparametric Statistical Methods

Fourth Edition

Peter Sprent and Nigel C. Smeeton

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20110713

International Standard Book Number-13: 978-1-4398-9401-9 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Applied Nonparametric Statistical Methods first appeared in 1989. Major developments in computing, especially for exact permutation tests, inspired a revised second edition in 1993. The third edition in 2001 reflected not only further advances in computing, but a widening of the scope of nonparametric or distribution-free methods and a tendency for these to merge with, or to be used in conjunction with, techniques such as exploratory data analysis, robust estimation and semiparametric methods. This trend has continued, being evident especially in computer intensive methods to deal with both intractable analytic problems and in processing large data sets.

This new edition reflects these developments while retaining features that have met a continuing positive response from readers and reviewers.

Nonparametric methods are basically tools for statistical analyses, but data collection and the interpretation of analyses are interrelated parts of the statistician's role. As in the third edition we comment, where appropriate, on all these aspects, some of which do not always receive the attention they deserve in undergraduate mainstream or in service courses in statistics.

Our approach is midway between a bare description of techniques and a detailed exposition of the theory, often illustrating key practical points by examples. We keep mathematics to the minimum needed for a clear understanding of scope and limitions.

We have two aims. One is to provide a textbook for those making first contact with nonparametric methods at the undergraduate level. This may be in mainstream statistics courses, or in service courses for students majoring in other disciplines. The second is to make the basic methods readily available to specialist workers, managers, research and development staff, consultants and others working in various fields. Many of them may have an understanding of basic statistics but only a limited acquaintance with nonparametric methods, yet feel these may prove useful in their work. The format we have adopted makes the book suitable not only as a class text, but also for self-study or use as a reference manual.

To meet our aims the treatment is broad rather than deep. We believe this to be a fruitful approach at the introductory stage. Once one has a broad overview of nonparametrics a more advanced study of topics of special interest becomes appropriate. Because fields of interest will vary from person to person, this second phase is best tackled by attending courses on, or referring to the literature that deals in depth with, aspects of particular interest. We give

references to books and papers where more advanced treatments of many topics can be found.

Popular features of earlier editions are retained, including a formal structure for most examples, lists of potential fields of application and a selection of exercises at the end of each chapter.

There has been a substantial reordering of topics and new material has been added. The former Chapter 1 has been split into two, with consequent renumbering of later chapters. Chapter 1 now gives a brief summary of some relevant general statistical concepts, while Chapter 2 introduces ideas basic to nonparametric or distribution-free methods. The new Chapters 3 to 7 broadly cover the content of the former Chapters 2 to 6, but with many changes in emphasis and removal of some material on designed experiments to a new Chapter 8, and some on analysis of survival data to an extended treatment in Chapter 9. Designed experiments, particularly those with a factorial treatment structure, are handled in an up-to-date way in Chapter 8. Chapters 10 to 14 are revisions of the former Chapters 7 to 11. Chapter 15 is new and introduces a few of the many important modern developments, most of the applications being computer intensive.

As in earlier editions we have not included tables of quantiles, critical values, etc. relevant to basic nonparametric procedures. Modern software has made many of these somewhat redundant, but those who need such tables will find them in many standard collections of statistical tables. We give references as needed to relevant specialized tables. Solutions to selected exercises are given in an appendix.

We are grateful to many readers and reviewers of the earlier editions who have made constructive comments about content and treatment of particular topics. Their input triggered several major changes in the present edition. We thank Edgar Brunner and Thomas P. Hettmansperger for drawing our attention to a number of papers dealing with interactions in a nonparametric context, and Joseph Gastwirth for alerting us to many recent practical developments and Nick Cox for providing a Stata program for simulating runs distributions. We renew thanks to those whose assistance was acknowledged in earlier editions — to Jim McGarrick for useful discussions on physiological measurements — to Professor Richard Hughes for advice on the Guillain–Barré syndrome — to Timothy P. Davis and Chris Theobald who supplied data sets for examples. We are grateful to Cyrus Mehta and Cytel Software for providing us with complementary software and manuals for StatXact 7 and to Shashi Kumar for help with technical problems with embedding fonts in diagrams.

P. Sprent
N.C. Smeeton

CHAPTER 1

SOME BASIC CONCEPTS

1.1 Basic Statistics

We assume most readers are familiar with the basic statistical notions met in introductory or service courses in statistics of some 20 hours duration. Nevertheless, those with no formal statistical training should be able to use this book in parallel with an introductory statistical text. Rees (2000) adopts a straightforward approach. Some may prefer a more advanced treatment, or an introduction that emphasizes applications in a discipline in which they are working.

Readers trained in general statistics, but who are new to nonparametric methods will be familiar with some of the background material in this chapter. However, we urge them at least to skim through it to see where we depart from conventional treatments, and to learn how nonparametric procedures relate to other approaches. We explain the difference between parametric and nonparametric methods and survey some general statistical notions that are relevant to nonparametric methods. We also comment on good practice in applied statistics.

In Chapter 2 we use simple examples to illustrate some basic nonparametric ideas and introduce some statistical notions and tools that are widely used in this field. Their application to a range of problems is covered in the remaining chapters.

1.1.1 Parametric and Nonparametric Methods

The word *statistics* has several meanings. It is used to describe a collection of data, and also to designate operations that may be performed with that primary data. The simplest of these is to form *descriptive* statistics. These include the mean, range, or other quantities to summarize primary data, as well as preparing tables or pictorial representations (e.g., graphs) to exhibit specific facets of the data. The scientific discipline called *statistics*, or *statistical inference*, uses observed data — in this context called a *sample* — to make inferences about a larger potentially observable collection of data called a *population*. We explain the terms *sample* and *population* more fully in Section 1.2

We associate *distributions* with populations. Early in their careers statistics students meet families of distributions such as the *normal* and *binomial* where

individual members of the family are distinguished by assigning specific values to entities called *parameters*.

The notation $\mathrm{N}(\mu, \sigma^2)$ denotes a member of the normal, or Gaussian, family with mean μ and variance σ^2. Here μ and σ are parameters.

The *binomial* family depends on two parameters, n and p, where n is the total number of observations and p is the probability associated with one of two possible outcomes at any observation. Subject to certain conditions, the number of occurrences, r, where $0 \leq r \leq n$, of that outcome among n observations, has a binomial distribution with parameters n and p. We call this a $\mathrm{B}(n, p)$ distribution.

Given a set of independent observations, called a random sample, from some population with a distribution that is a member of a family such as the normal or binomial, *parametric statistical inference* is often concerned with testing hypotheses about, or estimation of, unknown parameters.

For a sample from a normal distribution the sample mean is a point (i.e., a single value) estimate of the parameter μ. Here the well-known t-test provides a measure of the strength of the evidence provided by a sample in support of an *a priori* hypothesized value μ_0 for the distribution, or population, mean. We may also obtain a *confidence interval*, a term we explain in Section 1.4.1, for the "true" population mean.

When we have a sample of n observations from a $\mathrm{B}(n, p)$ distribution with p unknown, if the event with probability p is observed r times an appropriate estimate of p is $\hat{p} = r/n$. We may want to assess how strongly sample evidence supports an *a priori* hypothesized value p_0, say, for p, or to obtain a confidence interval for the population parameter p.

Other well-known families of distributions include the uniform (or rectangular), multinomial, Poisson, exponential, gamma, beta, Cauchy 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 such a use of the normal distribution especially in what are called *asymptotic approximations*. We often refer to these in this book.

Parametric inference may be inappropriate or even impossible. For example, records of examination results may only give the numbers of candidates in banded and ordered grades designated Grade A, Grade B, Grade C, etc. Given these numbers for pupils from two different schools, we may want to know if they indicate a difference in performance between those schools that might be attributed to different methods of teaching, or to the ability of one school to attract more able pupils. There is no obvious family of distributions that provides our data, and there are no clearly defined *parameters* about

which we can make inferences. Two terms are in common use for the type of inferences we may then make. They are either described as *nonparametric* or as *distribution-free*. There is sometimes an incorrect belief among nonstatisticians that these terms refer to the data.

There is not unanimity among statisticians in their use of the terms *nonparametric* and *distribution-free*. This is of no great consequence in practice, and to some extent simply reflects historical developments.

There is not even universal agreement about what constitutes a parameter. Quantities such as μ, σ^2 appearing in the density functions for the normal family are unquestionably parameters. The term is often used more widely to describe any population characteristic within a family such as a mean, median, moment, quantile, or range. In Chapter 8 and elsewhere we meet situations where we have observations composed of a deterministic and random element where we want to estimate constants occuring in the deterministic element. Such constants are also sometimes called parameters.

In nonparametric, or distribution-free, methods we often make inferences about parameters in this wider sense. The important point is that we do not assume our samples are associated with any prespecified *family* of distributions. In this situation the name *distribution-free* is more appropriate if we are still interested in parameters in the broader senses mentioned above.

Some procedures are both distribution-free and nonparametric in that they do not involve parameters even in the broader use of that term. The above example involving examination grades falls into this category.

