A Contingency Table Approach to Nonparametric Testing

J.C.W. Rayner
D.J. Best

CHAPMAN & HALL/CRC
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Rayner, J.C.W.
A contingency table approach to nonparametric testing / J.C.W. Rayner, D.J. Best.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-161-5 (alk. paper)
1. Nonparametric statistics. 2. Contingency tables. I. Best, D.J., II. Title.
QA278.8 .R39 2000
519.5—dc21 00-050443

International Standard Book Number 1-58488-161-5
Library of Congress Card Number 00-050443
Printed in the United States of America 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Contents

Preface

1 Introduction
- 1.1 Parametric or Nonparametric?
- 1.2 Instructors Example
- 1.3 Quadratic Differences and Ranking
- 1.4 Outline and Scope
- 1.5 Applications of Nonparametric Methods to Sensory Evaluation

2 Modelling Ties
- 2.1 Introduction
- 2.2 The Sign Test and Ties
- 2.3 Modelling Partitioned Ties in the Sign Test
- 2.4 Modelling Unpartitioned Ties in the Sign Test
- 2.5 McNemar's Test
- 2.6 Partitioning into Components
- 2.7 Ties in a Multinomial Test
- 2.8 Ties When Testing for Independence

3 Tests on One-Way Layout Data: Extensions to the Median and Kruskal-Wallis Tests
- 3.1 Introduction
- 3.2 A Model and Pearson's X^2 Test
- 3.3 Partitioning Pearson's Statistic
- 3.4 The Kruskal-Wallis Test with No Ties
- 3.5 The Kruskal-Wallis Test with Ties
- 3.6 Generalised Median Tests

4 Tests Based on a Product Multinomial Model: Yates' Test and its Extensions
- 4.1 Introduction
- 4.2 One-Way Tables
- 4.3 Two-Way Tables
- 4.4 Partitioning X_P^2 Using Score Statistics
- 4.5 Other Methods for Ordered Data
- 4.6 Small Sample Size and Power Comparisons
- 4.7 Examples

5 Further Tests Based on a Product Multinomial Model: Order in the Sign Test and Ordinal Categorical Data with a Factorial Response
- 5.1 Introduction
- 5.2 How Order Affects the Sign Test
- 5.3 The Sign Test and Gart's Tests
- 5.4 A New Model and Score Test
- 5.5 Comparison of the Sign and Score Tests
- 5.6 Sports Drink Example
- 5.7 Recommendations
- 5.8 Nonparametric Analysis of Ordinal Categorical Data with Factorial Response
- 5.9 Olives Data Example
- 5.10 Cross Cultural Study Example

6 Tests on Complete Randomised Blocks: Extensions to the Friedman and Cochran Tests
- 6.1 Peach Example
- 6.2 Friedman's Test and its Extensions
- 6.3 Derivations
- 6.4 Page's Test and its Relationship to Friedman's, Anderson's, and Pearson's Tests
- 6.5 An Alternative Partition of the Anderson Statistic: an Umbrella Test
- 6.6 Ties
- 6.7 Cochran's Test
- 6.8 Stuart's Test and its Extensions

7 Further Tests on Randomised Blocks: Extensions to Durbin's Test
- 7.1 Introduction
- 7.2 Durbin's Test and its Extensions
- 7.3 Derivations
- 7.4 A Page-Type Test
- 7.5 Paired Comparisons with a 2^n Factorial Structure

8 Extensions to a Nonparametric Correlation Test: Spearman's Test
- 8.1 Introduction
- 8.2 A Smooth Model and Tests for Independence
- 8.3 Smooth Extensions
- 8.4 Interpretation of the Components
- 8.5 Discussion
- 8.6 Multi-Way Tables

9 One and S-Sample Smooth Tests of Goodness of Fit
9.1 Introduction
9.2 One-Sample Testing for Uncategorised Distributions
9.3 One-Sample Testing for Categorised Distributions
9.4 S-Sample Testing
9.5 Derivations and Simulation Study

10 Conclusion
10.1 Introduction
10.2 Partitioning Pearson's Chi-Squared Statistic for at Least Partially Ordered Three-Way Contingency Tables
10.3 Generalised Correlations
10.4 Partially Parametric Testing
10.5 Concluding Comments

Appendices
A.1 Statistical Prerequisites
A.2 Orthogonal Matrices
A.3 Orthonormal Functions
A.4 Direct Sums and Products
A.5 Likelihood Ratio, Score, and Wald Tests
A.6 Assessing Univariate Normality
A.7 Multivariate Normality
A.8 Confidence Circles and Correspondence Analysis Plots
A.9 Permutation and Bootstrap Methods

References

Preface

The prime objective of this book is to provide a *unification and extension* of some popular nonparametric statistical tests by linking them to tests based on models for data that can be presented in contingency tables.

We asked several colleagues, "What, in your experience, are the most frequently used nonparametric techniques?" The answers focused on the sign, median, Wilcoxon, Kruskal-Wallis, Page, Friedman, Durbin, Cochran, Spearman and Kendall's tau tests. These tests all use ranks as data. The treatment of ties is a major concern.

For almost all of these tests, and for some other important tests also, we are able to present the data in contingency tables. We can then calculate a Pearson-type X^2 statistic, and its *components*. In the case of univariate data, the initial tests based on these components detect mean differences between treatments, and in the case of bivariate data, they detect correlations. The later components provide tests that detect variance, skewness and higher moment differences between treatments with univariate data, and higher bivariate moment differences with bivariate data. This approach provides a *unification* of much popular nonparametric statistical inference, and makes the traditional, most commonly performed nonparametric analyses much *more complete and informative*. In addition, the contingency-table approach means tied data are easily handled, and almost exact Monte Carlo p-values can be obtained. Modern statistical packages such as *StatXact* (1995) calculate p-values this way.

Sprent (1993, section 9.3) gives a description of the *StatXact* (1995) approach and also notes that many of the common nonparametric location tests can be given via a contingency table formulation. Neither *StatXact* (1995) nor Sprent (1993) looks at extensions based on higher order moments or at the extensions involving three-way contingency tables that we give. We also do not consider all marginal totals to be fixed as they do. Further, we consider tests for data from incomplete block designs, which neither *StatXact* (1995) nor Sprent (1993) considers.

Categorical data as well as ranked data can be presented in contingency tables, and we also look at nonparametric tests for certain

categorical data. Moreover ranks are but one choice of category scores. Our formulation permits other user-defined scores.

The methods advanced in this book are somewhat traditional in that practitioners trained 20 or more years ago will be comfortable with them. The power of modern computers is used to obtain accurate p-values using resampling and Monte Carlo methods that at least augment and sometimes replace p-values based on asymptotic distribution theory. When sample sizes are small, or the approach to the asymptotic limit is slow, the computer intensive methods are more reliable and meaningful than the older approaches. So the approach we recommend, while familiar, is more complete and more valid than was previously possible. However we have not considered modern methods in robustness and Bayesian analysis.

Some readers who are familiar with our published goodness of fit work may be surprised by this excursion into nonparametrics. We see it as totally consistent with our view of data analysis. If a parametric analysis is available and valid it will often, but not always, be more powerful and more efficient than the competitor nonparametric test. If so it should be used. Assessing validity will usually require, amongst other things, an assessment of the distributional assumptions. Smooth tests of goodness of fit, discussed in Rayner and Best (1989a), may be helpful in that assessment. If the parametric analysis is not valid, then nonparametric procedures, such as outlined here, may be appropriate.

Finally our thanks to

- Geoff Robinson and those other reviewers whose names we do not know, for all your constructive and insightful suggestions;
- James Chipperfield, Robert Denham, Don Findley, Jo Lea, Nell Stetner-Houweling, and Chris Valentine, students who worked through some of the material as part of their honours year studies at the University of Wollongong;
- Eric Beh who extended some of our work as part of his Ph.D. studies;
- Anthony Carolan, who shed light on several matters during his Ph.D. studies;
- Pam Davy, who assisted both Jo and Eric; and
- Chin Diew Lai and Geoff Jones, colleagues from Massey University in Palmerston North, New Zealand, for interesting discussions on various aspects of this and other work.

The book has benefited from their positive and insightful advice.

J.C.W. Rayner

D.J. Best

1
Introduction

1.1 Parametric or Nonparametric?

Parametric tests are based on parametric models. One such test is the traditional t-test for assessing if the mean of a population, assumed to come from a normal distribution, takes some specified value. A nonparametric alternative that has been traditionally employed in this situation is the Wilcoxon test. This book is about tests that have traditionally been labelled *nonparametric*. We avoid a formal definition of nonparametric, but such a definition wouldn't necessarily be helpful. Perhaps the essential point is that nonparametric tests are based on fewer assumptions than parametric tests. For an enlightening discussion see, for example, Sprent (1993, section 1.1). Our problem is that we can present some nonparametric tests in a way that is inconsistent with some of the favoured definitions of nonparametric. There is no ambiguity in saying this book extends and unifies some of the tests at the core of nonparametrics as it has been known and developed over the $20th$ century, including the sign, Mann-Whitney/Wilcoxon, Kruskal-Wallis, Page, Friedman, Durbin, Cochran, Stuart, and Spearman tests.

Sometimes the data analyst has no choice but to employ nonparametric methods; for example when the data are available only on the nominal or ordinal scales of measurement. (See A.1.6 for brief descriptions of nominal, ordinal, interval, and ratio scales.) Often there is a choice, as with, for example, in the application of either the Kruskal-Wallis or the one factor analysis of variance ANOVA F test. As is well known, under certain assumptions the ANOVA F test will be more powerful. However the Kruskal-Wallis test is available when these assumptions do not hold. Some statisticians retreat to nonparametric methods when there is the slightest doubt about the parametric

assumptions. Others rarely use nonparametric procedures. It is known, for example, that the ANOVA F test is robust to some of its assumptions, and this vague information is often taken as licence to apply the test when the assumptions are so obviously false that the analysis can have little validity. Like most statisticians, our view is between these extremes.

There is often a considerable parametric structure that can be used to one's advantage even when the usual parametric model clearly fails (see section 10.4 on *Partially Parametric Testing*). The idea is that normality may fail, but an alternative parametric model is appropriate, and this alternative model is the basis for a far more powerful analysis of the data than either the original parametric analysis or a nonparametric analysis. The alternative parametric analysis described in section 10.4 tends to be highly computer intensive, and its development is as yet far from complete.

Parametric model failure produces several problems, not all of which can be remedied by nonparametric analysis. Nonparametric analysis can assist if the originally proposed model is not, in fact, the correct one. We note that:

- the p-value or significance level based on the parametric model may be quite different under the correct model;
- a test that has high power under the parametric model may be quite poor under the correct model.

Moreover the proposed parametric model may be an important part in the analysis of the data. For example, in testing if lifetimes have a given mean, the assumption of exponentiality is an important part of the modelling. Failure of the exponential model may be more important than what the mean is. Retreating to a nonparametric analysis without careful consideration of the underlying model could be at best unfortunate.

The nonparametric tests of initial interest in this book detect, for univariate data, mean differences between treatments, or, in the case of bivariate data, correlations. Our approach produces additional *component*-based tests that detect variance, skewness and higher moment differences between treatments with univariate data, and higher bivariate moment differences with bivariate data. These additional tests make the traditional, most commonly performed nonparametric analyses much more informative. Our additional tests may all be derived by presenting the

data in contingency tables; they are all components of *Pearson's chi-squared test statistic*, X_P^2 (see A.1.3). This provides the rationale for the title of this book.

As we present the data in contingency tables, it is reasonable to ask why we do not advocate using tests based on log-linear models. As with the ANOVA test above, these models make strong assumptions and so do not necessarily apply as widely as nonparametric tests. Further, log-linear models need iterative methods which can present numerical problems if almost exact p-values are required. This is especially so with sparse data sets. It is nevertheless true that log-linear models are currently in widespread use; so in several cases we provide comparisons of our approach with that based on log-linear models. In all cases the p-values produced are *very* similar.

Other important features of our methods are that they easily cope with *tied* data and that exact or almost exact p-values can easily be obtained. Our nonparametric analyses also give least significant difference (LSD) assessments, contrasts and graphical presentations to assist the statistician to explain the analysis to those with little statistical background.

1.2 Instructors Example

We will now consider an example which illustrates the nature of our refinements to standard nonparametric tests. Our interest in this data set, and others we give, is more to illustrate points concerning the statistical methodology than to assess the application. In statistical practice, of course, the reverse is usually the case.

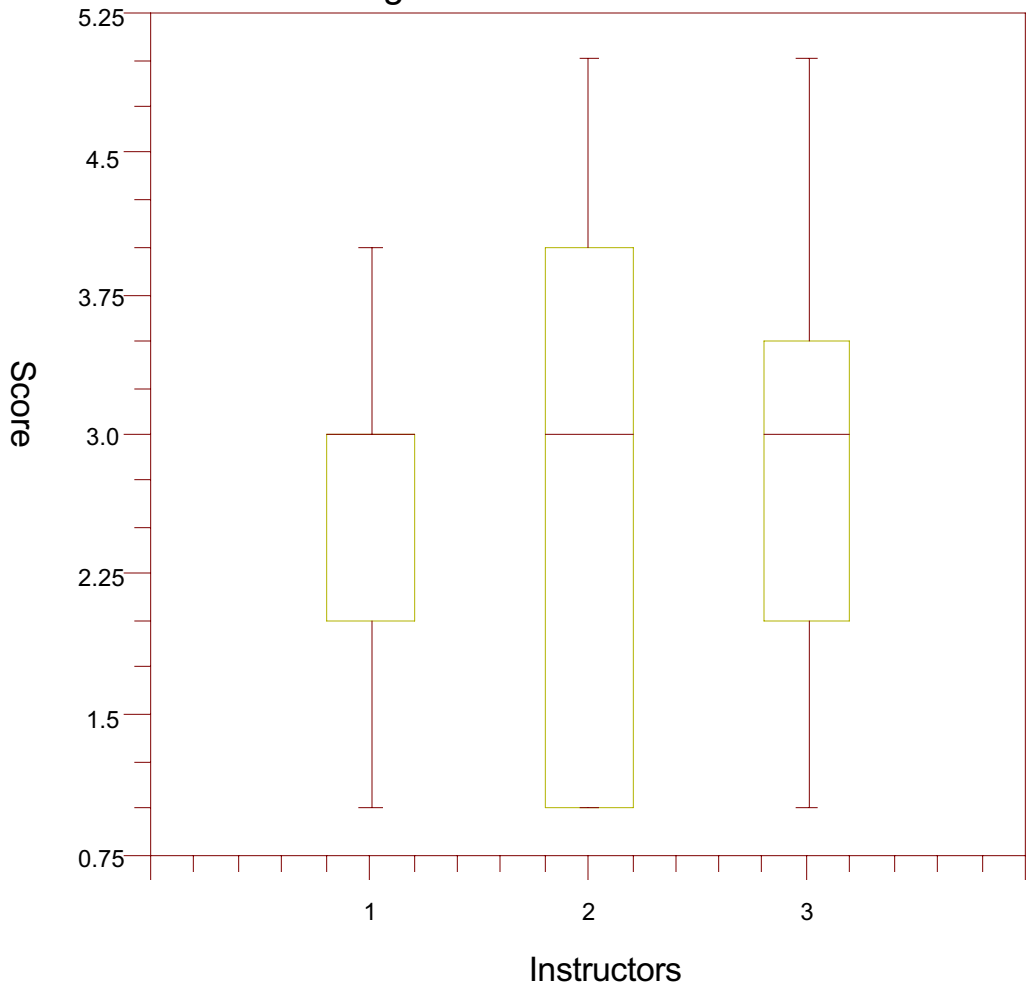
Instructors Example. Conover (1998, p.293) gave the following data of three instructors who assign grades in five categories according to Table 1.1. We assign the scores 1 to 5 to the categories, with A scored as 1, B as 2, and so on.

