

**Chapter 2 of “Applied Nonparametric Statistical Methods”, 5th Edition**

**FUNDAMENTALS OF NONPARAMETRIC METHODS**

**Final Draft**

Nigel C. Smeeton, Neil H. Spencer, Peter Sprent

Abstract

The concept of parametric inference is described, and the terms nonparametric, distribution-free and semiparametric compared in the context of statistical procedures. Permutation tests, which are based on the number of ways in which subgroups can be selected from the overall sample, are introduced. In null hypothesis testing,  $P$ -values are limited to a finite number of discrete values. These may exclude conventional benchmark values and this potential shortcoming is discussed. The reader is introduced to the binomial test with an example in which success and failure probabilities are equal (effectively the sign test). An explanation of extreme values is given from which the term order statistic follows. Exploratory data analysis is described, looking at descriptive statistics, the five-number summary, boxplots, histograms, frequency curves and cumulative distribution graphs. Superimposing a sample histogram on a population probability density function gives an indication of how well the data fit the proposed model. The efficiency of nonparametric procedures in terms of power and sample size is described, leading to the concept of asymptotic relative efficiency in the comparison of two tests. The statistical software available for nonparametric data analysis is discussed. The chapter finishes with suggestions for further reading.

## 2.1 Parametric and nonparametric methods

### 2.1.1 Approaches to inference

A probability density function, as described in Chapter 1, can be used to represent the distribution of a population. As hinted in Section 1.4.1, distributions are associated with parameters, and a particular distribution has a specific shape uniquely determined by the value(s) of its parameter(s). For instance, a Normal distribution must have a mean equal to zero and a standard deviation equal to 1 in order for it to take the shape of a standard Normal distribution. Other combinations of values for the mean and standard deviation lead to different specific shapes that have the general Normal form, giving rise to the *family* of Normal distributions. A particular distribution is often indicated by the notation  $N(\mu, \sigma^2)$ . Other families of distributions include the uniform (Section 1.1), binomial, Poisson and exponential distributions (Bury, 1999). Statistical deduction based on estimating the values of one or more parameters is known as *parametric inference*.

It may be reasonable based on past experience or current knowledge to assume that observations come from a particular family of distributions. Many measurements, including those for height and weight, generally have a distribution that, apart from a few extreme values, is close to Normal. For larger samples, the central limit theorem justifies the use of the Normal distribution in making deductions about a population from a sample, even if the population is non-Normal. These are described as *asymptotic approximations* as the accuracy of the approximation increases as the sample size becomes larger.

Parametric inference may not always be appropriate and in some situations it is impossible. For example, records of school examination results may only give candidate attainment in the

form of banded and ordered grades designated Grade A, Grade B, Grade C, etc. Given two different schools and information on the grade frequencies for each, it may be of interest to determine whether there is a genuine difference in performance that might be attributed to different methods of teaching, or to the ability of one school to attract more able students. There is no obvious family of distributions associated with these data, and there are no clearly defined parameters about which inferences can be made. Two terms are in common use for the type of inferences that can nevertheless be made. They are '*nonparametric*' and '*distribution-free*'. The description 'nonparametric data' is incorrect as these terms refer to the method of analysis.

Use of the terms nonparametric and distribution-free is not consistent and they are often interchanged. This is of no great consequence in practice, and to some extent it simply reflects historical developments. There is not even universal agreement about what constitutes a parameter. Quantities such as  $\mu$  and  $\sigma^2$  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, quantile, or range. Situations can arise where observations are composed of a deterministic and random element and constants occurring in the deterministic element are to be estimated. Such constants are also sometimes called parameters. In nonparametric, or distribution-free, methods inferences about parameters in this wider sense can be made.

In this situation the name distribution-free is more appropriate if interest is 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 widespread. 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*. As shown in Section 2.4 and elsewhere, sensible use of exploratory data analysis 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 that they are assumption free. In practice, nearly always some assumptions are made about the underlying population distribution. For example, a distribution may be assumed to be continuous and symmetric. These assumptions do not restrict reference to a particular family such as the Normal, but they exclude both discrete and asymmetric distributions. Given some data, exploratory data analysis will often indicate whether 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 distribution). 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.

### *2.1.2 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. These tests are described as *robust*, and findings are similar whether or not the underlying assumptions are satisfied.

Nonparametric methods are usually more robust than their parametric counterparts. They are widely used with data that can be ranked and are often the only ones available for data consisting of the number of individuals in various categories.

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

#### *Example 2.1*

Two machines produce metal rods. For each, 2.5 percent of all rods produced have a diameter exceeding 30 mm. 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 (Section 1.6.1), so 2.5 percent exceed  $27 + 1.96 \times 1.53 \approx 30$  mm. 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). This follows because any interval between 20.25 and 30.25 mm of width 0.25 mm (in particular from 30 to 30.25 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 the proportion of defectives may be the same in two populations, yet each has a different mean and their distributions do not even belong to the same family.

If an additional assumption is made that distributions of diameters for each machine differ, if at all, only in their means, then if the proportion over 30 mm in samples of, say, 200 from each machine is known, it would be possible to test whether the means could reasonably be supposed to be identical. Doing this would not make best use of the data. 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 centered. Both are formally described as *measures of location*. Not all distributions have a mean, but all have a median. For a symmetric distribution, if the mean exists, the mean and the median have the same value. Their values usually 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 are often concerned with location 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. Manufacturers of new cars or computers need 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 assessed by *variance* or by *standard deviation* (Section 1.1), but these measures 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 both parametric and nonparametric methods for assessing spread or variability. In other situations, how well data conform to some hypothesized population distribution function may be of interest, the appropriate 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 the information available is limited. For instance, in testing that weights of a large batch of items have a median equal to 2 mg, the only information available might be the number of items in a sample weighing 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 2.2.

*Example 2.2*

In medical studies the progress of patients is often monitored after treatment but this may be for a fixed period only due to resource constraints. Ling et al. (2017) reported the survival times in years for 14 retired footballers following a diagnosis of dementia. Assuming that follow-up is restricted to exactly 15 years the precise survival time is unknown for the longest surviving participant as he was alive at that point; this observation is referred to as being censored.

Survival time (years) were:

5 6 7 8 9 9 9 9 10 11 14 14 14 15\*

The asterisk denotes the censored observation.

Is it reasonable to suppose that these data are consistent with a median survival time of 12 years? Censored observations cause problems in many parametric tests, but in Example 2.4 a simple nonparametric test is used to show there is no strong evidence against the hypothesis that the median is 12. For that interpretation to be meaningful, and useful, it must be reasonable to assume that the data are a random sample from some population of patients with the disease.

\*\*\*

To confirm that the median might well be 12 without additional information is not particularly helpful. It would be more useful to have a 95% confidence interval for the median survival.

**Summary points:**

- A particular distribution has a shape this is uniquely determined by the value(s) of its parameter(s) (Section 2.1.1).
- In nonparametric inference, no assumption is made prior to the analysis that the samples are associated with any family of distributions (Section 2.1.1).
- A test is robust if it does not depend critically on the correctness of an assumption that samples come from a distribution in a particular family (Section 2.1.2).
- Nonparametric methods are usually more robust than analogous techniques that require assumptions about parameters (Section 2.1.2).
- Nonparametric techniques may be the only ones available if the information is limited (Section 2.1.2).