Historically, the term *nonparametric* was in use before *distribution-free* became popular. There are procedures for which one name is more appropriate than the other, but as in many areas of statistics, terminology does not always fit procedures into watertight compartments. A consequence is the spawning of hybrid descriptions such as *asymptotically distribution-free* and *semiparametric* methods. There is even some overlap between descriptive statistics and inferential statistics, evident in a practice described as *exploratory data analysis*, often abbreviated to EDA. We shall see in Section 2.4, and elsewhere, that sensible use of EDA may prove invaluable in selecting an appropriate technique, parametric or nonparametric, for making statistical inferences.

Designating procedures as *distribution-free* or *nonparametric* does not mean they are assumption free. In practice we nearly always make some assumptions about the underlying population distribution. For example, we may assume that it is continuous and symmetric. These assumptions do not restrict us to a particular family such as the normal, but they exclude both discrete and asymmetric distributions. Given some data, EDA will often indicate whether or not an assumption such as one of symmetry is justified.

Many *nonparametric* or *distribution-free* procedures involve, through the test statistic, distributions and parameters (often the normal or binomial distributions). This is because the terms refer *not* to the test statistic, but to the fact that the methods can be applied to samples from populations having distributions only specified in broad terms, e.g., as being continuous, symmetric,

identical, differing only in medians, means, etc. The distribution of the appropriate test statistic is the same no matter what the population distribution may be, providing only that it satisfies the broad-term specification. There is a grey area between what is clearly distribution-free and what is parametric inference. Some of the association tests described in Chapters 12 and 13 fall in this area.

1.1.2 The Use of Nonparametric Methods

Some parametric tests do not depend critically on the correctness of an assumption that samples come from a distribution in a particular family. They are then described as *robust*. Robustness is by no means a universal property of parametric tests. Because they require fewer assumptions for their validity, nonparametric methods are usually more robust than their parametric counterparts.

Nonparametric methods are often the only ones available for data that simply specify order (ranks) or counts of the number of events, or of individuals, in various categories.

In most statistical problems, no matter whether parametric or nonparametric methods are appropriate, what we can deduce depends on what assumptions can validly be made. An example illustrates this.

Example 1.1

Two machines produce metal rods. For each, 2.5 percent of all rods produced have a diameter exceeding 30 cm.

This condition is met if the first machine produces items having a normal distribution with mean 27 mm and standard deviation 1.53 mm. This is because, for any normal distribution, 2.5 percent of all items have a diameter at least 1.96 standard deviations above the mean, so 2.5 percent exceed $27 + 1.96 \times 1.53 \approx 30$.

The condition is also met if the second machine produces items with diameters uniformly distributed between 20.25 and 30.25 mm (i.e., with mean diameter 25.25 mm). Once again the condition that 2.5 percent have diameters exceeding 30 mm is met. This follows because any interval between 20.25 and 30.25 mm of width 0.25 mm contains a proportion 1/40 (i.e., 2.5 percent) of total production.

This uniform distribution is unlikely to be met in practice in this context, but the 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

If we make an additional assumption that distributions of diameters for each machine differ, if at all, only in their means, then if we know the proportion over 30 mm 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 then use an appropriate test.

Means and medians are widely used to indicate where distributions are centred. Both are formally described as *measures of centrality* or *measures of location*. Not all distributions have a mean, but all have a median. If the mean exists, the mean and the median have the same value for a symmetric distribution. Their values differ for asymmetric, or skew, distributions. The Cauchy distribution is a well-known example of a symmetric distribution that has no mean. It has a well-defined median, this being zero for the standard Cauchy distribution.

Tests and estimation procedures for measurement data are often about centrality measures, e.g.,

- Is it reasonable to suppose that a sample comes from a population with a prespecified 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?

Variation or *spread* or *dispersion* is often of interest also. Buyers of new cars or computers want not only a good average performance but also consistency, i.e., not too much variation in performance from item to item of the same brand. Each buyer expects his or her purchase to perform as well as those of other buyers 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.

Dispersion is often measured by *variance* or by *standard deviation*, but these may not exist for all distributions. Also, they are not well suited on their own to describe, or compare, the spread of skew distributions. There are parametric and nonparametric methods for assessing spread or variability.

In other situations we want to assess how well data conform to some hypothesized population distribution function, the approriate test being one for *goodness of fit*. Tests of association or correlation are also of considerable interest.

Nonparametric techniques may be the only ones available when we have limited information. We may want to test if it is reasonable to assume that weights of a large batch of items have a prespecified median, 2 mg, say, when 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, an available nonparametric approach may be cost effective.

Simple nonparametric methods are also useful when data are in some sense 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; this may be anything from a few weeks 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

$$49 \quad 58 \quad 75 \quad 110 \quad 112 \quad 132 \quad 151 \quad 276 \quad 281 \quad 362^*$$

The asterisk denotes the 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 Example 2.2 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. In the original study 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 second sample 12 of the 28 observations were censored at values of 300 or more. We show in Example 9.1 that, on the basis of his data, 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 Notes

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. Throughout this book we illustrate how we test evidence like this more formally to justify this commonsense conclusion. The biblical analysis is informal, but it contains the germ of a nonparametric, as opposed to a parametric, test. Be warned though that the commonsense conclusion may not be justified here. Daniel and his three brethren may already have been fairer and fatter before the "experiment" began!

John Arbuthnot (1710) observed that in each of the 82 years 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 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 *centisimal 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 10.1.3) 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. That problem has been overcome with appropriate statistical software, but we shall see later that the raw-data permutation tests proposed by these writers have practical limitations, although the concept is of considerable theoretical interest.

Other developments in the 1930s include work by Friedman (1937), Smirnov (1939) and others.

About the same time it was realized that observations consisting simply of preferences or ranks could be used in permutation tests 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. They often enjoy the characteristic we have already referred to as *robustness*, and which we describe more fully in Chapter 14.

From the 1940s nonparametric methods 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. At that time hypothesis testing was usually easy, the more important interval estimation described in Section 1.4 was not. This difficulty was overcome by Hodges and Lehmann (1963).

In parallel with the above advances, techniques relevant to counts were developed. Counts often represent numbers of items in categories that 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 because they involve repeated performance of simple calculations, something computers do well.

The dramatic postwar development of feasible, but until recently often tedious to carry out, nonparametric procedures was described in Noether (1984), but much has happened since then.

Another line of development has been in the use of the computer intensive procedures of *jack-knifing* introduced by Quenouille (1949), and the even-more widely used *bootstrapping* introduced by Efron (1979). The latter has both parametric and nonparametric versions.

Computers have revolutionized our approach to data analysis and to statistical inference. Hopes, often ill-founded, that data would fit a restricted mathematical model with few parameters, and emphasis on simplifying concepts such as linearity, have often been replaced by the use of robust methods and by EDA 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 (2001) at the practical level, often blend parametric and nonparametric approaches.

Whereas the initial developments in nonparametric methods were inspired by problems arising in small samples, interest in recent years has turned to their use in extracting information from very large data sets, an aspect we touch upon in Chapter 15.

Nonparametric procedures 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. Such trends may be difficult to detect due to the presence of disturbances usually referred to as "noise".

Recent important practical developments have been in computer software (Section 2.6) 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.

Since the 1990s there have been many important advances in the use of nonparametric methods in designed experiments, particularly those involving factorial structures or repeated observations on individuals or other experimental units. An introduction to these develpments is given in Chapter 8.

1.2 Populations and Samples

When making statistical inferences a key assumption is often that observations are a random sample from some population. That assumption is essential to the strict validity of many inferential procedures, although properties such as robustness may allow its relaxation in some, but not all, circumstances. Specification of the population may or may not be precise. If we select 20 books at random 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, nonfiction and reference works it is reasonable to assume that any inferences apply, at least approximately, to a wider population of books. This

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

If units are selected without replacement, a random sample of size n from a finite population is one where every possible sample of that size has an equal probability of selection. If sampling is with replacement a random sample is one where each item is independently selected with equal probability. This implies that if we arrange the sampled units in the order in which they are selected each possible ordered sample may be obtained with equal probability.

More often, 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 such as being 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 the 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. If the drug is beneficial, or better than that currently in use, the probability of better responses among those treated with it is increased. In Example 2.1 in Section 2.1 we formulate an appropriate nonparametric test for this situation.

Because of the many ways data may be obtained we need to consider carefully the validity and scope of inferences. 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 U.S., 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 who are alike in other relevant characteristics, e.g., with a similar age distribution. The patients in this study were all male, so 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 these inferences can be generalized. The implications of generalizing inferences is described for several specific tests by Lehmann (1975, 2006, Chapters 1–4).

In Section 1.1.1 we referred to a set of n independent observations from some normal distribution. In this situation independence implies, though this is not a definition of independence, that the value taken by any one observation tells us nothing about, and does not influence, the values of other observations. More formally, the probability that an observation lies in any small interval $(x, x + \delta x)$ is $f(x)\delta x$ where $f(x)$ is the probability density function of the relevant member of the normal distribution family. This notion extends to samples from any continuous distribution where $f(x)$ is the relevant probability density function.