Table 1.1 Grades assigned by three instructors

	Grade					
	A	B	C	D	E	Total
Instructor 1	4	14	17	6	2	43
Instructor 2	10	6	9	7	6	38
Instructor 3	6	7	8	6	1	28
Total	20	27	34	19	9	109

Assume that a standard computer package is available, and is used to explore these data. In testing for normality, the Kolmogorov-Smirnov normality test yields a p-value of greater than 0.20 for each of the three instructors, while the Shapiro-Wilk normality test yields p-values of at least 0.10 for all three distributions. These tests cannot reject the null hypothesis of normality, perhaps in part because of the small sample sizes.

Figure 1.1: Instructors' Data

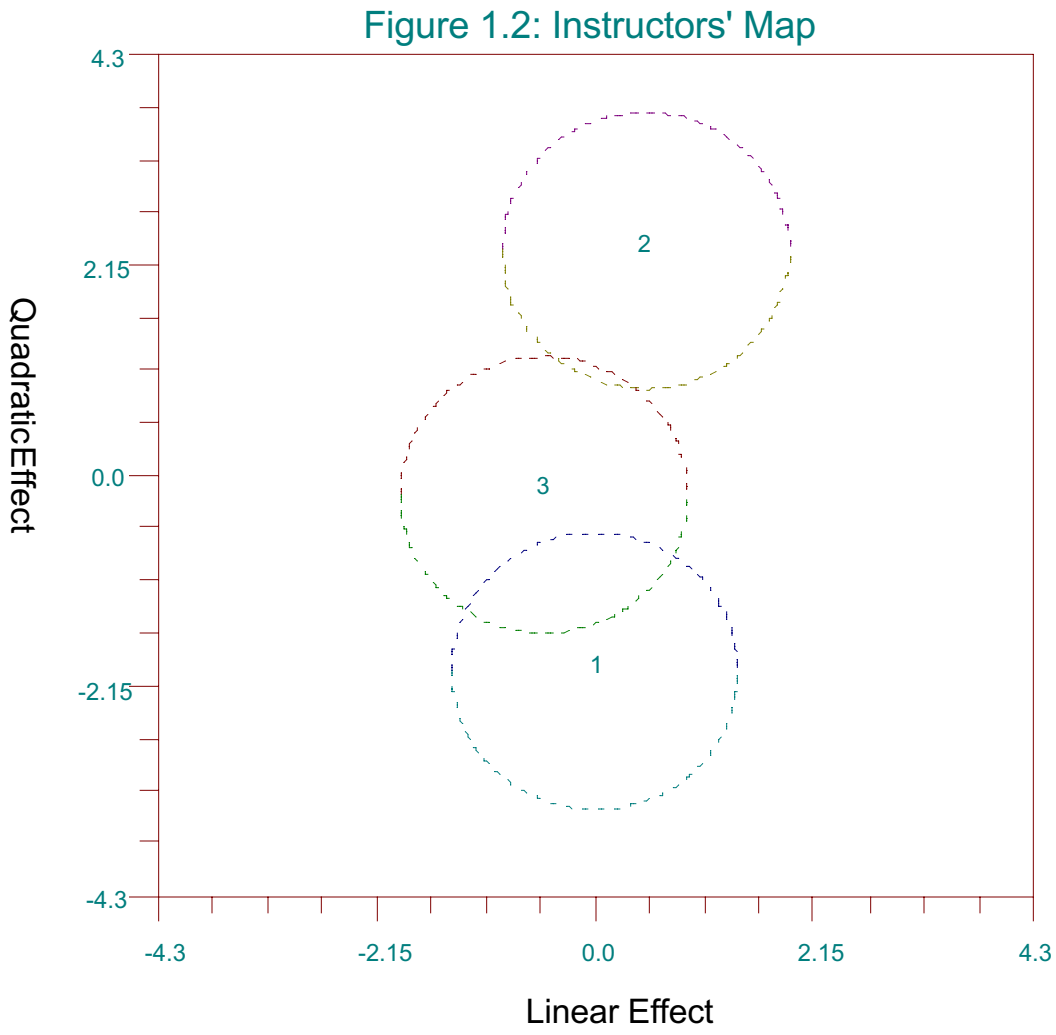


A one-way analysis of variance is in any case robust to normality and so may be confidently applied, yielding a p-value of 0.78. No assessment has yet been made of homogeneity of variance. In fact inspection of the data, via box plots, suggests that the variability differs between the instructors (see Figure 1.1). The failure of this assumption decreases our confidence in the p-value from the analysis of variance.

The Kruskal-Wallis test adjusted for ties has a p-value of 0.85. The less powerful median test has p-value 0.77, while the usual Pearson's X^2_P test for homogeneity of rows has p-value 0.15, all consistent with the conclusion that there seems to be no difference between the instructors in

regard to location differences.

The techniques we develop in this book enable us to decompose Pearson's X_P^2 into a location detecting component with p-value 0.85, a dispersion detecting component with p-value 0.01, and a residual with p-value 0.75. An LSD analysis of the dispersion detecting component shows that the first instructor is less variable than the third who is less variable than the second. The nonsignificant location component confirms the analyses of the location tests, while the nonsignificant residual indicates that there are no further effects in the data. This rather complete analysis makes no assumptions. Conover's analysis, like the analysis of variance and Kruskal-Wallis tests, did not identify the dispersion effect. Our analysis is very effectively presented graphically in what we call a product map. For an introduction to product maps see Appendix A.8. One axis gives information about linear effects, that we interpret as location differences between treatments, while a second gives information about quadratic effects, that we interpret as dispersion differences (see Figure 1.2). This example will be revisited later at the end of section 3.3.



1.3 Quadratic Differences and Ranking

The previous section gave an example in which a linear or location effect was not significant but a quadratic or dispersion effect was. Differences in quadratic effects are often associated with scale or dispersion effects. If the linear effect had been significant this may have masked the detection of the quadratic effect. To help see this consider the following data sets A and B:

data set A: -2.019, -0.802, -0.153, -1.151, -1.247, 0.293, 1.848, 1.430, 1.875, -0.192,
data set B: 12.838, 29.986, 16.284, 18.131, 23.386, 14.810, 41.251, 13.601, 10.940, 20.448.

Data set A consists of 10 random N(0, 1) values, while data set B consists of 10 random N(20, 10) values. Suppose we rank the data from smallest to largest.

Data Value	-2. 019	…	1.875	10.940	…	41.251
Rank	1	…	10	11	…	20
A	1	…	1	0	…	0
B	0	…	0	1	…	1

The A's occupy the first 10 ranks and B's the last 10. If the analysis of the previous section is used, the linear effect is very highly significant (p-value < 0.001) but the quadratic component is zero (p-value one). Clearly the ranking transformation has destroyed the quadratic differences between A and B. To compensate for this problem with ranking, a common procedure is to align the data sets by subtracting the A mean from each A value, and the B mean from each B value. See, for example, Sprent (1993, p. 124).

However when the raw data are available we suggest using these raw data as scores rather than using the ranks. If this is done here the analysis in Table 1.2 is obtained. Details will be given later. The quadratic effect is now significant at the 5% level, although a smaller p-value may have been obtained with an aligned ranks scale test.

In this text we will use the partition of X^2_P method with raw data as scores in preference to alignment or standardisation. We discuss this point again in the Boos example at the conclusion of section 3.4. The essential point here is that each choice of scores gives a different test, and every conclusion is conditional on the test chosen. Using ranks, natural scores, and the data as scores all may or may not result in similar

conclusions. The user should be aware that each choice casts the data in possibly different lights. Examples of the use of different scores with the same data set are given at the end of section 4.7.

Table 1.2 Partition of X_P^2 for random normal data

Statistic	Degrees of Freedom	Value	Monte Carlo p-value
Linear	1	14.40	< 0.001
Quadratic	1	4.41	0.036
Residual	17	1.19	-
Total (X_P^2)	19	20.000	-

1.4 Outline and Scope

For almost all of the tests considered in this book, and for some other important tests also, we are able to present the data in contingency tables. This provides a unification of these nonparametric tests. Order is usually important for at least one of the categorisations. For the sign test and one-sample goodness of fit tests for categorised data, one-way contingency tables suffice. Two-way contingency tables with fixed marginal totals are appropriate for the median and Kruskal-Wallis tests. Data for the relatively little known Yates' test can be presented in two-way contingency tables with one set of marginal totals fixed and one set random. Special two-way contingency tables that we call Anderson tables are appropriate for the Friedman and Durbin tests. If no decisions about ties and the category totals are made before collecting the data, then data for Spearman's test may be presented in two-way contingency tables with no marginal totals fixed. Marginal distributions of categorical data from matched pairs or from blocks can be given in contingency tables and homogeneity tested using statistics due to Cochran and Stuart, and

extensions of these.

Chapter 2 focuses on ties: how they have been traditionally treated and how different models lead to different treatments. The context is mainly what is for many the first nonparametric test, the sign test. There is also an introduction to the idea of decomposing a test statistic into its components to give a more detailed scrutiny of the data.

Chapters 3 to 8 focus on tests for which the data may be given in two-way contingency tables. The standard tests presented assess mean or correlation departures from the null. Since the traditional correlation is the (1, 1)*th* bivariate moment, all are essentially first order moment tests. We extend all these tests to detect higher moment departures from the null. This gives a detailed and informative scrutiny of the data.

Chapter 9 concerns one- and S-sample smooth tests of goodness of fit. One-sample goodness of fit was covered in detail in Rayner and Best (1989a), and this is an overview and update of that material. The S-sample material is relatively new, having been published only recently.

Chapter 10 includes a discussion of recent work on *partially parametric testing*. This has grown out of the work on one sample smooth tests of goodness of fit. The probability (density) function of the distribution tested for is nested in a rich family of probability (density) functions. If, for example, normality is rejected, a more complicated distribution from this family provides a better fit to the data. It seems reasonable to base inference on such a distribution. This distribution isn't that assumed by the classical parametric methods. However the methods are not nonparametric: they do assume considerable parametric structure. Also included in this chapter are some extensions of some current two-way table work to multiway tables.

The reader interested in methods only will need to refer carefully to the index, as some traditional topics appear in more than one chapter. Thus the sign test is treated in the chapter dealing with ties, and also later, where it is thought that the order in which the treatments are presented may be important. The appropriate model is then a two-way contingency table with fixed marginal totals, and so this is presented in a chapter dealing with the product multinomial model.

Note that while we do not often use traditional parametric models like the normal, we do use models. Since tables involve count data, we

extensively use the multinomial, and in our smooth testing for goodness of fit, quite complicated nested families are employed.

For almost all of the location detecting tests, Page-type tests for ordered alternatives can be constructed, and we usually do consider this. For almost all tests considered, normal and other scores may be used. However, we usually do not consider scores other than *natural* (the row or column number), mid-rank, and the raw data. Other options may be substituted if the user wishes.

The contingency-table approach means tied data are easily handled, and almost exact Monte Carlo p-values can be obtained. Modern statistical packages such as *StatXact* (1995) calculate p-values this way.

For data presented in contingency tables we can calculate Pearson's X_P^2 statistics for testing goodness of fit, independence or homogeneity as appropriate, and their components. As mentioned above, these components detect moment departures from the null, and are in many ways like contrasts used in the analysis of variance. Components are:

- additive, in that they have sum, or sum of squares, the original omnibus statistic;
- easy to use, as they are asymptotically mutually independent and asymptotically have the standard normal distribution;
- easy to interpret, detecting such important alternatives as a mean difference between treatments, or lesser variability between treatments than expected;
- the basis for either the familiar tests mentioned above, or *extensions* of them; these extensions are often new tests, not yet in the statistical literature; and
- the basis for new informative graphical displays.

We have made a major effort to keep the presentation of the technical material at an accessible level. Two main approaches are used. In one a vector of counts is constructed. The Pearson statistic is a quadratic form in this vector with matrix the inverse of the covariance matrix. Obtaining the components and the extensions of the traditional tests requires the diagonalisation of the covariance matrix. The covariance matrices of interest often involve direct or Kronecker products. These are described in the Appendix that gives some of the background

mathematics, statistics and computing used. Apart from direct or Kronecker sums and products, undergraduate algebra and statistics is largely sufficient. Multivariate normality underpins this approach, and a largely theoretical development of this topic is given in the Appendix.

The other approach we use is specifying a model and finding at least one of the asymptotically optimal tests, usually the score or Wald test. These are asymptotically equivalent to the likelihood ratio test, and, again, are described in the Appendix. The advantage of this approach is the tests so constructed are weakly optimal for specifiable models.

For some of the situations we consider the covariance matrix is prohibitively difficult to calculate. For others we have been unable to construct an appropriate model. Readers will usually find the covariance matrix approach easier to understand but mechanically more cumbersome. Our first preference usually is to use the asymptotically optimal tests because of their optimality properties. Every situation is different and flexibility is often rewarded.

By this point it will be apparent that our focus is on testing rather than estimation. This is more a matter of how we have chosen to present our work. Our methods, especially the smooth models, do permit consideration of estimation questions. However we leave such matters for others to develop. In many situations this will be routine.

There is an important caveat we should place on our approach, or more particularly, on how we interpret our components. There is a growing literature, mainly in the area of goodness of fit, on how to interpret the components. The evidence is that caution should be exercised, that a significant first order component may not always indicate a mean shift. But it is highly suggestive, a reasonable working hypothesis that perhaps should be examined before any others. We shall discuss this issue in greater detail later, but we note here that caution in the interpretation of components should be exercised for all tests we examine here.

Much of the material in this book has appeared in print in the statistical literature in the recent past, or is about to. See the references and in particular the papers of Beh, Best, Carolan, Davy, and Rayner. We hope that collecting this material here will enable the reader to better

appreciate the totality of the contingency table approach, and its power, simplicity, and applicability.

1.5 Applications of Nonparametric Methods to Sensory Evaluation

Popular texts for nonparametric methods such as Conover (1998), Hollander and Wolfe (1999), Daniel (1990), Sprent (1993, 1998) and Siegel and Castellan (1988) give many examples of applications of nonparametric methods. In this book we make special mention of important applications to sensory evaluation of foods. One of us (John Best) has been involved with the analysis of sensory data related to food science for over a quarter of a century. Many of the examples used in this book reflect that experience.

Important nonparametric tests used in sensory evaluation are the sign and triangle tests based on the binomial distribution and the Friedman ranking test. The triangle test, involving the ability to choose an odd sample from three, is commonly used in training expert taste-test panels, while the Friedman test is often used in consumer or market research to objectively compare competing brands.