## **2.2 Permutation tests**

There is a long history of studies that produce data suitable for analysis by *permutation tests*.

The earliest known example dates from around 600 BC. The book of Daniel in the Bible (Daniel 1 vv. 3-16) 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 indicates that the health of the study ‘participants’ could be ranked (Section 1.3.4), which is needed for a permutation test to be performed. Be warned though that the commonsense conclusion may not be justified here. Allocation to the two groups was anything but random. Daniel and his three brethren may already have been “fairer and fatter” before the experiment began. In addition, the method used for assessing the condition of the participants is not indicated. Given the lack of detail available from this study, data from another historic experiment will be used to illustrate these procedures.

### *Example 2.3*

Before the nineteenth century, scurvy was a common condition suffered by sailors on long voyages of more than a few weeks. Little was known about the factors associated with scurvy and how the disease could be controlled or cured. To address this issue, a study of sailors who had contracted scurvy was conducted in May/ June 1747 onboard the British vessel HMS *Salisbury* (Lind, 1753).

A group of 12 sailors suffering from scurvy was selected. The men were divided into six pairs, each pair receiving a different form of treatment. One pair received a diet that included citrus fruits in the form of two oranges and one lemon daily. Lind, as the ship’s surgeon, made clinical notes on each participant throughout the subsequent two weeks.

In assessing the benefits or otherwise of consuming citrus fruit, an appropriate analysis would

be a comparison of the pair in the 'citrus group' with the 10 participants who received a non-citrus diet. Using the clinical notes made by James Lind it is possible to rank the participants in terms of health following two weeks of intervention from the least sick (rank 1) to the most sick (rank 12). Of those receiving citrus fruits, one was fit for duty after six days and the other was employed in nursing the others, all of whom remained ill. These sailors could therefore be assigned the ranks 1 and 2. Following the publication of this study, most of the general public accepted this as clear evidence that citrus fruits should be included in the supplies when preparing a ship for a long voyage at sea. However, could Lind's results have been purely down to chance?

The reasoning involved in a permutation test is as follows. With no association between the course of scurvy and citrus fruit consumption, it is unlikely that either sailor would have been ranked as 1 or 2. Nevertheless, the possibility of the pair receiving ranks 1 and 2, as observed by Lind, cannot be ruled out. Assuming independence of the twelve observations, the probability of a specific outcome is based on the number of ways in which two individuals can be selected from a group of twelve.

There is a choice from 12 for the first individual, then 11 individuals to select from for the other member of the pair. First impressions might suggest that  $12 \times 11$  or 132 pairings are possible. However, the order of selection is not of importance here, only the individuals that make up the pair. The number of *distinct* pairings is therefore equal to  $132/2$  or 66.

If all six interventions are equally effective/ ineffective, the pair of ranks associated with the sailors receiving citrus fruits is equally likely to be any of the 66 pairs of numbers listed in Table 2.1. Thus, there is only 1 chance in 66 that the pair showing greatest improvement (ranked 1, 2 in order of condition) are the two sailors allocated to the citrus fruit diet. This

casts doubt on the claims of cynics, of which there were plenty in the eighteenth century, that citrus fruits are not beneficial.

From a hypothesis testing perspective, the study consists of a group of two treated with a diet that includes citrus fruits and a group of ten (the remainder) given no citrus fruits, a two independent sample experiment. The most favourable evidence for the use of citrus fruits would be that those receiving them are ranked 1 and 2 in health; the least favourable that they are ranked 11 and 12.

Table 2.1 *Possible selections of two individuals from twelve labelled 1 to 12*

---

1, 2	2, 3	3, 5	4, 8	5, 12	7, 12
1, 3	2, 4	3, 6	4, 9	6, 7	8, 9
1, 4	2, 5	3, 7	4, 10	6, 8	8, 10
1, 5	2, 6	3, 8	4, 11	6, 9	8, 11
1, 6	2, 7	3, 9	4, 12	6, 10	8, 12
1, 7	2, 8	3, 10	5, 6	6, 11	9, 10
1, 8	2, 9	3, 11	5, 7	6, 12	9, 11
1, 9	2, 10	3, 12	5, 8	7, 8	9, 12
1, 10	2, 11	4, 5	5, 9	7, 9	10, 11
1, 11	2, 12	4, 6	5, 10	7, 10	10, 12
1, 12	3, 4	4, 7	5, 11	7, 11	11, 12

---

Consider a test of:

$H_0$ : citrus fruit consumption has no effect

against the two-sided alternative

$H_1$ : citrus fruit consumption has an effect (beneficial or deleterious)

Given the lack of information on the treatment of scurvy prior to Lind's study a two-tailed approach would be perfectly reasonable.

In order to perform a hypothesis test, an appropriate statistic is required with possible values that reflect the degree of divergence from what might be expected were the null hypothesis to be true. An intuitively reasonable choice is the sum of the two ranks, denoted by  $S$ . Table 2.2 shows the possible pairs along with the summed ranks.

Table 2.2 Possible selections of two individuals from twelve labelled 1 to 12 with the sum of the labels (ranks) in parentheses.

---

1, 2 (3)	2, 3 (5)	3, 5 (8)	4, 8 (12)	5, 12 (17)	7, 12 (19)
1, 3 (4)	2, 4 (6)	3, 6 (9)	4, 9 (13)	6, 7 (13)	8, 9 (17)
1, 4 (5)	2, 5 (7)	3, 7 (10)	4, 10 (14)	6, 8 (14)	8, 10 (18)
1, 5 (6)	2, 6 (8)	3, 8 (11)	4, 11 (15)	6, 9 (15)	8, 11 (19)
1, 6 (7)	2, 7 (9)	3, 9 (12)	4, 12 (16)	6, 10 (16)	8, 12 (20)
1, 7 (8)	2, 8 (10)	3, 10 (13)	5, 6 (11)	6, 11 (17)	9, 10 (19)
1, 8 (9)	2, 9 (11)	3, 11 (14)	5, 7 (12)	6, 12 (18)	9, 11 (20)
1, 9 (10)	2, 10 (12)	3, 12 (15)	5, 8 (13)	7, 8 (15)	9, 12 (21)
1, 10 (11)	2, 11 (13)	4, 5 (9)	5, 9 (14)	7, 9 (16)	10, 11 (21)
1, 11 (12)	2, 12 (14)	4, 6 (10)	5, 10 (15)	7, 10 (17)	10, 12 (22)
1, 12 (13)	3, 4 (5)	4, 7 (11)	5, 11 (16)	7, 11 (18)	11, 12 (23)

---

Counting the number of times that each rank sum occurs in Table 2.2 gives the distribution of the test statistic (Table 2.3).

Table 2.3 *Number of occurrences for each sum of ranks of two items from twelve.*

---

<i>Rank sum (S)</i>	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
<i>Occurrences</i>	1	1	2	2	3	3	4	4	5	5	6	5	5	4	4	3	3	2	2	1	1

---

The outcomes 1, 2 (rank sum = 3) and 11, 12 (rank sum = 23) are extremes with a total associated probability of  $P = 2/66 \approx 0.0303$  if  $H_0$  is true.