The concept of observations being independent is an important, and by no means trivial, requirement for the validity of many statistical inference procedures.

1.3 Hypothesis Testing

Estimation, a topic we consider in Section 1.4, is often a key aim of a statistical analysis. Estimation may be explained in terms of testing a range of hypotheses, so we need to understand testing even though it is a technique that, in the view of many statisticians, tends to be overused and is even sometimes misused.

We assume familiarity with simple parametric hypothesis tests such as the t-test and chi-squared test, but we review some fundamentals and discuss changes in emphasis made possible by modern computer software. Until such software became widely available hypothesis testing nearly always required the use of tables.

Given a sample of n independent observations from a population having a normal distribution with unknown mean μ the t-test has a fundamental role in making inferences about μ. In the light of the *central limit theorem* the normality assumption may be somewhat relaxed.

We specify a null hypothesis, H_0, that μ has a specified value μ_0 and an alternative hypothesis, H_1, that it has some other value. Formally, this is stated as:

$$\text{Test } H_0\text{: } \mu = \mu_0 \text{ against } H_1\text{: } \mu \neq \mu_0. \tag{1.1}$$

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

$$\Pr(|t| \geq t_\alpha) = \alpha. \tag{1.2}$$

In practice α 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 α, or at the corresponding 100α percent level. The levels 0.05, 0.01 and 0.001 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.

To illustrate the point that traditional acceptance of a null hypothesis does not imply that H_0 is true, suppose that students over 18 years may study at a certain university. Most of the undergraduate courses run for three years, and the majority of the students go to university straight from school at age 18. However, on the courses there are some mature students aged 25 or more, making the overall mean age of the undergraduate students equal to 23. An external investigator who is unaware of the presence of these mature students specifies a null hypothesis H_0 that the mean undergraduate student age is 21; this null hypothesis is clearly untrue. However, suppose a sample of 50 undergraduates is selected and the mean age of these students is exactly 21 years. Because the sample mean is equal to the null hypothesis mean, the P-value for the test will be $P = 1$, the greatest possible value. It might then be tempting to deduce that H_0 is true, as the evidence is highly suggestive. However, this would be an incorrect conclusion.

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 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 α of these. Thus, 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 percent 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.

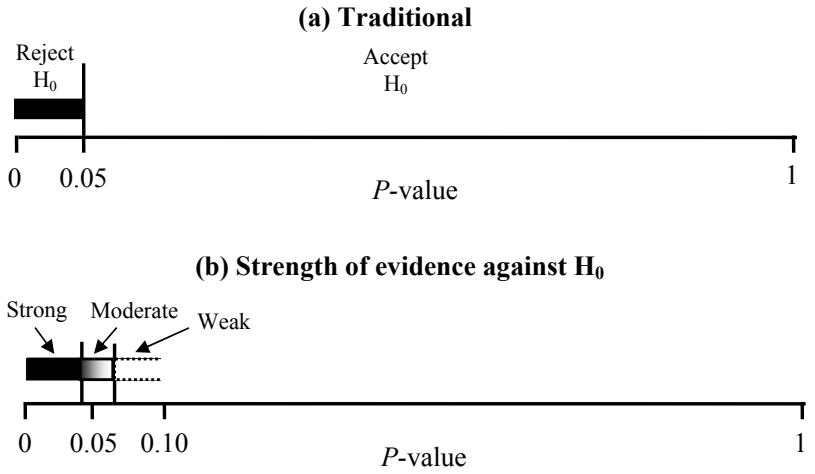


Figure 1.1 *(a) A fixed level approach to significance testing and (b) assessment based on strength of evidence where some flexibility in interpretation may be allowed for P-values close to 0.05.*

Modern statistical software lets us do something more sensible, though by itself still far from satisfactory. The output from any good computer program 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 called a P-value. It may be used as a measure of the strength of evidence against H_0 provided by the data — the smaller P is, the stronger is that evidence.

When we decide what P-values are sufficiently small for H_0 to seem implausible, we may speak formally of rejecting H_0 at the exact $100P$ percent significance level. This avoids the difficulty that rigid application of the 5 percent 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. Figure 1.1 compares the traditional "fixed level" approach to significance, with the more flexible assessment based on strength of evidence using a P-value.

A serious difficulty still remains. This is that with small data sets one may *never* observe sufficiently small P-values to justify rejection of H_0 even when it is not true. For instance, suppose a coin is tossed 5 times to test the hypothesis

$$H_0\text{: } \textit{the coin is fair},$$

against

$$H_1\text{: } \textit{the coin is biased}.$$

i.e., is either more likely to fall heads, or more likely to fall tails.

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, 0.5) distribution. The probability of each is $(0.5)^5$, so $P = 2 \times (0.5)^5 = 0.0625$. This is the smallest attainable P-value when $n = 5$ and $p = 0.5$, 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 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 that 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, the meaning of a P-value. We draw attention also to an ethical danger in Section 1.5.

Real-world policy decisions are often based on the outcome of statistical analyses. To appreciate the implications of either a formal rejection of, or a decision not to reject H_0, at a given significance level we need further concepts. Suppose we decide to reject H_0 whenever a P-value is less than some fixed value P_0, say. This means that 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. A P-value tells us the probability that we are making an error of the first kind by rejecting H_0.

In a t-test, if we reject H_0 whenever we observe a $P \leq P_0$ we do so when k is such that when H_0 holds, $\Pr(|t| \geq k) = P_0$. Such 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 even though, in fact, H_1 may be true. Continuing to accept H_0 when H_1 is true is an *error of the second kind*, or Type II error. Let β denote the probability of a Type II error.

The probability of a Type II error depends in part on the true value of μ, if indeed it is not equal to μ_0. Intuition correctly suggests that the more μ differs from μ_0, the more likely we are to get large values of $|t|$, i.e., values in the critical region, so that β 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 $\mu \neq \mu_0$ we increase β because the set of values of t for which we accept H_0 is larger. Another factor affecting β is the sample size, n. If we increase n we decrease β for a given μ and P_0. Thus β depends on the true value of μ (over which we have no control) and the value of n and of P_0 determining the size of the critical region. We often have some flexibility in the choice of n and P_0. We have framed our argument mainly in terms of the t-test statistics, but it generalizes to other test statistics.

Despite obvious limitations, P-values used constructively have a basic role in statistical inference. In Section 1.4 we show that a null hypothesis that specifies one single value of a parameter is usually 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 α, or the alternative P_0, does not determine β. We want β 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. The probability $1-\beta$ is the probability of getting a t-value in the critical region when H_0 is *not* true. It 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 μ, the t-test is more powerful than any other hypothesis test of the form specified in (1.1).

The historical choice of significance levels 5, 1 and 0.1 percent 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 would be silly to choose a significance level of 50 percent, 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 if a test has low power for one or more of these reasons:

- The true μ is close to the value specified in H_0.
- The sample size is small.
- We specify a very small P-value for significance.
- Assumptions required for the test to be valid are violated.

In later chapters we consider for some tests the often nontrivial problem of determining how big a sample is needed to ensure reasonable power to achieve given objectives.

Using small P-values in place of traditional 5, 1 and 0.1 percent significance levels gives more freedom in weighing evidence for or against a null hypothesis. Remembering that $P = 0.05$ corresponds to the traditional 5 percent significance level long used as a reasonable watershed, one should not feel there is strong evidence against a null hypothesis if P is substantially greater than 0.05. However, values of P not greatly exceeding 0.05 often point at least to a case for further studies. In particular, 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 α with an implication that we regard any observed P-value less than that α as sufficient evidence to prefer H_1 to H_0. Fixed levels comply with certain long-established conventions, and may be necessary for comparisons of the power of different procedures.

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. To be more specific, 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

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

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

Both lead 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 1000 g ingots of zinc produced by a standard process is normally distributed with a mean of 1.75 g 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. Either it is ineffective or it reduces the impurity. It is therefore appropriate to test

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

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

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

Use of a one-tail test is only justified if it is appropriate to the logic of the problem, as it is in the illustrations just given. It is not appropriate in the situation pertaining in (1.1).

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 P-value 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 relevant subregions are not symmetric about the mean. However, in many applications where relevant statistics have a chi-squared or an 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.

1.4 Estimation

1.4.1 Confidence Intervals

The sample mean is widely used as a point estimate of the population distribution mean if that mean exists. 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 describe a $100(1-\alpha)$ percent confidence interval for a parameter θ is to define it as the set of all values of θ for which, if any value in that set were specified in H_0, then the given data would lead to a $P > \alpha$. This implies that if a confidence interval includes the value of a parameter that is specified in H_0 there is no strong evidence against H_0. On the other hand a value specified in H_0 that lies well outside that confidence interval indicates strong evidence against H_0.

Another common interpretation of a $100(1-\alpha)$ percent confidence interval is in terms of the property that if we form such intervals for repeated samples, then in the long run $100(1-\alpha)$ percent of these intervals would contain (or cover) the true but unknown θ. 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 whether it is.