We will discuss both the binomial and Friedman tests in greater detail here than is often done in other texts, and we will also give extensions of these tests that are important in the sensory evaluation of food.

Categorical data are often obtained from sensory experiments or consumer market research. Analysis of variance would usually be employed for the analysis of such data, and while this may be satisfactory for seven- or nine-point *hedonic* or *liking* scales, its use for three- or five-point *just right* or *likely to purchase* data is more questionable. We will discuss alternative analyses for such data with very few categories. These analyses also have the advantage of allowing comparisons of *distributions*, not just means, which is all the analysis of variance gives. Comparison of distributions is important in identifying market segmentation in consumer research applications.

2
Modelling Ties

2.1 Introduction

Many nonparametric tests make the assumption that the data are continuous, and hence the probability of a tie is zero. Difficulties then arise when ties do occur, in which case ad hoc adjustments have to be made.

Gibbons and Chakraborti (1992, p. 135) gave seven possible treatments of ties for rank tests in general. A quick review of nonparametric texts indicates that the most commonly recommended procedure seems to be ignoring the tied observations; see for example Lehmann (1975, p. 123 and p. 144), Siegel and Castellan (1988) and Daniel (1990). Ignoring ties is an approach used in *StatXact* (1995), arguably a most comprehensive nonparametric statistics package. Consider a simple example: suppose that two chess players play nine matches, of which five are won by player A, none by B, and there are four draws. In testing if both players are equally matched, the draws are treated as ties, and clearly have positive probability. Moreover they are meaningful in comparing the players.

Almost none of the recommendations we have sighted are based on a model. In the comparison of the chess players, some of the many possible ad hoc procedures are:

- to ignore the tied observations, with p-value 0.0625 for the chess data above;
- distribute the ties evenly between A and B, giving a 7:2 preference for A, with p-value $2P\{X \leq 2 \mid X \text{ is binomial } (9, 0.5)\} = 0.1796$;
- distribute the ties randomly between A and B, giving five different possible p-values;
- distribute the probability of ties evenly between A and B, giving a

p-value of

$$2\{P[X = 0 \mid X \text{ is binomial } (9, 0.5)] + 0.5P(X = 1, 2, 3, \text{ or } 4)\} = 0.502.$$

Such wide variation, where quite conflicting assessments may be concluded from the one data set, is unacceptable.

This chapter focuses mainly on dealing with ties in the sign test, and is based on Rayner and Best (1999). We also briefly consider, in section 2.6, the partitioning of statistics into components, and use this machinery to further develop the modelling of ties. Partitioning is central to the approach of this book, and it is helpful to introduce the ideas early.

Central to our approach to ties is that underlying any analysis is a model. The analysis, and this includes the treatment of ties, depends heavily on that model. We demonstrate this in relation to the sign test.

2.2 The Sign Test and Ties

The sign test uses the number N_+ of positive counts and the number N_- of negative counts to test the null hypothesis H: $p_+ = p_-$ that the probability of a positive count is equal to the probability of a negative count, against both one and two sided alternatives. In practice the number N_0 of ties/undecideds/zeros is often positive, reflecting a positive probability p_0 of zeros.

If no zeros are observed, they are often assumed to have zero probability. Then p-values can be calculated using the binomial distribution for N_+ with parameters n, the sample size, and 0.5. Equivalently, $N_+ - N_-$ may be used as a test statistic. If zeros are observed, p-values should be calculated using the trinomial distribution, although in practice this is rarely done. To demonstrate, note that in the trinomial model there are $N_+ + N_0 + N_- = n$ observations in all. To test the null hypothesis H_0: $p_+ = p_-$ against a two sided alternative K: $p_+ \neq p_-$. We use the statistic $N_+ - N_-$, rejecting in a region R of both large and small values of $N_+ - N_-$. The p-value is then

$$\sup_{0 < p < 1} \sum_{(n_+, n_0, n_-) \in R} \frac{n!}{n_+!\, n_0!\, n_-!} \left(\frac{1 - p}{2}\right)^{n_+} p^{n_0} \left(\frac{1 - p}{2}\right)^{n_-}.$$

For the example above where five judges prefer A to B with four ties, this supremum occurs when p = 0, and is 0.1797.

An approximate p-value could be calculated by summing appropriate trinomial probabilities with p_+, p_0 and p_- replaced by the maximum likelihood estimators of $p_+ = p_- = (n_1 + n_3)/(2n)$ and $p_0 = n_2/n$. Again for the example above, the approximate p-value for a two sided alternative is 0.041. To see this, note that we observe $N_1 - N_3 = 5$ and need to first calculate $P(N_1 - N_3 \geq 5)$. Using multinomial probabilities with $p_+ = p_- = 5/18$ and $p_0 = 4/9$, the observed cell proportions, this is found to be 0.0205. This involved calculations such as

$$\begin{aligned} P(N_1 - N_3 = 5) &= P(N_1 = 5, N_2 = 4, N_3 = 0) + \\ &\quad + P(N_1 = 6, N_2 = 2, N_3 = 1) + P(N_1 = 7, N_2 = 0, N_3 = 2) \\ &= 0.014837 \end{aligned}$$

in which, for example,

$$P(N_1 = 5, N_2 = 4, N_3 = 0) = \frac{9!}{5!\, 4!\, 0!} \left(\frac{5}{18}\right)^5 \left(\frac{4}{9}\right)^4 \left(\frac{5}{18}\right)^0 = 0.0081307.$$

Since the sign test is often used for its speed of application, it is no wonder that this sort of approach hasn't become popular.

The procedure favoured by Coakley and Heise (1996) is the asymptotic uniformly most powerful nonrandomised test due to Putter (1955), which is based on

$$S = \frac{N_+ - N_-}{\sqrt{N_+ + N_-}}.$$

The statistic S is assumed to asymptotically have the standard normal distribution. Coakley and Heise's (1996) recommendation is based mainly on a size and power study, comparing the test based on this statistic with

other tests in the literature.

Colleagues have indicated that their intuition suggests that data consisting of 10 positives, zero ties and zero negatives convey quite different information to data consisting of 10 positives, 90 ties and zero negatives. Yet the test based on S comes to the same conclusion for both sets of data. The problem is that the probability of a zero is, in a sense, unmodelled, so that the zeros give no information about the null hypothesis. In practice, how to model the zeros will depend on the situation. We now explore some models that deal with zeros.

2.3 Modelling Partitioned Ties in the Sign Test

In opinion polls, it is very common for pollsters to press the undecideds, acknowledging that they may not wish to commit to either a positive or negative response to the question asked, but insisting on an indication of their *leaning* one way or the other. The tied responses are partitioned into leaning positive and leaning negative. One way of dealing with the outcomes that arise in this way is to score the responses, assigning a score +K for a positive, +1 for a positive leaning, -1 for a negative leaning, and -K for a negative. The constant K could be assigned a value such as 3, or 5.

Instead of assigning an arbitrary score in this way, we now consider some models for such data. We give several models and, mostly without derivation, one of the corresponding asymptotically optimal test statistics. For a review of asymptotically optimal statistics see Appendix A.5, which is based on Rayner (1997). Here we have a multinomial with four cells and counts N_1, N_2, N_3 and N_4, counting respectively the number of outright positive responses, the number leaning to positive, the number leaning to negative, and the number of outright negative responses. The corresponding cell probabilities are p_1, p_2, p_3 and p_4, and there are $N_1 + N_2 + N_3 + N_4 = n$ observations in all.

Consider the model that assigns

$$p_1 = (0.5 + \theta)(1 - \beta_1),\ p_2 = (0.5 + \theta)\beta_1,\ p_3 = (0.5 - \theta)\beta_2$$
$$\text{and } p_4 = (0.5 - \theta)(1 - \beta_2).$$

This reflects a *leakage* of a proportion β_1 of the positives to leaning positive, and β_2 of the negatives to leaning negative. A Wald-type statistic for testing H_0: $\theta = 0$ against K: $\theta \neq 0$ for this model is based on $(N_1 + N_2 - N_3 - N_4)^2/n$. This is a 1-1 function of $N_1 + N_2$.

This can be modified by taking $\beta_1 = \beta_2 = \beta$: leakages occur at the same rate in both directions. It follows that $p_1/p_4 = p_2/p_3$: the ratio of outright positives to outright negatives is the same as the ratio of positive leanings to negative leanings. Thus ties are modelled as occurring in the same proportion as untied observations. A Wald-type statistic for testing H_0: $\theta = 0$ against K: $\theta \neq 0$ is again $(N_1 + N_2 - N_3 - N_4)^2/n$, or, equivalently, $N_1 + N_2$. This could be called the *proportional allocation model*.

Because of its importance in our discussion, we pause to prove this result. Our model assigns cell probabilities

$$(0.5 + \theta)(1 - \beta),\ (0.5 + \theta)\beta,\ (0.5 - \theta)\beta,\ (0.5 - \theta)(1 - \beta).$$

In testing H_0 against K, β is an unspecified nuisance parameter. For this model the logarithm of the likelihood is

$$(n_1 + n_2) \log (0.5 + \theta) + (n_3 + n_4) \log (0.5 - \theta) + (n_2 + n_3) \log \beta + (n_1 + n_4) \log (1 - \beta),$$

so that

$$\frac{\partial \log L}{\partial \theta} = \frac{n_1 + n_2}{0.5 + \theta} - \frac{n_3 + n_4}{0.5 - \theta}.$$

Solving for θ gives the maximum likelihood estimator as

$$\hat{\theta} = (N_1 + N_2)/n - 0.5.$$

After a little calculation we find

$$\mathrm{var}(\hat{\theta}) = (p_1 + p_2)(p_3 + p_4)/n = (0.25 - \theta^2)/n.$$

The usual Wald test statistic is $\hat{\theta}^2/\widehat{\mathrm{var}(\hat{\theta})}$, where $\mathrm{var}(\hat{\theta})$ is estimated under the full model. Alternative Wald-type test statistics replace the parameters in the estimated variance by their value or estimated value under the null hypothesis. Clearly when $\theta = 0$, $\mathrm{var}(\hat{\theta}) = 1/(4n)$ and the Wald-type statistic is

$$4n\hat{\theta}^2 = \{(N_1 + N_2) - (N_3 + N_4)\}^2/n,$$

which is a one-one function of $N_1 + N_2$.

The proportional allocation model is the basis of a very appealing treatment of ties. Suppose that only the total number $N_2 + N_3$ of ties is reported. By assuming the proportional allocation model we can deduce N_2 and hence N_3. Since $p_1/p_4 = p_2/p_3$ it follows that $p_2 = p_1(p_2 + p_3)/(p_1 + p_4)$ and, by maximum likelihood estimation, $\hat{p}_2 = \hat{p}_1(\hat{p}_2 + \hat{p}_3)/(\hat{p}_1 + \hat{p}_4)$ or

$$N_2 = N_1(N_2 + N_3)/(N_1 + N_4).$$

Thus the ties are apportioned in the same ratio as the outright outcomes. The optimal procedure is based on S with this adjustment.

An alternative modification is to assume an *equal allocation model*: $p_2 = p_3$. Again the Wald-type test statistic is S, but maximum likelihood allows us to deduce that $N_2 = N_3$, both being half of the ties. Leakages occur at different rates, but in such a way that the ties clearly support the null hypothesis.

Reconsider the example at the end of the first section, with 10 positives, 90 ties and zero negatives. Ignoring ties as Coakley and Heise (1996) recommend, gives $S = \sqrt{10}$ and two-sided p-value 0.000. Proportional allocation gives S = 10, and even stronger rejection of the null hypothesis. Equal allocation gives S = 1, with two-sided p-value 0.317, leading to a quite different conclusion.

To see the variety of optimal tests corresponding to the variety of models, take

$$p_1 = (0.5 + \theta)(1 - \beta),\ p_2 = [1 - r\,(0.5 - \theta)]\beta,\ p_3 = (0.5 - \theta)r\beta$$
and $p_4 = (0.5 - \theta)(1 - \beta)$.

The constant r is assumed known; otherwise the model is over-parametrised. Relative to the previous model, some of the leaked negatives remain leaning negative, and some go to leaning positive. Through the variance of the maximum likelihood estimator, the Wald-type statistic involves the number of tied observations and through the efficient score, $N_1 + aN_2$. Here $a = (1 - r)/(2 - r)$. In terms of the model with $\theta = 0$ the restriction $0 < r < 1$ is reasonable, and then $0 < a < 0.5$.

Coakley and Heise's (1996) assessment was based on assessing size and power. We have proposed a number of models, and for each model we test using one of the asymptotically optimal test statistics. None of these tests is optimal for all models. For a given model the asymptotically optimal test will be superior to a competitor test for a different model. Coakley and Heise's (1996) recommended test statistic cannot compete with the collection of asymptotically optimal tests for all models.

2.4 Modelling Unpartitioned Ties in the Sign Test

We now consider the more usual situation for the sign test, of having just three outcomes: positive, tie or zero, and negative. The counts for these outcomes, respectively N_+, N_0 and N_-, follow a trinomial distribution with cell probabilities p_1, p_2 and p_3. Again there are $N_+ + N_0 + N_- = n$ observations in all.

First suppose

$$p_+ = (0.5 + \theta)(1 - \beta),\ p_0 = \beta, \text{ and } p_- = (0.5 - \theta)(1 - \beta).$$

The probability of a tie is just a nuisance parameter, and isn't central to the hypotheses of interest. The score statistic is Coakley and Heise's (1996) S, which, as noted before, does not involve the number of ties N_0.

Now take

$$p_+ = (0.5 + \theta)(1 - \beta),\ p_0 = (0.5 + \theta)\beta, \text{ and } p_- = 0.5 - \theta.$$

Ties are modelled as leakage from the positives. The score statistic is a 1-1 function of $N_+ + N_0$, or, equivalently, of N_-. Next, if ties are modelled as leakage from the negatives, with

$$p_+ = 0.5 + \theta,\ p_0 = (0.5 - \theta)\beta, \text{ and } p_- = (0.5 - \theta)(1 - \beta),$$

the score statistic is a 1-1 function of N_+.

Consider an alternative model that takes

$$p_+ = (0.5 + \theta)(1 - \beta) + (0.5 - \theta)(1 - s)\beta,$$
$$p_0 = (0.5 + \theta)r\beta + (0.5 - \theta)s\beta, \text{ and}$$
$$p_- = (0.5 + \theta)(1 - r)\beta + (0.5 - \theta)(1 - \beta).$$

This features leakages from the positives to the ties and negatives, and also leakages from the negatives to the ties and positives. If r = s, the score statistic is again S, but otherwise, it is quite complicated, with the efficient score a linear combination of N_+, N_0 and N_-, with the coefficients depending on r and s.

***Wholesale Comet* Example**. The first data set in Coakley and Heise's (1996) example relates to *Wholesale Comet*, a type of goldfish. In judging 59 such goldfish for symmetry, $N_+ = 4$ were judged asymmetric or 'positive', $N_0 = 54$ were judged symmetric or 'ties', and $N_- = 1$ was judged asymmetric or 'negative'. The number of ties is large, and any reasonable model would estimate p_0 as being quite substantial, but that is not of concern to the inference here. Coakley and Heise's recommended test statistic takes the value $(4 - 1)/\sqrt{(4 + 1)} = 1.342$, and referring this to the standard normal tables gives the reported two-sided p-value of 0.180.

