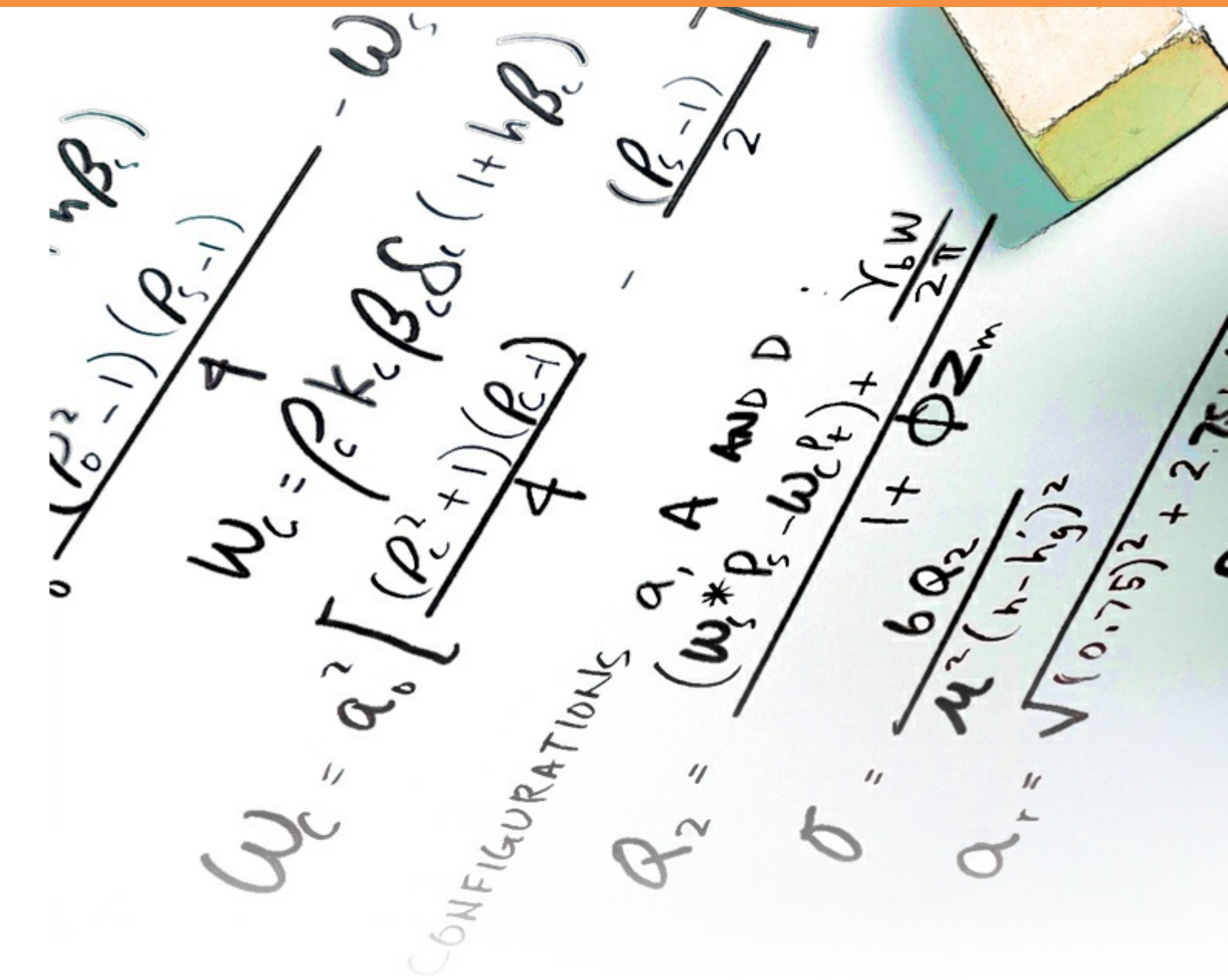

Introductory Nonparametrics

J.C.W. Rayner

bookboon.com
The eBook company

J.C.W. RAYNER

INTRODUCTORY NONPARAMETRICS

Introductory Nonparametrics
1st edition

ISBN 978-87-403-1475-5
Peer review: Dr Paul Rippon, Conjoint Lecturer in Statistics School of Mathematical and Physical Sciences, Faculty of Science and Information Technology

CONTENTS

ABOUT THE AUTHOR

John Rayner is currently Honorary Professorial Fellow at the Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW, Australia and Conjoint Professor of Statistics at the University of Newcastle in NSW, Australia. He served as Professor of Statistics and Head of Discipline at the University of Newcastle from 2006 to 2011 before retiring from full-time employment. Previously John worked full-time at the University of Otago in Dunedin, New Zealand from 1973 to 1992 and the University of Wollongong in NSW, Australia from 1992 to 2006.

John's prime research interests are goodness of fit (assessing statistical models) and nonparametric statistics. He is the lead author of *Smooth Tests of Goodness of Fit: Using R* and *A Contingency Table Approach to Nonparametric Testing*. He has written over 150 research articles and books, many with his long-time friend and colleague John Best.

Now in his 71st year, John exercises moderately every day with the aim of participating in a weekly parkrun. These are timed 5km runs at venues all over the world. Last year he ran under 25 mins several times and, once, under 24 mins!

PREFACE

Nonparametric procedures usually feature in undergraduate statistics courses

1) in parallel with the corresponding parametric procedures,
2) as a module in a larger course, and
3) as a sequence of modules, perhaps in a half course or short course.

In the first option it is common, for example, to treat the one-way analysis of variance in tandem with the Kruskal-Wallis test and the two-way analysis of variance in tandem with the Friedman test. However, it seems best that an introduction to nonparametric ideas should precede such a pairing.

The modular approach adopted here better serves the second and third options. The first module gives a gentle introduction to nonparametric ideas, assuming readers have already met the binomial and normal distributions and some parametric methods. The second module introduces nonparametric tests for the simpler and most commonly used experimental designs. The third focuses on permutation tests, a fundamental nonparametric tool. In an undergraduate statistics sequence it would be reasonable to include these modules in successive semesters in comprehensive courses or as a half course or short course towards the end of a statistics major.

Nonparametric tests should be applied when the parametric assumptions cannot reasonably be assumed. However there are users who believe that many parametric tests are so robust to their parametric assumptions that they can be applied almost always. Nevertheless there remain many parametric procedures, such as the Bartlett test of equality of variance, that are known to be highly non-robust. The contrary point of view is that it is always wise to use analyses that assume as little as possible. If normality is an assumption in a parametric test, how small a p-value should a test of normality be before the conclusions of the parametric test be regarded as suspect? In general there is no clear guidance on this. For a t-test a small p-value would not deter most analysts, but for Bartlett's test the same is not true. Perhaps the wisest course, even when dealing with robust procedures, is to apply both the parametric and the corresponding nonparametric procedure, if one is available, and to investigate further if they disagree.

The analysis of data is fundamental to statistics courses. Calculations can sometimes be done using calculators and tables, although elementary statistical packages would be more common. Favourites would be MINITAB, EXCEL, JMP and SPSS. Readers should work through the examples and exercises here using whatever software is familiar to them. However R is freeware and Dr Paul Rippon has written an *R Companion* for the material here: Rippon (2016). It is strongly recommended that the reader work through the material here and the *R Companion* simultaneously. For those developing their R skills verifying R output using the package with which they are most familiar would be a sensible way forward.

We assume that those readers initially without a background in R will acquire appropriate R skills over the time they are working through this material. Short courses are often available or appropriate resources are available at, for example, http://cran.r-project.org/other-docs.html.

Sometimes different software will produce different outcomes to each other and to hand calculations. This isn't necessarily wrong. For example there are different ways to treat ties and this is not always apparent in the output of various packages. If in doubt most packages have Help files that give details. However most users can cheerfully dismiss minor discrepancies as 'noise' but dig deeper when major discrepancies arise.

The first chapter gives a taste of nonparametric procedures appropriate in an introductory statistics course that assumes minimal mathematical background. The material will be covered towards the end of the course after the student has become familiar with random variables, the binomial and normal distributions and the one and two sample t tests. The appendix gives some challenge material that may be omitted at first reading.

The second chapter focuses on what are possibly the three most useful experimental designs: the completely randomised, randomised block and balanced incomplete block designs. The appropriate nonparametric tests, the Kruskal-Wallis, Friedman and Durbin tests are given, as are examples in which p-values are calculated using the asymptotic chi-squared approximations to the null distributions of the test statistics. Improved approximations are given using ANOVA F tests. Finally trend and umbrella tests are derived as orthogonal contrasts.

Many nonparametric tests find p-values using an asymptotic or approximate distribution that, in some circumstances, may be far from satisfactory. Resampling methods like the bootstrap or permutation testing may need a little coding, but given that, yield quick and accurate results. The third chapter introduces permutation testing. Knowledge of a computing language like R is assumed.

There are exercises for each chapter. Their solutions draw heavily on the R code in Rippon (2016). Parts of some of the analyses were done using JMP, the 'click and point' software with which I am most familiar nowadays. Readers should also use the software that is most familiar to them. Wherever possible the R programs in Rippon (2016) have been modified and applied to the exercises.

The following *Additional Supplementary Files* give some of the R code used in this book.

- Herbicide Example
- Vanilla flavour scores data entry
- Chapter Two Exercise 1 Solution
- Chapter Two Exercise 2 Solution
- Chapter Two Exercise 3 Solution
- Chapter Three Exercise 1 Solution Comprehensive Chocolate Analysis
- Chapter Three Exercise 2 Solution Comprehensive Word processors Analysis
- Chapter Three Exercise 3 Solution Comprehensive flavour score analysis

This is not all the R code in the text. You will need to type in the remainder.

My good friend and colleague Dr John Best wrote computer programs that supported the data analysis throughout this material, and has helped to clarify both my thinking and the text. Subsequently Dr Paul Rippon produced the *R Companion* (Rippon, 2016) that accompanies this effort. In doing so he read my text and made many useful suggestions.

My deepest thanks to Paul and John for their contributions.

Reference

RIPPON, Paul (2016). *An R Companion to Introductory Nonparametrics*.

1 A FIRST PERSPECTIVE ON NONPARAMETRIC TESTING

Learning Objectives

After successful completion of the material in this chapter the student will be able to

- discuss the nature of nonparametric methods and contrast them to parametric methods, and
- apply a number of nonparametric tests to appropriate data sets.

1.1 WHAT ARE NONPARAMETRIC METHODS?

Although there is good agreement on which tests are nonparametric, a definition is hard to give. Typically nonparametric tests assume less than their parametric competitors, but there is much more to distinguish the two than this.

In most introductory statistics courses most of the tests studied initially are parametric: they make reference to specific parameters of the population under study, or they are valid only if the population has some specific distribution, such as the normal, or the binomial. Consider, for example, the one-sample t-test: it assumes $X_1, \ldots, X_n$ are independent and normal with mean μ and variance σ^2. If $\overline{X} = \sum_i X_i / n$ and $S^2 = \sum_i \left(X_i - \overline{X}\right)^2 / (n-1)$ then the one-sample t-test tests for a particular population mean μ_0, formally testing H_0: $\mu = \mu_0$ against one sided (K_{11}: $\mu < \mu_0$ or K_{12}: $\mu > \mu_0$) or two-sided (K_2: $\mu \neq \mu_0$) alternatives, uses the test statistic $T = (\overline{X} - \mu_0)\sqrt{n}/S$. The test statistic T has the t distribution with $n-1$ degrees of freedom: t_{n-1}. This is very definitely a parametric test: it is testing for a particular value of the parameter μ, and it depends clearly on the assumption of normality. If the normality or independence assumptions are dubious, or if the first half of the observations is less variable than the second half, then this t-test cannot be validly applied. Then what is needed is a test that makes fewer assumptions. Typically a nonparametric test makes fewer assumptions than the corresponding parametric test. Sometimes so little is assumed in the scenario of interest that it is not reasonable to construct a parametric test. We give examples of both these situations.

Another reason why a parametric test may be inapplicable is that the data may not be in a suitable form. Four scales of measurement can be identified.

i) *Nominal*. Variables differ in kind rather than amount. The data are categorised, perhaps into colour, or gender.
ii) *Ordinal*. Ordinal scales may be based on qualitative rather than quantitative variables; however some ordering is also implied. This frequently involves ranks, as in ranking taste preferences. Alternatively, the variables may be categorical; for example, people may be categorised as short, medium height or tall.
iii) *Interval* and *ratio* scales. Measurements are quantitative, and the usual arithmetic operations can meaningfully be used. In the ratio scale the zero point reflects the absence of the attribute being observed. This is not the case with the interval scale. Examples are probabilities and temperatures.

Parametric procedures typically involve interval or ratio scales. Nonparametric procedures are available for all measurement scales, and are certainly possible for nominal and ordinal scales, where parametric procedures are not available. The wider applicability of nonparametric procedures may be offset by them having less *power* and *efficiency* when the parametric assumptions are valid. (We won't go into the technical meanings of power and efficiency here, but with both, more is good.) This is quite reasonable: if a parametric procedure can validly assume more, it is more likely to give deeper insights into the analysis. If some of the parametric assumptions are not valid, the nonparametric procedure is valid when the parametric is not.

In the following sections the reader is introduced to a number of nonparametric tests: the one and two-sample sign tests, a two-sample runs test and a runs test for randomness, the median test, the Wilcoxon signed ranks test and the Wilcoxon two sample test of location.

1.2 THE SIGN TESTS

1.2.1 A ONE-SAMPLE SIGN TEST

The one sample t-test is the parametric test for a particular population mean. We now develop a test for a particular *median*. Like the mean, the median of a distribution is a measure of central tendency. For the distribution of a random variable X the median satisfies $\mathrm{P}(X \leq \text{median}) = \mathrm{P}(X \geq \text{median})$. For symmetric distributions, the mean is equal to the median. The null hypothesis is H_0: median $= m_0$; the alternative could be one-sided, K_{11}: median $> m_0$ (or K_{12}: median $< m_0$) or two-sided, K_2: median $\neq m_0$. The test statistic is S, the number of observations greater than m_0. The name of the test comes from calling an observation greater than m_0 'positive', and an observation less than m_0 'negative'. The test statistic is then the number of positive (or negative) signs.

Under H_0, a single observation of X is equally likely to be at least or at most m_0. When X is continuous $P(X > m_0) = 0.5$ and under H_0 the statistic S is binomially distributed with parameters n, the total number of observations, and $p = 0.5$. For $n \geq 10$, p-values may be calculated using a normal approximation. When the distribution of X is assumed to be continuous observations equal to m_0 are usually discarded. There are many possible treatments for ties, that result when continuity cannot be assumed. See, for example, Rayner and Best (1999).

Aside. Subsequently we write bin (n, p) for the binomial distribution with n Bernoulli trials and probability of success p, while $\mathrm{b}(n, p, x) = {}^{n}\mathrm{C}_{x}\, p^{x}\, (1-p)^{n-x}$ for $x = 0, 1, \ldots, n$ is the binomial probability function.

Monkey Example. Adult female monkeys at a particular site are known to have median weight 8.41 kilograms. Can the same conclusion apply to monkeys from another site with weights, in kilograms,

8.30, 9.50, 9.60, 8.75, 8.40, 9.10, 9.25, 9.80,
10.05, 8.15, 10.00, 9.60, 9.80, 9.20, 9.30?

Of the 15 observations, 12 are greater than the hypothesised median. Calculation or tables give $P(S \leq 11) = b(15, 0.5, 0) + ... + b(15, 0.5, 11) = 0.982$. If the alternative specified larger monkeys at the new site, our p-value, the probability of observations at least as extreme as the observed, is $P(S \geq 12) = 1 - 0.982 = 0.018$. The null hypothesis can be rejected at the 0.05 level but not the 0.01 level. Since we were looking for a *different* median, P(data | null hypothesis) $= P(S \geq 12 \text{ or } S \leq 3) = 2\,P(S \geq 12) = 0.036$. Again there is evidence, at the 0.05 level but not the 0.01 level, against the null hypothesis.