We elaborate on these points in Section 1.4.2.

A useful way of looking at the distinction 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 θ, what is the probability (P-value) of getting a sample as or less likely than that obtained if θ_0 is indeed the true value of θ?

whereas estimation using a confidence interval answers the question:

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

1.4.2 Precision Significance and Practical Importance

Example 1.3

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) are carried out to test the efficacy of a new drug (A) for reducing blood pressure. A third trial (III) is carried out with a second new drug (B). Trial I involves only a small number of patients, but trials II and III involve larger numbers. The 95 percent confidence intervals for mean blood pressure reduction (mm Hg) after treatment at each trial are:

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 percent significance level since the confidence intervals do not include zero. This implies strong evidence against H_0. Trial I is imprecise; we would accept in a significance test at the 5 percent level any mean reduction between 3 and 35 units. The former is not of clinical importance; 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 clinical 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. It tells us the mean reduction is likely to be between 21 and 25 units, a difference of clinical 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 discussed with examples by Gibbons and Chakraborti (2004), by Hollander and Wolfe (1999), by Desu and Raghavarao (2004) and in many other books and articles. 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 frequentist approach to inference. The Bayesian school adopts 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 frequentist approach throughout. For a reasoned argument for and against each approach see Little (2006) and for a comprehensive and realistic review of these and other approaches to inference Cox (2006) is recommended.

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 medicine, but increasing attention is being given to ethical issues in environmentally sensitive fields

like biotechnology and 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 is 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 is unacceptable to study some issues by allocating individuals to possible groups at random. For instance, in a study of the effects of smoking on health, one could not instruct individuals to smoke or to abstain from smoking. This disregards the autonomy of the study participants. An individual's choice of whether to smoke or not must be respected, and an alternative type of study planned to make this possible.

It is generally good practice to incorporate early stopping rules into a study. If it becomes clear at an early stage of an investigation that a new treatment is much better than the established alternative the study should be closed. A similar decision should be made if the new treatment quickly shows a highly increased risk of harmful side effects. Continuing a study could deprive patients of a more effective treatment or expose participants to unnecessary risks from side effects. Early stopping recommendations are reached through a small number of planned interim analyses that take place throughout the period of the study. Cannistra (2004) discusses ethical issues around the application of early stopping rules.

When planning a new study a comprehensive search of findings from related work is important 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 time, money and scarce resources will not be used to best effect. Nevertheless, results of literature searches 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 helpful in choosing an appropriate number of participants for the main study. A sufficiently large number should be involved 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 outlining the main points about the study. All available information about the possible efficacy and side-effects of 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.

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.

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.

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 Acts. Data used for statistical purposes should not contain patients' names or addresses.

Limited availability of a treatment for experimental use may create ethical problems. Suppose that there were high hopes that a new drug might greatly relieve suffering in severe cases but only enough doses were available to treat six patients. The principles of beneficence and justice suggest that the six 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 as less well than patients receiving an alternative treatment because, although their condition may have improved, their symptoms may still be more severe than those of patients receiving the alternative treatment. In this situation any statistical analysis should be based on some measure of "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 show less severe symptoms. However, if it is likely that only patients in the early stages of the disease will benefit it is 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.

Statisticians may be guilty of unethical behaviour. A statistician who 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, is guilty of unethical suppression of evidence.

Ethical considerations 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 Exercises

Whether the exercises below seem difficult or trivial will depend on the extent of the reader's prior training in statistics. Although not directly concerned with nonparametric methods, they are relevant to the background material covered in this chapter.

Solutions to exercises marked with an asterisk (*) in this and later chapters are discussed briefly in Appendix 2.

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 percent of the rods have diameter 30 mm or more. A second machine is known to produce rods with diameters normally distributed with mean 24 mm and 2.5 percent of rods it produces have diameter 30 mm or more. What is the standard deviation of rods produced by the second machine?

1.2 In a group of 145 patients admitted to hospital with a stroke, weekly alcohol consumption in standard units had a mean of 17 and a standard deviation of 22. Explain why their alcohol consumption does not follow a normal distribution. Is this finding surprising?

1.3 Following a television campaign about the risks of smoking tobacco, the cigarette consumption of a group of 50 smokers decreases by a mean of 5 cigarettes per day with a standard deviation of 8. Explain why the reasoning of Exercise 1.2 cannot be used to show that this distribution is not normal.

***1.4** In Section 1.3 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 percent significance level. What is the least number of tosses needed to provide such evidence using a two-tail test, and what is then the exact P-value?

1.5 A biased coin is such that $\Pr(\text{heads}) = 2/3$. If this coin is tossed the least number of times calculated in Exercise 1.4, what is the probability of an error of

the second kind associated with the 5 percent significance level? What is the power of the test? Does the discrete nature of possible P-values cause any problems in calculating the power?

***1.6** If a random variable X_i is distributed N(μ, σ^2) and all X_i are independent it is well known that the variable

$$Y = \sum_{i=1}^{n} X_i$$

is distributed N($n\mu$, $n\sigma^2$). Use this result to answer the following:

The times in minutes a farmer takes to place any fence post are each independently distributed N(10, 2). He starts placing posts at 9 a.m one morning, and immediately one post is placed he proceeds to place another, continuing until he has placed 9 posts. What is the probability that he has placed all 9 posts by (1) 10.25 a.m, (ii) 10.30 a.m and (iii) 10.40 a.m?

***1.7** The following two sample data sets both have sample mean 6.

Set I	13.9	2.7	0.8	11.3	1.3
Set II	2.7	8.3	5.2	7.1	6.7

If μ is the population mean perform for each set t-tests of (i) H_0: $\mu = 8$ against H_0: $\mu \neq 8$ and (ii) H_0: $\mu = 10$ against H_0: $\mu \neq 10$, Do you consider the conclusions of the tests reasonable? Have you any reservations about using a t-test for either of these data sets?

***1.8** Use an available standard statistical software package, or one of the many published tables of binomial probabilities to determine, for samples of 12 from binomial distributions with $p = 0.5$ and with $p = 0.75$, the probabilities of observing each possible number of outcomes for each of these values of p. In a two-tail test of the hypotheses $H_0 : p = 0.5$ against H_1: $p = 0.75$ what is the largest attainable P-value less than 0.05? What is the critical region for a test based on this P-value? What is the power of the test?

CHAPTER 2

FUNDAMENTALS OF NONPARAMETRIC METHODS

2.1 A Permutation Test

Parametric inference assumes observations are samples from populations with distributions belonging to a specified family. We pointed out in the previous chapter that for nonparametric inference we make only weaker assumptions such as one of symmetry, or where two or more populations are involved, that their distributions differ, if at all, only in some measure of location such as their *medians*. This calls for a new approach to hypothesis testing and estimation.

We introduce some basic ideas and illustrate their use primarily by examples. Our first illustration describes the procedure called a *permutation test*.

Example 2.1

Four from nine patients are selected at random to receive a new drug. The remaining five are treated with a standard 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 most satisfactory (rank 1) to least satisfactory (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. It is not impossible, but less likely, that those chosen would be ranked 1, 2, 3, 4 or 6, 7, 8, 9 after treatment.

There are 126 ways of selecting a set of four from nine patients. This may be verified using the well-known mathematical result that the number of ways of selecting r objects from n is $n!/[r!(n-r)!]$. For any integer m, the expression $m!$, called *factorial* m, is the product of all integers between 1 and m. We also define $0! = 1$. Table 2.1 gives all 126 selections. Ignore for the moment the numbers in parentheses 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 2.1. 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 reflects a beneficial effect of the drug.

Table 2.1 *Possible selections of four individuals from nine labelled 1 to 9 with the sum of the labels (ranks) in parentheses.*

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)				

In a hypothesis testing framework we have a group of four treated with the new drug and a group of five (the remainder) given a standard drug in what is called a *two independent sample experiment*. We discuss such experiments in detail in Chapter 6. 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: *new drug has no effect*

against the two-sided alternative

H_1: *new drug has an effect (beneficial or deleterious)*

the outcomes 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.

Table 2.2 *Number of occurences for each sum of ranks of four items from nine.*

Rank sum	10	11	12	13	14	15	16	17	18	19	20
Occurences	1	1	2	3	5	6	8	9	11	11	12
Rank sum	21	22	23	24	25	26	27	28	29	30	
Occurrences	11	11	9	8	6	5	3	2	1	1	

In classic hypothesis testing terms we speak of rejecting H_0 at an exact 1.59 percent significance level if we observed either of these extreme 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. 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 2.1, and count how many times each sum occurs we may 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 2.1 the number in parentheses 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 2.2 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. 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. It is easily seen from Table 2.2 that if we select in the lower tail $S = 10$ and $S = 11$, 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 $S \leq 11$ giving $P = 2/126 \approx 0.0159$. By symmetry, the upper-tail region is $S \geq 29$ also with $P \approx 0.0159$. 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., Gibbons and Chakraborti (2004) in their Table J and Hollander and Wolfe (1999) in their 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 are strong reasons before the experiment is started to believe 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 2.2.