The proportional allocation model distributes the 54 ties in the ratio of 4:1, giving adjusted N_+ and N_0 values of 47.2 and 11.8 and an adjusted S value of $(47.2 - 11.8)/\sqrt{(47.2 + 11.8)} = 4.609$. If there is complete leakage to the positives or to the negatives, the adjusted N_+ values are 58 and 4

respectively, with corresponding S values 7.421 and -6.640 respectively. In all these cases the p-values are zero, reversing the conclusion above.

Which of these p-values, if any, is to be believed? This depends on what is the most appropriate model, and that is perhaps best answered by those who posed the original question. It seems to us that this case relates to symmetry, and the large number of ties, reflecting symmetry, supports the null hypothesis $p_+ = p_-$. The equal allocation model seems a most reasonable approach. This gives S = 0.391 and two-sided p-value 0.696.

Recommendation. Many different ways of handling ties have been suggested in the literature; in addition to the references already cited see Wittkowski (1989, 1998). A colleague has identified measurement limitation ties (due to rounding the data), ties induced by grouping, uninformative ties (not modelled as important to the current inference) and indecisive ties (as in opinion polling). It is certainly possible to have different sorts of ties in the one data set. The Coakley and Heise (1996) recommendation is appropriate if we assume all the ties are uninformative. However models can be constructed so that the probability of a tie *is* central to the hypotheses being tested, and is thus informative. Then the number of ties is relevant. The treatment of ties depends very strongly on just what we are prepared to assume. The best choice is what is most appropriate for the situation. The treatment of the ties should always depend on the underlying model rather than a uniform ad hoc approach.

2.5 McNemar's Test

Voting Intention Example. McNemar (1947) derived a test for data such as the following, given by Sprent (1998, Example 6.3, p. 144). The data are usually presented in a table. Here 200 potential voters in a referendum are asked their voting intentions (vote 'yes' or 'no'), both before and after listening to a broadcast. The question is, has the broadcast influenced the voting intentions of these 200 voters?

The data are multinomial with four cells, that may be labelled

A and B, A and not B, not A and B, not A and not B.

Voting Intentions of 200 Voters

		Pre-broadcast intention	
		Yes	No
Post-broadcast intention	Yes	93	12
	No	6	89

As Zar (1984) points out, the aim is to test, in standard notation, $p_{1.} = p_{.1}$, or P(A) = P(B). Clearly P(A and B) = β_1, say, is common to both $p_{1.}$ and $p_{.1}$, and we put, under the null hypothesis, P(A and not B) = P(not A and B) = β_2. We now introduce the parameter θ to model increased frequency in the absence of A and presence of B. The cell probabilities are now given by β_1, $-\theta + \beta_2$, $\theta + \beta_2$, and $1 - \beta_1 - 2\beta_2$. This notation is given in the table following.

Cell Probabilities for McNemar's Test

	A	not A	Total
B	β_1	$\theta + \beta_2$	$p_{1.}$
not B	$-\theta + \beta_2$	$1 - \beta_1 - 2\beta_2$	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

Routine calculations show that the score test statistic of H_0: $\theta = 0$ against K: $\theta \neq 0$ is

$$S^2 = (N_2 - N_3)^2/(N_2 + N_3),$$

which is the statistic recommended by McNemar (1947). This test is a version of the sign test for grouped data. The LOGXACT Users Manual (1996, section 8.6) shows how to obtain S^2 using logistic regression. In later chapters we will also note how similar parametric and nonparametric tests can be.

Sprent (1998, p. 144) applies S^2 to the voting intention data. He uses the exact binomial distribution associated with S rather than the asymptotic χ^2_1 distribution associated with S^2. At commonly used levels he accepts the null hypothesis that switches are equally likely to be either way.

Our presentation of data for McNemar's test, and the usual one, both interpret the diagonal terms as uninformative ties. Hence the optimal treatment ignores them. Ignoring ties is *not* our preferred philosophy. However the model is not to be taken lightly. So if, in the proportional allocation model, we have 10 tied and 90 untied observations, proportional allocation may well be reasonable. But if we have 90 tied and 10 untied observations, proportional allocation is a *very* strong assumption, and not one to be adopted lightly. For McNemar data, modelling the diagonal cells in a manner similar to proportional allocation does not seem justified for the data sets we have investigated. That may not be universally true. But when we can find no other way to model the data, McNemar's test is a score test and hence is weakly optimal, and is thus to be recommended.

Note that if we were interested in a different hypotheses, say testing H: P(not A and B | B) = P(A and not B | A) against K: not H, then the ties may well be informative. We have not yet investigated this situation. There are certainly many other questions of interest in the Voting Intention Example.

2.6 Partitioning into Components

In the preceding sections we have discussed ties associated with binomial data. This discussion will now be extended to multinomial data. But first we look briefly at the classical Pearson goodness of fit test based on the statistic X^2_P. This statistic may be partitioned into useful and informative

components. In Appendix A.1.4 we discuss components and contrasts; both

- are at least asymptotically mutually independent,
- have sum, or sum of squares, the original omnibus statistic,
- have, at least asymptotically, convenient distributions, and
- have an immediate and useful interpretation.

The parallel with the partitioning of a set into disjoint sets is immediate. Partitioning a statistic into components isn't helpful unless the components have both a convenient distribution and a ready interpretation. We now demonstrate with an example.

Cordials Example. In a market research trial 40 consumers were asked for their colour preference of five orange-mango cordial drinks. The five drinks had different orange colours ranging from *pale* orange to *deep* orange. Arranged in order of increasing orange colour the preference counts were 5, 10, 11, 10, 4. In testing if the five cells are equiprobable, Pearson's X_P^2 takes the value 5.25 with χ_4^2 and Monte Carlo p-values 0.26 and 0.27 respectively. There is no evidence of a difference between cells. However X_P^2 provides an omnibus test, simultaneously assessing departures from the null hypothesis in four (the degrees of freedom) dimensions. When using omnibus tests, significant departures in one or two dimensions may be masked by non-departures in the remaining dimensions. We shall see that this is the case here.

To partition X_P^2, first define the components V_r by

$$V_r = \sum_{j=1}^{m} N_j g_r(x_j)/\sqrt{n}\ ,\ r = 1, \ldots , m - 1,$$

in which $\{g_r(x_j)\}$ are the orthogonal polynomials of the random variable that takes the values x_j with probabilities p_j, j = 1, ... , m. In A.3, for general m, $g_0(x_j)$, $g_1(x_j)$ and $g_2(x_j)$ are given explicitly. It was shown in Rayner and Best (1989a, Corollary 5.1.1) that

$$V_1^2 + \ldots + V_{m-1}^2 = X_P^2$$

and that asymptotically the V_r are mutually independent and

asymptotically each has the standard normal distribution. Note that we sometimes call the V_r^2, r = 1, ... , m the components of X_P^2.

Interpretation of the Components. By its definition, an observed value v_r of V_r is proportional to the sample covariance, and hence the sample correlation between the observed counts n_j and $g_r(x_j)$. It follows that a large value of v_r will occur when the n_r are almost linearly related to the $g_r(x_j)$. Since $g_1(x_j)$ is linearly related to x_j, if the x_j are 1, ... , m, v_1 is proportional to the correlation between the counts n_1, ... , n_m and the class scores 1, ... , m. It is then reasonable to interpret a large v_1 as indicating the counts n_j increase (or decrease) with the scores x_j = j, so the mean is more (or less) than the null uniform distribution would suggest. Similarly v_2 is proportional to the correlation between the counts n_j and the scores 1^2, ... , m^2, so that a large v_2 indicates greater (or less) variability in the n_j than the null distribution suggests. The direction of these differences is best determined by assessing the data.

Cordials Example Continued. We calculate V_1 and V_2, noting that $g_1(j)$ and $g_2(j)$ are given explicitly by

$$\sqrt{2}\ g_1(j) = -2, -1, 0, 1, 2 \text{ for } j = 1, \ldots, 5, \text{ and}$$
$$\sqrt{(14/5)}\ g_2(j) = 2, -1, -2, -1, 2 \text{ for } j = 1, \ldots, 5.$$

We find v_1 = -0.2236 with two tailed Monte Carlo p-value 0.87 and v_2 = -2.2678 with two tailed Monte Carlo p-value 0.02. Since the sum of the squares of the components is X_P^2, a residual may be calculated via

$$v_3^2 + v_4^2 = 5.25 - 0.2236^2 - 2.2678^2 = 0.0571;$$

without further calculation this clearly has a large (nonsignificant) p-value.

We conclude there is no departure in mean (or location) but there is a departure in variance (or dispersion) from that expected for a discrete uniform distribution. There is a significant lack of preference for pale orange or deep orange coloured cordial while at the same time there is a significant preference for mid-range orange colours. There are no other

effects as the residual is nonsignificant.

2.7 Ties in a Multinomial Test

We will now consider the effect of ties in the cordials example. Suppose that some judges could not decide between some of the adjacent categories. The counts were 5, **2**, 10, **0**, 11, **1**, 10, **1**, 4, where the counts in bold are ties between adjacent categories. If we use a smooth model that makes no assumptions about the tied categories, an optimal test gives the values and p-values of the components V_1 and V_2, and of X^2_P that are unchanged from those above, and hence yield the same conclusions.

To show this we need some machinery, and not unnaturally, we turn to smooth models for the multinomial proposed in Rayner and Best (1989a). Under the null hypothesis the cell probabilities depend on a q by 1 vector of nuisance parameters β, and are denoted by $p_j = p_j(\beta)$, j = 1, ... , m. The order k smooth alternative gives cell probabilities

$$\pi_j = C(\theta, \beta) \exp\{\sum_{i=1}^{k} \theta_i \, h_{ij}(\beta)\} \, p_j(\beta), \; j = 1, \ldots, m,$$

where the k by m matrix $H = (h_{ij}(\beta))$, whose elements are constants, may be chosen to achieve a variety of tests. The order k must satisfy $k \leq m - 1 - q$.

Suppose we have cell counts $N = (N_1, \ldots, N_m)$, there are $n = N_1 + \ldots + N_m$ counts in all, $p = (p_1, \ldots, p_m)$, and $\hat{\beta}$ is the maximum likelihood estimator of β under the null hypothesis.

Theorem 2.1. The score statistic for testing $H_0: \theta = \theta_0$ against $K: \theta \neq \theta_0$ is

$$\hat{S}_k = (N - n\hat{p})^T \, \hat{H}^T \, \hat{\Sigma}^{-1} \, \hat{H} \, (N - n\hat{p})/n \, ,$$

in which $\hat{p} = (p_j(\hat{\beta}))$, $\hat{H} = (h_{ij}(\hat{\beta}))$ and the elements of Σ, given below, are

also evaluated under the null hypothesis. The asymptotic covariance matrix of $H(N - np)/\sqrt{n}$ is Σ, which is given by

$$\Sigma = H\{D - pp^T - W^T(WD^{-1}W^T)^{-1}W\}H^T, \text{ in which}$$
$$D = \text{diag}(p_1, \ldots, p_m) \text{ and } W = (\partial p_j/\partial \beta_u).$$

Proof. See Rayner and Best (1989a, Theorem 7.1.1) and its corollaries.

We now show that, for certain models within this class of smooth tests, the score tests ignore the number of ties.

Suppose that we wish to test for equiprobability amongst certain categories, but that ties may occur between any pair of categories. Suppose that there are k categories with positive probabilities of ties, and m categories in all. The categories are numbered so that we wish to test H: $p_1 = \ldots = p_{(m-k)} = \rho$ say, against K: not H, with $p_{(m-k+1)} = \beta_1, \ldots, p_m = \beta_k$ nuisance parameters. If we write 1_a for the a by 1 vector with every element 1, I_a for the a by a identity matrix $\beta = (\beta_1, \ldots, \beta_k)^T$ and $\Lambda = \text{diag}(\beta_1, \ldots, \beta_k)$, we find

$$W = \left(-\frac{1}{m-k} 1_k 1_{(m-k)}^T \mid 1_k \right)$$

$$WD^{-1}W^T = \frac{1}{(m-k)\rho} 1_k 1_k^T + \text{diag}\left(\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_k}\right)$$

$$(WD^{-1}W^T)^{-1} = \text{diag}(\beta_1, \ldots, \beta_k) - (\beta_i\beta_j) = \Lambda - \beta\beta^T$$

$$W^T(WD^{-1}W^T)^{-1}W =$$
$$\begin{pmatrix} \frac{1}{(m-k)^2} 1_{(m-k)} 1_k^T (\Lambda - \beta\beta^T) 1_k 1_{(m-k)}^T & -\frac{1}{(m-k)} 1_{(m-k)} 1_k^T (\Lambda - \beta\beta^T) \\ -\frac{1}{(m-k)} (\Lambda - \beta\beta^T) 1_k 1_{(m-k)}^T) & \Lambda - \beta\beta^T \end{pmatrix},$$

and, if "$\oplus$" means direct sum and 0_k the k by k matrix of zeros,

$$D - pp^T - W^T(WD^{-1}W^T)^{-1}W = \rho\left\{I_{(m-k)} - \frac{1}{(m-k)} 1_{(m-k)} 1_{(m-k)}^T\right\} \oplus 0_k.$$

It is routine to show that $\rho\left\{I_{(m-k)} - \frac{1}{(m-k)} 1_{(m-k)} 1_{(m-k)}^T\right\}$ has rank (m - k - 1), with all nonzero eigenvalues ρ and with corresponding eigenvectors orthonormal to $1_{(m-k)}$. The matrix H is at our disposal, but must be chosen so that $\Sigma = H\{D - pp^T - W^T(WD^{-1}W^T)^{-1}W\}H^T$ has an inverse. A choice that simplifies the calculations is to take H to be (m - k - 1) by m with rows $h_1^T, \ldots, h_{(m-k-1)}^T$ orthonormal to the row vector with first m - k elements 1, and the remaining k elements arbitrary. Then we find

$$\Sigma = \rho I_{(m-k)}.$$

If we note that

$$\hat{\rho} = (1 - \hat{p}_{(m-k+1)} - \ldots - \hat{p}_m)/(m-k) = (N_1 + \ldots + N_{(m-k)})/(m-k),$$

the score test statistic is

$$\hat{S}_{(m-k-1)} = \hat{V}_1^2 + \ldots + \hat{V}_{(m-k-1)}^2 \text{ in which } \hat{V}_r = h_r^T N/\sqrt{(n\hat{\rho})}.$$

Since the final k elements of each h_r are arbitrary, they may be chosen to be zero, thereby ignoring the counts for the numbers of ties. We note that with m = 3 and k = 1, the situation for the sign test with ties, this score statistic is the statistic $S = \frac{N_+ - N_-}{\sqrt{N_+ + N_-}}$ due to Putter (1955) that we discussed in section 2.2.