In classic hypothesis testing terms  $H_0$  is rejected at an exact 3.03 percent significance level if either of these extreme outcomes is observed. This small  $P$ -value provides strong evidence that the citrus fruit diet has an effect.

To find what other outcomes would be consistent with a  $P$ -value not exceeding 0.05, a region in each tail is selected (since  $H_1$  implies a two-tailed test) with a total associated probability not exceeding 0.025. From Table 2.3, selecting in the lower tail  $S = 3$  and  $S = 4$  gives an associated probability of  $\Pr(S \leq 4) = 2/66 \approx 0.0303$ . This exceeds 0.025, so the lower-tail critical region should be  $S \leq 3$  giving  $P = 1/66 \approx 0.0152$ . By symmetry, the upper-tail region is  $S \geq 23$  also with  $P \approx 0.0152$ . Thus, for a two-tailed test the largest symmetric critical region with  $P \leq 0.05$  is  $S = 3, 23$  and the exact  $P = 2/66 \approx 0.0303$ . This critical region consists only of the two extreme values, showing the paucity of information available from a small study.

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  $S = 4$  and the symmetric  $S = 22$  to the critical region gives a two-tailed  $P = 4/66 \approx 0.0606$ . This is closer to 0.05 than the size (0.0303) of the region chosen above. It is best to quote the exact  $P$ -value where possible. The practical argument (though there are further theoretical ones) for quoting nominal sizes such as 0.05 is that many tables give only these. A few, e.g., Gibbons and Chakraborti (2020) give relevant exact  $P$ -values for many sample size combinations and different values of  $S$ . Computer programs giving exact  $P$ -values may remove the need to use such tables.

In the preliminary testing of drugs for treating a rare disease the population may be in a strict sense the only cases available. However, if these patients are typical of all who might have the disease, it is not unreasonable to assume that findings from this 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 an 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 as discussed in Section 1.7.1. For a small study, a marginal  $P$ -value associated with what looks to be an intuitively encouraging result may indicate that a larger experiment might give more convincing positive findings.

\*\*\*

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 2.3, the permutations giving all possible assignments. These tests have an intuitive appeal and comply with well-established theoretical criteria for sound inference. Hettmansperger and McKean (2011) summarize this theoretical basis for many different procedures.

**Summary points:**

- A permutation test considers the number of ways in which subgroups can be selected from the overall sample.
- Permutation tests can be applied to ranked data to investigate whether higher ranks are concentrated in a particular subgroup.
- When an experiment involves what is essentially the whole population, and the only data are ranks, a permutation test is the best test available.

## **2.3 Binomial tests**

### *2.3.1 Number of ways of selecting a sample from a larger group*

In Example 2.3 it was demonstrated that the number of ways in which a pair can be selected from 12 individuals is equal to  $12 \times 11/2 = 66$ . Another way of representing this is as:

$$\frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) (10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)}$$

This can be written more concisely as:

$$\frac{12!}{2! \times 10!}$$

The term 12! is often referred to as '12 factorial' other numbers written in this form being described in a similar way. Highlighting the fact that 2 objects are being selected from a group of size 12, the above quantity can be written as:

$$\frac{12!}{2! \times (12-2)!} \text{ or } \binom{12}{2}$$

In general, the number of ways in which  $r$  observations can be selected from a larger group of  $n$  observations is:

$$\frac{n!}{r! \times (n-r)!} \text{ or } \binom{n}{r}$$

### 2.3.2 Independent events

An assumption required for most of the methods described in this text is that the observations are independent. If this is the case, the multiplication rule of probability can be applied. This states that for independent events (e.g. head and tail outcomes when a coin is tossed repeatedly) the probability of obtaining the series observed is given by multiplying together the probabilities of the individual outcomes. In particular, the probability of obtaining  $n$

successes in a series of length  $n$  is given by the probability of a success ( $p$ ) raised to the  $n$ th power, i.e.  $p^n$ .

In a somewhat lighthearted style, the satirical writer 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. By considering the outcomes of throws with two-sided dice, he noted that if male and female births were equally likely the observed finding would have an extremely low probability of  $0.5^{82}$ . This finding, he reasoned, demonstrates strong evidence against the commonly believed assumption. Note that he did not distinguish between births and christenings; his conclusion is only true if the probability of an infant being christened was equal for males and females.

### *2.3.3 An application of the Binomial test*

One observation was censored in the data for the survival of retired footballers in Example 2.2. Despite its exact value being unknown, it could be shown that it was reasonable, given that data, to retain a hypothesis that the population median was 12. An appropriate test to justify that conclusion is now considered.

#### *Example 2.4*

Survival times (years) in order of retired footballer enrollment were:

9 9 8 5 9 6 14 10 14 9 11 7 15\* 14

The asterisk denotes the censored observation.

The appropriate null hypothesis here is that the median,  $\theta$ , of survival times for the population from which the sample was obtained is 12 against the alternative of some other value:

$$H_0: \theta = 12 \text{ against } H_1: \theta \neq 12 \quad (2.1)$$

A simple test needs only a count of the number of sample values exceeding 12. In the data above, values over 12 are replaced by a “plus” other values being replaced by a “minus” to give the sequence consisting of 4 “plus” and 10 “minus” signs:

- - - - - + - + - - - + +

Under the null hypothesis, since  $H_0$  is concerned with the median the probability of the next observation being a plus is 0.5. As only two outcomes are possible, the probability of the next observation being a minus is also 0.5. Assuming that the 14 retired footballers form a random sample, and order of enrollment is taken into account, the probability of observing the sequence above is obtained by multiplying the individual probabilities together giving  $0.5^{14}$ .

Of more relevance for testing  $H_0$  is the probability of obtaining a combination of 4 plus and 10 minus signs irrespective of ordering. This gives the overall chance of observing 4 values in excess of the median in a sample of size 14. The probability is calculated by multiplying the probability of obtaining a specific sequence of plus and minus signs ( $0.5^{14}$ ) by the number of ways in which a sample of 4 observations can be selected from a group of 14:

$$\frac{1}{4! \times 10!} \text{ or } \binom{14}{4}$$

Writing the number of pluses as  $X$ , the probability of 4 pluses in a sample of size 14 can be written as:

$$\Pr(X=4) = \binom{14}{4} \times 0.5^{14}$$

It follows that the probability of obtaining  $r$  pluses from 14 observations is:

$$\Pr(X=r) = \binom{14}{r} \times 0.5^{14}$$

Probabilities for the possible values of  $r$  are given below (Table 2.4).