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 discussed in Section 1.5.

Tests based on permutation of ranks or on permutation of certain functions of ranks (including the original measurements on a continuous scale when these are available) are central to many nonparametric methods. They are called *permutation* or *randomization* tests. The latter term applies when the permutation process is based on the randomization procedure used to assign treatments to units. That was the situation in Example 2.1, the permutations giving all possible assignments. These tests have an intuitive appeal and comply with well-established theoretical criteria for sound inference. This theoretical basis is summarized by Hettmansperger and McKean (1998) for many different procedures.

Small scale tests of a drug like that in Example 2.1 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 these 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.

2.2 Binomial Tests

One observation was censored in the data in Example 1.2. We mentioned that it could be shown that it was not unreasonable, given that data, to accept a hypothesis that the population median was 200. We now consider an appropriate test to justify that conclusion.

Example 2.2

The data for survival times in weeks given in Example 1.2 were

$$49 \quad 58 \quad 75 \quad 110 \quad 112 \quad 132 \quad 151 \quad 276 \quad 281 \quad 362^*$$

The asterisk denotes the censored observation.

We want to test the hypothesis that the median, θ, of survival times for the population from which the sample was obtained is 200 against the alternative of some other value, i.e., to test

$$\mathrm{H}_0\colon \theta = 200 \text{ against } \mathrm{H}_1\colon \theta \neq 200 \tag{2.1}$$

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 continuous 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, 0.5) distribution.

The probability of observing r plus signs in 10 observations when $p = 0.5$ is given by the binomial formula

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

where

$$\binom{10}{r} = \frac{10!}{r!(10-r)!}$$

and is called the *binomial coefficient*.

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
p_r	0.0010	0.0098	0.0439	0.1172	0.2051	0.2461
r	6	7	8	9	10	
p_r	0.2051	0.1172	0.0439	0.0098	0.0010	

In the data 3 observations, including the censored one, exceed 200 so there are 3 plus signs and, from the table above, we see that when H_0 is true the probability of 3 or less plus signs in a sample of 10 is $0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.1719$. There is no strong evidence against H_0, tail probabilities for our observed statistic, the number of plus signs) is $2 \times 0.1719 = 0.3438$. This implies 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 3.3. The test is *distribution-free* because we have made no assumption about the form of the continuous distribution of the underlying observations. We have only formulated and tested hypotheses concerning possible values of the population median.

When applying a t-test, or most other parametric tests, all values of P between 0 and 1 are possible. For the sign test, however, only certain discrete P-values occur. In this example, for a two-tail test the three smallest are $P = 2 \times (0.0010) = 0.0020$ corresponding to 0 or 10 plus; $P = 2 \times (0.0010 + 0.0098) = 0.0216$ 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 $P = 0.3438$. In all cases probabilities have been rounded to four decimal places. For a one-tail test these P-values are all halved. Our statistic — the number of plus signs — has a discrete distribution. This means that, as in Example 2.1, 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.

Once they are recognized, and the consequences appreciated, discontinuities in possible P-values do not cause serious interpretational problems in the analysis of a particular data set. However, these discontinuities do lead to some theoretical difficulties in comparing performance of competing tests.

A device called a *randomized decision rule* has been proposed with the property that in the long run an error of the first kind has, in repeated testing, a probability at a prechosen nominal level, e.g., at 5 percent. 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. An account of how a randomized decision rule works is given by Gibbons and Chakraborti (2004) (pp. 28–29). They rightly comment that such devices may seem artificial and are "probably seldom employed by experimenters". We suggest they should never be used in real-world applications.

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.

The sign test provides a basis for forming a confidence interval for the population median.

Example 2.3

We saw in Example 2.2, when using a sign test for a median with a sample of 10, we would, in a two-tail test at the 2.16 percent level, accept H_0 if we got between 2 and 8 plus signs.

Consider again the data in that example, i.e.,

$$49 \quad 58 \quad 75 \quad 110 \quad 112 \quad 132 \quad 151 \quad 276 \quad 281 \quad 362^*$$

where the asterisk represents a censored observation. 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. This implies that the interval (58, 281) is a $100(1 - 0.0216) = 97.84$ percent confidence interval for θ, the population median survival time. Since we would accept any H_0 that specified a 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 the estimate lacks precision.

Care is needed in interpreting P-values especially in one-tail tests. 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 0.5, 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 2.4 and the remarks following it. If the test statistic has a symmetric distribution, doubling is equivalent to considering equal deviations from the median value of the statistic in either direction. 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 2.4 exposes 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 that hypothesis 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 2.4

In a dental practice, experience has shown that 75 percent of adult patients require treatment following a routine inspection. So the number of individuals requiring treatment, S, in a sample of 10 independent patients has a binomial B(10, 0.75) 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

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

The relevant probabilities are

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

If we had data for another practice and wanted, for that practice, to test H_0: $p = 0.75$ against H_1: $p > 0.75$, the smallest P-value for testing is in the upper tail and is associated with $r = 10$, i.e., $P = 0.0563$. This means that if we only regard $P \leq 0.05$ as sufficiently strong evidence to discredit H_0 such values are never obtained.

There would be no problem here for a one-tail test of H_0: $p = 0.75$ against H_1: $p < 0.75$ 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 a 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 $P \leq 0.05$. Even if we did, it would be for the wrong reason. 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.

An implication is that in this example we need a larger sample to detect departures of the form $H_1 : p > 0.75$. Again, the fairly large P-value associated with the possible critical region for the one-tail test 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. We consider situations of this kind in more detail in Section 15.7

There is no universal agreement 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 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, as already pointed out, this is equivalent to doubling the probability, but it seems inappropriate with a nonsymmetric distribution. In Example 2.4 the region $r \leq 4$ is appropriate for a lower-tail test. The mean of the test statistic (the binomial mean np) is here 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.

2.3 Order Statistics and Ranks

Many nonparametric procedures are based on the ordering, or ranking, of detailed observations. In Examples 2.2, we did not use ranks, but ordering was inherent in our procedure. We took order into account in determining the number of survival times that exceeded the hypothesized median.

Ordering data is important in more general statistical contexts, both parametric and nonparametric. We may be interested in the distribution of the largest or smallest observations in a sample to answer questions such as

- On the basis of maximum flood levels recorded in a river over a number of years, what is the probability of the level exceeding, say, 5 m, in future?
- Given a sample of times to first breakdown of a certain brand of computer, what is the probability of a first breakdown being observed within 6 months in one machine in a production run of 1000 machines?

In a parametric context such questions are often answered using families of distributions called *extreme value distributions*. A simple example of the role of order statistics in a parametric context is given in Exercise 2.10.

Greatest and least values in samples are just two examples of *order statistics*. The sample median is also an order statistic.

A detailed account of order statistics and their properties is given by Gibbons and Chakraborti (2004, Chapter 2). Here we only indicate the relevance of these statistics to nonparametric inference, and quote some key results without proof.

Consider a sample of n observations $x_1, x_2, \ldots, x_n$ from a continuous distribution. Continuity implies that there should be no ties and thus observations may be uniquely ordered from smallest to largest. We denote the smallest observation by $x_{(1)}$, the second smallest by $x_{(2)}$ and so on, finally the largest by $x_{(n)}$. It follows that

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

The $x_{(i)}, i = 1, 2, \ldots, n$ are called the *order statistics*. The minimum order statistic, $x_{(1)}$, is relevant to the study of minimum extremes such as the distribution of shortest times to a machine breakdown, or minimum survival times after some treatment. The largest, $x_{(n)}$, is relevant to the study of floods, or maximum time to failure of a certain type of lightbulb.

The median is widely used as a measure of location in nonparametric inference, and the sample median is defined in terms of order statistics. For a sample of n the median is $x_{[(n+1)/2]}$ if n is odd, and is usually defined as $[x_{(m)} + x_{(m+1)}]/2$ if $n = 2m$ is even. A possible measure of dispersion is the sample range $x_{(n)} - x_{(1)}$. More satisfactory measures are the *interquartile range* or *semi-interquartile range* defined in Section 2.4. One reason for preferring one of the latter is that the extreme order statistics are often strongly influenced by suspect observations associated with the terms *outliers* and *dirty data*.

In nonparametric inference an important concept based on order statistics is the *sample*, or *empirical*, distribution function. For a random sample of size n from a population having cumulative distribution function $F(x)$, the sample, or empirical, distribution function is defined as

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

For any x the value of $S_n(x)$ is expressible in terms of the order statistics.