The $\hat{V}_r$ are components of $\hat{S}_{(m-k-1)}$. By the Central Limit Theorem $(\hat{V}_r)$ is asymptotically (m - k - 1)-variate normal, and since the covariance matrix $\Sigma = \rho I_{(m-k)}$ is diagonal, the uncorrelated $\hat{V}_r$ are asymptotically

independent and asymptotically standard normal. If the h_r are chosen to be rows of a Helmert matrix (see Lancaster, 1965), and hence

$$(1, 1, 0, \ldots, 0)/\sqrt{2},$$
$$(1, 1, -2, 0, \ldots, 0)/\sqrt{6},$$
$$(1, 1, 1, -3, 0, \ldots, 0)/\sqrt{12},$$

and so on, the components are contrasts between the average of the first r categories and the $(r + 1)th$. This is a very useful decomposition of the score statistic.

A reasonable interpretation of these results is that if the tied categories are ignored in the model, they are ignored in some asymptotically optimal tests based on a smooth model. On reflection this is quite reasonable. Although a tied category may have many counts, if it cannot be modelled as being relevant to the hypotheses of interest, an asymptotically optimal test ignores the tied counts.

2.8 Ties When Testing for Independence

In any two-way contingency table, neither, one, or both of the marginal totals may be known before collecting the data. Different models are appropriate in each case. It seems inappropriate to us to calculate permutation test p-values based on both marginal totals being known before collecting the data, if in fact neither or only one marginal total was known.

Suppose for now that no marginal totals are known before collecting the data, so that the model is multinomial (see, for example, Chapter 8). We may test for independence by calculating one of the asymptotically optimal tests. If this is done the marginal probabilities enter the calculations as nuisance parameters. Those categories that arise because of tying are treated exactly the same as the remaining categories. In this situation there is no need to model the ties specifically. In a sense, the ties model is embedded in the independence model. The advice of section 2.5, to model ties, needs to be addressed for each situation we model. Sometimes nothing specific needs to be done.

If components are to be calculated, then it will be necessary to assign scores for all categories, and the score assigned could reflect the fact that a particular category arises as a result of a tie (1, 2, 2.5, 3) rather than as a continuation of the ordered category (1, 2, 3, 4). In practice the analysis often varies little with the assigned score, and if this is the case, we will usually omit comment. Midrank scores are a common choice of category assigned scores. Again, this assignment reflects a choice of model, and a means of modelling ties.

3
Tests on One-Way Layout Data: Extensions to the Median and Kruskal-Wallis Tests

3.1 Introduction

The idea of decomposing a test into orthogonal contrasts, as in the analysis of variance, has long been appreciated by statisticians as a way of making hypothesis tests more informative. We have done this in our smooth goodness of fit work (see Rayner and Best, 1989a), and now, over the next several chapters, we do this in a variety of settings relevant to nonparametric testing. We use components. The definitions change with the setting, but all have the properties previously mentioned. These are that they are at least asymptotically mutually independent, they have convenient distributions, they reconstitute the original omnibus statistic, and they have immediate and useful interpretations. They therefore provide powerful directional tests and permit a convenient and informative scrutiny of the data. In addition, in each setting, the first one or two components of our omnibus statistic are familiar nonparametric test statistics, such as the Kruskal-Wallis, Friedman, and Spearman's rho statistics. The remaining components can be viewed as extensions of these familiar statistics.

Data for the tests we discuss subsequently may be presented in contingency tables and assessed using Pearson's X_P^2. In this chapter we consider the Kruskal-Wallis test both with and without ties and a generalisation of the median test. Data for these tests may be presented in the form of two-way tables with fixed marginal totals. We derive the covariance matrix of entries in such tables and then partition (a multiple of) X_P^2 into components that detect location and higher moment differences

between rows. Chapter 4 deals with two-way tables where only one margin is fixed.

For Kruskal-Wallis-type data when there are no ties, the location component of an appropriate scalar multiple of Pearson's X^2_P statistic is the Kruskal-Wallis test. Our approach enables us to generalise to when there are ties, and to when there is a fixed number of categories and a large number of observations. For the generalisation of the well known median test, we show that the location detecting first component of X^2_P reduces to the usual median test statistic when there are only two categories. Using more categories allows components other than this location component to be calculated. These additional components, that detect dispersion and higher moment effects, are not available when using the usual median test.

The structure of this chapter is as follows. In the next section the model for two-way contingency tables with fixed marginal totals is given, and Pearson's X^2_P is derived as a test statistic for the null hypothesis of like rows. In section 3.3 a multiple of X^2_P is partitioned into components. The material in section 3.2 and 3.3 will be familiar to many readers, but is necessary background for the new work. In section 3.4 it is shown that when there are no ties the first component is the usual Kruskal-Wallis statistic. The non-location detecting components are our extensions. Section 3.5 generalises the treatment to deal with ties; tied data are also considered in Chapter 4. Section 3.6 introduces a generalisation of the usual median X^2 test, which is thus identified as a location detecting test; the extensions permit dispersion and other effects to be detected.

The reader more interested in applications or examples may wish to go straight to these at the ends of sections 3.3 and 3.4.

3.2 A Model and Pearson's X^2 Test

Suppose we have a two-way table of counts N_{ij}, with i = 1, ... , r and j = 1, ... , c. The row and column totals, respectively $n_{i.}$, i = 1, ... , r and $n_{.j}$, j = 1, ... , c, are known constants. Under the null hypothesis of simple random sampling, the likelihood was given by Roy and Mitra (1956) as

$$\left\{\prod_{i=1}^{r} n_{i.}!\right\}\left\{\prod_{j=1}^{c} n_{.j}!\right\} \Big/ \left\{n_{..}! \prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!\right\},$$

in which $n_{..} = \sum_i n_{i.} = \sum_j n_{.j}$ is the grand total of the observations. The models for tables with just one set of marginal totals fixed, or only the grand total fixed, are quite different from our model in which all row and column totals are fixed. See Lancaster (1969, chapter XI section 2, pp. 212-217). This likelihood can be expressed as a product of extended or multivariate hypergeometric probability functions:

$$\prod_{i=2}^{r} \left\{\left[\prod_{j=1}^{c} n_{1j} + \ldots + n_{ij} C_{n_{ij}}\right] / n_{1.} + \ldots + n_{i.} C_{n_{i.}}\right\}.$$

To find moments of the N_{ij}, expectations may be taken with respect to the distribution of the second row conditional on knowledge of the column sums of the first two rows, then conditional on the column sums of the first three rows, and so on. It suffices to know the moments of the extended hypergeometric distribution. We use this approach to find the means, variances and covariances of the N_{ij}.

Theorem 3.1. Write $N_i = (N_{i1}, \ldots, N_{ic})^T$, $i = 1, \ldots, r$. The expectations of the N_{ij} are given by

$$E[N_{ij}] = n_{i.}\, n_{.j}/n_{..},\ i = 1, \ldots, r \text{ and } j = 1, \ldots, c,$$

and the joint covariance matrix of N_i and N_j is, for $i \neq j$,

$$\operatorname{cov}(N_i, N_j) = -\frac{n_{i.}}{n_{..}}\frac{n_{j.}}{n_{..}}\left\{\operatorname{diag}\left(\frac{n_{.r}\, n_{..}}{(n_{..} - 1)}\right) - \left(\frac{n_{.r}\, n_{.s}}{n_{..} - 1}\right)\right\}.$$

Write $f_j = n_{.j}/n_{..}$, $j = 1, \ldots, c$, and $R = \operatorname{diag}\left(\frac{n_{.r}\, n_{..}}{(n_{..} - 1)}\right) - \left(\frac{n_{.r}\, n_{.s}}{n_{..} - 1}\right)$. Then the covariance matrix of N_i is

$$\text{cov}(N_i) = -\sum_{i \neq j} \text{cov}\,(N_i, N_j) = \sum_{i \neq j} f_i f_j\, R = f_i\,(1 - f_i)\, R.$$

Proof. To find $E[N_{21}]$, take $E[N_{21} \mid N_{1j} + N_{2j},\ j = 1, \ldots, c]$, then the conditional expectation of this expression with the sum of the first three columns being known, and so on. The successive expectations are

$$\begin{aligned}
&n_{2.}(N_{11} + N_{21})/(n_{1.} + n_{2.}),\\
&\{n_{2.}/(n_{1.} + n_{2.})\}*\{(n_{1.} + n_{2.})(N_{11} + N_{21} + N_{31})/(n_{1.} + n_{2.} + n_{3.})\},\\
&\qquad\qquad\vdots\\
&\{n_{2.}/(n_{1.} + n_{2.})\}*\{(n_{1.} + n_{2.})/(n_{1.} + n_{2.} + n_{3.})\}* \ldots\\
&\qquad\qquad\qquad\qquad *\{(n_{1.} + \ldots + n_{(c-1).})/(n_{1.} + \ldots + n_{c.})\}*\{n_{.1}\}
\end{aligned}$$

$$= n_{2.}*n_{.1}/n_{..}.$$

By symmetry $E[N_{ij}] = n_{i.}*n_{.j}/n_{..}$, $i = 2, \ldots, r$ and $j = 1, \ldots, c$. By difference the expectations for the first row may be obtained, giving the familiar

$$E[N_{ij}] = n_{i.}*n_{.j}/n_{..},\ i = 1, \ldots, r \text{ and } j = 1, \ldots, c.$$

In the same way

$$E[N_{21}(N_{21} - 1)] = n_{2.}(n_{2.} - 1)n_{.1}(n_{.1} - 1)/\{n_{..}(n_{..} - 1)\},$$

from which we obtain $\text{var}(N_{21})$, and

$$\text{var}(N_{ij}) = n_{i.}\,\frac{n_{.j}}{n_{..}}\left(1 - \frac{n_{.j}}{n_{..}}\right)\left(\frac{n_{..} - n_{i.}}{n_{..} - 1}\right),\ i = 1, \ldots, r \text{ and } j = 1, \ldots, c.$$

Similarly

$$\text{cov}(N_{ij}, N_{ik}) = -\,n_{i.}\,\frac{n_{.j}}{n_{..}}\,\frac{n_{.k}}{n_{..}}\left(\frac{n_{..} - n_{i.}}{n_{..} - 1}\right),\ i = 1, \ldots, r \text{ and } j \neq k = 1, \ldots, c.$$

By symmetry

$$\text{cov}(N_{rj}, N_{sj}) = - n_{.j} \frac{n_{r.}}{n_{..}} \frac{n_{s.}}{n_{..}} \left(\frac{n_{..} - n_{j}}{n_{..} - 1} \right), \; i = 1, \ldots, r \text{ and } j \neq k = 1, \ldots, c,$$

and by the expectation argument again

$$\text{cov}(N_{ir}, N_{js}) = \frac{n_{i.} n_{.r}}{n_{..}} \frac{n_{j.} n_{.s}}{n_{..}} \left(\frac{1}{n_{..} - 1} \right), \; i \neq j = 1, \ldots, r, \text{ and } r \neq s = 1, \ldots, c.$$

Thus, as required, the joint covariance matrix of N_i and N_j is, for $i \neq j$,

$$\text{cov}(N_i, N_j) = - \frac{n_{i.}}{n_{..}} \frac{n_{j.}}{n_{..}} \left\{ \text{diag} \left(\frac{n_{.r} n_{..}}{(n_{..} - 1)} \right) - \left(\frac{n_{.r} n_{.s}}{n_{..} - 1} \right) \right\}.$$

Since the $\{N_{ij}\}$ are such that the row and column totals are known constants, $\text{cov}(N_i, N_1 + \ldots + N_r) = 0$ for $i = 1, \ldots, r$. The covariance matrix of N_i is

$$\text{cov}(N_i) = - \sum_{i \neq j} \text{cov} \, (N_i, N_j) = \sum_{i \neq j} f_i f_j \, R = f_i \, (1 - f_i) \, R,$$

which agrees with direct calculation.

Now write $N^T = (N_1^T, \ldots, N_r^T)$ and $\otimes$ for the Kronecker product (see Appendix A.4), the covariance matrix of N is

$$\text{cov}(N) = \{\text{diag}(f_j) - (f_i f_j)\} \otimes R,$$

and define the standardised cell counts Z_{ij} by

$$Z_{ij} = (N_{ij} - E[N_{ij}]) / \sqrt{E[N_{ij}]}, \; i = 1, \ldots, r \text{ and } j = 1, \ldots, c, \text{ and}$$

$$Z = (Z_{11}, \ldots, Z_{1c}, \ldots, Z_{r1}, \ldots, Z_{rc})^T.$$

It follows that

$$\mathrm{cov}(Z) = \{I_r - (\sqrt{[f_i f_j]})\} \otimes R.$$

The matrix $\{I_r - (\sqrt{[f_i f_j]})\}$ has r - 1 eigenvalues one and one eigenvalue zero. The eigenvalues of R are difficult to find in general, but their asymptotic limits follow from Lancaster (1969, Chapter V.3). Lancaster showed that the quadratic form with vector the standardised cell counts and matrix essentially R is the familiar Pearson goodness of fit statistic, with asymptotic distribution χ^2_{c-1}. Hence the eigenvalues of R are asymptotically one c - 1 times and zero once. Hence under the null hypothesis of simple random sampling, Z has zero mean and covariance matrix cov(Z), which asymptotically has (r - 1)(c - 1) eigenvalues one, and the remaining r + c - 1 eigenvalues zero.

In the well known and often used 'classical' model, r and c are fixed and the total count $n_{..} \to \infty$. The test statistic is, X^2_P, given by

$$X^2_P = \sum_{i=1}^{r} \sum_{j=1}^{c} (N_{ij} - \frac{n_{i.} n_{.j}}{n_{..}})^2 / (\frac{n_{i.} n_{.j}}{n_{..}}) = Z^T Z.$$

We now confirm that our model leads to this test statistic. Suppose H is orthogonal and diagonalises cov(Z). Asymptotically we then have

$$H^T \mathrm{cov}(Z) H = I_{(r-1)(c-1)} \oplus 0_{(r+c-1)},$$

where $\oplus$ means direct or Kronecker sum. Define $Y = H^T Z$. Now $Z^T Z = Y^T Y$, in which Y, by the multivariate Central Limit Theorem, is asymptotically $N_{rc}(0, [I_{(r-1)(c-1)} \oplus 0_{(r+c-1)}])$ under the null hypothesis of simple random sampling. It follows that under the null hypothesis, $X^2_P = Z^T Z = Y^T Y$ asymptotically has the $\chi^2_{(r-1)(c-1)}$ distribution.

3.3 Partitioning Pearson's Statistic

We now show that X_P^2 may be partitioned into components, the s*th* of which may be interpreted as detecting the departure in the s*th* moment in what may be expected under the null hypothesis of similarly distributed rows (treatments).

The elements Y_i of Y are such that $X_P^2 = \sum_{i=1}^{rc} Y_i^2$. As H is not yet fully specified there is some choice in defining Y. In doing so, our aim is to find Y_i that can be easily and usefully interpreted. To achieve one such partition, first suppose that the $g_s(j)$ are the polynomials orthonormal on $\{n_{.j}/n_{..}\}$. See Appendix A.3. For reasons to be made clear shortly, we prefer to work with $\left(\frac{n-1}{n}\right)X_P^2$. When using natural scores $x_j = j$, $j = 1, \ldots, c$, and when there are no ties, our approach results in the first component being the Kruskal-Wallis statistic.