Table 2.4 *Probabilities of obtaining  $r$  pluses from 14 observations under  $H_0$*

---

|                      |        |        |        |        |        |        |        |        |        |
|----------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| <i>Number of + :</i> | 0      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      |
| <i>Probability :</i> | 0.0001 | 0.0009 | 0.0056 | 0.0222 | 0.0611 | 0.1222 | 0.1833 | 0.2095 | 0.1833 |
| <i>Number of + :</i> | 9      | 10     | 11     | 12     | 13     | 14     |        |        |        |
| <i>Probability :</i> | 0.1222 | 0.0611 | 0.0222 | 0.0056 | 0.0009 | 0.0001 |        |        |        |

---

Note that the number of pluses takes a binomial distribution  $B(n,p)$ :

$$\Pr(X=r) = \binom{n}{r} p^r (1-p)^{n-r}$$

The sample size  $n = 14$ , the probability of a positive outcome (plus sign) of  $p = 0.5$ , the probability of a negative outcome (minus sign) of  $(1 - p) = 0.5$ , the number of runs being  $r$ . Hence, this distribution is denoted by  $B(14,0.5)$

If the null hypothesis is true, the expected number of plus signs is  $14 \times 0.5$  or 7. From the definition of a  $P$ -value, the relevant tail probabilities are those for  $r \leq 4$  pluses and  $r \geq 10$  pluses. These add up to  $(0.0001 + 0.0009 + 0.0056 + 0.0222 + 0.0611) \times 2 = 0.1798$ .

Hence there is no strong evidence against  $H_0$ . This implies that departures from the expected number of plus signs, 7, as large, or larger, than that observed will occur in slightly more than one-sixth of all samples when  $H_0$  is true. This simple test, called the *sign test*, is discussed more fully in Section 3.2. The test is *distribution-free* because no assumption about the form of the continuous distribution of the underlying observations is made. Formulation and testing of hypotheses are only concerned with possible values of the population median.

For most parametric tests (e.g. the  $t$ -test), all values of  $P$  between 0 and 1 are possible.

However, for many nonparametric methods  $P$ -values are limited to a finite number of discrete values. These do not always correspond to conventional thresholds for statistical significance.

This is the case for the sign test. In this example, for a two-tailed test the four smallest are  $P = 2 \times 0.0001 = 0.0002$  corresponding to 0 or 14 plus signs;  $P = 2 \times (0.0001 + 0.0009) = 0.002$  corresponding to no more than 1 or at least 13 plus signs;  $P = 2 \times (0.0001 + 0.0009 + 0.0056) = 0.0132$  corresponding to no more than 2 or at least 12 plus signs; then  $P = 2 \times (0.0001 + 0.0009 + 0.0056 + 0.0222) = 0.0576$  corresponding to no more than 3 or at least 11 plus signs. Next comes the observed  $P = 0.1798$  (no more than 4 or at least 10 plus signs). In all cases probabilities have been rounded to four decimal places.

The statistic used in this example - the number of plus signs - has a discrete distribution. This means that, as in Example 2.3, there is no direct way of obtaining a critical region of exact size 0.05 for a two-tailed test; the only choice is between regions of size 0.0132 or 0.0576.

\*\*\*

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 pre-chosen nominal level, e.g., at 5 percent. Gibbons and Chakraborti (2020) describe how a randomized decision rule works (pp. 24–25). They comment that such devices may seem artificial and are “probably seldom employed by experimenters”. Hence, it is better, when known, to use exact levels, rather than worry about nominal arbitrary levels. When there are discontinuities, however, there is 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. It is not used 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 2.5*

In Example 2.4, when using a sign test for a median with a sample of 14, in a two-tailed test

at the 1.32 percent level,  $H_0$  is retained if between 2 and 12 plus signs are observed.

Consider again the data in that example ordered by size, i.e.,

5 6 7 8 9 9 9 9 10 11 14 14 14 15\*

where the asterisk represents a censored observation. There are between 2 and 12 plus signs if the median specified in  $H_0$  has any value greater than 6 but less than 14. This implies that the interval (6, 14) is a  $100(1 - 0.0132) = 98.68$  percent confidence interval for  $\theta$ , the population median survival time. Since any  $H_0$  that specified a value for the median greater than 6 but less than 14 would be retained, there is considerable doubt about the population median value. It is stating the obvious to say the estimate lacks precision.

\*\*\*

#### 2.3.4 A Binomial distribution with unequal success and failure probabilities

The next illustration involves a more general Binomial distribution.

##### *Example 2.6*

Inchley et al. (2020) found that for 15-year-old boys in England, 75 percent brush their teeth frequently (more than once daily). Hence, the number who brush their teeth frequently,  $S$ , in a sample of 10 independent boys aged 15 years has a binomial  $B(10, 0.75)$  distribution. Here the probabilities for the various values,  $r$ , of the statistic  $S$ , where  $r$  takes integer values between 0 and 10, are given by:

$$\Pr(X=r) = \binom{10}{r} 0.75^r 0.25^{n-r}$$

The relevant probabilities are (Table 2.5):

Table 2.5 *Probabilities of obtaining  $r$  individuals who brush their teeth more than once daily out of 10 boys aged 15 under  $H_0$*

---

|                      |        |        |        |        |        |        |        |        |        |
|----------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| <i>Number</i> :      | 0      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      |
| <i>Probability</i> : | 0.0001 | 0.0000 | 0.0004 | 0.0031 | 0.0162 | 0.0584 | 0.1460 | 0.2503 | 0.2816 |
| <i>Number</i> :      | 9      | 10     |        |        |        |        |        |        |        |
| <i>Probability</i> : | 0.1877 | 0.0563 |        |        |        |        |        |        |        |

---

If based on data from previous years it was thought that the proportion should be higher than 0.75 (e.g. it has never been less than 0.9 before), the relevant test would be:

$$H_0: p = 0.75 \text{ against } H_1: p > 0.75$$

for which the smallest  $P$ -value for testing is in the upper tail and is associated with  $r = 10$ , i.e.,  $P = 0.0563$ . Applying the convention  $P \leq 0.05$  as sufficiently strong evidence to discredit  $H_0$ , such values are never obtained.

This is not a problem for the one-tailed 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 suggested in the literature that the appropriate level for a two-tailed test is twice that for a one-tailed test, for if the outcome is  $S = 4$  the two-tailed 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-tailed test is appropriate, departures from the null hypothesis can only be detected in one direction. There may well be a departure in the other direction, but if so detection is highly unlikely at the conventional level  $P \leq 0.05$ . Even if a departure were to be detected, it would be for the wrong reason. This is not surprising when, as shown above, the appropriate one-tailed test must fail to detect it, for generally a one-tailed test at a given significance level is more powerful for detecting departures in the appropriate direction than is a two-tailed test at the same level.

An implication is that in this example a larger sample is needed to detect departures of the form  $H_1: p > 0.75$ . Again, the unconvincing  $P$ -value associated with the possible critical region for the one-tailed test only indicates that the sample is too small. The stipulation that the study participants be independent is important. If the boys included members from the same household (e.g. twins) it is quite likely that their teeth brushing habits would be similar if not identical. Situations of this kind are considered in more detail in Agresti (2013).

\*\*\*

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, one should, for a two-tailed test, add the probabilities associated with an opposite tail situated

equidistant from the mean value of the test statistic to that associated with the observed statistic value. In the symmetric case, as already pointed out, this is equivalent to doubling the probability, but it seems inappropriate with an asymmetric distribution. In Example 2.6 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-tailed test based on equidistance from the mean would have the same exact significance level as the one-tailed test.