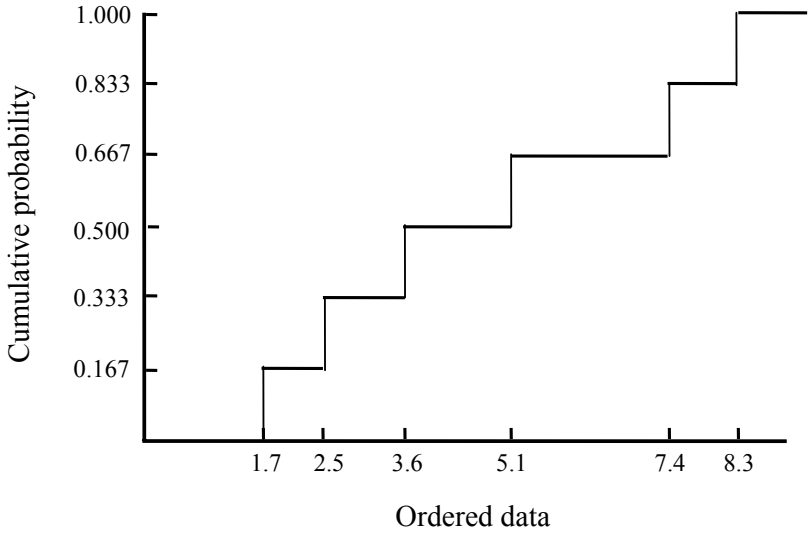


Figure 2.1 *Sample distribution function for a sample of six.*

It is easy to see that

$$\begin{aligned} S_n(x) &= 0 \quad \text{if } x < x_{(1)}, \\ S_n(x) &= i/n \quad \text{if } x_{(i)} \le x < x_{(i+1)},\ i = 1, 2, \ldots, n-1, \\ S_n(x) &= 1 \quad \text{if } x \ge x_{(n)}. \end{aligned} \tag{2.2}$$

The function $S_n(x)$ is a step function with a step of size $1/n$ at $x = x_{(i)}$, $i = 1, 2, \ldots, n$. This is illustrated in Figure 2.1 for six observations

$$1.7,\ 2.5,\ 3.6,\ 5.1,\ 7.4,\ 8.3\,.$$

The sample cumulative distribution function $S_n(x)$ is important because it is closely related to the population cumulative distribution function $F(x)$. This is reflected in the following properties:

- The mean value of $S_n(x)$ is $\mathrm{E}[S_n(x)] = F(x)$.
- The variance of $S_n(x)$ is $\mathrm{Var}[S_n(x)] = F(x)[1 - F(x)]/n$.
- $S_n(x)$ is a consistent estimator of $F(x)$ for any fixed x.

The term *consistent estimator* implies that $S_n(x)$ converges in probability to $F(x)$ as n tends to infinity. For a proof of these properties see Gibbons and Chakraborti (2004, Section 2.3).

The assumption that $F(x)$ is continuous rules out, in theory, the possibility of tied observations. In practice tied values in sample data are not uncommon. This may be due to rounding in the recording of data, or to the population

distribution being not strictly continuous. We see the practical implications of tied data in specific techniques in later chapters.

2.4 Exploring Data

The addition of nonparametric or distribution-free methods to the procedures for making statistical inferences widens the choice of techniques appreciably, An invaluable first step in selecting an appropriate technique in any given situation is to use *exploratory data analysis* or EDA. Some basic tools of EDA are

- Descriptive statistics.
- Boxplots.
- Histograms and frequency curves.
- Empirical and theoretical cumulative distribution graphs.

Descriptive statistics are commonly presented in lists or tables. The other tools above are by nature graphical. Commonly met descriptive statistics that summarize key features of sample data are the sample mean, median, maximum value, minimum value, standard deviation and quartiles. Slightly less well-known ones, but often of interest when questions of robustness arise, are the *trimmed mean* and the *Winsorized mean*. We introduce the last two in Chapter 14. Statistics such as the mean, median, standard deviation, or other derived quantities, are often referred to as *secondary* data to distinguish them from the original raw or observational data called *primary data*. Most general statistical software packages have a facility for computing a wide range of descriptive statistics.

A study of relevant descriptive statistics may give a quick indication of, for example, whether an assumption of normality appears to be seriously invalidated, or whether it is reasonable to suppose the sample comes from a symmetric or a skew distribution; and if the latter, whether the long tail is to the left or right. Such basic characteristics may be explored more fully by graphical techniques. These are often useful to indicate how well samples reflect population features. We have already indicated in Section 2.3 that the sample, or empirical, distribution function is a consistent estimator of the population cumulative distribution function.

We indicate the use of some basic EDA tools by examples.

Example 2.5

In Appendix 1 we give four small data sets and indicate how they were collected. Table 2.3 gives a set of descriptive statistics useful for summarizing and comparing the data for each of the four sets.

The first row tells us the number of data for each set. The small sample of 13 observations for the *McDelta* clan might be expected to be less informative than the sample of 59 *McAlphas*.

Table 2.3 *Descriptive, or summary, statistics for Badenscallie data given in Appendix 1.*

Clan	*McAlpha*	*McBeta*	*McGamma*	*McDelta*
Number	59	24	21	13
Mean	61.8	61.1	62.9	48.1
Median	74.0	67.5	77.0	65.0
St. dev	27.52	24.92	26.77	33.45
SE mean	3.58	5.08	5.84	9.28
Minimum	0	0	13	1
Maximum	95	96	88	87
1st quartile	44.0	41.5	33.0	13.0
3rd quartile	81.00	78.75	83.50	80.00
Range	95	96	75	86
IQ range	37.00	37.25	50.50	67.00

The sample mean for *McDelta* is markedly lower than that for the other clans. We may be interested in whether this indicates a shorter average life expectancy for that clan, or whether the difference represents some sampling quirk that might disappear if we had a larger sample.

The medians are all appreciably higher than the means, suggesting that the distributions of ages are asymmetric. This follows because we expect samples to reflect broadly the population characteristics, and for symmetric population distributions the mean and median coincide.

The abbreviation *St. dev* is used in the table for the *standard deviation*, the well-known measure of spread that in the case of a sample from a normal population is an appropriate estimator of the parameter σ. Once again the clan *McDelta* is the odd one out.

SE mean is an abbreviation for *standard error of the mean*. If we denote the sample standard deviation by s, then the standard error of the mean is computed as $s/\sqrt{n}$. Thus, the standard error decreases with sample size for a given standard deviation.

The maximum and minimum ages at death indicate at least one case of infant mortality for each clan except *McGamma*, and at least one nonogenarian survivor for two of the clans.

The *quartiles* divide each ordered sample into four groups of equal size. If we consider the median as dividing the sample into two groups of equal size, the first quartile is in effect the median of the group of lower values and the third quartile is the median of the group of higher values. More formally the first quartile is the median of $x_{(1)}, x_{(2)}, \cdots x_{((n-1)/2)}$ if n is odd and is the median of $x_{(1)}, x_{(2)}, \cdots x_{(n/2)}$ if n is even, with corresponding definitions for the third quartile. The *second quartile* is the sample median. While the third quartiles are similar for all clans, the first

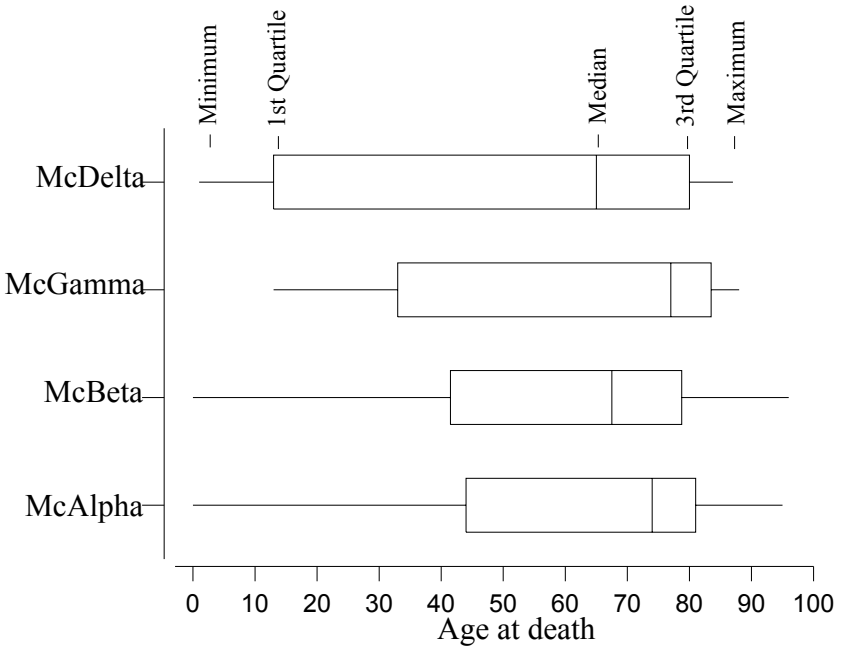


Figure 2.2 *Boxplots for Badenscallie data given in Appendix 1.*

quartile is strikingly low for *McDelta*. In a more formal analysis we may want to know if this can be accounted for by a quirk of the relatively small sample, or if it represents a different age distribution from that of the other clans.