For $n \geq 10$ and p not extreme the normal distribution with the same mean and variance as the binomial is a good approximation to the binomial. The more extreme the probability of success the larger n needs to be for a good approximation. It is counterproductive to be more precise.

In this example $n \geq 10$ and since this is the sign test $p = 0.5$. We therefore apply the normal approximation: $P(S \geq 12 \mid S \text{ is bin } (15, 0.5))$ is approximately equal to $P(Z > (11.5 - 7.5)/\sqrt{(15/4)} = 2.0656 | Z \text{ is } N(0, 1)) = 0.019$. From the above the exact value is 0.018: the approximation is excellent. This calculation uses the binomial mean np and variance $np(1 - p)$ and the continuity correction. Note that whenever a discrete distribution is approximated by a continuous one, the continuity correction is required. So if a discrete distribution takes the values x_1, x_2, and is approximated by a continuous random variable Y then $P(X = x_i)$ is approximately $P([x_{i-1} + x_i]/2 < Y < [x_i + x_{i+1}]/2)$. For the normal approximation to the binomial $P(X \geq x | X \text{ is bin}(n, p))$ is approximately $P(Y > x - 0.5 | Y$ is normal with mean np and variance $np(1 - p))$.

Compared to the t-test, the sign test has asymptotic relative efficiency (ARE) $2/\pi = 63.66\%$. Roughly this means that when the data are normally distributed, to achieve the same power as the sign test does with 100 observations, the t-test asymptotically requires only 63.66 observations. However the sign test is more generally applicable even if it is generally less powerful. When the data are not normal the ARE is greater than $2/\pi$, and for some distributions the ARE is greater than 100%.

Exercise. Analyse the monkey data above using a convenient and appropriate computer package. You should find that normality is a valid assumption. For example the Shapiro-Wilk test for normality has p-value 0.25. Any convenient test of normality should give a p-value well in excess of 0.05. The t-test for $H_0: \mu = 8.41$ against $K: \mu \neq 8.41$ has p-value 0.000. The signed-ranks test, a nonparametric test we will consider later in this chapter, also has p-value 0.000. When normality holds the signed-ranks test is more efficient than the sign test and less efficient than the t-test.

1.2.2 A TWO-SAMPLE SIGN TEST

Suppose now we have paired subjects. These might be the same individual with identical skin conditions on both forearms. In some studies the paired subjects may be twins, or individuals matched on a number of factors. Different treatments are then applied to each individual in the pair: treatment X say to one, and treatment Y say to the other. The aim is to assess which treatment is preferable.

More formally suppose we have n independent pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, and we wish to test that the X's and Y's have the same distribution against the alternative that they differ in location. The parametric paired t-test could be used if it is valid to assume that the differences are normally distributed and all have the same variance.

To perform this sign test first calculate the differences $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$. Second, count S, the number of positive differences, or *signs*. The distribution of S is binomial $\mathrm{b}(n, p)$, and we wish to test $\mathrm{H}_0: p = 0.5$ against one or two sided alternatives such as $\mathrm{K}: p \neq 0.5$. This test doesn't assume normality but it does involve the binomial parameter. It is widely regarded as a nonparametric test, but it might more accurately be said to be less parametric than the paired t-test.

Heart Rates Example. In Table 1.1 the heart rates in beats per minute of 10 rats alone and in the presence of another are given. We now test at the 0.05 level if togetherness increases heart rate. There are eight negative differences or signs and two positive differences or signs. We calculate $P(S \leq 2|\mathrm{bin}(10, 0.5)) = 0.0547$. At the 0.05 level the null hypothesis of no increase cannot be rejected, but given this borderline acceptance it may well be prudent to conduct another study with more than 10 subjects.

Interestingly the Shapiro-Wilk test for normality has a large p-value and the t-test p-value is 0.001. As is not uncommon, the parametric test is far more critical of the null hypothesis than the nonparametric test. One possible explanation of this phenomenon is that the t-test makes more assumptions, which may not be true, and hence is more critical of the data *and* the assumptions. Conversely the nonparametric test doesn't make use of assumptions that may not be true.

Rat number	1	2	3	4	5	6	7	8	9	10
Rate alone (x_i)	463	462	462	456	450	426	418	415	409	402
Rate together (y_i)	523	494	461	535	476	454	448	408	470	437
Difference	–60	–32	1	–79	–26	–28	–30	7	–61	–35

TABLE 1.1. Rat data

1.3 RUNS TESTS

Suppose we have observations of two types: A and $\overline{A}$. These might be defective and non-defective, male and female or, when candidates are interviewed for a position, university graduate and non-graduate. A typical sequence could be

$$A, \overline{A}, \overline{A}, A, A, A, \overline{A}, \overline{A}, A.$$

Any sequence of like observations, bounded by observations of a different type, is called a *run*. Alternatively a run is defined to be the greatest subsequence of like elements. The number of observations in the run is called its *length*. In the above sequence there are five runs, of lengths 1, 2, 3, 2, 1.

Runs tests are given in the next two subsections: both are nonparametric. Although approximate distributions are given for the test statistics, no distributional assumptions are made about the data.

1.3.1 A TWO-SAMPLE RUNS TEST

Suppose we have $X_1, \ldots, X_m$, a random sample of size m from the X population, and $Y_1, \ldots, Y_n$, an independent random sample of size n from the Y population. These samples are combined, ordered, and classified as either an X or a Y. A typical result would be similar to the following:

$$X < X < Y < Y < Y < X < Y < X < Y.$$

From such data we can calculate the number of runs. Under the null hypothesis that the X and Y populations are identical a moderate or large number of runs could be expected. A small number of runs could result under an alternative hypothesis of differences in location (such as mainly Xs followed by mainly Ys):

$$\_\_X\_XX\_X\_\_\_\_X\_\_\_\_\_\_\_\_\_\_\_\_\_YY\_\_\_\_\_Y\_\_\_Y$$

or under an alternative hypothesis of differences in dispersion (such as mainly Xs between clusters of mainly Ys in the tails):

$$\_\_Y\_\_\_\_\_Y\_\_\_Y\_\_\_\_X\_\_\_Y\_XX\_X\_X\_XX\_YX\_\_\_\_\_YY\_\_\_Y\_\_\_\_Y$$

The test statistic is therefore taken to be one-tailed.

Flints Example. Four pieces of flint were collected from area A, and five pieces from area B. By scratching against each other, the flints were ranked in order of hardness:

A, A, A, B, A, B, B, B, B.

There are four runs. Is this significantly small? We need the distribution of T, the number of runs. No proof is given for the following result.

Probability function of T, the number of runs
Under the null hypothesis that the m Xs and n Ys come from the same population, the probability function of T is given by

$$\mathrm{P}(T = 2k) = \frac{2\ {}^{m-1}C_{k-1}\ {}^{n-1}C_{k-1}}{{}^{m+n}C_n} \text{ and } \mathrm{P}(T = 2k+1) = \frac{{}^{m-1}C_k\ {}^{n-1}C_{k-1} + {}^{m-1}C_{k-1}\ {}^{n-1}C_k}{{}^{m+n}C_n}.$$

Flints Example continued. For the flints we have $m = 4$, $n = 5$ and $t = 4$. So with $k = 1$,

$$\mathrm{P}(T = 2) = \frac{2\ {}^3C_0\ {}^4C_0}{{}^9C_5} = 2/126 \text{ and } \mathrm{P}(T = 3) = \frac{{}^3C_1\ {}^4C_0 + {}^3C_0\ {}^4C_1}{{}^9C_5} = 7/126.$$

With $k = 2$,

$$\mathrm{P}(T = 4) = \frac{2\ {}^3C_1\ {}^4C_1}{{}^9C_5} = 24/126.$$

Thus the probability of observations at least as extreme as the observed is $\mathrm{P}(T \le 4) = 33/126 = 0.2619$.

Normal Approximation

Use of the exact formula is tedious for moderate m and n. However if both are greater than 10 an excellent normal approximation is available. This needs the mean and variance of T:

$$\mathrm{E}[T] = 1 + \frac{2mn}{m+n} \text{ and } \mathrm{var}(T) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}.$$

Again no proof is given.

Flints Example continued. Direct calculation gives $\mathrm{E}[T] = 1 + 2 \times 4 \times 5/9 = 49/9 = 5.4444$ and $\mathrm{var}(T) = 2\times4\times5\times(40 - 9)/(81\times8) = 155/81 = 1.9136 = 1.3833^2$. Using the continuity correction, for the flints data

$$\begin{aligned}\mathrm{P}(T \le 4) &= P(Z \le (4.5 - 5.4444)/1.3833 \mid Z \text{ is } N(0, 1)) \\ &= P(Z \le -0.6827) = 0.2474.\end{aligned}$$

This is remarkable agreement since neither m nor n is greater than 10. However the approximation is not as good in the tails, the important region, unless both m and n are greater than 10.

1.3.2 A RUNS TEST FOR RANDOMNESS

The data must be of two types, such as success or failure as in the binomial situation, or above or below the median (or upper quartile) with continuous data. Usually, though not necessarily, we try to ensure approximately the same numbers of each type, as this improves the normal approximation. Unlike the two-sample test described in 1.3.1, the alternative here is two-sided. A large number of runs suggests some alternating mechanism (continuously overcorrecting) and a small number of runs is consistent with a mechanism in control until a fault occurs. The previous formulas apply.

Examination Example. A true-false examination produces the following sequence of correct answers:

T, F, F, T, F, T, F, T, T, F, T, F, F, T, F, T, F, T, T, F.

Here we have $m = 10, n = 10$, and $t = 16$. We have $\mathrm{E}[T] = 11$ and $\mathrm{var}(T) = 4.7368$, so the normal approximation gives $P(T \geq 16) = P(Z \geq (15.5 - 11)/2.1764 | Z \text{ is } N(0, 1)) = P(Z \geq 2.0676) = 0.0193$. Since a two-tailed test is called for, the p-value is 0.0386. At the 0.05 level there is some evidence of non-randomness in the answers.

Speed Example. The (unordered) speeds of every fifth passenger car past a check-point, in miles per hour, were:

46, 58, 60, 56, 70, 66, 48, 54, 62, 41, 39, 52, 45, 62, 53, 69, 65,
65, 67, 76, 52, 52, 59, 59, 67, 51, 46, 61, 40, 43, 42, 77, 67, 63,
59, 63, 63, 72, 57, 59, 42, 56, 47, 62, 67, 70, 63, 66, 69, 73.

The median is 59.5 and if observations are classified as either above or below the median you should obtain $m = n = 25$ and $t = 20$. Since $E[T] = 26$, $\text{var}(T) = 12.2449$ a p-value (for two tails) of 0.116 results. If speeds above and below 55.1 are assessed, we obtain $m = 33$, $n = 17$, $t = 18$, so that $E[T] = 23.44$, $\text{var}(T) = 9.8186$ and the p-value is $2P(T \leq 18) = 2P(Z < -1.5765) = 0.115$.

Trends of similar speeds could be expected: in high density traffic almost everyone travels at the same speed, but on the open road drivers tend to treat the speed limit as both an upper and a lower limit. This is equivalent to using one-sided tests; the alternative would specify low values of T. This means we are more critical of these data.

1.4 THE MEDIAN TEST

Experience has shown that the runs test is sensitive to differences both in shape and location. The median test is sensitive to differences in location, but not to differences in shape. When normality holds both the median test and the runs test have test efficiency (ARE) $2/\pi = 63.66\%$ compared to their parametric competitors.

From the ith of c populations a random sample of size n_i is drawn. The observations are pooled and a predetermined quantile, usually the median, is found. The numbers of observations in the ith sample that are above (A_i) and below (B_i) the combined quantile are then as in Table 1.2. All row and column totals are known before sighting the data.

Sample	1	2	…		c	Totals
Above	A_1	A_2	…		A_c	$A_\bullet$
Below	B_1	B_2	…		B_c	$B_\bullet$
Total	n_1	n_2	…		n_c	$n_\bullet$

TABLE 1.2. Layout of data for the median test

We assume that all samples are independent random samples, that measurement is at least ordinal and that if all populations have the same quantile, then all populations have the same probability of an observation exceeding that quantile. Then we may test H: all c populations have the same quantile against K: at least two of the populations have different quantiles. The probability of the observed table is an *extended hypergeometric*:

$$ {}^{n_1}C_{A_1}\,{}^{n_2}C_{A_2}\ldots{}^{n_c}C_{A_c} \,/\, {}^{n_\bullet}C_{A_\bullet}\,, $$

in which $A_i = 0, \ldots, n_i$, $i = 1, \ldots, c$, $A_\bullet = A_1 + \ldots + A_c$ and $n_\bullet = n_1 + \ldots + n_c$.

To test H against K find and sum the probabilities of every table with a value of $X^2 = \sum_{all\ cells}(\text{observed} - \text{expected})^2 /$ expected at least as large as the observed.

When the quantile chosen is (close to) the median, the distribution of X^2 is generally well approximated by χ^2_{c-1}. In hand calculations the extended hypergeometric is rarely used.

Corn Example. Four different methods of growing corn were randomly assigned a large number of different plots of land, and the yield per acre computed for each plot (Conover, 1999, p. 173).

Method 1: 83, 91, 94, 89, 89, 96, 91, 92, 90: $n_1 = 9$;
Method 2: 91, 90, 81, 83, 84, 83, 88, 91, 89, 84: $n_2 = 10$;
Method 3: 101, 100, 91, 93, 96, 95, 94: $n_3 = 7$;
Method 4: 78, 82, 81, 77, 79, 81, 80, 81: $n_4 = 8$.

The median is 89 (three values). In Table 1.3, expected cell values are found using row total × column total/grand total are given in brackets. We find $X^2 = 17.54$ with a χ^2_3 p-value of 0.0005. There is very strong evidence that the methods are different.

Method	1	2	3	4	Totals
> 89	6 (4.24)	3 (4.71)	7 (3.29)	0 (3.76)	16
< 89	3 (4.76)	7 (5.29)	0 (3.71)	8 (4.24)	18
Total	9	10	7	8	34

TABLE 1.3. Corn data

Achievement Test Example. An achievement test was given to comparable classes in two different schools. The scores were

School 1: 43, 80, 99, 86, 68, 70, 85, 93, 98, 96, 75, 81, 32, 92, 96, 64, 79, 97, 76, 80;

School 2: 76, 65, 73, 95, 77, 99, 55, 35, 72, 83, 70, 65, 86, 60, 62, 90, 71, 65, 89, 71, 80, 76, 93, 94.

School	1	2	Totals
< 78	7 (10)	15 (12)	22
> 78	13 (10)	9 (12)	22
Total	20	24	44

TABLE 1.4. Achievement test data

The median is 78, a score not achieved by any student. We find $X^2 = 3.3$ and using the χ_1^2 approximation results in a p-value of 0.0693. Using the exact distribution,

$$\mathrm{P}(A_1 = 7) = {}^{20}C_7\,{}^{24}C_{15}/{}^{44}C_{22} = 0.0482,\ \mathrm{P}(A_1 = 6) = {}^{20}C_6\,{}^{24}C_{16}/{}^{44}C_{22} = 0.0135,$$
$$\mathrm{P}(A_1 = 5) = 0.0026,\ \mathrm{P}(A_1 = 4) = 0.0003,\ \ldots\ ,$$

giving $\mathrm{P}(A_1 \le 7) = 0.0646$. The required p-value is double this, 0.1292. To be sure of the meaning of 'at least as extreme as the observed', plot X^2 against A_1. The χ^2 approximation is poor here because a continuity correction is needed.