**Summary points:**

- The number of ways in which  $r$  observations can be selected from a larger group of size  $n$  may be calculated using a formula based on factorial terms (Section 2.3.1).
- If events are independent, the probability of obtaining the series of outcomes observed is given by multiplying together the probabilities of the individual outcomes (Section 2.3.2).
- For many nonparametric methods  $P$ -values are limited to a finite number of discrete values (Section 2.3.3).
- If a binary outcome has unequal success and failure probabilities, the distribution of the number of successes in a sample is asymmetric. This can cause problems in defining appropriate critical regions for two-tailed tests (Section 2.3.4).

## 2.4 Order statistics and ranks

Many nonparametric procedures are based on the ordering, or ranking, of detailed observations. In Example 2.4, ranks were not used, but ordering was inherent in the procedure. Order was considered 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. There may be interest 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, 5m, 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 2.8. Greatest and least values in samples are just two examples of *order statistics*. The sample median is also an order statistic.

Consider a sample of  $n$  observations  $x_1, x_2, \dots, x_n$  from a continuous distribution. Continuity implies that there should be no ties and thus observations may be uniquely ordered from

smallest to largest. 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

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}.$$

The  $x_{(i)}$ ,  $i = 1, 2, \dots, n$  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 light bulb.

The measures introduced in Sections 1.3.3 and 1.3.4 can be written algebraically using 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  $x_{(n)} - x_{(1)}$ . A more satisfactory measure is the *interquartile range*. One reason for preferring the latter is that the extreme order statistics are often strongly influenced by suspect observations associated with the terms *outliers* (Section 1.3.3) and *dirty data*.

**Summary points:**

- The  $i$ th order statistic in a sample is the observation having a rank of  $i$ .
- The largest and smallest values in a sample are examples of order statistics that can be modeled using extreme value distributions.

## 2.5 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* (EDA). Some basic tools of exploratory data analysis 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 a group of observations are the sample mean, median, maximum value, minimum value, standard deviation and quartiles.

When questions of robustness arise due to the presence of extreme values, one approach is to remove or adjust the outliers. With the *trimmed mean*, the most extreme observations (typically amounting to 10 percent) are removed and inferences are based on the remaining observations. For the *Winsorized mean*, small outliers are replaced by the value of the smallest plausible observation and large outliers are replaced by the value of the largest plausible observation.

Statistics such as the mean, median and standard deviation 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. Typically, the five descriptive statistics presented in the order *minimum, 1st quartile, median, 3rd quartile, maximum*, (or *five-number summary*) are used for this purpose.

Such basic characteristics may be explored more fully by graphical techniques. For instance, the descriptive statistics from a five-number summary can be used to construct a *boxplot* or a *box and whisker plot*. These plots can be used to assess how well samples reflect population features.

### *Example 2.7*

Appendix 1 gives four small data sets indicating how they were collected. Table 2.6 gives a set of descriptive statistics useful for summarizing and comparing these data for each of the four sets. The first row gives 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 2.6 *Descriptive, or summary, statistics for Badenscallie data given in Appendix 1.*

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

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

The medians are all appreciably higher than the means, suggesting that the distributions of ages are asymmetric and skewed to the left (Section 1.3.4). This follows because samples are expected to reflect broadly the population characteristics, and for symmetric population

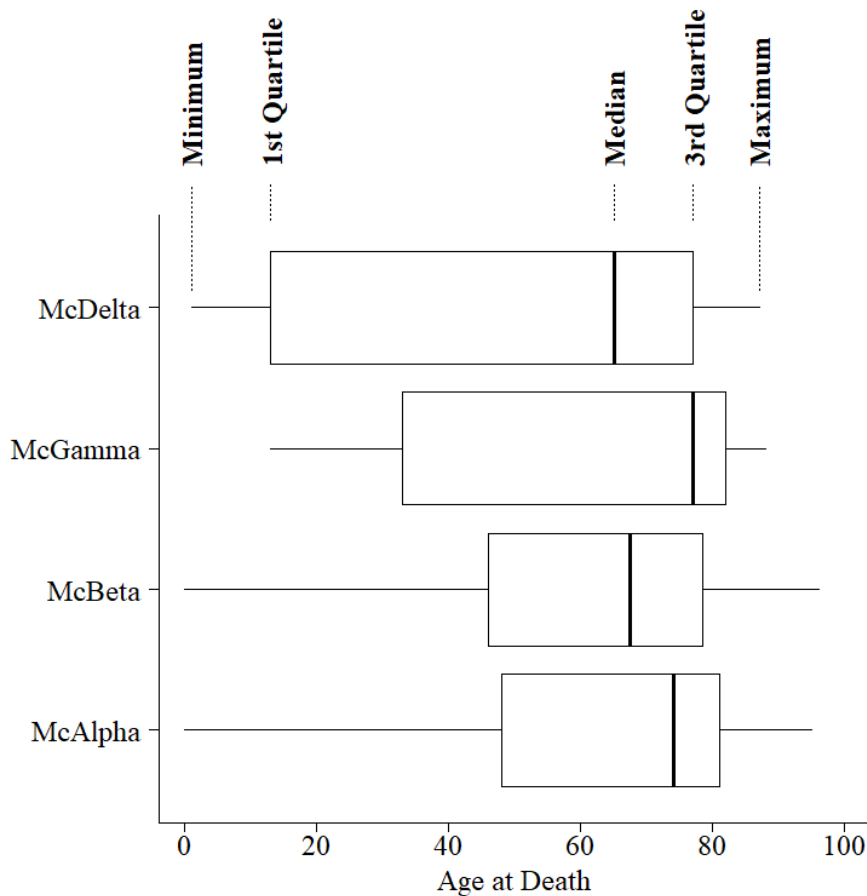
distributions the mean and median coincide.

The abbreviation *SD* is used in the table for the standard deviation, a commonly encountered measure of spread that in the case of a sample from a Normal population is an appropriate estimator of the parameter  $\sigma$ . Once again, the clan *McDelta* is the odd one out.

*SE mean* is an abbreviation for *standard error of the mean*. If the sample standard deviation is denoted 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 nonagenarian survivor for two of the clans.

The *quartiles* divide each ordered sample into four groups of equal size. If the median is considered 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  $x_{(1)}, x_{(2)}, \dots, x_{((n-1)/2)}$  if  $n$  is odd and is the median of  $x_{(1)}, x_{(2)}, \dots, x_{(n/2)}$  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 quartile is much smaller for *McDelta*.