Write g_s for the c by 1 vector with elements $g_s(j)$. Define G by

$$G = \left[\begin{array}{ccc} G_1 & \cdots & G_c \end{array} \right] / \sqrt{c}$$

in which G_s is the rc by r matrix

$$G_s = \begin{bmatrix} g_s & 0 & \cdots & 0 \\ 0 & g_s & & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_s \end{bmatrix}, \; s = 1, \ldots, c-1, \text{ and}$$

$$G_c = \begin{bmatrix} 1_c & 0 & \cdots & 0 \\ 0 & 1_c & & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_c \end{bmatrix}$$ is also rc by r.

Define $Y = \sqrt{\frac{n - 1}{n}}\, G^T Z$. The factor (n - 1)/n appears here because we are working with the exact eigenvalues. In section 3.2 we used the asymptotic eigenvalues. The elements of Y may be considered in blocks of r, the s*th* block corresponding to the polynomial of order s. These blocks are asymptotically mutually independent. Write $Y^T = (V_1^T, \ldots, V_c^T)$, in which

$$V_1 = (Y_1, \ldots, Y_r)^T, \ldots, V_{c-1} = (Y_{(c-2)r+1}, \ldots, Y_{(c-1)r})^T,$$

and $V_c = 0$ (all the V_s are r by 1) so that

$$\left(\frac{n - 1}{n}\right) X_P^2 = \left(\frac{n - 1}{n}\right) Z^T Z = Y^T Y = V_1^T V_1 + \ldots + V_{c-1}^T V_{c-1}.$$

This partitions $\left(\frac{n - 1}{n}\right) X_P^2$ into components $V_s^T V_s$, s = 1, ... , c - 1. The V_s are asymptotically mutually independent and asymptotically $N_r(0, I_{(r-1)} \oplus 0)$. It follows that the $V_s^T V_s$ are asymptotically mutually independent χ^2_{r-1}. Explicitly, we have, for s = 1, ... , c - 1,

$$V_s = \frac{\sqrt{n - 1}}{n}\, G_s^T Z = \frac{\sqrt{n - 1}}{n} \sum_{j=1}^{c} g_s(j)\, Z_{ij}.$$

Because V_s involves, through g_s, a polynomial of order s, the elements of V_s are polynomials of order s in the elements of N. Under the null hypothesis E[Z] = 0, but when this is not true E[V_s] involves moments up to order s of Z. Thus for s = 1, ... , r - 1, $V_s^T V_s$ typically detects s*th*

moment departures from the null hypothesis of similarly distributed rows (treatments).

It can be shown that this discussion also applies for general scores $\{x_j\}$ as well as the simpler scores $\{j\}$ used above. See Appendix A.3 for definitions of the g's. If meaningful scores $\{x_j\}$ are available then we would usually suggest they be used in the nonparametric test rather than risk loss of information by taking ranks. Often different choices of scores, including ranks, will not alter conclusions, but sometimes they do. See, for example, Graubard and Korn (1987).

Instructors Example. We now return to this example from Conover (1998, p. 293), previously discussed in the Introduction. Three instructors assign grades in five categories according to Table 1.1, for convenience reproduced here as Table 3.1.

Table 3.1 Grades assigned by three instructors

	Grade					
	A	B	C	D	E	Total
Instructor 1	4	14	17	6	2	43
Instructor 2	10	6	9	7	6	38
Instructor 3	6	7	8	6	1	28
Total	20	27	34	19	9	109

The mid-rank scores are 10.5, 34, 64.5, 91 and 105. If these are used, the theory of section 3.5 suggests that apart from a factor (n - 1)/n, the first component of X_P^2 is the Kruskal-Wallis statistic corrected for ties. However this assumes a t by n* table, with n* *large*. With n* = 5 here, *large* seems an inappropriate description. In this section it was assumed that the numbers of rows and columns were fixed, and the grand total of observations became large. While the instructor totals are known before sighting the data, the same cannot be claimed for the grade totals.

Nevertheless, for illustrative purposes we will now proceed assuming such a model.

Conover (1998) found the Kruskal-Wallis statistic adjusted for ties to be 0.3209, which is to be compared with the $\chi^2_2(5\%)$ point of 5.991. We find the location detecting component is $V_1^T V_1$ to have p-value 0.85, confirming, as Conover reported, that "none of the instructors can be said to grade higher or lower than the others on the basis of the evidence presented". However the dispersion detecting component $V_2^T V_2$ takes the value 9.643, with corresponding p-value 0.01, indicating a significant variability difference. From the data it appears that the first instructor is less variable than the other two. In fact, $9.643 = (-2.113)^2 + (2.274)^2 + (-0.031)^2$, with the elements of $v_2 = (-2.113, 2.274, -0.031)$ being values of approximately standard normally distributed contributions from instructors 1, 2 and 3 respectively. The first instructor is less variable than the third who is less variable than the second. This can be formalised by a LSD analysis. The residual $X_P^2 - V_1^T V_1 - V_2^T V_2$ has p-value 0.75, indicating no further effects in the data.

For completeness we note that although we and Conover (1998, p. 293) approach statistical testing for these data in a nonparametric fashion, many statisticians would use a parametric log-linear model. A test based on a row-effects log-linear model would be an alternative to the Kruskal-Wallis test suggested by Conover (1998, p. 293). Agresti (1984, section 5.2) gives a description of row-effects log-linear models and the iterative techniques needed to fit them. For the instructors data the row-effects log-linear model gives a test statistic value of 0.279, very close to the 0.324 value we obtained in Table 3.2. This uses the same mid-rank scores for the grades as we used to obtain Table 3.2, namely 10.5, 34, 64.5, 91 and 105. We have also observed this close agreement between the Kruskal-Wallis statistic and the row-effects log-linear model statistic with other tables. See, for example, the ulcer example in section 4.7.

Table 3.2 Partition of X_P^2 for instructors data

Statistic	Degrees of Freedom	Value	p-value
$V_1^T V_1$	2	0.324	0.85
$V_2^T V_2$	2	9.643	0.01
$X_P^2 - V_1^T V_1 - V_2^T V_2$	4	2.021	0.75
X_P^2	8	11.985	0.15

In regard to the row-effects log-linear model, Beh and Davy (2000) show that if $V_1 = (V_{1i})$, then for i = 1, ... , r $V_{1i}/\sqrt{n_{i.}\mu_2}$ approximates the log-linear row-effects model parameters. The parameter μ_2 is defined in Appendix A.3. We suspect this approximation works best for the $n_{i.}$ at least approximately equal and the V_{1i} not too large.

3.4 The Kruskal-Wallis Test with No Ties

We now consider models that lead to the Kruskal-Wallis test when there are no ties. The eigenvalues of cov(Z) will be found explicitly as in section 3.3 rather than asymptotically as in section 3.2. We show that X_P^2 is not an appropriate test statistic, but, nevertheless, its components are. The first component is the Kruskal-Wallis test statistic, and the subsequent components provide informative extensions.

Suppose we have distinct observations x_{ij}, being the j*th* of n_i observations on the i*th* of t treatments. All $n = n_1 + ... + n_t$ observations are combined, ordered, ranked, and the sums R_i of the ranks obtained by the i*th* treatment calculated. The Kruskal-Wallis statistic is

$$\{12/[n(n + 1)]\} \sum_i R_i^2/n_i - 3(n + 1).$$

See, for example, Conover (1998, section 5.2). The data may be presented as a t by n contingency table of counts $\{N_{ij}\}$, with $N_{ij} = 1$ if rank j is allotted to treatment i, and $N_{ij} = 0$ if rank j is allotted to some other treatment. The row and column totals are all fixed: the row totals are the treatment sample sizes, so that $n_{i.} = n_i$ for i = 1, ... , t, while the column totals are all one: $n_{.j} = 1$ for j = 1, ... , n. Such a table has $X_P^2 = (t - 1)n$ no matter what the $\{N_{ij}\}$. Since X_P^2 is constant, it is not a suitable test statistic.

The model of section 3.2 holds, except that now

$$R = \frac{n}{n - 1} I_n - \frac{1}{n - 1} 1_n 1_n^T.$$

This matrix has one eigenvalue zero and n - 1 eigenvalues n/(n - 1). It follows that cov(Z) has (t - 1)(n - 1) eigenvalues n/(n - 1), and the remaining t + n - 1 eigenvalues zero. Now if H is orthogonal and diagonalises cov(Z),

$$H^T \mathrm{cov}(Z) H = \frac{n}{(n - 1)} \left[I_{(t - 1)(n - 1)} \oplus 0_{(t + n - 1)} \right].$$

Define $Y = \sqrt{\left(\frac{n - 1}{n}\right)}\, H^T Z$. Then

$$Y^T Y = \left(\frac{n - 1}{n}\right) Z^T Z = \left(\frac{n - 1}{n}\right) X_P^2.$$

With r replaced by t and c replaced by n, this is the same as section 3.3.

As in section 3.2, we are interested in the distribution theory as n $\to \infty$. However there Z was an rc by 1 vector of fixed length; here Z is a tn by 1 vector. Fortunately, it is not the asymptotic distribution of Z that is required. First recall that X^2_P has a fixed value, (t - 1)n, for all tables, and so is not available as a test statistic. Second, as in section 3.3, the multivariate Central Limit Theorem shows that each V_s is asymptotically $N_t(0, I_{(t-1)} \oplus 0)$. Moreover consideration of any pair V_s, V_t shows that they are asymptotically jointly multivariate normal, and since their cov(V_s, V_t) = 0, any pair V_s and V_t are asymptotically independent. The $V_s^T V_s$ still partition $\left(\frac{n-1}{n}\right) X^2_P$. It is the pairwise independence and convenient χ^2_{t-1} distribution of each $V_s^T V_s$ that makes data analysis so informative and convenient. What is lost by the unavailability of X^2_P is demonstrated in the Employees Example below: there is no residual available to assess if there are higher moment differences between the treatments.

We now show that provided there are no ties, $V_1^T V_1$ is the Kruskal-Wallis statistic, so that the subsequent $V_s^T V_s$ provide extensions to the Kruskal-Wallis test. First note that the $\{g_s(j)\}$ are the polynomials orthonormal on the discrete uniform distribution, so that $g_1(j) = aj + b$, j = 1, ... , n, in which

$$a = \sqrt{\{12/(n^2 - 1)\}} \text{ and } b = -\sqrt{\{3(n+1)/(n-1)\}} = -\{(n+1)/2\}a.$$

The rank sum for treatment i, R_i, is $\sum_{j=1}^{n} jN_{ij}$, i = 1, ... , t. Now since $n_{.j} = 1$ for j = 1, ... , n,

$$\sum_j g_1(j)\sqrt{(E[N_{ij}])} = \sqrt{(n_i\, n)} \sum_j g_1(j)g_0(j)\,(1/n) = 0$$

and

$$\sum_j Z_{ij}g_1(j) = \sqrt{[n/n_i]} \sum_j N_{ij}(aj + b) = \sqrt{[n/n_i]}\,\{aR_i + bn_i\}$$
$$= a\,\sqrt{[n/n_i]}\,\{R_i - \tfrac{n+1}{2}n_i\}.$$

Now $V_1^T V_1 = Y_1^2 + \ldots + Y_t^2$

$$= \frac{n-1}{n^2} \sum_{i=1}^{t} \left(\sum_{j=1}^{n} g_1(j)\, Z_{ij} \right)^2$$

$$= \frac{n-1}{n^2}\, a^2 n \sum_{i=1}^{t} \left\{ \frac{R_i}{\sqrt{n_i}} - \frac{n+1}{2} \sqrt{n_i} \right\}^2$$

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

after some manipulation. This is the Kruskal-Wallis statistic defined previously. It is well known to be sensitive to location departures from the null hypothesis. Since V_s assesses s*th* moment departures between treatments, we have partitioned the statistic $\left(\frac{n-1}{n}\right) X_P^2$ into asymptotically pairwise independent components, $V_s^T V_s$, s = 1, ... , n - 1, each with the χ^2_{t-1} distribution, and such that the s*th* detects s*th* moment departures from the hypothesis of similarly distributed rows (treatments). Since the first of these is the Kruskal-Wallis statistic, the subsequent components provide extensions to the Kruskal-Wallis test.

Employees Example. Conover (1998, p. 298, exercise 2) gave an exercise in which 20 new employees are randomly assigned to four different job training programmes. At the end of their training the employees are ranked, with a low ranking reflecting a low job ability.

Table 3.3 Employees' rankings after job training via four programmes

Programme	Ranks
1	2, 4, 6, 7, 10
2	1, 3, 8, 11, 12
3	5, 14, 16, 19, 20
4	9, 13, 15, 17, 18

The value of the Kruskal-Wallis statistic is 9.72, and $V_1^T V_1$ = 9.72*20/19 = 10.23. The linear location detecting component has Monte Carlo permutation test p-value 0.010, while the quadratic dispersion detecting component has p-value 0.715. As the sample size is small, these are more likely to be accurate than the p-values based on the asymptotic χ^2 distribution. The analysis follows from expressing the data as a 4 by 20 contingency table of 0s and 1s where each marginal column total is fixed at one. The column scores are the ranks. From the calculations given we cannot assess the residual, as X_P^2 is independent of the data. Of course we could calculate a cubic component and calculate a new residual, but for this analysis we take the view that the existence or not of higher order effects was not of interest.

Table 3.4 Partition of X_P^2 for employees data

Statistic	Degrees of Freedom	Value	Monte Carlo p-value
$V_1^T V_1$	3	10.232	0.010
$V_2^T V_2$	3	1.541	0.715
$X_P^2 - V_1^T V_1 - V_2^T V_2$	51	48.227	-
X_P^2	57	60.000	-

The values of the elements v_{1u} of v_1 are -1.823, -1.357, 1.667 and 1.512. An approximate 5% LSD is $2\sqrt{2} = 2.828$, suggesting that programmes 1 and 2 are equally effective, as are programmes 3 and 4, but 3 and 4 being superior to 1 and 2.

Boos Example. Boos (1986) suggests an analysis for one-way data that looks at location, scale, skewness and kurtosis effects. He takes the ranks of the data and if location or scale effects are significant he then uses standardised data before ranking. The analysis for the employees data was based on the ranks as only the ranks were available. For the Boos data we suggest using the actual data as scores for our components. The data given by Boos (1986) and originally by Nation et al. (1984) consist of the number of platform descents by rats randomly assigned to three groups. These were:
(i) 82, 80, 77, 75, 72, 68, 59, 47, and 42 for the control group,
(ii) 86, 66, 60, 51, 44, 41, 38, 29, 10 for rats on a 1 mg. cadmium diet, and
(iii) 81, 67, 38, 36, 32, 29, 20, 17, 14 for rats on a 5 mg. cadmium diet.

Boos (1986) finds location and skewness effects. Our partition of X_P^2 gives the results in Table 3.5. Our analysis indicates significant location and kurtosis effects. It is difficult to tell by plotting the data whether our analysis or that of Boos (1986) gives a better summary. Our analysis is based on a 3 by 25 contingency table of 0s and 1s where two of the

column totals are two and the remainder one.