1.5 THE WILCOXON TESTS

The Wilcoxon tests are the nonparametric equivalents of the *t*-tests. We consider both one and two sample tests of location.

1.5.1 THE SIGNED RANKS TESTS

Suppose that $x_1, \ldots, x_n$ are observations from a distribution assumed to be symmetric and with mean and hence median hypothesised to be m_0. In practice symmetry might be informally assessed using a histogram, although there are formal tests for symmetry. We calculate the $x_i - m_0$, their absolute values, the ranks of these, and then W_- and W_+, the sum of the ranks corresponding to the negative and positive $x_i - m_0$. If there is a tie, the rank assigned is the mean of the ranks that would otherwise have been assigned.

A useful check is to note that $W_+ + W_- = 1 + \ldots + n = n(n+1)/2$.

Some books give exact tables of the W_+ distribution, but for sample sizes of at least 15 under the null hypothesis W_+ and W_- are both approximately normal with

$$\text{mean} = \frac{n(n+1)}{4} \text{ and variance} = \frac{n(n+1)(2n+1)}{24}.$$

Tyre Example. A testing company finds that 16 tyres of a certain make have provided miles of service as in Table 1.5. Do the results support the claim that on average this kind of tyre provides at least 30,000 miles of service?

Note that the change of scale in Table 1.5 doesn't affect the ranks. We find

$$W_+ = 13.5 + 2 + 7.5 + 11.5 + 3 = 37.5,$$
$$W_- = 5 + 1 + 6 + 11.5 + 10 + 15 + 9 + 13.5 + 16 + 4 + 7.5 = 98.5 \text{ and}$$
$$W_+ + W_- = 37.5 + 98.5 = 136 = 16\times 17/2, \text{ verifying the identity } W_+ + W_- = n(n+1)/2.$$

In this example H_0: $m_0 = 30{,}000$ and K: $m_0 < 30{,}000$. We have

$$Z = \frac{W_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{37.5 - 68}{\sqrt{374}} = -1.577.$$

This gives a p-value of approximately 0.0574. So at the 0.05 level there is no evidence against the null hypothesis of an average of (at least) 30,000 miles of service. However this conclusion is marginal.

Miles of service (x_i)	$y_i = (x_i - 30{,}000)/1{,}000$	ranks of the $\lvert y_i \rvert$
27900	–2.1	5
35100	5.1	13.5
29800	–0.2	1
27700	–2.3	6
26700	–3.3	11.5
30700	0.7	2
26900	–3.1	10
32400	2.4	7.5
24800	–5.2	15
27400	–2.6	9
24900	–5.1	13.5
33300	3.3	11.5
31600	1.6	3
24300	–5.7	16
28300	–1.7	4
27600	–2.4	7.5

TABLE 1.5. Type service data

It is interesting to note that of the 16 signs 11 are negative and since $P(S \geq 11|S$ is bin $(16, 0.5)) = 0.105$, the one-sided sign test is not significant at the 0.1 level. The t-test has p-value 0.062, and since the Shapiro-Wilk test of normality has p-value 0.45, the t-test may be validly applied. There is remarkable consistency in the conclusions from these tests, so perhaps the message is that the null hypothesis is neither confirmed nor rejected at the usual levels of significance, and this suggests a study with a larger sample size could be of value.

1.5.2 THE TWO-SAMPLE WILCOXON TEST

Suppose we have a random sample of size m from a population labelled X, and an independent random sample of size n from a population labelled Y. We test the null hypothesis that the populations are identical against what are called *slippage alternatives* in the literature. This means that the X's tend to be smaller/larger/different from the Y's. These are *sometimes* equivalent to differences in the medians or the means: $\mathrm{E}[X] <, >, \neq \mathrm{E}[Y]$. The Wilcoxon test in this situation involves a sum of ranks.

The two samples are combined, ordered and ranked. The sum of the X ranks, W_X, and the sum of the Y ranks, W_Y, are calculated. If there is a tie, each of the tied observations is assigned the mean of the ranks they would otherwise have received. The following result gives an arithmetic check when calculating the test statistic.

Result: $W_X + W_Y = 1 + 2 + \ldots + (m + n) = (m + n)(m + n + 1)/2$.

If both m and n are greater than 8 then under the null hypothesis the distribution of W_X is approximately $N(m(m + n + 1)/2, mn(m + n + 1)/12)$. Similarly under the null hypothesis W_Y is approximately $N(n(m + n + 1)/2, mn(m + n + 1)/12)$.

Flints Example Continued. The combined, ordered and ranked data are as follows:

	A	A	A	B	A	B	B	B	B	
A ranks	1	2	3		5					$W_A = 11$
B ranks				4		6	7	8	9	$W_B = 34$

TABLE 1.6. Flints data

We have $W_A = 11$, $W_B = 34$.
Check: $W_A + W_B = 45$ and $(m + n)(m + n + 1)/2 = 9\times10/2 = 45$.

We are interested in testing for equality of the A and B distributions. If the alternative is that the B ranks tend to be greater than the A ranks, the rejection region is large values of W_B. Since neither m nor n is greater than 8 the normal approximation isn't recommended. However if it is used, the approximating mean is 25 and variance is 50/3. Applying a continuity correction results in a p-value of

$$P(W_B \geq 34) = P(Z > (33.5 - 25)\sqrt{(3/50)} = 2.0821) = 0.0187.$$

Here we won't discuss the (exact) small sample distribution of W_B. However that is available through R, which gives a p-value of 0.0159.

Teaching Methods Example. Two groups of students were taught by two different methods. We use the Wilcoxon test to see if the methods are equally effective. The combined, ordered and ranked scores are as follows.

X					70		72		74				79	80
Y	65	66	68	69		71		73		75	76	78		
rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14

X	82		86		91	93	95
Y		84		90			
rank	15	16	17	18	19	20	21

TABLE 1.7. Teaching methods data

The sum of the X ranks is $5 + 7 + 9 + 13 + 14 + 15 + 17 + 19 + 20 + 21 = 140$, and the sum of the Y ranks is $1 + 2 + 3 + 4 + 6 + 8 + 10 + 11 + 12 + 16 + 18 = 91$.

Check: $140 + 91 = 231 = 21\times 22/2$.
We calculate

$$\frac{W_X - \frac{m(m+n+1)}{2} - 0.5}{\sqrt{\frac{mn(m+n+1)}{12}}} = \frac{139.5-110}{\sqrt{605/3}} = 2.07733.$$

and $2P(Z >= 2.07733) = 0.0378$. Using a two-tailed test the null hypothesis that the two methods are equally effective is rejected at the 0.05 level.

The two sample Wilcoxon test is sometimes referred to as the Wilcoxon Mann-Whitney Test. At about the same time (the 1940s) as the Wilcoxon test was proposed, so was an alternative test, the Mann-Whitney test. Although it is not immediately apparent, the tests are equivalent. This is shown in the appendix that can be viewed as challenge material.

Appendix: Wilcoxon Mann-Whitney Tests

As in section 1.5.2 suppose we have a random sample of size m from a population labelled X, and an independent random sample of size n from a population labelled Y. To test for equality of the X and Y distributions the Wilcoxon test in this situation involves a sum of the X (or Y) ranks.

As before, suppose the two samples are combined, ordered and ranked. If there is a tie, the rank assigned is the mean of the ranks that would otherwise have been assigned. The following development assumes there are no ties but can be adjusted to cope with ties. For each pair i, j, define $h_{ij} = 1$ if $X_i > Y_j$, 0 otherwise. So h_{ij} counts 1 if Y_j is smaller than X_i, $\sum_j h_{ij}$ counts the number of Y's that are smaller than X_i, and $\sum_i \sum_j h_{ij} = U_Y$ say, counts the number of Y's that are smaller than the X's. Equivalently U_Y counts the number of X's that are bigger than the Y's. We may similarly define U_X as the number of X's that are smaller than the Y's, or, equivalently, the number of Y's that are bigger than the X's.

Recall that we have defined the sum of the X ranks to be W_X, and the sum of the Y ranks to be W_Y. We now state and prove results relating W_X, W_Y, U_X and U_Y.

Result 1: $W_X + W_Y = (m + n)(m + n + 1)/2$.

Proof. $W_X + W_Y = 1 + 2 + 3 + \ldots + (m + n) = (m + n)(m + n + 1)/2$, using the result for the sum of an arithmetic progression.

Result 2. $U_X + U_Y = mn$.
Proof. Consider any adjacent pair XY in the combined ordered sample. If XY is transposed to YX, U_X decreases by 1, while U_Y increases by 1. Thus $U_X + U_Y$ is unaffected by such changes. Perform these transpositions until the combined sample becomes $Y_1\, Y_2 \ldots Y_n X_1 X_2 \ldots X_m$. Then $U_X = 0$ and $U_Y = m + m + \ldots + m$ (n times) $= mn$.

Exercise. Verify the result if the transformed sample is $X_1 X_2 \ldots X_m\, Y_1\, Y_2 \ldots Y_n$.

These results are useful to check calculations done by hand. If W_X and W_Y are calculated independently, then Result 1 should be satisfied. Similarly if U_X and U_Y are calculated independently, then Result 2 should be satisfied.

Result 3: $W_Y + U_Y = mn + n(n + 1)/2$.
Proof. Consider transposing XY to YX as in the proof of Result 2. Each such exchange reduces W_Y by 1 and increases U_Y by 1. Thus $W_Y + U_Y$ is unaffected by such changes. Perform these transpositions until the combined sample becomes $Y_1\ Y_2\ \ldots\ Y_nX_1X_2\ \ldots\ X_m$. Then $W_Y = 1 + 2 + \ldots + n = n(n + 1)/2$ and $U_Y = m + m + \ldots + m$ (n times) = mn. Hence the result.

Exercises.
(i) Verify the Result 3 if the transformed sample is $X_1X_2\ \ldots\ X_m\ Y_1\ Y_2\ \ldots\ Y_n$.
(ii) Using transpositions show that $W_X + U_X = mn + m(m + 1)/2$.
(iii) Using Results 2 and 3 show both algebraically and using transpositions that

$$W_X = U_Y + m(m + 1)/2 \text{ and } W_Y = U_X + n(n + 1)/2.$$

These three results establish that the four statistics W_X, W_Y, U_X and U_Y are equivalent in the sense that from any one all the others can be calculated. For larger data sets when calculation of U_X and U_Y can be tedious, we could use the data to calculate W_X and W_Y and then

$$U_Y = W_X - m(m + 1)/2 \text{ and } U_X = W_Y - n(n + 1)/2.$$

Then either U_X or U_Y can then be referred to the $N(mn/2,\ mn(m + n + 1)/12)$ distribution.

In addition taking expectations in $U_Y = W_X - m(m + 1)/2$ gives

$$\mathrm{E}[W_X] = \mathrm{E}[U_Y] + m(m + 1)/2 = m(m + n + 1)/2,$$

while taking variances gives $\mathrm{var}(W_X) = \mathrm{var}(U_Y)$, so W_X may be referred to the $N(m(m + n + 1)/2, mn(m + n + 1)/12)$ distribution. Similarly W_Y may be referred to the $N(n(m + n + 1)/2, mn(m + n + 1)/12)$ distribution. Tests based on the normal approximations to all of W_X, W_Y, U_X and U_Y will all give the same p-values and conclusions.

Flints Example Continued.
It is routine to show that for these data $W_A = 11$, $W_B = 34$, $U_B = 1$ and $U_A = 5 + 5 + 5 + 4 = 19$.

Checks:
(i) $W_A + W_B = 45$ and $(m + n)(m + n + 1)/2 = 9\times10/2 = 45$.
(ii) $W_A + U_A = 30$ and $mn + m(m + 1)/2 = 20 + 4\times5/2 = 30$.
(iii) $W_B + U_B = 35$ and $mn + n(n + 1)/2 = 20 + 5\times6/2 = 35$.
(iv) $U_A + U_B = 20$ and $mn = 20$.

Tables of U_X (and U_Y) are available in most nonparametric texts. However there is a convenient normal approximation, adequate if both m and n are greater than 8. Then both U_X and U_Y are normal with mean $mn/2$ and variance $mn(m + n + 1)/12$. Note we are testing for equality of the X and Y means. If the alternative is that larger values of Y are expected the rejection region consists of large values of U_X. To analyse the flint data we need the (exact) small sample distribution of U_X. This is given in some texts, but will not be discussed here.

Teaching Methods Example.
Hand calculation of U_X and U_Y is tedious, but from previously $W_X = 140$ and $W_Y = 91$ and incidentally the first result is verified. In addition

$$U_Y = W_X - m(m + 1)/2 = 140 - 55 = 85 \text{ and } U_X = W_Y - n(n + 1)/2 = 91 - 66 = 25.$$

Check: $U_X + U_Y = 25 + 85 = 110 = 10\times 11$.

We calculate $Z = \left(U_Y - 0.5 - mn/2\right)/\sqrt{mn\left(m+n+1\right)/12} = 2.0773$. Using a two-tailed test and the 0.05 significance level, the null hypothesis that the two methods are equally effective is rejected (the p-value is 0.0346). This is exactly as in section 1.5.2 using W_X.

2 NONPARAMETRIC TESTING IN THE COMPLETELY RANDOMISED, RANDOMISED BLOCKS AND BALANCED INCOMPLETE BLOCK DESIGNS

Learning Objectives

After successful completion of the material in this chapter the student will be able to

- apply the Kruskal-Wallis, Friedman and Durbin tests using both chi-squared and F distribution approximations to the appropriate test statistics and
- explain the meaning of component statistics and find linear (Page-type) and quadratic (umbrella) components of the Kruskal-Wallis, Friedman and Durbin test statistics.

2.1 INTRODUCTION AND OUTLINE

Among the elementary statistical designs, probably the most widely used are the completely randomised, randomised blocks and balanced incomplete block designs. There exist parametric ANOVA F tests for treatment effects for these designs, and these tests are known to be *robust*: the assumptions underpinning the test may not hold but the analysis is hardly affected by this. For example a test of the normality of the residuals may be significant at the 0.05 level but the p-values found using the F test and a permutation test (to be discussed in the next chapter) are often very similar. However if the assumptions for these parametric tests are seriously flawed, then the validity of the conclusions will be in doubt. In a similar vein tests of equality of variance are known to be very sensitive to the assumption of normality. For these tests a small deviation from normality will seriously affect inference. When parametric inference is in any way dubious nonparametric tests are required.

In sections 2.2, 2.3 and 2.4 we introduce the Kruskal-Wallis test in the case of the completely randomised design, the Friedman test for the randomised block design and the Durbin test for the balanced incomplete block design. These nonparametric test statistics have asymptotic chi-squared distributions.

All statistics and analyses in this chapter assume there are no ties. When ties occur the usual practice is to give tied data the mean of the ranks they would otherwise have received. This introduces an extra element of approximation to the distributions of the test statistics.

In section 2.5 it is shown that ANOVA F test statistics on the ranks are one to one functions of the appropriate nonparametric test statistics. This leads to improved approximations to the null distribution of the treatment test statistics.

In section 2.6 orthogonal contrasts are used to decompose the nonparametric test statistics into linear, quadratic etc. components. The linear components are the basis of Page or Page-type tests while the quadratic component is the basis of umbrella tests.

2.2 THE KRUSKAL-WALLIS TEST

Suppose we have distinct (untied) observations, with y_{ij} being the jth of n_i observations on the ith of t treatments. The model assumed is the completely randomized design, sometimes called the one-way layout and sometimes the one-way analysis of variance. All $n_1 + \ldots + n_t = n$ observations are combined, ordered and ranked. For $i = 1, \ldots, t$ the sums of the ranks for treatment i, R_i, is calculated. The Kruskal-Wallis test statistic KW is given by

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{t} \frac{R_i^2}{n_i} - 3(n+1).$$

Under the null hypothesis of no treatment effects the test statistic has the χ^2_{t-1} distribution. The alternative hypothesis is sometimes simplistically presented as the treatment medians being different. A more general and much more correct presentation of the alternative is that the distributions for the different treatments are different. Although the Kruskal-Wallis test is well known to be sensitive to location differences, the simplistic presentation above requires an additional assumption that the only difference in the distributions is in their medians.