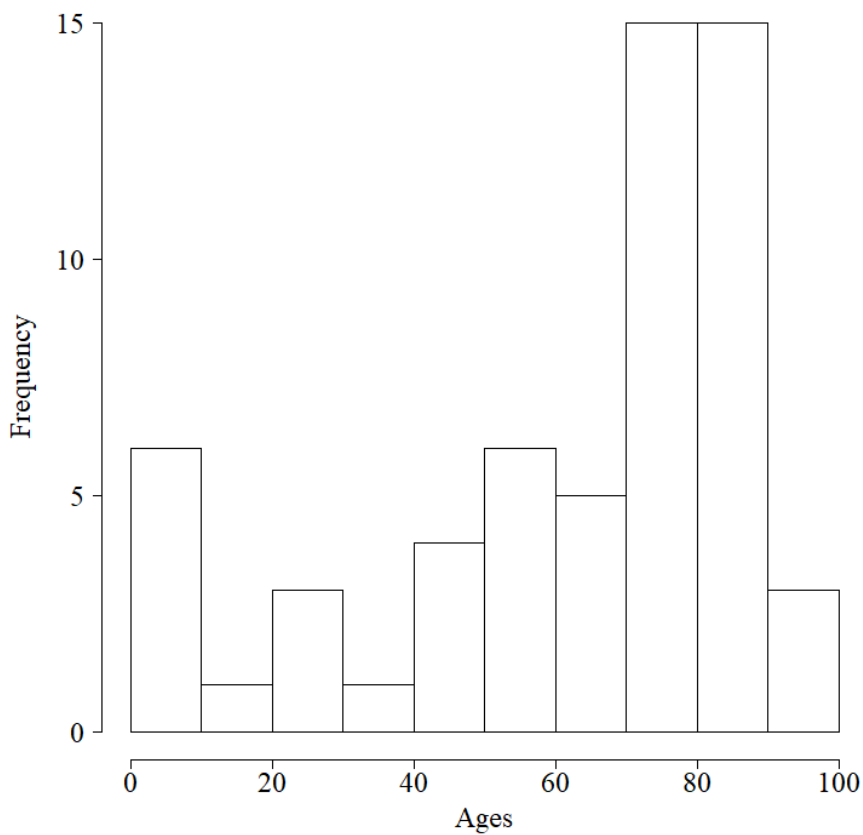


**Figure 2.1.** Boxplots for Badenscallie data given in Appendix 1.

The first quartiles raise the issue of whether these differences 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 the 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 occasionally 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.

Figure 2.1 gives boxplots for each clan for the Badenscallie data based on five-number summaries presented in Table 2.6.



**Figure 2.2.** Histogram for clan McAlpha data given in Appendix 1.

\*\*\*

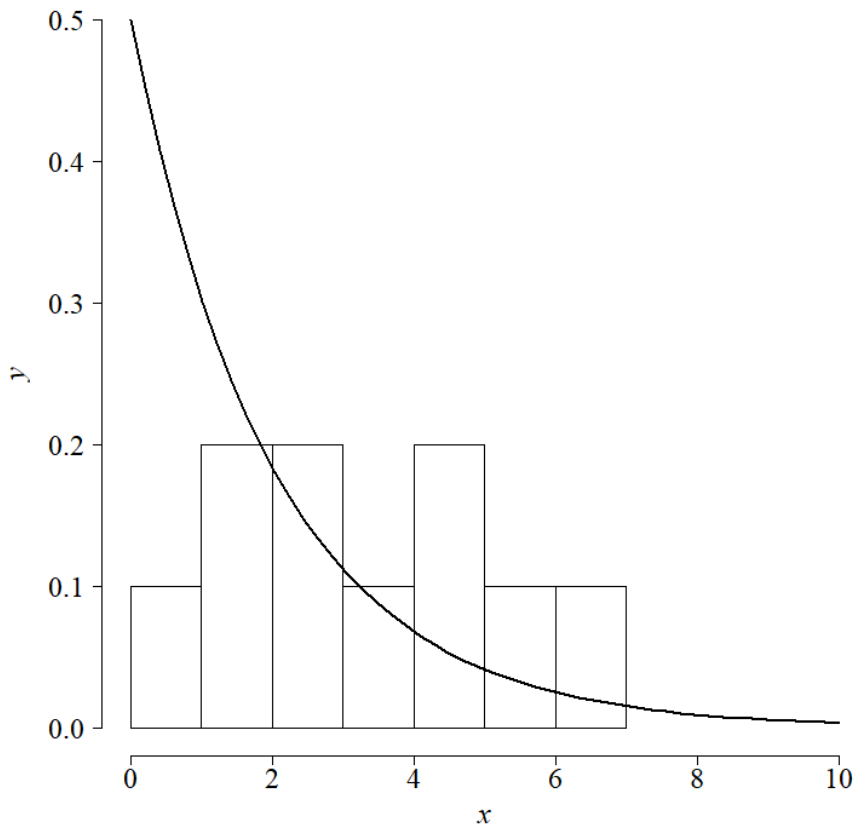
The labels attached to the boxplot for the *McDelta* clan apply to any boxplot and indicate that the box section extends from the first to the third quartile. The vertical line dividing this box

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, it is seen 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 2.2 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 around 80 years.

As the size of a random sample increases so it mirrors 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.



**Figure 2.3.** Histogram for a sample of 50 from an exponential distribution with mean 2. The fitted curve is that of the probability density function.

*Example 2.8*

A computer-generated sample of 50 observations from an exponential distribution with mean 2 using Minitab gave the histogram in Figure 2.3. 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.

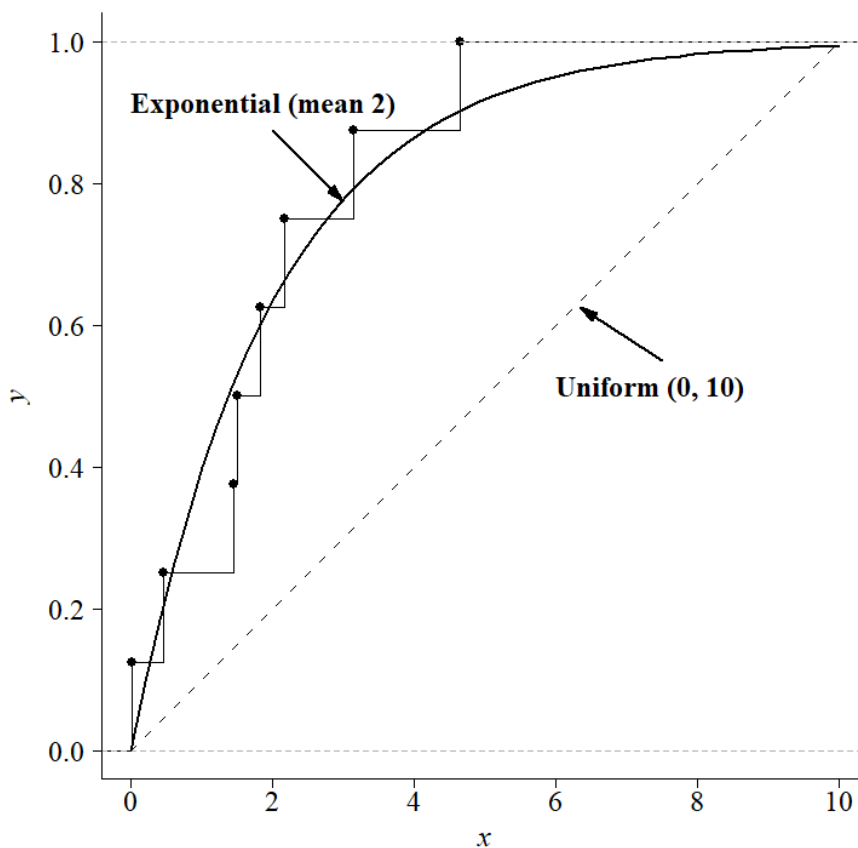
The curve superimposed on the histogram is that of the *frequency function* or *probability density function* of the exponential distribution with mean 2, which has the form:

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

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 2.4.** 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 2.9*

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

0.25 0.53 0.91 0.94 1.56 1.73 4.71 5.50

where these have been arranged in ascending order. Figure 2.4 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

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

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 the sample was not taken from that distribution.