Range and interquartile range, the latter abbreviated in the table to *IQ range*, are respectively the differences *maximum–minimum* and *third quartile–first quartile*. Each is a measure of spread alternative to standard deviation. Of the two, the interquartile range is preferred because range depends only on two observations x_1 and x_n, either of which may represent some unusual, or even a rogue, observation. On the other hand the interquartile range covers an interval containing the central 50 percent of the observations. Intuitively, this may be expected to be a more stable estimate of general variability. As an alternative to the interquartile range the *semi-interquartile range* is often used. As its name implies, it is obtained by dividing the interquartile range by 2. For the clan data the striking differences in interquartile range might be an aspect of the data requiring further analysis.

The five descriptive statistics presented in the order *minimum, 1st quartile, median, 3rd quartile, maximum* consititute a *five number summary*. This is the basis of what is called a *boxplot* or a *box and whisker plot*. Figure 2.2 gives boxplots for each clan for the Badenscallie data based on 5-number summaries easily obtained from Table 2.3.

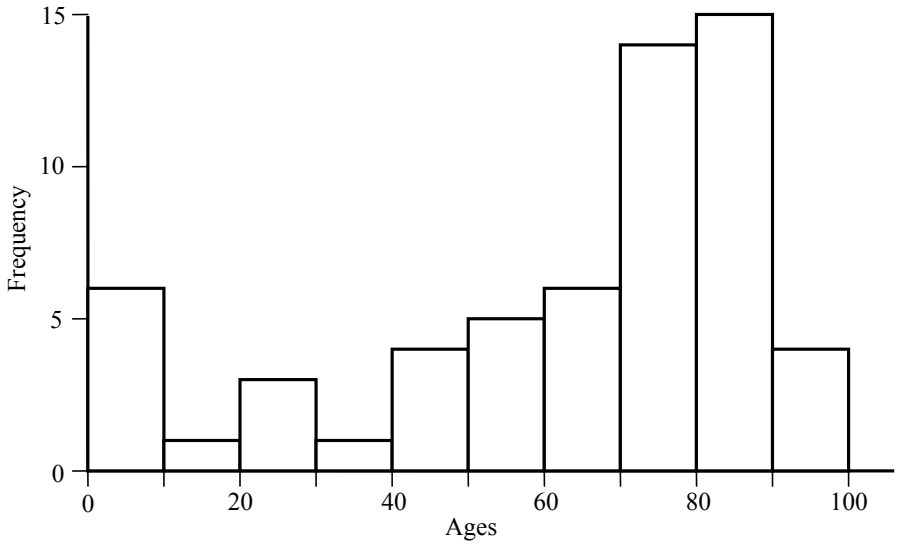


Figure 2.3 *Histogram for clan McAlpha data given in Appendix 1.*

The labels attached to the boxplot for the McDelta clan apply to any boxplot and indicate that the *box* section extends from the first to the third quartile. The vertical line dividing this box into two portions represents the median. The horizontal line outside the boxes extends from the minimum to the maximum.

Including box and whisker plots for all four clans on the one diagram enables useful comparisons of the kind outlined above to be made very easily. In particular, remembering that half the observations lie at or above the median, and half lie at or below the median, we see that for all clans the distribution of ages at death is skewed to the left or lower tail. This is made very clear by the median in all cases being nearer to the third quartile than to the first quartile. Recall that the quartiles are effectively the medians of the lower and upper halves of the data respectively.

Histograms are another widely used graphical device to exhibit key data characteristics. Figure 2.3 is a histogram based on the clan McAlpha data for ages at death with a class interval of 10 years. The long tail to the left is evident. There is also an indication of a mixture of distributions, with a smaller portion of the data indicating *infant mortality* or death before reaching adulthood, while the larger portion represents a more normal (in the physiological but not necessarily in the statistical sense) lifespan peaking at an age close to 80 years.

We pointed out in Section 2.3 that the sample or empirical distribution function was a consistent estimator of the population distribution function. This reflects the fact that as the size of a random sample increases it mirrors

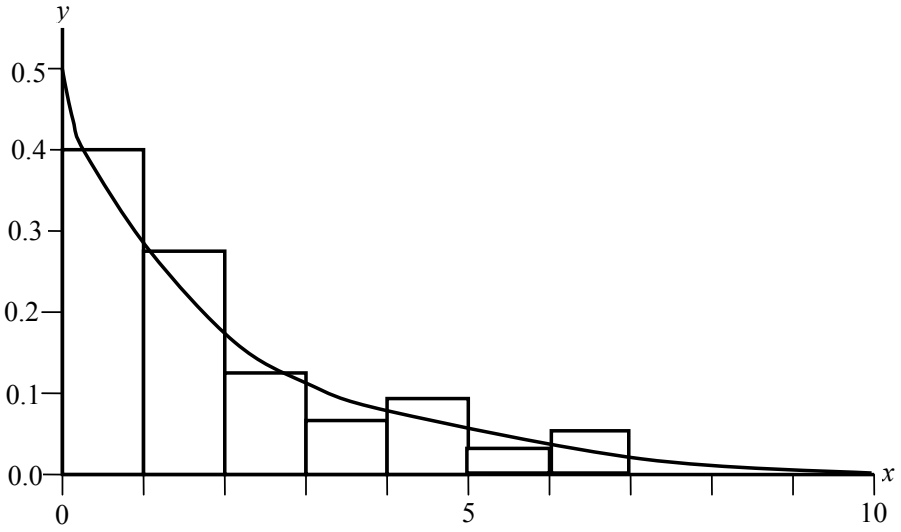


Figure 2.4 *Histogram for a sample of 50 from an exponential distribution with mean 2. The fitted curve is that of the distribution frequency function.*

the population characteristics ever more closely. Modern statistical software packages allow one to draw random samples of any chosen size from a wide range of distributions. For reasonably large samples, i.e., those of at least 50 observations, constructing appropriate histograms and superimposing these on the relevant population distribution frequency function gives a good impression of how effective these matches are.

Example 2.6

A computer-generated sample of 50 observations from an exponential distribution with mean 2 using Minitab gave the histogram in Figure 2.4. All sample values were less than 10, and 20 of them lay in the interval [0, 1), 14 in the interval [1, 2), 6 in the interval [2, 3), and so on. The curve superimposed on the histogram is that of the *frequency function* or *probability density function* of the exponential distribution with mean 2, which has the form:

$$f(x) = \frac{1}{2}e^{-x/2}, \qquad x \geq 0.$$

Statisticians would regard the closeness of the curve to the histogram as an indication that the data might be a sample from this distribution.

For samples smaller than 50 the grouping required to form a histogram may result in a rather poor fit to the population frequency function. However, even for small samples the sample distribution step function usually lies fairly close to the population cumulative distribution function.

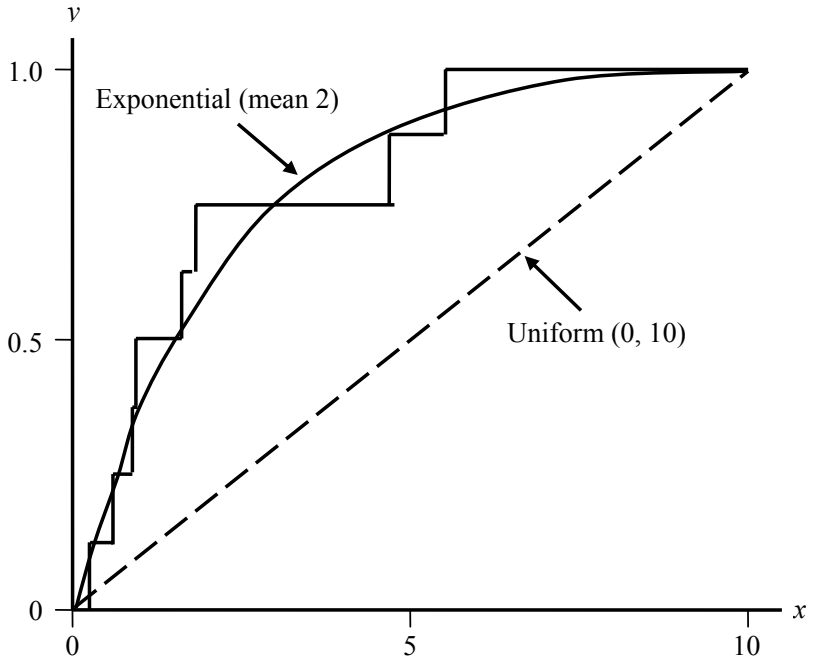


Figure 2.5 *Sample cumulative distribution function (stepped) for a sample of eight from an exponential distribution with mean 2. The curve is the population cumulative distribution function and the straight line is that for a uniform distribution over (0,10).*

Example 2.7

A computer generated sample of eight from an exponential distribution with mean 2 gave the values

$$0.25 \quad 0.53 \quad 0.91 \quad 0.94 \quad 1.56 \quad 1.73 \quad 4.71 \quad 5.50$$

where these have been arranged in ascending order. Figure 2.5 shows the sample cumulative distribution function for these data (stepped function) and the cumulative distribution function for an exponential function with mean 2. This takes the form

$$F(x) = 1 - e^{-x/2}, \qquad x \geq 0.$$

The step function lies close to this population cumulative distribution function. For illustrative purposes the straight line joining the points (0, 0) and (10, 1) on the graph is the cumulative distribution function for a uniform distribution over (0, 10). It is almost self-evident that our sample was not taken from that distribution.