Tomato Example. The data in Table 2.1 are 'scores' – lengths in millimetres measured along a line with one end marked 'poor' and the other 'good'. It is assumed there were 24 independent tasters for each of the four tomato varieties: Floradade, Momotara, Summit and Rutgers. The analysis in Rayner and Best (1989, Section 8.2.3) found normality to be a marginal assumption.

The one-way ANOVA F test for treatment effects for this data set yields a p-value of 0.4978. The null hypothesis of equality of tomato means is accepted at all reasonable significance levels, and certainly at the 0.05 level. However when the Shapiro-Wilk test of normality is applied to the residuals a p-value of 0.0052 is obtained. The validity of the ANOVA F test is questionable, although the test is known to be robust to the normality assumption.

To assess the null hypothesis of no treatment effects using the Kruskal-Wallis test hand calculation requires the samples to be combined, ordered and ranked. Here that results in rank sums of 1058, 1200, 1303.5 and 1094.5. Of course this is tedious for larger data sets such as this, and using R as in Rippon (2016) is to be preferred. Using the formula $KW = 1.9772$ with a χ^2_3 p-value of 0.5772. At the 0.05 level the null hypothesis that the tomato distributions are similar is accepted. Note that this value of the Kruskal-Wallis test statistic ignores the fact that there are a few tied values in the data. Here an adjustment for ties makes little difference to the value of the test statistic.

Tomato variety	Flavour scores
Floradade	43, 5, 74, 64, 10, 16, 75, 20, 36, 76, 60, 57, 55, 29, 82, 91, 66, 27, 72, 108, 84, 50, 82, 39
Momotara	74, 112, 64, 101, 105, 12, 33, 90, 129, 37, 50, 44, 18, 24, 48, 62, 88, 50, 73, 119, 109, 50, 12, 37
Summit	109, 25, 48, 91, 52, 35, 42, 100, 22, 122, 105, 119, 29, 26, 102, 48, 108, 53, 57, 82, 105, 108, 13, 74
Rutgers	39, 82, 100, 62, 126, 26, 24, 35, 74, 19, 113, 56, 61, 21, 6, 13, 118, 91, 60, 88, 15, 32, 134, 29

TABLE 2.1. Tomato data

2.3 THE FRIEDMAN TEST

For the randomised block design suppose we have distinct (untied) observations, y_{ij}, this being the ith of t treatments on the jth of b blocks. The observations are ranked within each block and R_i, the sum of the ranks for treatment i over all blocks is calculated for $i = 1, \ldots, t$. The Friedman test statistic is

$$S = \frac{12}{bt(t+1)} \sum_{i=1}^{t} R_i^2 - 3b(t+1).$$

Under the null hypothesis of no treatment effects the test statistic has the χ^2_{t-1} distribution. Note that here ranking is within blocks; for the Kruskal-Wallis test overall ranking is used. As with the Kruskal-Wallis test the Friedman is not testing for equality of means (or medians). That would only be the case if initially it could be assumed that the distributions sampled differed only in their means (or medians). If that is not the case, testing assesses whether or not the treatment distributions are similar. However it is well known that the Friedman test is sensitive to location differences.

Lemonade Example. Five lemonades with increasing sugar content are ranked by each of ten judges. They were not permitted to give tied outcomes. The results are in Table 2.2 below. We wish to assess what, if any, differences there are between the lemonades.

Routine calculation finds $S = 9.04$. The χ^2_4 p-value is 0.0601. There is weak evidence, at the 0.10 level, of a difference in lemonades. While there is no evidence of a difference at the 0.05 level, further investigation would seem to be warranted. We will return to this data set later in this and subsequent chapters.

	Judge										Product
Product	1	2	3	4	5	6	7	8	9	10	mean
1	5	3	4	5	3	4	5	3	1	3	3.6
2	2	5	3	2	5	3	2	5	4	1	3.2
3	1	2	2	1	2	2	1	2	2	2	1.7
4	3	1	5	3	1	5	3	1	5	4	3.1
5	4	4	1	4	4	1	4	4	3	5	3.4

TABLE 2.2. Lemonades ranked by ten judges

2.4 THE DURBIN TEST

In the balanced incomplete block design each of the b blocks contains k experimental units, each of the t treatments appears in r blocks, and every treatment appears with every other treatment precisely λ times. Necessarily

$$k < t,\ r < b,\ bk = rt, \text{ and } \lambda(t-1) = r(k-1).$$

Treatments are ranked on each block and Durbin's statistic, D, is given by

$$D = \frac{12(t-1)}{bk(k^2-1)}\sum_{i=1}^{t} R_i^2 - \frac{3r(t-1)(k+1)}{(k-1)}$$

in which R_i is the sum of the ranks given to treatment i, $i = 1, \ldots, t$. Without further assumptions the test statistic D tests the null hypothesis of equality of the treatment distributions. The asymptotic distribution of D is χ^2_{t-1}. If $k = t$ then $D = S$: the design is no longer incomplete; it is, in effect, complete.

Ice Cream Example. Conover (1999, p. 390) gave an example of a taste test involving seven ice cream varieties, coded S, U, V, W, X, Y and Z, and presented three at a time. Table 2.3 shows, for each judge, the rank given for each variety.

For this design we see that $b = t = 7, k = r = 3$ and $\lambda = 1$. We calculate that $\{R_1, R_2, R_3, R_4, R_5, R_6, R_7\} = \{8, 9, 4, 3, 5, 6, 7\}$ and $D = 12$. The χ^2_6 p-value is 0.0620. At the 0.05 level there is no evidence of a difference in the distributions of the ice creams; however there is weak evidence, at the 0.10 level, of a difference. It is known that the χ^2_{t-1} approximation to the distribution of D is not particularly accurate, so it would be wise to pursue a better approximation in a marginal case like this. See Section 2.5.

	Variety						
Judge	S	U	V	W	X	Y	Z
1	2	3		1			
2		3	1		2		
3			2	1		3	
4				1	2		3
5	3				1	2	
6		3				1	2
7	3		1				2

TABLE 2.3. Ranks of seven judges of seven ice cream varieties

In section 2.6 D will be decomposed into orthogonal contrasts that will show that for these data there is evidence of what will be called an umbrella effect. Since D is associated with six degrees of freedom, the Durbin test is attempting to detect a quite complex alternative: the parameter space is six dimensional. The orthogonal contrast tests are each associated with a single degree of freedom, and are seeking to detect simpler effects: the parameter spaces are each one dimensional.

2.5 RELATIONSHIPS OF KRUSKAL-WALLIS, FRIEDMAN AND DURBIN TESTS WITH ANOVA F TESTS

In this section we look at the test statistics KW, S and D and their relationships with ANOVA F test statistics for their design. These relationships lead to improved approximations to the sampling distributions of these test statistics.

2.5.1 THE COMPLETELY RANDOMISED DESIGN

Suppose, as in section 2.2, we have distinct observations, y_{ij}, in the completely randomized design. Write $T_{i\bullet} = \sum_j y_{ij}$ and $T_{\bullet\bullet} = \sum_i T_{i\bullet} = \sum_{i,j} y_{ij}$. The ANOVA F test statistic F for this design is given by

$$F = \frac{\left\{\sum_{i=1}^{t} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{n}\right\}/(t-1)}{\left\{\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{t} \frac{T_{i\bullet}^2}{n_i}\right\}/(n-t)}.$$

If the observations are the ranks then $T_{i\bullet} = R_i$ and if also there are no ties then $T_{\bullet\bullet} = 1 + \ldots + n = n(n+1)/2$ and $\sum_{i,j} Y_{ij}^2 = 1^2 + \ldots + n^2 = n(n+1)(2n+1)/6$. Moreover, reorganising

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{t} \frac{R_i^2}{n_i} - 3(n+1)$$

gives $\sum_i R_i^2 / n_i = \{KW + 3(n+1)\}n(n+1)/12$. Substituting into F gives

$$F = \left\{ \frac{\{KW + 3(n+1)\}\frac{n(n+1)}{12} - \frac{n(n+1)^2}{4}}{\frac{n(n+1)(2n+1)}{6} - \{KW + 3(n+1)\}\frac{n(n+1)}{12}} \right\} \left\{ \frac{(n-t)}{(t-1)} \right\}$$

$$= \frac{KW(n-t)}{(n-1-KW)(t-1)}$$

after simplification. Note that if there are ties this relationship doesn't hold. The usual procedure for a group of tied observations is to assign to each the mean of the ranks they would otherwise be given. Then $\sum_i R_{i\bullet} = n(n+1)/2$ as before, but no longer is $\sum_{i,j} Y_{ij}^2 = 1^2 + \ldots + n^2 = n(n+1)(2n+1)/6$. It is not possible to treat ties in such a way that both the sum of the ranks and the sum of the squares of the ranks are constants known before sighting the data.

While not particularly good, the asymptotic chi-squared approximation to the distribution of KW is generally superior to the $F_{t-1,n-t}$ approximation to the distribution of F. However a simple adjustment retrieves the situation. If F is referred to the F distribution with $d(t-1)$ and $d(n-t)$ degrees of freedom where

$$d = 1 - 6(n+1)/\{(n-1)(5n+6)\},$$

then this approximation is generally superior to the chi-squared approximation to the distribution of KW. Of course these degrees of freedom are no longer integers, but this can be easily handled by most modern computer packages. See, for example, Spurrier (2003).

Tomato Example. In section 2.2 we found that for the tomato data the Kruskal-Wallis test statistic takes the value 1.9772 with a χ_3^2 p-value of 0.5772. Using R gives a p-value of 0.577, using its own ties correction. Using the relationship above, the F test statistic F takes the value 0.6518. The $F_{3,92}$ p-value is 0.5838; using the improved approximation the $F_{2.9622,90.8403}$ p-value is 0.5820. For these data, the p-values are very similar.

2.5.2 THE RANDOMISED BLOCK DESIGN

Suppose, as in section 2.3, that in the completely randomized design we have distinct observations, y_{ij}. The observations are ranked within each block and R_i, the sum of the ranks for treatment i over all blocks is calculated for $i = 1, \ldots, t$. The Friedman test statistic is

$$S = \frac{12}{bt(t+1)} \sum_{i=1}^{t} R_i^2 - 3b(t+1).$$

The notation is as before, but additionally $T_{\bullet j} = \sum_i y_{ij}$. The ANOVA F test statistic F for this design is given by

$$F = \frac{\left\{ \sum_{i=1}^{t} \frac{T_{i\bullet}^2}{b} - \frac{T_{\bullet\bullet}^2}{bt} \right\} / (t-1)}{\left\{ \sum_{i=1}^{t} \sum_{j=1}^{b} Y_{ij}^2 - \sum_{i=1}^{t} \frac{T_{i\bullet}^2}{b} - \sum_{j=1}^{b} \frac{T_{\bullet j}^2}{t} + \frac{T_{\bullet\bullet}^2}{bt} \right\} / (b-1)(t-1)}.$$

If the observations are ranks and if there are no ties then

$$T_{i\bullet} = R_i,\ T_{\bullet j} = 1 + \ldots + t = t(t+1)/2,\ T_{\bullet\bullet} = bt(t+1)/2 \text{ and}$$
$$\sum_{i,j} Y_{ij}^2 = b(1^2 + \ldots + t^2) = bt(t+1)(2t+1)/6.$$

Moreover, reorganising the equation for S gives $\sum_i R_i^2 = bt(t+1)\{S + 3b(t+1)\}/12$. Substituting and simplifying gives

$$F = \frac{S(b-1)}{b(t-1) - S}.$$

A slightly less algebraic approach is to note that with no ties every rank appears on each block, so there is no block effect, and the block sum of squares is zero. The treatment mean square is as in the numerator of the first equation for F above and the error mean square is the total sum of squares minus the treatment sum of squares.

It cannot be assumed that the ranks are normally distributed, which is one of the assumptions needed for F to have an F distribution. Thus only approximately does the ANOVA F statistic have the $F_{t-1,(b-1)(t-1)}$ distribution.

Lemonade Example. In the Lemonade example in section 2.3 we found $S = 9.04$ with a χ_4^2 p-value of 0.0601. The F test statistic using the formula above takes the value 2.6279 with $F_{4,36}$ p-value 0.0504.

2.5.3 THE BALANCED INCOMPLETE BLOCK DESIGN

Suppose, as in section 2.4 above, that in the balanced incomplete block design we have distinct observations, y_{ij}. The observations are ranked within each block and R_i, the sum of the ranks for treatment i over all blocks is calculated for $i = 1, \ldots, n$. The Durbin test statistic is

$$D = \frac{12(t-1)}{rt(k^2-1)} \sum_{i=1}^{t} \{R_i - \frac{r(k+1)}{2}\}^2 .$$

Aside. There are different possible ANOVA analyses for balanced incomplete block designs. See, for example, Kuehl (2000, section 8.5); note that a balanced incomplete block design can be regarded as a block design with missing data. Such designs are not *orthogonal* and so it is possible to calculate treatment sums of squares both adjusted and not adjusted for blocks. We use the former.

Here we assume there are no ties and so the ranks assigned on each block are $1, 2, \ldots, k$. There are no block effects, and the block sum of squares, no matter how it is calculated, is zero.

We now show that the ANOVA F test statistic F when the data are the untied ranks is related to D by

$$F = \frac{D}{\{b(k-1)-D\}} \frac{(bk-b-t+1)}{(t-1)}.$$

The total sum of squares is $SS_{\text{Tot}} = \sum_{i,j} y_{ij}^2 - y_{\bullet\bullet}^2 / (bk)$. As above, the ranks assigned on each block are $1, 2, \ldots, k$, so

$$y_{\bullet\bullet} = b(1 + \ldots + k) = bk(k+1)/2 \text{ and}$$
$$\sum_{i,j} y_{ij}^2 = b(1^2 + \ldots + k^2) = bk(k+1)(2k+1)/6$$

giving $SS_{\text{Tot}} = bk(k+1)(k-1)/12$ after simplifying.

Moreover the treatment sum of squares, no matter how it is calculated, is

$$SS_{\text{Treat}} = \frac{k}{\lambda t} \sum_{i=1}^{t} \{y_{i\bullet} - \text{block average across blocks that contain treatment } i\}^2.$$

Here the $y_{i\bullet}$ are rank sums. For any given i, there are r blocks containing treatment i and each block total is $k(k+1)/2$, so the block average across blocks that contain treatment i is $r(k+1)/2$. It follows that

$$SS_{\text{Treat}} = \frac{k}{\lambda t} \sum_{i=1}^{t} \{R_i - \frac{r(k+1)}{2}\}^2 = \frac{k(k+1)}{12} D.$$

This uses $\sum_i \{R_i - r(k+1)/2\}^2 = rt(k^2-1)/\{12(t-1)\}$ from the equation for D and both $bk = rt$, and $\lambda(t-1) = r(k-1)$.

The error sum of squares is

$$SS_{\text{Error}} = SS_{\text{Tot}} - SS_{\text{Treat}} = bk(k+1)(k-1)/12 - k(k+1)D/12$$
$$= k(k+1)\{b(k-1) - D\}/12$$

on simplifying. Substitution gives the relationship between F and D given above.
The randomised block design is recovered if blocks are 'complete'; that is, $k = t$ and $r = b$.

As the observations here are ranks and not normally distributed, only approximately does the ANOVA F statistic have the distribution $F_{t-1,bk-b-t+1}$.

As with the randomised block design the approximate F distribution of F generally improves on the asymptotic χ^2 distribution of D.