Table 3.5 Partition of X_P^2 for Boos' data

Statistic	Degrees of Freedom	Value	Monte Carlo p-value
$V_1^T V_1$	2	8.057	0.018
$V_2^T V_2$	2	0.732	0.714
$V_3^T V_3$	2	1.618	0.445
$V_4^T V_4$	2	7.874	0.019
$X_P^2 - V_1^T V_1 - ... - V_4^T V_4$	40	39.211	-
X_P^2	48	48.000	-

The alignment and standardisation used by Boos (1986) has also been suggested by other authors. As we demonstrated in the example in section 1.3, if the ranks of the data are used as scores then the ranking can diminish or even destroy the dispersion effects. However if the original data, rather than the ranks, are used as scores, then this diminution of the scale or dispersion effects is lessened. Note that even with the use of the original data as scores, it is still possible for a significant or merely large lower order effect to mask a higher order effect. If it is considered that detection of the largest effect is more important than detecting all effects, then such masking may not be thought too important. Although the data under discussion here are essentially continuous, for ordinal categorical data Nair (1986) stated that detecting dispersion effects in the presence of strong location effects is difficult.

3.5 The Kruskal-Wallis Test with Ties

If there are ties, the data may be presented as an t by n* contingency table of counts $\{N_{ij}\}$, with the row totals fixed at the treatment sample sizes; so again $n_{i.} = n_i$, i = 1, ... , t, while the column totals are no longer all one. The covariance matrix of Z is

$$\text{cov}(Z) = \{I_t - (\sqrt{[f_i f_j]})\} \otimes R \text{ with } R = \text{diag}\left(\frac{n_{.u}\, n_{..}}{(n_{..} - 1)}\right) - \left(\frac{n_{.u}\, n_{.v}}{n_{..} - 1}\right).$$

As in section 3.2, the eigenvalues of R are zero once and asymptotically one n* - 1 times. It follows that cov(Z) has (t - 1)(n* - 1) eigenvalues asymptotically one, and the remaining t + n* - 1 eigenvalues asymptotically zero. With suitable modifications the partitioning of section 3.3 holds. For s = 1, ... , n* - 1,

$$V_s = G_s^T Z/\sqrt{n^*} = \left(\sum_{j=1}^{n^*} g_s(x_j)\, Z_{ij}\right)/\sqrt{n^*}.$$

Note that $\{g_s(x_j)\}$ is the set of polynomials orthonormal on $\{n_{.j}/n_{..}\}$, not on the discrete uniform as in the previous section where there were no ties. This is the partition derived in section 3.3 for X_P^2. The subsequent components are extensions to the Kruskal-Wallis test adjusted for ties. As observed in the previous chapter, this choice of orthonormal polynomials effectively models the ties. If this model is not appropriate, then a different choice of orthonormal polynomials must be made. In applications we have looked at, the inference is not greatly affected by alternative choices, but the effort involved is. The Instructors Example of sections 1.2 and 3.3 illustrates the application of a Kruskal-Wallis test with ties. Also see the examples in the next chapter where we consider only one set of margins to be fixed.

3.6 Generalised Median Tests

Conover (1998, section 4.3) described the median test, in which random samples are taken from each of c populations. Each observation is classified as above and below the grand median (the median of the combined random samples), forming an r by 2 contingency table with fixed marginal totals. The usual chi-squared test, based on X_P^2, is then applied to this contingency table.

If instead of the grand median, a 'grand quantile' is used, the resulting test is described as a quantile test: see Conover (1998, p. 222). These tests can be generalised by choosing c instead of two categories for the combined random samples, and so forming an r by c contingency table of counts N_{ij} of the number of observations for the i*th* sample in the j*th* category. This table has all row and column totals fixed and can be tested for row consistency using the results of the sections 3.2 and 3.3. The first three, say, components of X_P^2 are of particular interest, indicating location, dispersion and skewness differences between treatments.

It is routine to show that the location component $V_1^T V_1$ of X_P^2 reduces to the median test statistic, T, when observations are classified into just two categories.

Theorem 3.2. $T = V_1^T V_1$.

Proof. If there are b observations below a predetermined point in the combined sample, and a above it, then Conover (1998, p. 219) gave the X^2 Median test statistic as

$$T = \frac{n_{..}^2}{ab} \sum_{i=1}^{r} (N_{i1} - \frac{n_{i.}b}{n_{..}})^2 / n_{i.}.$$

As in Appendix A.3, $g_{1j} = (j - \mu)/\sigma$, j = 1, ... , c, in which μ and σ are the mean and standard deviation of the distribution defined by $P(X = 1) = b/n_{..}$ and $P(X = 2) = a/n_{..}$. It follows that $\mu = 1 + a/n_{..}$ and $\sigma^2 = ab/n_{..}^2$. Now

$$\sigma\sqrt{n_{i.}}V_{1i} = \sum_{j=1}^{2} N_{ij}g_{1j} = N_{i1}\,(-\,a/n_{..}) + (n_{i.} - N_{i1})(1 - a/n_{..})$$
$$= (n_{i.} - an_{i.}/n_{..}) - N_{i1} = n_{i.}b/n_{..} - N_{i1}.$$

It follows that, as required, $V_1^T V_1 = T$.

This result identifies the median test as a location detecting test. To detect up to s*th* moment differences between the populations requires categorisation into (s + 1) categories and the use of the $V_2, \ldots, V_s$ components. If there are as many categories as observations and each category has one observation, the test based on the location component is the Kruskal-Wallis test, which is known to be more powerful than the median test. Using more than two categories will result in less loss of information due to categorisation compared to the median test, and will permit assessment of higher moment differences between the treatments.

Corn Example. Conover (1998, p. 220) gave the example of four different methods of growing corn. He classified the data as greater than 89 and up to 88 and applied the median test. In this form this does not conform to the fixed margins model. If the objective were to divide the data into groups of the lowest 18 and highest 16 observations, it would conform to the fixed margins model. We now classify the data into four approximately equal groups. Strictly speaking, to fit our model we need to decide, before seeing the data, to divide the data into groups of consecutive sizes 9, 9, 8 and 8. See Table 3.6.

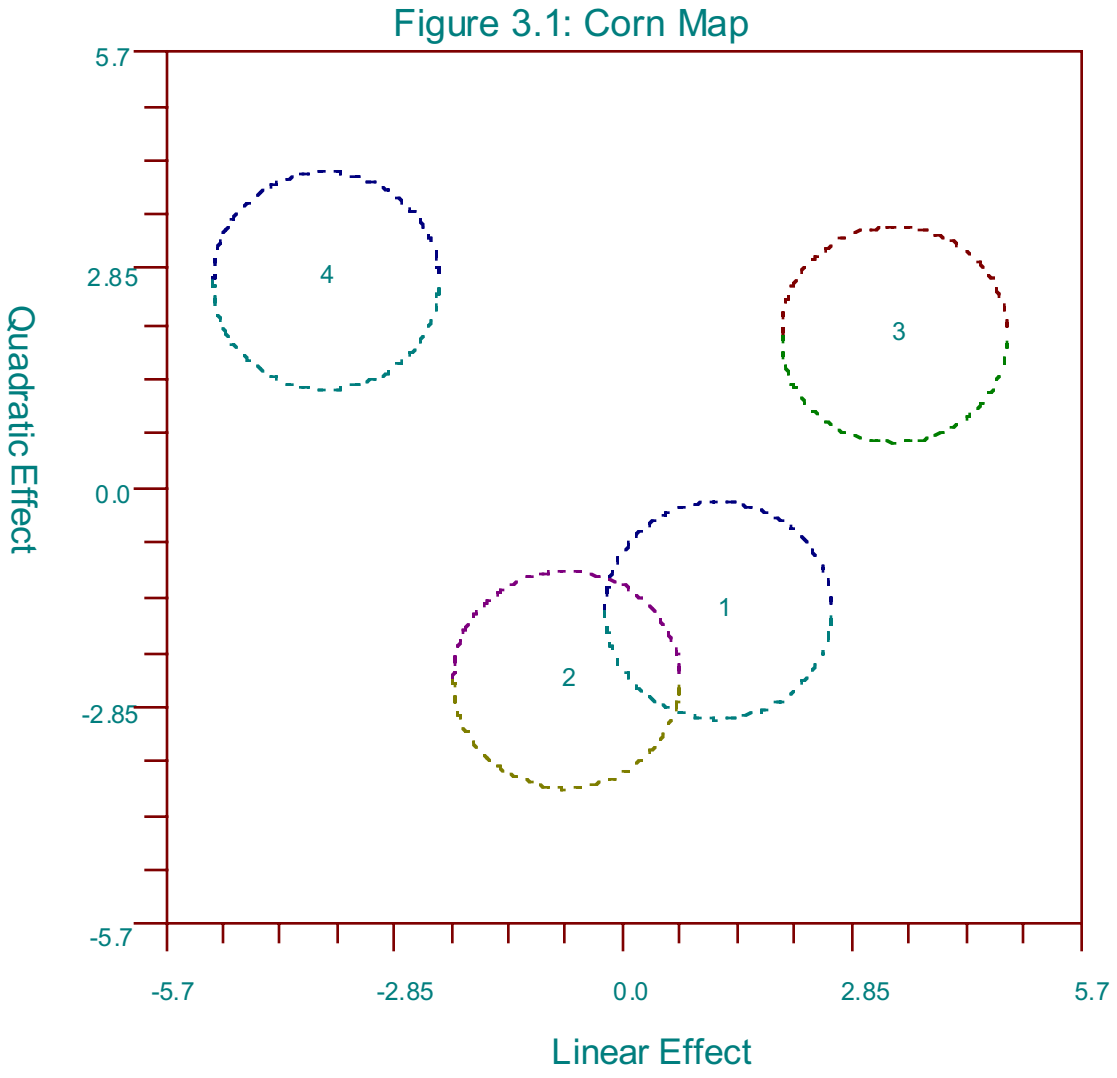
Table 3.6 Yields for corn grown by four different methods

	First Quartile	Second Quartile	Third Quartile	Fourth Quartile	Total
Method 1	0	3	4	2	9
Method 2	1	6	3	0	10
Method 3	0	0	1	6	7
Method 4	8	0	0	0	8
Total	9	9	8	8	34

Using the median test, Conover reported a p-value "slightly less than 0.001": the method median yields are clearly different. We calculate $X_P^2 = 49.712$ on 9 degrees of freedom. In addition $V_1^T V_1 = 25.723$, $V_2^T V_2 = 19.972$ and $V_3^T V_3 = 2.574$, all on 3 degrees of freedom. The location and dispersion components and X_P^2 are all significant, with p-values all zero to three decimal places. The residual or skewness component has χ_3^2 p-value 0.45. The finer classification, compared to that employed by the median test, has uncovered a variability difference between the methods.

In fact more information is available. Since, for example, $V_2^T V_2 = V_{12}^2 + V_{22}^2 + V_{32}^2 + V_{42}^2$, the contribution to the significant variability from each treatment can be identified. Moreover, an approximate LSD analysis is available if we treat the V_{is} as independent standard normal random variables. This is not quite accurate, as the V_{is} are dependent (otherwise $V_2^T V_2$ would have the χ_r^2 distribution, when in fact the distribution is χ_{r-1}^2), and the normality of the V_{is} is only asymptotically true, and may not be accurate in small samples. We calculate $v_{12} = -1.6355$, $v_{22} = -2.5487$, $v_{32} = 1.9789$, $v_{42} = 2.7332$. The LSD is $z\sqrt{2}$, where, for an overall level α comparison, z is the standard normal point with $P(Z > z \mid Z$

is N(0, 1)) = $\alpha/(2\ {}^{r}C_2) = \alpha/12$ here. With α of 6% or more, the LSD is at most 3.6 and methods 3 and 4 are identified as significantly less variable than methods 1 and 2.



Similarly v_{r1} = 1.2030, -0.6831, 3.3554, -3.6510 for r = 1, ... , 4. Methods 4 and 2, 2 and 1, and 1 and 3 appear to be similar. Method 4 has the greatest yield and method 3 the least. Figure 3.1 gives a product map of these data.

Note that this example is included to illustrate a case where both margins are fixed. Usually we would suggest a nonparametric analysis with the actual data as scores $\{x_j\}$ without assuming both margins fixed.

4
Tests Based on a Product Multinomial Model: Yates' Test and its Extensions

4.1 Introduction

This chapter continues with the nonparametric examination of one-way layout data except, unlike Chapter 3, we now take only one margin of the associated contingency table to be fixed. This is a more common situation than that discussed in Chapter 3.

In the social sciences and in more specialized areas such as sensory evaluation, it is common to obtain categorized ratings for a number of items. For example, if eyesight is under consideration, one might have five categories of eyesight ranging from poor to excellent, with observations on both males and females, giving a two by five contingency table of responses. As one of the categorizations is ordered, it is possible to do a more thorough analysis than that given by the usual X_P^2 test for a two-way contingency table. Sometimes, although the X_P^2 test may not be significant, an effect may be suggested by one of a number of analyses suggested in the literature, including:

(i) giving the categories equi-spaced scores and using a regression analysis as in Yates (1948);
(ii) using nonequi-spaced scores based on mid-rank values as in Bross (1958), Conover (1998, p. 281) and Nair (1986);
(iii) linear logistic models as in McCullagh (1980);
(iv) log-linear models and user defined assigned scores as in Agresti (1984, p. 84);
(v) the cumulative chi-square method of Taguchi (1966); see also Nair (1986) and Hamada and Wu (1990) for a discussion of this method and reasons for not using it;

(vi) analysis of variance and user defined assigned scores as in Box and Jones (1986) or Nair (1990).

We demonstrate here that a method using the components of X_P^2 compares well with more recent methods and has the appeal of being weakly optimal and simple both conceptually and arithmetically. The method generalizes readily to other models, and to multi-way tables; see Chapters 8 and 10 that review Beh and Davy (1998 and 1999).

Our approach involves a family of simple parametric distributions. With sufficiently many parameters our models will fit the data exactly, but in practice highly parametrised models are not needed. Our parameters are related to moments, and we usually find it is sufficient to include only location and dispersion parameters; rarely is it necessary to include skewness and kurtosis parameters in addition. The test statistics we recommend for testing are asymptotically independent, assessing if the rows agree with regard to their location, dispersion, skewness, etc. We suggest simultaneously assessing location and dispersion, and combining the remainder into a residual unless there are *a priori* reasons for doing otherwise.

The reader more interested in examples may wish to proceed straight to section 4.7.

4.2 One-Way Tables

We now describe some results for one-way tables because our two-way results are strongly motivated by the corresponding one-way results.

Best and Rayner (1987) gave formulae for obtaining components of the usual X_P^2 goodness of fit statistic for the multinomial, where there are n observations categorized into c classes with known class probabilities p_1, p_2, ... , p_c. The numbers of observations in the c classes are N_1, N_2, ... , N_c, constrained by $n = N_1 + N_2 + ... + N_c$. The components have the sum of their squares equal to X_P^2 and may be correlated in small samples but are asymptotically uncorrelated. *Asymptotic* means $n \to \infty$. If the categories are ordered and the components are based on orthogonal polynomials, then, for example, the first two components identify linear and quadratic

effects: loosely location and dispersion effects.