More sophisticated EDA methods include the so-called P–P and Q–Q plots, abbreviations for plots of probabilities and of quantiles respectively associated with two distributions or with a hypothesized distribution and a sample believed to be from a population having that distribution. A description of how

these are used and interpreted is given by Gibbons and Chakraborti (2004, Section 4.7).

The examples in this section only touch on the potential of an EDA approach. Further examples are given throughout this book.

2.5 Efficiency of Nonparametric Procedures

We pointed out in Section 1.3 that the power of a test depends upon (i) the sample size, n, (ii) the choice of the largest P-value to indicate significance (usually denoted in power studies by α), (iii) the magnitude of any departure from H_0 and (iv) whether assumptions that are 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. This is because such tests are usually also good at detecting larger departures, a desirable state of affairs.

If α is the probability of a Type I error, and β 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 α, β. In practice, we usually fix α at some P-value appropriate to the problem at hand. Then β 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 α, β.

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 α and then allow the alternative in H_1 to vary in such a way that β remains constant as the sample size n_1 increases. For each n_1 we determine n_2 such that T_2 has the same β for the particular alternative considered.

Increasing sample size usually increases the power for alternatives closer to H_0. Therefore, 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 $n_1 \to \infty$. More importantly, this limit, which he called the *asymptotic relative efficiency* (ARE) was the same for all choices of α, β. A full discussion of asymptotic relative efficiency is given by Gibbons and Chakraborti (2004, Chapter 13).

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 a rather low $2/\pi \approx 0.64$. Lehmann (1975, 2006) shows that for samples of size 10 and a range of values of the median θ relative to the value θ_0 specified in H_0 with α fixed, the relative efficiency exceeds 0.7. 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 small sample sizes.

We have already mentioned that when it is relevant and valid the t-test 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 (at least for large samples) is as efficient as a t-test applied to a sample of size $2n$. There are, however, situations where asymptotic relative efficiency may give an unduly optimistic picture of small sample behaviour.

2.6 Computers and Nonparametric Methods

Computer software packages suitable for nonparametric analysis fall into three main categories. The first is specialist menu-driven packages that use exact permutation or related methods for small to medium sized samples and provide Monte Carlo and/or asymptotic tests for larger samples.

The second category are the mainstream menu-driven statistical software packages that allow exact inferences for some, but by no means all, widely used nonparametric tests, or are user-friendly in the sense that they allow the user to write programs to carry out such procedures.

The final category is comprised of versatile interactive statistical packages that have a variety of options, or tools, to perform various data manipulations and statistical operations. These are not menu driven. The user combines relevant tools, often with further options of his or her own creation, to achieve some desired objective. Such programs are by their nature generally less user-friendly than menu-driven packages, but they are often more powerful.

In the first category widely used packages are StatXact 7.0, distributed by Cytel Software Corporation, Cambridge, Ma, and Testimate, distributed by IDV Daten-analyse und Versuchs-planung, Munich, Germany. StatXact gives exact permutation P-values for small samples together with Monte Carlo estimates of these, for a large 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 tests in StatXact are also available in SAS. Testimate has considerable overlap with StatXact, but some methods are included in one but not both these packages and there are minor differences between the packages in detail for some procedures. There are also specialized programs dealing with particular aspects of the broad fields of nonparametric and semiparametric inference. These include LogXact, which is especially

relevant to logistic regression, a topic only covered briefly in Chapter 15 in this book.

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 Tsiatis (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.

General statistical packages such as SAS, Minitab, SPSS, and Stata include some nonparametric procedures. In some of these exact tests are given, but many rely heavily on asymptotic results, sometimes with little warning about when, particularly with small or unbalanced sample sizes, these may be misleading.

In the third category the increasingly popular R, and the closely related S-PLUS are particularly useful for the bootstrap described in Chapter 14, as well as for some of the semiparametric procedures discussed in Chapter 15.

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. Packages such as R and SAS have a versatility that makes combining of approaches such as EDA and more formal analyses quick and easy.

Users should test nonparametric procedures in any package programs they use with examples from this book and other sources. In some cases the output will 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 percent) significance levels rather than exact P-values. Sometimes the convention of doubling a one-tail P-value may be used to obtain a two-tail test value, but as indicated in Example 2.4, this may not always be appropriate. Particular care should be taken to check whether exact or asymptotic results are given.

This book is largely about well-established methods, but only modern computing facilities allow us to use them in the way we describe. Solutions to examples, or illustrations in this book using statistical packages, are usually based on StatXacT, Minitab or R, but in many cases it would be equally appropriate to use other well-known packages such as SAS, SPSS, Stata, etc., providing these packages contain relevant programs.

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.

2.7 Further Reading

Hollander and Wolfe (1999), Conover (1999), Gibbons and Chakraborti (2004), Higgins (2004) and Desu and Raghavarao (2004) give, in some cases, more background for some of the procedures described here. Each book covers a slightly different range of topics, and at varying depths, but all are suitable references for those who want to get a broad picture of the many aspects of basic nonparametrics. Daniel (1990) is a general book on applied nonparametric methods.

A moderately advanced mathematical treatment of the theory behind 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. A classic is the book by Lehmann (1975), a revised edition of which appeared in 2006. This book repays careful reading for those who want to pursue the logic of the subject in more depth without too much mathematical detail. Applications in the social sciences are covered by Leach (1979) and by Siegel and Castellan (1988), the latter 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. Rayner and Best (2001) give a wide ranging treatment of many standard and a few specialist procedures using methods based largely on partitioning of the chi-squared statistic. Wasserman (2006), despite its title, deals mainly with more advanced modern topics in nonparametric statistics, a few of which we touch upon in Chapters 14 and 15. He gives a lucid introduction to those topics he covers.

Agresti (1984, 1996, 2002) 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 (2006). Good (2005) and Edgington (1995) cover randomization and permutation tests. The theory behind rank tests is given by Hájek, Sidák and Sen (1999).

Books dealing with the bootstrap include Efron and Tibshirani (1993), Davison and Hinkley (1997) and Chernick (1999).

2.8 Exercises

2.1 A new type of intensive physiotherapy is developed for individuals who have undergone spinal surgery. Due to limited hospital resources it can only be given to 3 out of 10 patients. The patients are aged:

15 21 26 32 39 45 52 60 70 82

Explain how a permutation test could be used to investigate whether use of the physiotherapy is related to patient age, (i.e., whether there is a policy to give the

treatment to younger as opposed to older groups or *vice versa*). If the patients aged 15, 26 and 32 have the intensive physiotherapy find the P-value for a two-tailed test of an appropriate null hypothesis. Comment on your findings.

***2.2** Suppose that the new drug under test in Example 2.1 has all the ingredients of a standard drug at present in use and an additional ingredient that 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.

2.3 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?

2.4 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, θ, 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 an 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?

***2.5** The numbers of pages in the sample of 12 books in Exercise 2.4 were:

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

Find a confidence interval at a level not less than 95 percent for the median θ.

***2.6** In Sect.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 percent based on the sign test criteria for the 12 book sample values given in Exercise 2.5 relevant to a test of $H_0 : \theta = \theta_0$ against a one-sided alternative $H_1 : \theta > \theta_0$.

2.7 From 6 consenting patients requiring a medical scan, 3 are chosen at random to undergo a positron emission tomography (PET) scan, the others receiving a magnetic resonance imaging (MRI) scan. Image quality is ranked in order by a hospital consultant from 1 (best) to 6 (worst). Describe how you would test H_0: *scan quality is unrelated to scan method* against (i) H_1: *PET scans are better* (ii) H_1: *the scans differ in quality depending on whether they are from PET or MRI.* Interpret the finding that the consultant rates the three PET scans as the three highest quality images.

2.8 In Example 2.4 we remarked that a situation could arise where we might reject H_0 for the wrong reason. Explain how this is possible in that example.

***2.9** State appropriate null and alternative hypotheses for the example from the book of Daniel about diet in Section 1.1.3. How could you use ranks to calculate the probability that the four receiving the diet of pulses were ranked 1, 2, 3, 4? Calculate this probability assuming that there were 20 young men involved altogether.

***2.10** A sample of 12 is taken from a continuous uniform distribution over the interval (0, 1). What is the probability that the largest sample value exceeds 0.95? (Hint: Determine the probability that any sample value exceeds 0.95. The condition is met if at least one value exceeds 0.95.)

2.11 A sample of 24 is known to come either from a uniform distribution over the interval (0, 10) or else from a symmetric triangular distribution over the same interval (0, 10). The sample values are

4.17	8.42	3.02	2.89	9.77	6.06	2.72	5.12	6.00	4.78	2.62	7.20
1.61	5.92	7.25	8.01	4.76	5.36	5.34	7.59	0.66	7.27	3.39	1.40

Use appropriate graphical or other EDA techniques to get an indication as to which of these distributions is the more likely source of the sample.