Ice Cream Example. In section 2.4 we found the Durbin test statistic took the value 12 with χ^2_6 p-value 0.0620. Using the formula above the *F* test statistic takes the value 8 with $F_{6,8}$ p-value 0.0049. From not quite significant at the 0.05 level using the χ^2_6 approximation the F approximation results in significance at the 0.01 level. Which is more precise? We will return to this question in the next chapter.

2.6 ORTHOGONAL CONTRASTS: PAGE AND UMBRELLA TESTS

We now show how to decompose the test statistics *KW*, *S* and *D* into component statistics called *orthogonal contrasts*. These contrasts give *focused* tests for particular aspects of the hypotheses under consideration. They assume there is a meaningful ordering of the treatments being compared. The first component is usually described as a test for linear trend, while the second is usually described as a test for a quadratic effect. The effects assess whether, as the ordered treatments pass from first to last, polynomial effects of order one, two etc. are observed in the responses. The component tests typically are associated with a single degree of freedom, and seek alternatives in a one dimensional parameter space. Test statistics such as *KW*, *S* and *D* are more *omnibus*, seeking alternatives in a higher dimensional (typically $t-1$) parameter space. The focused tests have higher power than the corresponding omnibus tests when the alternative falls within their one dimensional parameter space. But outside that space they are totally insensitive, with power close to the test size.

Aside. The test size is the probability of rejecting the null hypothesis when the null hypothesis is true. It is usually close to the nominated significance level.

The following treatment requires linear algebra that many readers will not have met yet. These readers should focus on the results rather than the mathematical details.

We begin by showing that *KW*, *S* and *D* may be represented in a similar way. Tedious but routine algebra shows that the Kruskal-Wallis *KW* statistic is given by

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{t} \frac{R_i^2}{n_i} - 3(n+1) = \frac{12}{n(n+1)} \sum_{i=1}^{t} \{\frac{R_i}{\sqrt{n_i}} - \frac{n+1}{2}\sqrt{n_i}\}^2$$
$$= \sum_{i=1}^{t} (Y_i - E[Y_i])^2 / \mathrm{var}(Y_i)$$

in which $Y_i = R_i$, the sum of the ranks for treatment i, $E[Y_i] = (n+1)n_i/2$ and $\mathrm{var}(Y_i) = n_i\, n(n+1)/12$.

Similarly it may be shown that the Friedman test statistic S is given by

$$S = \frac{12}{bt(t+1)} \sum_{i=1}^{t} R_i^2 - 3b(t+1) = \frac{12}{bt(t+1)} \sum_{i=1}^{t} \{R_i - \frac{b(t+1)}{2}\}^2$$
$$= \sum_{i=1}^{t} (Y_i - E[Y_i])^2 / \mathrm{var}(Y_i)$$

in which $Y_i = R_i$, $\mathrm{E}[Y_i] = b(t+1)/2$ and $\mathrm{var}(Y_i) = bt(t+1)/12$. Note that this uses $\sum_i R_i = b(1 + \ldots + t) = bt(t+1)/2$.

Finally the Durbin test statistic D is given by

$$D = \frac{12(t-1)}{rt(k^2-1)} \sum_{i=1}^{t} R_i^2 - \frac{3r(t-1)(k+1)}{(k-1)} = \frac{12(t-1)}{rt(k^2-1)} \sum_{i=1}^{t} \{R_i - \frac{r(k+1)}{2}\}^2 =$$
$$\sum_{i=1}^{t} (Y_i - E[Y_i])^2 / \mathrm{var}(Y_i)$$

in which $Y_i = R_i$, $\mathrm{E}[R_i] = r(k+1)/2$ and $\mathrm{var}(Y_i) = rt(k^2-1)/\{12(t-1)\}$.

Thus if $V_i = (Y_i - \mathrm{E}[Y_i])/\sqrt{\mathrm{var}(Y_i)}$ and $V = (V_i)$ then with $T = KW$, S or D,

$$T = \sum_{i=1}^{t} V_i^2 = V^{\mathrm{T}}V.$$

It is well known that in all three cases asymptotically T has the χ^2_{t-1} distribution. To give this a little substance note that for each i, since Y_i is a rank sum, by the central limit theorem it is asymptotically normal. Thus V is multivariate normal. Clearly $\mathrm{E}[V] = 0$ and $\mathrm{var}\,(V_i) = 1$. However there are linear constraints on the V_i: $V_t = 0$ for $T = S$ and D, while for KW, $\sum_{i=1}^{t}\sqrt{n_i}\{R_i/\sqrt{n_i} - \sqrt{n_i}(n+1)/2\} = 0$. Thus the V_i are correlated and $\Sigma = \mathrm{cov}(V)$ is of rank at most $t - 1$.

Aside. Write $\Sigma = \mathrm{cov}(V)$. Since Σ is positive semi-definite there exists a matrix $\Sigma^{0.5}$ such that $\Sigma^{0.5}\Sigma^{0.5} = \Sigma$. Define Z by $V = \Sigma^{0.5}\ Z$ in which the elements Z_i of Z are asymptotically independent and standard normal, written IN(0, 1). Then $T = V^{\mathrm{T}}V = Z^{\mathrm{T}}\Sigma Z$. By a well-known theorem on quadratic forms $Z^{\mathrm{T}}\Sigma Z$ asymptotically has the χ^2_{t-1} distribution if and only if Σ is idempotent of rank $t - 1$. We will not pursue that further here.

Suppose now that 1_t denotes the $t \times 1$ column of units as opposed to I_t that denotes a $t \times t$ identity matrix. Suppose that H is an orthogonal $t \times t$ matrix with last row $1_t^{\mathrm{T}}/\sqrt{t}$ and $Z = \mathrm{H}V$. Then

$$T = V^{\mathrm{T}}V = V^{\mathrm{T}}H^{\mathrm{T}}HV = Z^{\mathrm{T}}Z \text{ since } H^{\mathrm{T}}H = \mathrm{I}_t.$$

Thus $T = Z_1^2 + \ldots + Z_t^2$. If the ith row of H is denoted by h_i^{T}, the $Z_i = h_i^{\mathrm{T}}V$ are our so-called orthogonal contrasts; orthogonal because the rows of H are orthogonal and contrasts because the choice of last row means the elements of every other row sum to zero, using the orthogonality $h_i^{\mathrm{T}}1_t = 0$.

The rows of H may be based on the orthonormal polynomials. The completely randomised design with equal treatment numbers, the randomised block and the balanced incomplete block designs are all *balanced* in the sense that the n_i are all equal. For unbalanced designs the orthonormal polynomials will vary with $\{n_i\}$, but for balanced designs the orthonormal polynomials can be given explicitly. Thus the order one polynomial and the first row of H will consist of 1, 2, …, t with the mean $(t+1)/2$ subtracted from each element and then each of the resulting numbers divided by the square root of the sum of their squares, which can readily be shown to be $t(t^2-1)/12$. Thus in Table 2.4 corresponding to the linear coefficients for $t = 5$ we start with 1, 2, 3, 4 and 5, subtract off their mean, 3, giving –2, –1, 0, 1, 2 and then divide by the square root of their sum of squares, $\sqrt{10}$. This gives the row $-2/\sqrt{10}, -1/\sqrt{10}, 0, 1/\sqrt{10}, 2/\sqrt{10}$. Similarly for the quadratic coefficients start with $1^2, 2^2, \ldots, t^2$, subtract their mean and divide by the square root of their sum of squares. For $t = 3, 4, \ldots, 7$ the elements of h_1 and h_2, that give the linear and quadratic contrasts, are given in Table 2.4.

For the randomised block design the test statistic Z_1 is the Page test statistic, while for the completely randomised and balanced incomplete blocks designs Z_1 could be called Page-type test statistics. For the randomised block and balanced incomplete block designs Z_1 is of the form $\sum_i h_{1i}\{R_i - \mu\}/\sigma = \sum_i h_{1i}R_i/\sigma$ since $\sum_i h_{1i} = 0$ using the orthogonality of the first and last rows of H. Thus the Page and Page-type test statistics are specifically,

- $\sum_i h_{1i}\{R_i - (n+1)n_i/2\}/\sqrt{n(n+1)n_i/12}$ for the completely randomised design
- $\sum_i h_{1i}R_i/\sqrt{bt(t+1)/12}$ for the randomised block design
- $\sum_i h_{1i}R_i/\sqrt{bk(k^2-1)/[12(t-1)]}$ for the balanced incomplete block design.

a) Linear Coefficients

t	$h_{11}, h_{12}, \ldots, h_{1t}$
3	$-1/\sqrt{2}, 0, 1/\sqrt{2},$
4	$-3/\sqrt{20}, -1/\sqrt{20}, 1/\sqrt{20}, 3/\sqrt{20}$
5	$-2/\sqrt{10}, -1/\sqrt{10}, 0, 1/\sqrt{10}, 2/\sqrt{10}$
6	$-5/\sqrt{70}, -3/\sqrt{70}, -1/\sqrt{70}, 1/\sqrt{70}, 3/\sqrt{70}, 5/\sqrt{70}$
7	$-3/\sqrt{28}, -2/\sqrt{28}, -1/\sqrt{28}, 0, 1/\sqrt{28}, 2/\sqrt{28}, 3/\sqrt{28}$

b) *Quadratic Coefficients*

t	$h_{21}, h_{22}, \ldots, h_{2t}$
3	$1/\sqrt{6}, -2/\sqrt{6}, 1/\sqrt{6},$
4	$1/\sqrt{4}, -1/\sqrt{4}, -1/\sqrt{4}, 1/\sqrt{4},$
5	$2/\sqrt{14}, -1/\sqrt{14}, -2/\sqrt{14}, -1/\sqrt{14}, 2/\sqrt{14},$
6	$5/\sqrt{84}, -1/\sqrt{84}, -4/\sqrt{84}, -4/\sqrt{84}, -1/\sqrt{84}, 5/\sqrt{84}$
7	$5/\sqrt{84}, 0, -3/\sqrt{84}, -4/\sqrt{84}, -3/\sqrt{84}, 0, 5/\sqrt{84}$

TABLE 2.4. Linear and Quadratic Coefficients for balanced designs

The test statistics Z_2 are umbrella test statistics. Obviously they take the same form as the Z_1 statistics, but with h_{1i} replaced by h_{2i}.

Although it is possible to calculate and base inference upon all of the Z_i it is more usual to aggregate $Z_3, \ldots, Z_t$ into a residual, $Z_3^2 + \ldots + Z_t^2 = T - Z_1^2 - Z_2^2$.

The asymptotic distributions of the contrasts are standard normal. This follows routinely because the R_i are rank sums and being sums, are asymptotically normal by the central limit theorem. They are clearly standardised, and so each has mean zero and variance one. Thus for all three designs both Z_1 and Z_2 are asymptotically standard normally distributed.

For the randomised block and balanced incomplete blocks designs $Z_t = 0$. Note that for $T = S$ or D, Z_t is proportional to $\sum_i (R_i - E[R_i])$ which is zero because $\sum_i R_i$ is the sum of all of the ranks assigned, namely $bt(t + 1)/2$ for randomised block design, and this is just $\sum_i E[R_i]$. The reasoning for balanced incomplete blocks is similar. The same argument does not apply to the completely randomised design because of the factors $\sqrt{n_i}$.

The discussion above does not prove that asymptotically T has the χ^2_{t-1} distribution, nor that the orthogonal contrasts are asymptotically independent. To prove these results requires relatively advanced distribution theory and linear algebra that we will not pursue here.

Tomato Example. For the balanced completely randomised design if $n_i = m$ for all i then

- $Z_1 = \sqrt{12/\{n(n+1)m\}} \sum_i h_{1i} R_i$.

The rank sums for the four treatments are 1058, 1200, 1303.5 and 1094.5. For these data the Page-type test statistic takes the value – 0.3490 with a two-sided p-value using the normal distribution of 0.7271. The umbrella test statistic Z_2 takes the value – 1.2860 with two-sided p-value 0.1984. There is no statistical evidence of linear or quadratic effects at all reasonable levels.

Again, no adjustment for ties has been made as it makes little difference, when, as here, there are few ties.

Lemonade Example. For this example using the randomised block design the Page test statistic takes the value – 0.3162 with a two-sided p-value using the normal distribution of 0.7518. The umbrella test statistic takes the value 2.2984 with a two-sided p-value using the normal distribution of 0.0215. There is evidence of an umbrella effect at the 0.05 level but no evidence of a linear trend at the same level. The rank sums for treatments 1 to 5 are 36, 32, 17, 31 and 34. A by-eye inspection of the data supports the conclusions: there is no evidence of linear trend but the rank sums clearly decrease then increase – an umbrella effect.

Ice Cream Example. For these data using the balanced incomplete block design the Page-type test statistic takes the value – 0.9897 with a two-sided p-value using the normal distribution of 0.3223. The umbrella test statistic takes the value 2.5714 with a two-sided p-value using the normal distribution of 0.0101. There is evidence of an umbrella effect at the 0.05 level but no evidence of a linear trend at the 0.05 level.

The rank sums for treatments 1 to 7 are 8, 9, 4, 3, 5, 6 and 7. Again a by-eye inspection of these rank sums supports the conclusions: there is no evidence of linear trend but the rank sums tend to decrease then increase – an umbrella effect.

3 PERMUTATION TESTING

Learning Objectives

After successful completion of the material in this chapter the student will be able to

- explain to peers the concept of permutation testing of statistical hypotheses
- implement permutation tests for designs such as the completely randomised, randomised block and balanced incomplete block designs.

3.1 WHAT IS PERMUTATION TESTING AND WHY IT IS IMPORTANT?

In Chapters 1 and 2 we usually found approximate p-values using the asymptotic distributions of the test statistics. Only for some test statistics and for small samples is it possible to find exact p-values. However it is possible to find almost exact p-values using permutation tests, based on random samples of permutations of the data. To illustrate how to carry out these permutation tests below we consider herbicide data from Higgins (2004, pp. 38–39). Here and elsewhere we use our own R code for permutation tests; we will use the almost exact p-values based on a random sample of all possible permutations. Random samples are used because for larger data sets exact p-values may take too long to calculate.

Herbicide Example. A study was conducted to assess the damage to strawberry plants caused by a particular type of herbicide used for controlling weeds. The dry weight of nine plants treated with the herbicide was compared with the dry weights of seven untreated plants. It is expected that the untreated plants will have larger dry weights than treated plants. Data and ranks are given in Table 3.1. We use the test statistic the rank sum of the untreated plants. Since the ranks of the untreated plants are expected to be larger than those of the treated plants, an upper-tailed test is used.

Rank	1	2	3	4	5	6	7	8	9
Untreated plants					0.55				0.63
Treated plants	0.44	0.47	0.51	0.52		0.58	0.59	0.60	

Rank	10	11	12	13	14	15	16
Untreated plants			0.67	0.68	0.79	0.81	0.85
Treated plants	0.65	0.66					

TABLE 3.1. Dry weights in kg of strawberry plants

The raw data and the combined, ordered and ranked data are in the Table 3.1. The rank sum of the untreated plants is 5 + 9 + 12 + 13 + 14 + 15 + 16 = 84; the rank sum of the treated plants is 1 + 2 + 3 + 4 + 6 + 7 + 8 + 10 + 11 = 52.

If there is no difference between the treated and untreated strawberry plants then all data sets obtained by randomly assigning seven of the 16 ranks to the untreated plants and nine to the treated plants would have an equal chance of being observed. In this case the average of the ranks for the treated and untreated plants would be similar. If the untreated weights were higher though, then the ranks of these weights should be greater, as rank one is assigned to the smallest weight and rank 16 to the largest. We will carry out a Wilcoxon test with almost exact p-value by seeing if among 10,000 random permutations of the 16 ranks the rank sum of the untreated weights, namely 84, is one of the largest rank sums. Our p-value will be the proportion of the untreated rank sums greater than or equal to 84.