\*\*\*

More sophisticated exploratory data analysis 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. Thas (2010) gives a description of how these are used and interpreted.

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

**Summary points:**

- Before choosing the statistical methods to be applied, samples should be investigated using exploratory data analysis.
- Descriptive statistics in common use include the sample mean, median, maximum value, minimum value, standard deviation and quartiles.
- The boxplot provides a graphical representation of the data based on the five-number summary (minimum, 1st quartile, median, 3rd quartile, maximum).
- Histograms can be used to investigate characteristics of continuous data grouped into intervals.
- Superimposing a sample histogram on a population probability density function gives an indication of how well the data fit the proposed model.

## **2.6 Efficiency of nonparametric procedures**

As discussed in Section 1.5, 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  $\alpha$ ), (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. Tests are sometimes required that 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.

A statistic is regarded as efficient (put simply) if the sample size required to obtain low probabilities for the Type I and Type II errors is 'small'. *Efficiency* is usually described in terms of one statistic relative to another. If  $\alpha$  is the probability of a Type I error, and  $\beta$  is the probability of a Type II error (the power is  $1 - \beta$ ), then the efficiency of a test T2 relative to a test T1 is the ratio  $n_1/n_2$  of the sample sizes needed to obtain the same power for the two tests with these values of  $\alpha, \beta$ . In practice,  $\alpha$  is usually fixed at some  $P$ -value appropriate to the problem at hand. Then  $\beta$  depends on the particular alternative as well as the sample sizes. Fresh calculations of relative efficiency are required for each particular value of the parameter or parameters of interest in  $H_1$  and for each choice of  $\alpha, \beta$ .

Pitman (1949) introduced the concept of *asymptotic relative efficiency* for comparing two tests (p.54). He considered sequences of tests T1, T2 in which  $\alpha$  is fixed. The alternative in  $H_1$  is then allowed to vary in such a way that  $\beta$  remains constant as the sample size  $n_1$  increases. For each  $n_1$  the value  $n_2$  is determined such that T2 has the same  $\beta$  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  $n_1 \rightarrow \infty$ . More importantly, this limit, which he called the asymptotic relative efficiency (ARE), was the same for all choices of  $\alpha, \beta$ . A detailed discussion of asymptotic relative efficiency is given by Mood (1954) and Gibbons and Chakraborti (2020, Chapter 13).

Bahadur (1967) proposed an alternative definition that is less widely used, so for clarity and

brevity Pitman's concept is here referred to 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 (2006) shows that for samples of size 10 and a range of values of the median  $\theta$  relative to the value  $\theta_0$  specified in  $H_0$  with  $\alpha$  fixed, the relative efficiency exceeds 0.7. 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.

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 a 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.

**Summary points:**

- A test is regarded as efficient if the sample size needed to obtain low probabilities for the Type I and Type II errors is small, relative to other available tests.
- For tests T2 and T1 having the same probability of a Type I error, the efficiency of T2 relative to T1 is the ratio  $n_1/n_2$  of the sample sizes needed to obtain the same power for the two tests.
- As the sample size for T1 is increased, the ratio for T2 relative to T1 approaches the asymptotic relative efficiency.

**2.7 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 consists of 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 macros to carry out such procedures. The final category consists 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 self-designed options, 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 StatXact is perhaps the most widely used. It 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. 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.

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. Facilities for creating Monte Carlo approximations to exact  $P$ -values, or for bootstrap estimation, are often also available in these standard packages.

In the third category the open source R is dominant. A number of exact and asymptotic nonparametric tests are available via standard packages that are maintained by the R Core Team while others are available through packages created by R users. Packages such as R and SAS have a versatility that makes combining of approaches such as exploratory data analysis 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-tailed  $P$ -value may be used to obtain a two-tailed test value, but as indicated in Example 2.6, 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, although powerful computing facilities are needed for the application of some of them. Solutions to examples, or illustrations in this book largely use the ANSM5 package in R, written to accompany the book. 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. In creating the ANSM5 package we do not claim to have produced software which is necessarily the most computationally efficient nor the most comprehensive. At the time of writing there are other R packages which also undertake nonparametric statistical analyses, and which the reader may wish to investigate. We do not list them here because over time various packages may cease to be maintained and other packages may emerge. However, the ANSM5 package has been designed to cover all the techniques covered in this book and be dependent only on other R packages which are maintained by the R Core Team. It can be downloaded from CRAN and associated mirror sites, or from the “neilhspencer/ANSM5” repository on GitHub. The package also contains data used in the book. Readers who replicate the analyses in the book which use the ANSM5 package should obtain the same results as presented in the examples and exercises. A possible exception to this is where Monte Carlo methods are used. These are based on repeatedly taking random samples of data and, of course, the random samples chosen by the readers’ computers may be different from those chosen by the authors’ computers. As a result, the Monte Carlo results may differ to a small degree.

Developments in statistical computer software are rapid and mention should also be made of the programming language Python which is hugely flexible and has packages that can perform some nonparametric tests.

## **2.8 Further reading**

Hollander, Wolfe and Chicken (2014), Conover (1999), Gibbons and Chakraborti (2020), 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 (2000) is a general book on applied nonparametric methods.

Hettmansperger and McKean (2011) give a moderately advanced mathematical treatment of the theory behind nonparametric methods. Randles and Wolfe (1991) 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 (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). Siegel and Castellan (1988) and Corder and Foreman (2014) are readable introductory texts.

Noether (1991) uses a nonparametric approach to introduce 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 are described in Chapter 14. He gives a lucid introduction to those topics addressed. Li and Racine (2006) is an advanced text that gives detailed coverage of modern nonparametric methods in the context of econometrics. Agresti (2010, 2013, 2019) and Everitt (1992) give detailed accounts of various parametric and nonparametric models 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 and Onghena (2007) cover randomization and permutation tests as do Berry, Johnston and Mielke (2014). Hájek, Šidák and Sen (1999) give the theory behind rank tests. Books dealing with the bootstrap include Efron and Tibshirani (1993), Davison and Hinkley (1997), Efron and Hastie (2016), Chernick and LaBudde (2011) and Dickhaus (2018).

## 2.9 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 a new drug under test 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. Explain why a one-tailed test might be justifiable here and formulate appropriate hypotheses for the test. If 9 patients are involved in the study and the post-treatment ranking of the patients receiving the new drug is 1, 2, 4, 6 (a low rank representing a better outcome) assess the strength of the evidence against the relevant null hypothesis. Would a two-tailed test give similar findings?

2.3 An education authority is responsible for 114 primary schools. A random sample of 12 schools is selected to test the hypothesis that the median number of children,  $\theta$ , in all 114 schools is 225. In the sample of 12, it is found that 3 schools have less than 225 children. Does this justify retention of the hypothesis that  $\theta = 225$ ? What would be an appropriate alternative hypothesis? What is the largest critical region for a test with  $P \leq 0.05$  and what is the corresponding exact  $P$ -value?