We now give some R code to obtain 10,000 random permutations and so 10,000 rank sums for seven untreated ranks. Note that in R to get a random permutation we use the `sample` command. The other R commands should be obvious. If the R commands are typed into a text editor such as notepad in Windows then they may be copied and pasted into the R console window and pressing enter will give the p-value. Possible R code is

```
d<-c (.55,.67,.63,.79,.81,.85,.68,.65,.59,.44,.6,.47,.58, .66,.52,.51)
r<-rank(d)
nperm<-10000
teststat.obs<-sum(r[c(1:7)])
print("test statistic")
teststat.obs
teststat<-rep(NA,nperm)
for (i in 1:nperm){
rankperm=sample(r)
rp=rankperm[c(1:7)]
teststat[i]=sum(rp)
}
print("p-value")
sum(teststat>=teststat.obs)/nperm
```

Press enter after pasting this code into the R console window and get a p-value similar to 0.0034. This is close to the exact p-value of 0.0039 found using the R routine `wilcox.test`. Repeated running of this code will produce a cluster of p-values around 0.0039. The number of permutations used is `nperm` in the R code. If that is changed to 100,000 a value similar to 0.004 is obtained almost instantly. A wait of a few seconds is required with `nperm` set to 1,000,000, which returns a value similar to 0.00392. Experiment by running the code with different values of `nperm`.

Note. Alternative R code and a fuller discussion is given in Rippon (2016).

The extension of permutation tests for comparing two treatments to studies with k treatments follows in similar fashion. Suppose the number of observations or ranks in the ith treatment group is n_i with the total number of observations being $\sum_{i=1}^{k} n_i = n$. Form a vector of size n containing all the observations and generate random permutations of the elements of this vector. For each of these random permutations take the first n_1 elements to belong to the first treatment group, the second n_2 elements to belong to the second treatment group and so on, with the last n_k elements belonging to the kth treatment group. For each random permutation of the observations calculate the Kruskal-Wallis test statistic and see what proportion of these statistics are greater than or equal to the Kruskal-Wallis test statistic for the original observations. This is the 'almost exact' p-value. An exact p-value requires all possible $n!/(n_1!\, n_2! \ldots n_k!)$ permutations to be considered but for larger n this becomes time consuming. In general it is preferable to calculate almost exact p-values with user defined number of permutations. Clearly increasing the number of permutations improves the precision of the p-value estimated while increasing the time taken to calculate it.

The permutation test for the Friedman randomised blocks test for the comparison of t treatments is calculated in a slightly different manner. There are t ranks in each of the b blocks. For each of the b blocks random permutations of the ranks 1,2, …, t are generated. The Friedman statistic is calculated for each such set of $n = bt$ random permutations and the proportion of these statistics greater than or equal to the Friedman statistic for the original rankings is found. This is the almost exact p-value. An exact p-value requires all $(t!)^b$ permutations to be considered, but again this is too time consuming for a large number of rankings.

We will return to permutation tests in Section 3.3. In Section 3.2 we consider a new topic, nonparametric ANOVA. This appears to be very similar to parametric ANOVA but is based on very weak assumptions rather than assuming normality of the residuals. P-values may be found using F tests, but these are only approximations, albeit very good approximations in general. As the true distribution of the test statistics is not known permutation tests are needed to find almost exact p-values. These are found and presented in the final section, where previous p-values and permutation test p-values are collected and compared.

3.2 NONPARAMETRIC MULTIFACTOR ANOVA WHEN THE LEVELS OF THE FACTORS ARE UNORDERED

Nonparametric ANOVA is a technique applicable to multifactor ANOVA. In this section we will apply the technique to designs in which the levels of all factors are unordered or any ordering is ignored.

Only a brief sketch of the theory underpinning the technique will be given here. For more details see Rayner and Best (2013). A model for a multifactor ANOVA may be constructed using product multinomial distributions. These models essentially label cells in a relevant table rather than impose strong assumptions that may not be true, such as the error distributions all being normally distributed. Then a 1-1 transformation is given that transforms the multinomial cell probabilities to ANOVA-like parameters. We may then test if each ANOVA-like parameter is zero against the negation of this. The usual ANOVA F test statistics can be shown to be appropriate test statistics. Because the residuals are not assumed to be normally distributed these statistics do not have F distributions. However using permutation tests almost exact p-values can be found and these are remarkably similar to p-values obtained from F distributions, even when the residuals are not well approximated by normality. There are other options for test statistics, but they do not perform as well as the F test statistics.

The essence of the technique is to transform the responses using successive order r orthonormal polynomials on the responses. We call the ANOVA F tests on the responses transformed by the order r orthonormal polynomial order r tests, and these tests assess treatment effects in the moments up to order r in the responses. Test statistics of different orders are uncorrelated with each other, and so the significance or not of tests for one order does not influence significance or not of tests for any other order.

It will rarely be useful to consider effects beyond those of order three. Even so, many tests on treatment effects on several levels are being done at the same time, and if testing is done at the 0.05 level then we should expect approximately 5% of them will be significant, even if there are no effects present in the model. This suggests nonparametric ANOVA should be considered to be exploratory data analysis.

To apply the nonparametric ANOVA the first few orthonormal polynomials of a random variable are needed. Here we give the orthonormal polynomials of a random variable X up to order three. Write μ for the mean of X and $\mu_r, r = 2, 3, \ldots$ for the central moments of X. Ambiguity is avoided by setting $a_0(x) = 1$ for all x. Then

$$a_1(x) = (x - \mu)/\sqrt{\mu_2},$$

$$a_2(x) = \{(x - \mu)^2 - \mu_3(x - \mu)/\mu_2 - \mu_2\}/\sqrt{d}$$

in which $d = \mu_4 - \mu_3^2 / \mu_2 - \mu_2^2$, and

$$a_3(x) = \{(x - \mu)^3 - a(x - \mu)^2 - b(x - \mu) - c\}/\sqrt{e},$$

in which $a = (\mu_5 - \mu_3\mu_4 / \mu_2 - \mu_2\mu_3)/d$,

$$b = (\mu_4^2 / \mu_2 - \mu_2\mu_4 - \mu_3\mu_5 / \mu_2 + \mu_3^2)/d,$$

$$c = (2\mu_3\mu_4 - \mu_3^3 / \mu_2 - \mu_2\mu_5)/d, \text{ and}$$

$$e = \mu_6 - 2a\mu_5 + (a^2 - 2b)\,\mu_4 + 2(ab - c)\,\mu_3 + (b^2 + 2ac)\,\mu_2 + c^2$$

Should further orthonormal polynomials be required the recurrence relation in Rayner et al. (2008) can be used.

Given a data set $\{y_{ij}\}$ say, the unordered nonparametric multifactor ANOVA applies the intended ANOVA to $\{a_r(y_{ij})\}$ for $r = 1, 2, \ldots k$, say, for some predetermined k: usually 3. For each specified r the ANOVA tests whether or not the $E[a_r(Y_{ij})]$ are consistent across the levels of the factors. Thus the ANOVA applied to $E[a_1(Y_{ij})]$ tests whether or not the $E[a_1(Y_{ij})] = (E[Y_{ij}] - \mu)/\sigma$ are consistent across the levels of the factors. Here the moments defining the orthonormal polynomials refer to the distribution of the responses. Since ANOVA is location-scale invariant the first order analysis is equivalent to testing whether or not the $E[Y_{ij}]$ are consistent across the levels of the factors; this is the traditional ANOVA null hypothesis.

The second order ANOVA that tests whether or not the $E[a_2(Y_{ij})] = E[(Y_{ij} - \mu)^2 - \mu_3(Y_{ij} - \mu)/\mu_2 - \mu_2]$ are consistent across the levels of the factors. If the first order null hypothesis is accepted, $E[Y_{ij}] = \mu$ and $E[a_2(Y_{ij})] = \mathrm{var}(Y_{ij}) - \mu_2$, and the second order null hypothesis is testing whether or not the $\mathrm{var}(Y_{ij})$ are consistent across the levels of the factors. If the first order null hypothesis is not accepted the test concerns a measure of dispersion, generally not the variance, and assesses whether or not this measure is consistent across levels. The same argument applies to the higher order tests.

Tomato Example. To apply the suggested analysis the orthonormal polynomials are required, and these in turn require the moments of the responses. For n responses each response is given probability $1/n$. For example here the responses 5, 6 and 10 each have probability 1/96 as do *each* of the responses 12. To construct the orthonormal polynomials of order up to three requires moments up to order six. These are

$$\mu = 62.09375,\ \mu_2 = 1205.24375,\ \mu_3 = 9539.0866089,$$
$$\mu_4 = 2747648.0196,\ \mu_5 = 51562852.798 \text{ and } \mu_6 = 8071672926.3.$$

Note. If the sample variance is used in place of the population variance the result here is 1192.689 and the subsequent analysis is identical.

Response	Tomato variety	First order orthonormal polynomial	Second order orthonormal polynomial	Third order orthonormal polynomial
5	1	-1.6532	2.2583	-2.7114
6	4	-1.6242	2.1498	-2.4677
10	1	-1.5084	1.7339	-1.5893
12	2	-1.4505	1.5367	-1.2059
12	2	-1.4505	1.5367	-1.2059
13	3	-1.4216	1.4408	-1.0275
13	4	-1.4216	1.4408	-1.0275
15	4	-1.3636	1.2544	-0.6968

TABLE 3.2. Eight smallest transformed responses of orders one, two and three

The values taken by the eight smallest responses transformed by the polynomials of order one, two and three are given in the Table 3.2. These calculations will enable users to check their calculations should they prefer to program in a language other than R.

A summary of the analysis is given in the following Table 3.3. The ANOVA F-tests applied to the responses and to the responses transformed by the first order orthonormal polynomial are, as discussed above, identical.

	First order	Second order	Third order
Tomato F test statistic	0.7986	1.4708	0.5025
Tomato F test p-value	0.4978	0.2277	0.6815
Shapiro-Wilk Normality test p-value	0.0052	< 0.0001	0.0014

TABLE 3.3. Summary of the nonparametric unordered analysis for the tomato data

The ANOVA F-tests applied to the responses transformed by the first, second and third order orthonormal polynomials all produced large p-values, giving non-significant results at all reasonable levels. In line with the discussion above, since first order tomato effects are consistent across varieties, the test for second order effects is, in fact, a test for consistency of variety variances. Since first and second order effects are consistent across varieties, the test for third order effects is, in fact, a test for equality of third order moments across varieties. So the tomato varieties have similar first, second and third order moments across varieties. The first order p-value here is a little smaller than the Kruskal-Wallis p-value reported in Section 5.1 of Chapter 2. However the Kruskal-Wallis test is based on the ranked responses; the analysis here is based on the raw data, the unranked responses.

A check on the normality of the residuals for the three ANOVAs using the Shapiro-Wilk test of normality gave small p-values in all cases. Since the normality assumption underpinning the F-tests is in doubt it is desirable to check the p-values by calculating permutation test p-values. This will be done in the next section.

The tests applied here are unable to find any differences between the tomato varieties.

Lemonade Example. The nonparametric ANOVA ignoring order applies the randomised blocks ANOVA to the data transformed by the orthonormal polynomials of orders one, two and three. The p-values are summarised in the Table 3.4.

p-value	First order	Second order	Third order
Lemonade F test	0.0504	0.9232	0.0047
Shapiro-Wilk Normality test	0.0211	< 0.0001	0.0125

TABLE 3.4. Summary of the nonparametric unordered analysis for the lemonade data

Although normality is dubious it seems there are first and third order effects – roughly indicating a mean effect and an effect due to moments up to third order. The p-value here for the first order transformation agrees with that given in Section 5.2 of Chapter 2.

The polynomial means for the lemonade varieties are given in Table 3.5. The first order effects are almost significant at the 0.05 level. In section 2.6 the orthogonal components of the first order test statistic identified an umbrella effect, so this is the cause of the borderline first order significance. The umbrella effect is apparent in Table 3.5. The second order mean differences are just natural variation: they are not significant at all reasonable levels. The third order effect is significant at the 0.01 level. It is not interesting to decompose this effect into linear, quadratic etc. components, although from 'eyeballing' the third order means in Table 3.5 there is a possible umbrella effect. However it is not clear that such an effect is useful in interpreting the data.

Lemonade variety	1	2	3	4	5
First order mean	0.424	0.141	-0.919	0.071	0.283
Second order mean	-0.120	0.000	-0.060	0.299	-0.120
Third order mean	-0.141	0.424	0.778	-0.141	-0.919

TABLE 3.5. Lemonade polynomial means

Ice Cream Example. Using the usual parametric F test the responses are significant at the 0.01 level with a p-value of 0.0049. This is as reported in Section 5.3 of Chapter 2. Using the Shapiro-Wilk test for normality yields a p-value of 0.4445, so the residuals are consistent with normality. The second order orthonormal polynomial responses are not significant with a p-value of 0.6118 and a Shapiro-Wilk normality test p-value of 0.5992. There can be no third order nonparametric ANOVA analysis as there are only three responses, and so only two orthonormal polynomials can be constructed.

It seems there are first order, that is, mean effects, but no order two effects. The polynomial means of first and second order are given in Table 3.6. There is no pattern apparent in the second order means, the differences being natural variation. However the first order means increase, and although the first order effect is of standardised ranks, this reflects the significant linear trend discussed in section 2.6.

Ice Cream	1	2	3	4	5	6
First order mean	-1.012	-0.559	-0.106	0.348	0.529	0.801
Second order mean	0.265	0.051	-0.564	-0.578	0.618	0.209

TABLE 3.6. Ice cream polynomial means

3.3 REVISITING SOME PREVIOUS EXAMPLES

In this section we collect the treatment p-values based on distribution theory for our three main examples, the tomato, lemonade and ice cream examples, and compare them with permutation test almost exact p-values. Note that the permutation test p-values will be different each time they are calculated, but they will cluster around the true value, since they are based on the true distribution of the test statistic. P-values based on χ^2 and F distributions are only approximations, relying on the assumption of normality. In some cases this assumption is dubious.

In particular note that in Tables 3.7, 3.8 and 3.9 the p-values for the ANOVA F and first order NP ANOVA are the same. The corresponding permutation test p-values vary slightly because different permutations are being generated.

Tomato Example.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{3,92}$	0.4978	0.4966
Kruskal-Wallis	χ^2_3	0.5772	0.5824
	$F_{3,92}$	0.5838	-
	$F_{2.9622,90.8403}$	0.5820	-
Page-type	N(0, 1)	0.7271	0.7284
Umbrella	N(0, 1)	0.1984	0.1994
Cubic	N(0, 1)	0.6535	0.6552
First order NP ANOVA	$F_{3,92}$	0.4978	0.4968
Second order NP ANOVA	$F_{3,92}$	0.2277	0.2266
Third order NP ANOVA	$F_{3,92}$	0.6815	0.6833

TABLE 3.7. Summary of the analyses of the Tomato data

The permutation test p-values for nonparametric ANOVA use method 1 suggested in Manly (2007, p. 145). This freely randomises observations and uses F statistics as opposed to restricted randomisation and/or using mean squares or other test statistics.

The permutation test p-values are very similar to the p-values based on the nominated asymptotic and approximate distributions, confirming the validity of these tests.

Lemonade Example.