2.4 The numbers of children in the sample of 12 schools in Exercise 2.3 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  $\theta$ .

2.5 In Section 1.6.1 a confidence interval was associated with a two-tailed 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-tailed test. Use this argument to obtain a confidence interval at level not less than 95 percent based on the sign test criteria for the 12 schools given in Exercise 2.4 relevant to a test of  $H_0: \theta = \theta_0$  against a one-sided alternative  $H_1: \theta > \theta_0$ .

2.6 From 6 consenting patients requiring a medical scan, 3 are chosen at random to undergo positron emission tomography (PET), the others receiving magnetic resonance imaging (MRI). 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.

2.7 In Example 2.6 it was remarked that a situation could arise where the rejection of the null hypothesis might be for the wrong reason. Explain how this is possible in that example.

2.8 A sample of 12 observations 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.)

2.9 A sample of 24 observations 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 exploratory data analysis techniques to obtain an indication as to which of these distributions is the more likely source of the sample.

### References

- Agresti, A. (2010) *Analysis of Ordinal Categorical Data*. 2nd edn. Hoboken: John Wiley & Sons.
- Agresti, A. (2013) *Categorical Data Analysis*. 3rd edn. Hoboken: John Wiley & Sons.
- Agresti, A. (2019) *An Introduction to Categorical Data Analysis*. 3rd edn. Hoboken: John Wiley & Sons.
- Arbuthnot, J. (1710) An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Phil. Trans. Roy. Soc.*, **27**, 186–190.
- Bahadur, R.R. (1967) Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, **38**, 303–324.
- Berry, K.J., Johnston, J.E. and Mielke, P.W. (2014) *A Chronicle of Permutation Statistical Methods: 1920-2000, and Beyond*. New York: Springer.
- Bury, K. (1999) *Statistical Distributions in Engineering*. Cambridge: Cambridge University Press.
- Chernick, M.R. and LaBudde, R.A. (2011) *An Introduction to Bootstrap Methods with Applications to R*. Hoboken: John Wiley and Sons.
- Conover, W.J. (1999) *Practical Nonparametric Statistics*. 3rd edn. New York: John Wiley & Sons.
- Corder, G.W. and Foreman, D.I. (2014) *Nonparametric Statistics: a Step-by-Step Approach*. 2nd edn. Hoboken: John Wiley & Sons.

- Daniel, W.W. (2000) *Applied Nonparametric Statistics*. 2nd edn. Belmont: Duxbury.
- Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- Desu, M.M. and Raghavarao, D. (2004) *Nonparametric Statistical Methods for Complete and Censored Data*. Boca Raton: Chapman & Hall/ CRC.
- Dickhaus, T. (2018) *Theory of Nonparametric Tests*. Cham: Springer.
- Edgington, E.S. and Onghena, P. (2007) *Randomization Tests*. 4th edn. Boca Raton: Chapman & Hall/ CRC.
- Efron, B. and Hastie, T. (2016) *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. New York: Cambridge University Press.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Everitt, B.S. (1992) *The Analysis of Contingency Tables*. 2nd edn. London: Chapman & Hall.
- Gibbons J.D. and Chakraborti, S. (2020) *Nonparametric Statistical Inference*. 6th edn. Boca Raton: Chapman & Hall/ CRC.
- Good, P. (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer.
- Hájek, J., Sidák, Z. and Sen, P.K. (1999) *Theory of Rank Tests*. 2nd edn. San Diego: Academic Press.
- Hettmansperger, T.P. and McKean, J.W. (2011) *Robust Nonparametric Statistical Methods*. 2nd edn. Boca Raton: Chapman & Hall/ CRC.
- Higgins J.J. (2004) *Introduction to Modern Nonparametric Statistics*. Belmont: Duxbury.
- Hollander, M., Wolfe, D.A. and Chicken, E. (2014) *Nonparametric Statistical Methods*. 3rd edn. Hoboken: John Wiley & Sons.

- Inchley, J., Currie, D., Budisavljevic S., Torsheim, T., Jåstad A., Cosma, A, Kelly, C., Már Arnasson Á., Samdal, O. (eds) (2020) *Spotlight on Adolescent Health and Well-being. Findings from the 2017/2018 Health Behaviour in School-aged Children (HBSC) Study. International Report. Volume 2. Key Data.* Copenhagen: WHO Health Organization.
- Kendall, M.G. and Gibbons, J.D. (1990) *Rank Correlation Methods.* 5th edn. London: Edward Arnold.
- Leach, C. (1979) *Introduction to Statistics. A Nonparametric Approach for the Social Sciences.* Chichester: John Wiley & Sons.
- Lehmann, E.L. (2006) *Nonparametrics: Statistical Methods Based on Ranks.* Revised edn. Berlin: Springer.
- Li, Q. and Racine, J.S. (2006) *Nonparametric Econometrics: Theory and Practice.* Princeton: Princeton University Press.
- Lind, J. (1753) *A Treatise of the Scurvy. In Three Parts. Containing an Inquiry into the Nature, Causes and Cure, of that Disease. Together with a Critical and Chronological View of what has been Published on the Subject.* Edinburgh: Printed by Sands, Murray and Cochran for A. Kincaid and A. Donaldson.
- Ling, H., Morris, H.R., Neal, J.W., Lees, A.J., Hardy, J., Holton, J.L., Revesz, T. and Williams, D.D.R. (2017) Mixed pathologies including chronic traumatic encephalopathy account for dementia in retired association football (soccer) players. *Acta Neuropathologica*, **133**, 337-352.
- Manly, B.F.J. (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology.* 3rd edn. Boca Raton: Chapman & Hall/CRC.

- Maritz, J.S. (1995) *Distribution-free Statistical Methods*. 2nd edn. London: Chapman & Hall.
- Mood, A.M. (1954) On the asymptotic efficiency of certain nonparametric two-sample tests. *Ann. Math. Statist.*, **25**, 514–522.
- Noether, G.E. (1991) *Introduction to Statistics: The Nonparametric Way*. New York: Springer–Verlag.
- Pitman, E.J.G. (1949) *Notes on Non-parametric Statistical Inference. Institute of Statistics Mimeo Series 27*. North Carolina State University. Dept of Statistics: Raleigh: NC.  
<https://www.lib.ncsu.edu/resolver/1840.4/2430> (accessed 24 March 2024).
- Randles, R.H. and Wolfe, D.A. (1991) *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley & Sons.
- Rayner, J.C.W. and Best, D.J. (2001) *A Contingency Table Approach to Nonparametric Testing*. Boca Raton: Chapman & Hall/CRC.
- Siegel, S. and Castellan, N.J. (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd edn. New York: McGraw-Hill.
- Thas, O. (2010) *Comparing Distributions*. New York: Springer.
- Wasserman, L. (2006) *All of Nonparametric Statistics*. New York: Springer.
- Yates, F. (1984) Tests of significance for  $2 \times 2$  contingency tables. *J. Roy. Statist. Soc. A*, **147**, 426–463.