The permutation test p-values here and in the next example involve permuting within blocks. Again the permutation test p-values are very similar to the p-values based on the nominated asymptotic and approximate distributions, confirming the validity of these tests.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{4,36}$	0.0504	0.0531
Friedman	χ^2_4	0.0601	0.0554
	$F_{4,36}$	0.0504	-
Page	N(0, 1)	0.7518	0.7771
Umbrella	N(0, 1)	0.0215	0.0204
First order NP ANOVA	$F_{4,36}$	0.0504	0.0523
Second order NP ANOVA	$F_{4,36}$	0.9232	0.9175
Third order NP ANOVA	$F_{4,36}$	0.0047	0.0061

TABLE 3.8. Summary of the analyses of the Lemonade data

Ice Cream Example.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{6,8}$	0.0049	0.0031
Durbin	χ^2_6	0.0620	0.0181
	$F_{6,8}$	0.0049	-
Page	N(0, 1)	0.3223	0.3208
Umbrella	N(0, 1)	0.0101	0.0068
First order NP ANOVA	$F_{6,8}$	0.0049	0.0081
Second order NP ANOVA	$F_{6,8}$	0.6118	0.6284

TABLE 3.9. Summary of the analyses of the Ice cream data

Note there is no third order NP ANOVA analysis as each judge only assess three ice creams. The permutation test p-value for the Durbin test is similar to that for the $F_{6,8}$ approximation, but is somewhat different to that based on the χ^2_6 distribution. The other permutation test p-values agree well with the nominated asymptotic and approximate distributions.

CONCLUDING REMARKS

Introductory Nonparametrics is intended as a gentle introduction to nonparametric methods. Having worked through this material the reader should have the ability to apply several tests generally acknowledged as 'nonparametric'. The Kruskal-Wallis, Friedman and Durbin tests are important because they arise in experimental designs that are often applied in practice. P-values are often calculated using asymptotic distributions of the test statistics. However it is possible to improve on using these chi-squared distributions by taking certain transformations of the test statistics and using their F distributions. Better still is to use permutation tests. Uncritical use of any recipe is poor science. If a user is not in a position to calculate permutation test p-values then use the F statistics and check by also using the chi-squared statistics; the answers should be similar.

Some of the material here, specifically that on nonparametric multifactor ANOVA, is relatively recent research. It should be a comfort to the reader that statistics is a vibrant science with better methodology constantly emerging.

My intention is to produce follow-up material in *Advanced Nonparametrics*. This will focus on more advanced methods, with substantial content in recent research papers. It will include chapters on correlation and independence, the Cochran-Mantel-Haenszel tests, goodness of fit testing and powerful new methods based on probability index models.

REFERENCES

CONOVER, W.J. (1999). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.

HIGGINS, J.J. (2004). *Introduction to Modern Nonparametric Statistics*. Belmont, CA: Duxbury Press.

KUEHL, R.O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. Pacific Grove, CA: Duxbury Press.

RAYNER, J.C.W. and BEST, D.J. (1989). *Smooth Tests of Goodness of Fit*. New York: Oxford University Press.

RAYNER, J.C.W. and BEST, D.J. (1999). Modelling ties in the sign test. *Biometrics*, 55, 2, 663–666.

RAYNER, J.C.W. and BEST, D.J. (2013). Extended ANOVA and rank transform procedures. *Australian and NZ Journal of Statistics*, 55(3), 305–319.

RAYNER, J.C.W., THAS, O. and De BOECK, B. (2008). A generalised Emerson recurrence relation. *Australian and NZ Journal of Statistics* 50(3), 235–240.

RIPPON, Paul (2016). *An R Companion to Introductory Nonparametrics*.

SPURRIER, J.D. (2003). On the null distribution of the Kruskal-Wallis statistic, *Journal of Nonparametric Statistics*, 15:6, 685–691, DOl: 10.1080/10485250310001634719.

SUBJECT INDEX

EXAMPLES INDEX

EXERCISES

CHAPTER ONE EXERCISES

1. In about 100 words contrast some of the essential features of parametric and nonparametric methods.

   Questions 2 and 3 don't really require the use of R; a standard package such as JMP or SPSS will be sufficient. However the R code in Rippon (2016, Chapter 1) may usefully be modified. This has been done in the solutions.

2. (i) Use the sign test to test the hypothesis that zero is the median of the following 16 observations:
   0.3, 6.3, 3.7, 2.8, 5.8, –1.4, 1.7, 2.3, –1.7, 1.6, –1.8, 0.6, 4.5, 1.9, 2.4, 6.8.
   (ii) An approximate confidence interval for a binomial proportion p based on an observed proportion $\hat{p}$ from a sample of size n is $\hat{p} \pm a_\alpha \sqrt{\hat{p}(1-\hat{p})/n}$. Here a_α is the point that gives probability $\alpha/2$ in each tail of the standard normal distribution. Construct approximate 95% and 99% confidence intervals for the proportion of negative observations.
   (iii) Are your answers in (i) and (ii) consistent? If not, why not?

3. Coded potencies of a series of lots of a pharmaceutical product as measured by two different methods were:

   Method I: 3.3, 2.3, 3.7, 2.8, 2.8, –1.4, 1.7, 2.3,
   Method II: –4.7, 4.6, –1.8, –2.6, 4.5, 3.9, 2.4, 6.8.

   We wish to assess if the two methods can be regarded as being the same.

   (i) Analyse the data both parametrically and nonparametrically. Use any convenient software with which you are familiar. Comment on the output. In particular comment on the assumptions for the pooled t-test.
   (ii) Combine, order and rank the data. This will make it easier to do the tests following.
   (iii) Apply the runs test to the data above to assess if there is a difference in methods. Find the exact probability of differences at least as extreme as the observed, and compare this with the normal approximation. State your conclusion carefully.

(iv) Apply the median test, classifying the data as above and below both
   a) the median (show this is 2.6), and
   b) the lower quartile (show this is 0.15).
   In b) calculate the exact p-value using the extended hypergeometric distribution.

(v) Apply the two-sample Wilcoxon test to these data. In the process of doing this, verify that $W_1 + W_2 = (m + n)(m + n + 1)/2$.

(vi) Summarise your conclusions.

CHAPTER TWO EXERCISES

1. At the CSIRO Food Research Laboratory in the early 1990s, the Japan project was set up to look at, among other things, how Japanese and Australian consumers rated various sweet foods on seven point scales. These scales had the anchors 'Dislike Extremely' and 'Like Extremely'. The scores shown were assigned later to assist with analysis. Note that the comparison is not between products, but between consumers from two cultures.

	Score						
Consumers	1	2	3	4	5	6	7
Australian	2	1	6	1	8	9	6
Japanese	0	1	3	4	15	7	1

Japanese chocolate responses

Analyse these data both parametrically and nonparametrically to see if there are location differences. Your parametric analysis should include the pooled t-test, the Welch test, an assessment of which is preferred by checking the sample variances for consistency, and an assessment of the normality of the residuals. (As there are only two treatments these analyses are essentially an orthogonal contrast, so do not attempt to apply the techniques of section 2.6.) Discuss your results. We will return to these data in a later chapter.

R Help. As there are different numbers of Australian and Japanese consumers a data frame is not appropriate; use the list instead, as in the following:

```
choc1 <- list(Australian=c(1, 1, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5,
5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7),

                Japanese=c(2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5,
5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7))
```

The Wilcoxon test requires ranks rather than scores. As in the following, the rank command may be helpful

```
choc2$r <- rank(choc2$score, ties.method = "average")
```

2. Four experts compared five word processors. The data are the time (in minutes) taken to prepare a report on each machine. The data come from Freund (2004, Exercise 15.30).

	Experts			
Word Processors	**1**	**2**	**3**	**4**
A	49.1 (2)	48.2 (4)	52.3 (4)	57.0 (4)
B	47.5 (1)	40.9 (1)	44.6 (1)	49.5 (1)
C	76.2 (5)	46.8 (3)	50.1 (3)	55.3 (3)
D	50.7 (3)	43.4 (2)	47.0 (2)	52.6 (2)
E	55.8 (4)	48.3 (5)	82.6 (5)	57.8 (5)

This is a randomised blocks design. Analyse the data both parametrically and nonparametrically for location effects between word processors. To do the latter modify the R code in Rippon (2016). Input the ranks within blocks, given in brackets in the table. Remember to load `reshape2`.

3. Rayner et al. (2005, p.93) analyse data from the former Dairy Manufacturing Department at North Carolina State College, USA, for ice creams rated on a six-point scale with one meaning no vanilla flavour and six meaning the highest amount of vanilla flavour. Balanced incomplete block (BIB) designs are used in sensory evaluation due to palate paralysis or sensory fatigue. Previous experience shows that often only four products can reliably be rated at one time by a judge. The data below use the ranks of the original data with ties broken at random.

Judge/Ice cream	A	B	C	D	E	F
1	2	4	1	3		
2	1	3	2		4	
3	1	2	3			4
4	1	2		3	4	
5	1	2		3		4
6	1	2			3	4
7	2		3	1	4	
8	1		2	3		4
9	1		2		4	3
10	2			3	1	4
11		1	3	4	2	
12		1	2	3		4
13		1	2		4	3
14		1		3	4	2
15			4	3	1	2

Vanilla flavour ratings for six ice creams

Analyse the data both parametrically and nonparametrically. For the latter you may assume the ice creams are ordered from A to F and calculate the Page-type and umbrella statistics. Discuss your findings.

R Help. The data entry is a little tedious; use the text file on the book web page.

CHAPTER THREE EXERCISES

1. For the Japan project chocolate data from the first exercise for Chapter Two copy and complete the following table.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{1,62}$		
Kruskal-Wallis	χ_1^2		
Page-type	N(0, 1)		
First order NP ANOVA	$F_{1,62}$		
Second order NP ANOVA	$F_{1,62}$		
Third order NP ANOVA	$F_{1,62}$		

Comment.

2. For the word processors data from the second exercise for Chapter Two copy and complete the following table.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{4,12}$		
Friedman	χ^2_4 $F_{4,12}$		-
Page	N(0, 1)		
Umbrella	N(0, 1)		
Cubic	N(0, 1)		
First order NP ANOVA	$F_{4,12}$		
Second order NP ANOVA	$F_{4,12}$		
Third order NP ANOVA	$F_{4,12}$		

Comment.

3. For the ice cream data from the third exercise for Chapter Two copy and complete the following table.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{5,40}$		
Durbin	χ^2_5		
	$F_{5,40}$		-
Page	N(0, 1)		
Umbrella	N(0, 1)		
Cubic	N(0, 1)		
First order NP ANOVA	$F_{5,40}$		
Second order NP ANOVA	$F_{5,40}$		
Third order NP ANOVA	$F_{5,40}$		

Comment.

SOLUTIONS

SOLUTIONS TO THE CHAPTER ONE EXERCISES

1. Parametric methods are not available for data on the nominal or ordinal measurement scales, only on ratio and interval scales. Nonparametric methods are available on all measurement scales.

   Nonparametric methods make minimal assumptions whereas parametric methods make more assumptions and are more powerful when these assumptions are valid. Nonparametric methods are available when the assumptions needed for parametric methods may not be valid. When the parametric assumptions do not hold nonparametric methods is usually have greater power and efficiency.

   Often parametric methods are about distributions (are these data consistent with normality?) or parameters of distributions (is the population mean zero?), whereas nonparametric methods may be more nebulous (are these data random?).

2. (i) We are testing if the 16 observations have median 0. Since there are three negative differences, a one-sided p-value would be $\{{}^{13}C_3 + {}^{13}C_2 + {}^{13}C_1 + {}^{13}C_0\}/216$ = 697/216 = 0.011. A two-tailed test is appropriate, so the p-value is double this, 0.021. The null hypothesis that the median is zero is rejected at the 0.05 level but not the 0.01 level; there is some evidence that the median is not zero. Here is some R code, modified from Rippon (2016). It supports the calculations above.

```
# vector of question 2 data
y <- c(0.3,6.3,3.7,2.8,5.8,-1.4,1.7,2.3,-1.7,1.6,-1.8,0.6,4.5,1.9,2.4,6.8)
n <- length(y) # number of measurements
hmed <- 0 # hypothesized median
S <- length(y[y>hmed]) # number of measurements greater than
hypothesized median
left.tail <- pbinom(q=S-1, size=n, prob=0.5)
right.tail <- pbinom(q=S-1, size=n, prob=0.5, lower.tail=FALSE)
cat("pval(less)=",left.tail, "; pval(greater)=",right.tail, ";
pval(both)=", 2*right.tail)
```

(ii) Since $\hat{p}$ = 3/16, the approximate 95% confidence interval is (–0.004, 0.379), or, since p must be non-negative, (0, 0.379). The negative part of the confidence interval reflects the fact that the normal approximation to the binomial isn't adequate in the tails. The 99% confidence interval is (0.066, 0.430).

(iii) The 95% confidence interval in (ii) excludes 0.5, which is consistent with concluding, at the 0.05 level, that $p \neq 0.5$, and that the median is not 0. The 99% confidence interval also excludes 0.5, so that the test for $p = 0.5$ would also be rejected, now at the 0.01 level. This is in conflict with (i), reflecting the fact that the test based on the approximate confidence interval uses the normal approximation to the binomial, whereas the test in (i) does not. The two tests are making different assumptions, so it is not surprising they come to different conclusions.

3. (i) The Shapiro-Wilk test of normality has p-value 0.0226 for method 1 and 0.3728 for method 2. Tests assuming normality are therefore dubious. For the normal theory tests, tests of equality of variance, such as the Bartlett and the Levene, have p-values less than 0.05. Thus the pooled t-test, which has p-value 0.857, is dubious. The Welch test, that does not assume equality of variances, has p-value 0.858.

The Wilcoxon test has p-value 0.713 or 0.674, depending on the approximation used, and the median test has p-value 1.0. Of these two tests the Wilcoxon test is the more powerful location test.

However none of these tests gives evidence of a location difference between methods at any commonly used significance level.

(ii) The ordered data, keeping track of the methods and ranks, is

Method 1				–1.4	1.7	2.3	2.3		2.8	2.8
Method 2	–4.7	–2.6	–1.8					2.4		
Rank	1	2	3	4	5	6	7	8	9	10

Method 1	3.3	3.7				
Method 2			3.9	4.5	4.6	6.8
Rank	11	12	13	14	15	16

(iii) There are $t = 5$ runs, $E[T] = 9$, and $\text{var}(T) = 56/15 = 1.9322$. We find $P(5 \text{ or fewer runs}) = P(Z < (5.5 - 9)/1.932 = -1.811) = 0.035$. Exact calculations give

$$P(T = 2) = 2\,{}^{7}C_{0}\,{}^{7}C_{0}/{}^{16}C_{8},\ \ldots,\ P(T = 5) = 2\,{}^{7}C_{2}\,{}^{7}C_{1}/{}^{16}C_{8} \text{ and}$$
$$P(T \leq 5) = 2(1 + 7 + 49 + 147)/{}^{16}C_{8} = 408/{}^{16}C_{8} = 0.0317.$$

There is good agreement between the exact p-value and that based on the normal approximation. Both show significance at the 0.05 level but not at the 0.01 level. There is some evidence of a difference in methods.

Here is some R code, modified from Rippon (2016). It supports the calculations above.

```
y <- c("A", "A", "A", "B", "B", "B", "B", "A", "B", "B", "B",
"B","A","A","A","A") # data vector
yt <- y == y[1] # convert to TRUE and FALSE, ie 1 and 0
yd <- diff(yt) # non-zero elements of the difference vector
indicate the end of a run
T <- length(yd[yd != 0])+1
ty <- table(y) # table of counts
M <- ty[1]
N <- ty[2]
PT <- function (t, m, n) {# probability calculations
      k <- t %/% 2 # note use of the integer division operator
      if (t %% 2 == 0) # t is even
            prob <- 2 * choose(m-1, k-1) * choose(n-1, k-1) / choose(m+n, n)
else # t is odd
```

```
            prob <- (choose(m-1, k) * choose(n-1, k-1) + choose(m-1,
k-1) * choose(n-1, k)
            ) / choose(m+n, n)
      return(prob)
}

cat("p-value = P(T =< ", T, ") = ", PT(t=2,m=M,n=N) + PT(t=3,m=M,n=N)
+ PT(t=4,m=M,n=N)+ PT(t=5,m=M,n=N), sep="")
```

(iv) (a) The median is (2.4 + 2.8)/2 = 2.6. This gives a table

	Method 1	Method 2	Total
Above 2.1	4 (4)	4 (4)	8
Below 2.1	4 (4)	4 (4)	8
	8	8	16

Expected values are in parentheses.

This is exactly as expected, so $X^2 = 0$ with p-value $P(X^2 \geq 0) = 1$. The data are not significant at *any* level. In fact the agreement with the null hypothesis is suspiciously good. We conclude that at all 'reasonable' levels, the method *medians* are consistent.
Again, here is some supporting R code.

```
# data vectors
m1 <- c(3.3, 2.3, 3.7, 2.8, 2.8, -1.4, 1.7, 2.3)
m2 <- c(-4.7, 4.6, -1.8, -2.6, 4.5, 3.9, 2.4, 6.8)

methods <- data.frame ( )
for (m in paste("m",1:2,sep="")) {
      methods <- rbind(methods, data.frame(yield=get(m), method=m))
}
om <- median(methods$yield) # observed median of combined methods yields

# use the cut function to create a factor based on whether yield
# is above or below the median
methods$cut <- cut(methods$yield, breaks=quantile(methods$yield,
probs=c(0,0.5,1)),
                 labels=paste(c("<=", ">"), om), include.lowest=TRUE)
```

```
# create and print contingency table
methods.xt <- xtabs(~cut+method,data= methods)
methods.xt
# perform chi-squared test
methods.chisq <- chisq.test(methods.xt)
methods.chisq
# show expected values
methods.chisq$expected
chisq.test(methods.xt, simulate.p.value=TRUE)
```

(b) The lower quartile is (–1.4 + 1.7)/2 = 0.15. This gives the following table.

	Method 1	Method 2	Total
Above 0.15	7 (6)	5 (6)	12
Below 0.15	1 (2)	3 (2)	4
	8	8	16

Expected values are in parentheses.

$X^2 = 2(1/6 + 1/2) = 4/3$. This is not significant at the 0.1 level as $\chi_1^2(0.1) = 2.706$. To find the exact p-value, note that the only table not more extreme than the observed (in terms of X^2 values), is that with all entries exactly as expected, and this table has probability ${}^8C_2\,{}^8C_2/{}^{16}C_4 = 0.4308$. The exact p-value is thus $P(X^2 \geq 4/3) = 1 - P(X^2 < 4/3) = 1 - 0.4308 = 0.5692$. All other tables result X^2 values of at least 4/3. We conclude that at the 0.05 level (and all reasonable levels), the method *lower quartiles* are consistent.

Using a quantile other than the median requires slight adjustments to the R code. In general such code won't be given subsequently.

```
m1 <- c(3.3, 2.3, 3.7, 2.8, 2.8, -1.4, 1.7, 2.3)
m2 <- c(-4.7, 4.6, -1.8, -2.6, 4.5, 3.9, 2.4, 6.8)

methods <- data.frame()
for (m in paste("m",1:2,sep="")) {
      methods <- rbind(methods, data.frame(yield=get(m), method=m))
}
oq <- quantile(methods$yield, probs=0.25) # observed quantile of
combined methods yields
```

```
# use the cut function to create a factor based on whether yield
# is above or below the observed quantile
methods$cut <- cut(methods$yield, breaks=quantile(methods$yield,
probs=c(0,0.25,1)),
                 labels=paste(c("<=", ">"), oq), include.lowest=TRUE)

# create and print contingency table
methods.xt <- xtabs(~cut+method,data= methods)
methods.xt
# perform chi-squared test
methods.chisq <- chisq.test(methods.xt)
methods.chisq
# show expected values
methods.chisq$expected
chisq.test(methods.xt, simulate.p.value=TRUE)
```

(v) Using the ranked data we find $W_1 = 64$, $W_2 = 72$.
Check: $W_1 + W_2 = 1 + \ldots + 16 = 8*17 = 136 = 64 + 72$.

With $m = n = 8$ we find $E[W_2] = 68$, $\text{var}(W_2) = 90.6667$, giving $P(W_2 \geq 72) = P(Z > (71.5 - 68)/\sqrt{90.6666} = 0.3676) = 0.357$. Since both large and small W_2 values are inconsistent with the null hypothesis, the p-value is double this, 0.714. At the 0.05 level, and indeed, at all reasonable levels, there is no evidence against the null hypothesis that the methods are consistent. The following R code supports the above.

```
methods <- data.frame(area= c("A", "A", "A", "B", "B", "B", "B",
"A", "B", "B", "B", "B","A","A","A","A"))
methods$r <- 1:nrow(methods) # ranks
methods
W.A <- sum(methods$r[methods$area=="A"]) # sum of ranks for method 1
W.B <- sum(methods$r[methods$area=="B"]) # sum of ranks for method 2

counts <- table(methods$area)
m <- counts[1] # method 1
n <- counts[2] # method 2

# determine parameters for normal approximation to null
distribution of W.A
mu.A <- m*(m+n+1)/2
sig2.A <- m*n*(m+n+1)/12

# calculate z score and p value corresponding to W.A (note
continuity correction)
z <- (W.A - 0.5 - mu.A)/sqrt(sig2.A)
p.val <- 1-pnorm(z)
cat("p-value = P(Z < ", z, ") = ", p.val, sep="")
```

(vi) In addition to the p-values reported in 3(i), we have the median test p-values of 0.57 and 1.0, and the runs test 0.03.

It is not valid to apply many different tests and take the most or least extreme. Nevertheless what seems to be happening here is that the runs test has detected an alternative to the null hypothesis that none of the other tests has been able to detect. The runs test is sensitive to differences in both location and shape, while the Wilcoxon and median tests are sensitive to location differences only. It seems the runs test is picking up dispersion differences between the methods that the tests for parametric equality of variances detected. Recall that all had low p-values, suggesting the variances were inconsistent. Although the other tests of equality of variance are parametric, Levene's test does not assume normality, and is traditionally labelled as nonparametric.

SOLUTIONS TO THE CHAPTER TWO EXERCISES

1. The two sample t-test gives a p-value of 0.9178. The Welch test gives a p-value of 0.9167, remarkably consistent, but JMP also gives p-values for tests of consistency of the population variances: 0.0098 for the Bartlett test that assumes the data are normally distributed and which is known to be sensitive to non-normality. The nonparametric Levene test gives a p-value of 0.0097. This is remarkable agreement given that the Shapiro-Wilk test for normality of the residuals has p-value 0.0001. So in spite of the agreement that the Australian and Japanese consumers have consistent means but different variances, none of the tests here are valid as they assume normality, and at the 0.001 level the data are not consistent with normality.

   Using the asymptotic chi-squared approximation the Kruskal-Wallis test has p-value 0.4445. Referring $F = KW(n - t)/\{n - 1 - KW)(t - 1)\}$ to the $F_{t-1,n-t}$ distribution gives a p-value of 0.4603; using the improved $F_{d(t-1),d(n-t)}$ distribution gives a p-value of 0.4572.

   At all reasonable levels there is no evidence of a location difference between the Australian and Japanese consumers. However any plot of the data, or even just 'eyeballing' the data suggests there are no location differences but that the Japanese may be less variable than the Australian consumers. This is consistent with the Levene test result, which does not assume normality. Tests sensitive to more than differences in location are needed to unlock what is happening here.

   The following R code gives the Wilcoxon p-value.

```
rm(list=ls())
choc1 <- list(Australian=c(1, 1, 2, 3, 3, 3, 3, 3, 3, 4, 5, 5, 5,
5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7),
Japanese=c(2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5,
5,5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7))
choc2 <- stack(choc1)
names(choc2) <- c("score", "nationality")
wilcox.test(score~nationality, data=choc2)
choc2$r <- rank(choc2$score, ties.method = "average")

A <- sum(choc2$r[choc2$nationality=="Australian"])
J <- sum(choc2$r[choc2$nationality=="Japanese"])

t <- 2
```

```
n <- nrow(choc2)
KW <- 12/n/(n+1) * (A^2/33 + J^2/31) - 3*(n+1)
d <- 1 - 6*(n+1)/(n-1)/(5*n+6)
F <- KW / (n-1-KW) * (n-t) / (t-1)
pval <- pf(F, df1=d*(t-1), df2=d*(n-t), lower.tail=FALSE) # note
adjusted df
F
pval
```

2. A parametric analysis of the raw unranked data yields a p-value of 0.1626 for word processors (and 0.2717 for experts, who are blocks). At even the 0.1 level word processors are not significantly different. However the Shapiro-Wilk test of normality has p-value 0.0083. At the 0.01 level the data are not consistent with normality and although ANOVA in general is robust to the assumption of normality, the analysis is problematic.

One way of analysing the data nonparametrically is to analyse the ranks within blocks. The randomised blocks analysis of the ranks yields a p-value of 0.0003 for word processors. The blocks p-value is 1.0; because there are no ties every block receives ranks 1, 2, 3 and 4, and hence there are no differences between blocks. Now word processor ranks are significant at the 0.001 level, which can be interpreted as there are significant differences in the means of the word processor ranks. The Shapiro-Wilk test of normality has p-value 0.6209, so on this score at least, the inference is valid. One multiple comparisons analysis gives the following diagram, with rank means in parentheses:

B (1) D (2.25) A = C (3.5) E (4.75)

At an overall 0.05 level, there is evidence that the mean ranks for A and C are similar, and otherwise the word processor mean ranks are different.

Modifying the R code gives a value of the Friedman statistic of 13 with a χ^2_4 p-value of 0.01128 and a $F_{4,12}$ p-value 0.0002553. From the latter (which is more reliable) at the 0.001 level there is evidence of a difference in the word processor ranks. To assess the nature of this difference the orthogonal components are investigated. The Page test statistic takes the value 1.5 and p-value 0.1336, while the umbrella statistic takes the value 2.1129 with p-value 0.0346. The mean ranks in order A, B, C, D and E are 3.5, 1, 3.5, 2.25 and 4.75. So as we pass from A to E there is evidence (at the 0.05 level but not the 0.01 level) of an umbrella effect: here that the mean ranks decrease and then increase.

The R code is given in the text file on the book web page.

3. The ANOVA F test shows that the (ranked and ties broken) responses are significant at the 0.001 level with p-value less than 0.0001. Using the Shapiro-Wilk test for normality yields a p-value of 0.319, so the residuals are consistent with normality. One multiple comparisons analysis at the 0.05 level gives the means for ice creams one to six as 1.3, 1.9, 2.4, 2.9, 3.1 and 3.4 respectively, with corresponding diagram.

1.3 1.9 2.4 2.9 3.1 3.4

It seems that at this level ice creams A and B are consistent, as are the ice creams B and C, C to E and ice creams D to E and; all others are significantly different.

The nonparametric analysis finds the Durbin statistic takes the value 20.9333 with χ^2_5 p-value 0.0008 and $F_{5,40}$ p-value 0.0001. The linear contrast takes the value 4.5057 with corresponding p-value 0.0000; the quadratic contrast takes the value –0.7066 with corresponding p-value 0.4469.

There is strong evidence, at the 0.001 level, of a treatment effect. This is consistent with the F test conclusion, although for the Durbin analysis the conclusion is weaker: that the treatment distributions differ. The quadratic contrast is not significant at the 0.05 level but the linear contrast is significant at the 0.05 level but not the 0.01 level. Again this is consistent with the F test analysis and with simply eyeballing the data: as we pass from ice cream one to six that mean ranks strictly increase.

The R code is given in the text file on the book web page.

SOLUTIONS TO THE CHAPTER THREE EXERCISES

1. Here is the completed table for the Japan project chocolate analysis. Permutation test p-values are based on 10,000 permutations and of course will differ slightly each time they are calculated.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{1,62}$	0.9178	0.9324
Kruskal-Wallis	χ^2_1	0.4403	0.4446
Page-type	N(0, 1)	0.4560	0.4446
First order NP ANOVA	$F_{1,62}$	0.9178	0.9324
Second order NP ANOVA	$F_{1,62}$	0.0037	0.0030
Third order NP ANOVA	$F_{1,62}$	0.5079	0.5110

Comment. In Exercises 2, question 1 a p-value for the Wilcoxon rank sum test with continuity correction is given. This is slightly different from the Kruskal-Wallis p-value given here, presumably because R either does not use a continuity correction for the Kruskal-Wallis test, or uses one that doesn't reduce to that for two treatments.

The permutation test p-values agree well with the distribution theory p-values. At the 0.05 level the ANOVA F test reveals no difference in mean scores and the Kruskal-Wallis test, known to be sensitive to differences in medians, reveals no difference in distributions. As there are only two factors there is only one orthogonal contrast, the linear. It will also be sensitive to location differences, and it, too, shows no evidence of same. The nonparametric ANOVAs show no evidence of first and third order effects but at the 0.01 level gives evidence of a second order effect. As there is no evidence of a location effect, this is a variance effect.

The R code is given in the text file on the book web page.

2. For the word processors ranked data from the second exercise for Chapter Two the completed table is as follows.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{4,12}$	0.0003	0.0011
Friedman	χ^2_4	0.0113	0.0012
	$F_{4,12}$	0.0003	0.0011
Page	N(0, 1)	0.1336	0.1438
Umbrella	N(0, 1)	0.0346	0.0321
Cubic	N(0, 1)	0.6171	0.5936
First order NP ANOVA	$F_{4,12}$	0.0003	0.0010
Second order NP ANOVA	$F_{4,12}$	0.0088	0.0201
Third order NP ANOVA	$F_{4,12}$	0.1148	0.1270

Comment. Again the agreement between the distribution theory p-values and the permutation test p-values is good. At the 0.001 level the word processor mean ranks are significantly different. For word processors A, B, C and D these are are 3.5, 1.0, 3.5, 2.25 and 4.75 respectively. An LSD analysis at an overall 0.05 level gives

B D A C E

Word processors A and C have similar mean ranks but otherwise all word processor mean ranks are significantly different. At the 0.05 level there is no evidence of a linear trend but there is of an umbrella effect. This is consistent with the significance, at the 0.05 level, of the first and second order nonparametric ANOVAs.

The R code is given in the text file on the book web page.

3. For the ice cream flavour data from the third exercise for Chapter Two the completed table is as follows.

Test Statistic	Distribution theory		Permutation test
	Distribution	p-value	p-value
ANOVA F	$F_{5,40}$	0.0001	0.0003
Durbin	χ^2_5	0.0008	0.0002
	$F_{5,40}$	0.0001	
Page	N(0, 1)	0.0000	0.0000
Umbrella	N(0, 1)	0.4469	0.4680
Cubic	N(0, 1)	0.9846	0.9892
First order NP ANOVA	$F_{5,40}$	0.0001	0.0002
Second order NP ANOVA	$F_{5,40}$	0.0473	0.0581
Third order NP ANOVA	$F_{5,40}$	0.0283	0.0318

Comment. At the 0.001 level there is a significant difference in ice cream flavours. For ice creams A to F the flavour means are 1.4, 1.9, 2.4, 2.9, 3.1 and 3.4. Using the comparison command in R gives the following LSD analysis.

A B C D E F

There are five degrees of freedom associated with the location effect. These can be decomposed into orthogonal components of degree one to five. Here we only give the first three. Of these it seems only the first component, reflecting a linear trend in the mean ranks, is important. The Page-type test indicates, at the 0.05 level, that as we pass successively through the ice creams from A to F the flavour ranks increase.

The significance of the first order NP ANOVA reflects the location effect. The (near) significance of the second order NP ANOVA reflects moment effects up to order two. At the 0.05 level there are location effects; there may be second order effects as well.

The R code is given in the text file on the book web page.