Both the linear and the quadratic components are asymptotically distributed as standard normal variables, and power studies in Best and Rayner (1987) indicated these components compete well with a variety of other statistics when alternatives involve location and dispersion effects. The orthogonal polynomials are given in Appendix A.3, but for convenience sake are repeated here. For j = 1, ... , c,

$$g_0(x_j) = 1,\ g_1(x_j) = (x_j - \mu)/\sqrt{\mu_2} \text{ and}$$
$$g_2(x_j) = a\{(x_j - \mu)^2 - \mu_3(x_j - \mu)/\mu_2 - \mu_2\},$$

in which

$$\mu = \sum_{j=1}^{c} x_j p_j,\ \mu_r = \sum_{j=1}^{c} (x_j - \mu)^r p_j \text{ and } a = \left(\mu_4 + \mu_3^2/\mu_2 - \mu_2^2\right)^{-0.5}.$$

The components, previously discussed in section 2.6, are given explicitly by

$$\hat{V}_u = \sum_{j=1}^{c} N_j\, g_u(x_j)/\sqrt{n},\ u = 1, \ldots , m - 1.$$

They depend on the orthogonal polynomials, which are most conveniently given by using the explicit formulae for the g_1 and g_2, and then the recurrence relations of Emerson (1968). The $\hat{V}_u^2$ are score statistics in their own right, and hence provide weakly optimal directional tests (each seeks to detect alternatives in a one dimensional parameter space), so supplementing the omnibus nature of the X_P^2 test (that seeks to detect alternatives in a c - 1 dimensional parameter space). The latter is based on the statistic

$$X_P^2 = \sum_{j=1}^{c} (N_j - np_j)^2/(np_j) = \hat{V}_1^2 + \ldots + \hat{V}_{c-1}^2.$$

Lancaster (1969, p. 134) demonstrated such a partition of X_P^2 into

components for the particular case $p_1 = ... = p_c$.

4.3 Two-Way Tables

Consider the following product multinomial model. We have an r by c contingency table with cell probabilities p_{ij}, i = 1, ... , r, and j = 1, ... , c, such that $p_{i1} + ... + p_{ic} = 1$ for i = 1, ... , r. Observations N_{i1}, ... , N_{ic} are taken on the cells of each of the i rows, yielding row totals $n_{i.}$, i = 1, ... , r that were known before the collection of the data. Column totals are random variables and are denoted by $N_{.j}$, j = 1, ... , c; the total count is $n_{..}$. Suppose that the columns are ordered categories, and it is of interest to compare rows for similarity of location and dispersion effects. The null hypothesis is equality of the corresponding row probabilities. If $p_{.j} = (p_{1j} + ... + p_{rj})/r$ for j = 1, ... , c, we test the null hypothesis $p_{ij} = p_{.j}$ for i = 1, ... , r, and j = 1, ... , c, against the alternative hypothesis, not the null. The usual X_P^2 statistic is derived in, for example, Conover (1998, section 4.2) to be

$$X_P^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (N_{ij} - n_{..}\hat{p}_{ij})^2/(n_{..}\hat{p}_{ij}),$$

where $\hat{p}_{ij} = (n_{i.}/n_{..})(N_{.j}/n_{..})$. This X_P^2 statistic examines all deviations from what is expected under a homogeneity model. As in the one-way table, it is appropriate to decompose X_P^2 to obtain more informative directional tests. The summands of the decomposition ($\hat{V}_{ui}$ subsequently) are not quite components as we usually use the term, as they are not independent, not even asymptotically.

To decompose X_P^2 statistics for the two-way table, $\hat{V}_1$ and $\hat{V}_2$ defined earlier can be calculated for each row (provided $c \geq 3$), yielding $\hat{V}_{1i}$ and $\hat{V}_{2i}$, for $1 \leq i \leq r$. In these $\hat{V}_{ui}$ the p_j are now taken to be $(N_{.j}/n_{..})$ for $1 \leq j \leq c$, and n is taken as $n_{i.}$, $1 \leq i \leq r$, and N_j becomes N_{ij}. Using the subsequent $g_u(j)$, further statistics $\hat{V}_{ui}$ can be defined by

$$\hat{V}_{ui} = \sum_{j=1}^{c} N_{ij}\, \hat{g}_u(j)/\sqrt{n_{i.}},\ u = 1, \ldots, c-1 \text{ and } i = 1, \ldots, r$$

where the hats indicate that the maximum likelihood estimators $\hat{p}_{.j} = N_{.j}/n_{..}$, $j = 1, \ldots, c$ have been used in the construction of the orthonormal functions.

We will show in section 4.4 following that

$$\sum_{u=1}^{c-1} \sum_{i=1}^{r} \hat{V}_{ui}^2 = X_P^2,$$

and that this is an alternative decomposition of X_P^2 to that given by Lancaster (1969, Theorem 6.2). We also show in section 4.4 that the $\hat{V}_{ui}$ are score statistics for an appropriate model; this implies weak optimality. See Rayner and Best (1989a, section 3.4) for a discussion of this optimality.

A measure of the location effect for the whole table is $\hat{V}_{11}^2 + \ldots + \hat{V}_{1r}^2$, which, when $x_j = j$ for $j = 1, \ldots, c$, is just the statistic Q of Yates (1948). Similarly the overall dispersion effect can be assessed by $\hat{V}_{21}^2 + \ldots + \hat{V}_{2r}^2$, provided $c \geq 3$. Provided $c \geq u + 1$, a measure of the u*th* moment departure from the null hypothesis is $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$. This interpretation is called *diagnostic*. Departure from the null *could* be due to moments between the $(u + 1)th$ to the $(2u)th$. However, if the model is correct then in large samples significance will be due to moments up to the u*th*, and we believe that is where most attention should focus.

If $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$ is significant then 2 by c tables could be examined in a multiple comparisons fashion. These statistics are asymptotically independent and each is well approximated by the χ_{r-1}^2 distribution. Again *asymptotic* means $n_{..} \to \infty$. Effectively we have a decomposition of X_P^2 into components that assesses, under the hypothesis of homogeneity of rows, the agreement of the rows of the table in regard to specific moment effects, up to the $(c - 1)th$ moment. The statistics are asymptotically

independent, and hence so are the assessments. By analogy with our goodness of fit work reported in Rayner and Best (1989a), we expect that the most significant effects will be in the first two to four moments.

4.4 Partitioning X_P^2 Using Score Statistics

We test for equality of the corresponding row probabilities by first setting, for j = 1, ... , c and i = 1, ... , r,

$$p_{ij} = \left\{1 + \sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}}\right\} p_{.j}. \qquad (4.1)$$

In (4.1) we note the following.

- The θ_{ui} are real valued parameters.
- The $\{g_{uj}\}$ is taken to be orthonormal:

$$\sum_{j=1}^{c} g_{uj}\, g_{vj}\, p_{.j} = \delta_{uv} \text{ for } u, v = 1, \ldots, c - 1,$$

where δ_{uv} is the Kronecker delta, $\delta_{uv} = 1$ for $u = v$, and $= 0$ for $u \neq v$. Typically the g_{uj} depend on the $p_{.j}$; we require that they do not depend on the row, since otherwise row comparison would be virtually impossible. A number of choices for g_{uj} are possible, but for ordered categories a ready interpretation is available if we use the $g_u(j)$ of section 4.2. Then each θ_{ui} reflects the u*th* moment shift of the distribution defined by the i*th* row from that defined on the $\{p_{.j}\}$.

- In the goodness of fit context we call k the *order* of the model. It can be at most c - 1, when the model becomes saturated, and an identity similar to Fisher's identity (given, for example, by Lancaster 1969, Theorem 2.1, Corollary 2) would result. Normally k would be chosen to be at most four, and more usually two.

Recall from sections 1.4 and A.1.4 and that components are at least asymptotically mutually independent, and have sum, or sum of squares, the original omnibus statistic. Although the sum of the $\hat{V}_{ui}^2$ is X_P^2, the $\hat{V}_{ui}^2$

are not even asymptotically independent, and the partition of X^2_P using the $\hat{V}^2_{ui}$ could be thought of as an arithmetic rather than a statistical partition. On the other hand, the sums $\hat{V}^2_{u1} + ... + \hat{V}^2_{ur}$ are asymptotically independent and provide a partition in our usual sense.

In using the $\{g_{uj}\}$, we are effectively assigning scores $\{j\}$ to the ordered categories. The subsequent derivations generalise to user-assigned scores $\{x_j\}$. It should, however, be emphasised that our approach assumes user-assigned and not estimated scores.

Colleagues have commented that although models of the form (4.1) are well known to sometimes give excellent results, they can also produce negative probabilities. However (4.1) is asymptotically equivalent to

$$p_{ij} = C(\theta) \left\{ \exp\left[\sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}} \right] \right\} p_{.j} ,$$

where $\theta = (\theta_{11}, ... , \theta_{1r}, ... , \theta_{k1}, ... , \theta_{kr})^T$. This model of course cannot produce estimates of probabilities that are negative. The score tests from the two different models are asymptotically equivalent, but the derivations for (4.1) given here, messy as they are, are much simpler than for the exponential model above. Whatever the asymptotic optimality probabilities of the test statistics from both models, they will ultimately be judged on their practicality, convenience and small sample properties. By these standards, the statistics derived here are not inferior to *any* of the competitor tests known to us.

To test for equality of the corresponding row probabilities, take θ_{ui} to be the $[(u - 1)r + i]$*th* element of a vector θ. We test H_0: $\theta = 0$ against K: $\theta \neq 0$ with $p_{.1}, ... , p_{.(c-1)}$ as nuisance parameters; $p_{.c}$ is omitted from the set of nuisance parameters because the constraints $p_{i1} + ... + p_{ic} = 1$, i = 1, ... , r imply $p_{.1} + ... + p_{.c} = 1$.

Subsequently we write $f = (\sqrt{n_{1.}}, ... , \sqrt{n_{r.}})^T$, $f^* = (\sqrt{n_{1.}}, ... , \sqrt{n_{(r-1).}})^T$ and I_n for the n by n identity matrix. Note that observed values of the N_{ij} are written n_{ij}, etc. The logarithm of the likelihood for our model is

$$\ell = \text{constant} + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log p_{ij}.$$

If θ_a and p_b are typical elements of θ and p respectively, then the efficient score is defined by $V = (\partial \ell / \partial \theta_a)$, and the asymptotic covariance matrix by

$$\Sigma = I_{\theta\theta} - I_{\theta p} I_{pp}^{-1} I_{p\theta},$$

in which

$$I_{\theta\theta} = (- E[\partial^2 \ell / \partial \theta_a \partial \theta_b]), \; I_{\theta p} = (- E[\partial^2 \ell / \partial \theta_a \partial p_b]),$$
$$I_{p\theta} = I_{\theta p}^{T}, \text{ and } I_{pp} = (- E[\partial^2 \ell / \partial p_a \partial p_b]).$$

See A.5.2. The score statistic is of the form $V_0^T \Sigma_0^{-1} V_0$, in which the subscript zero indicates evaluation under the null hypothesis. Subsequently we will need to find the inverse of I_{pp} for our model. For this purpose the following lemma will be needed; it may be easily verified.

Lemma 4.1. Let a be a constant, D an n by n diagonal matrix, and w an n by 1 vector. Provided $1 + a\, w^T D^{-1} w \neq 0$, put $b = - a/\{1 + a\, w^T D^{-1} w\}$. Then

$$(D + a\, ww^T)^{-1} = D^{-1} + b\, D^{-1} ww^T D^{-1}.$$

In our model $p = (p_{.1}, \ldots, p_{.(c-1)})$ is a $(c - 1)$ by 1 vector and $\theta = (\theta_{11}, \ldots, \theta_{1r}, \ldots, \theta_{k1}, \ldots, \theta_{kr})$ is kr by 1. We write θ_{wa} for a typical element of θ, where θ_{wa} is the $[(w - 1)r + a]th$ element of θ. The same convention is used for the efficient score and elsewhere. Note that for $u = 1, \ldots, k$

$$\sum_{j=1}^{c} g_{uj}\, p_{.j} = 0.$$

We call these the zero mean conditions, and they follow from our choice of orthonormal functions.

Subsequently we write 1_n for the n by 1 vector with every element 1. In the derivations that follow, the p_{ic}, i = 1, ... , r, will be treated as algebraically dependent variables.

Theorem 4.1. For the model (4.1), the information matrix evaluated at $\hat{p}_{.j} = N_{.j}/n_{..}$, j = 1, ... , c, is given by the direct sum of k matrices $I_r - ff^T/n_{..}$. This information matrix is singular.

Proof. Using

$$\ell = \text{constant} + \sum_i \sum_j n_{ij}\{\log(1 + \sum_u \theta_{ui}\, g_{uj}\, /\sqrt{n_{i.}}) + \log p_{.j}\},$$

it follows that

$$\partial\ell/\partial\theta_{wa} = \sum_j n_{aj} g_{wj}/\{\sqrt{n_{a.}}(1 + \sum_u \theta_{ua}\, g_{uj}\, /\sqrt{n_{a.}})\},$$

$$\partial\ell/\partial p_{.s} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}\sum_u \theta_{ui}(\partial g_{uj}/\partial p_{.s})}{\sqrt{n_{i.}}\,(1 + \sum_u \theta_{ui} g_{uj}/\sqrt{n_{i.}})} + \frac{n_{.s}}{p_{.s}} - \frac{n_{.c}}{p_{.c}}$$

since $\sum_i n_{ij} = N_{.j}$, and

$$\partial^2\ell/\partial\theta_{wa}\partial\theta_{zb} = -\,\delta_{ab} \sum_j n_{aj}\, g_{wj}\, g_{zj}\, (\sqrt{n_{a.}} + \sum_u \theta_{ua}\, g_{uj})^{-2}.$$

After further differentiation and some manipulation

$$\partial^2\ell/\partial p_{.s}\partial\theta_{wa} = \sum_j (n_{aj}/\sqrt{n_{a.}})(\partial g_{wj}/\partial p_{.s}) + \text{terms zero when } \theta = 0$$

and

$$\partial^2\ell/\partial p_{.s}\partial p_{.t} = -\,\delta_{st}\, n_{.t}/p_{.t}^2 - n_{.c}/p_{.c}^2 + \text{terms zero when } \theta = 0.$$

Taking E_0 of the second order derivatives and evaluating at $\hat{p}_{.j} = N_{.j}/n_{..}$, j = 1, ... , c, gives $I_{\theta\theta} = I_{kr}$ using the orthonormality conditions. Also

$$I_{pp} = (n_{..}/p_{.c} + \delta_{st} n_{..}/p_{.t}) = \text{diag}(n_{..}/p_{.s}) + (n_{..}/p_{.c})\, 1_{(c-1)} 1_{(c-1)}^T.$$

The matrix $I_{\theta p}$ has typical element $(\sqrt{n_{a.}} \sum_j p_{.j} [\partial g_{uj}/\partial p_{.s}])$. To simplify this, differentiate the zero mean conditions $\sum_j g_{uj}\, p_{.j} = 0$ with respect to $p_{.s}$, to give

$$0 = g_{us} - g_{uc} + \sum_j p_{.j} [\partial g_{uj}/\partial p_{.s}],$$

so that $\sum_j p_{.j} [\partial g_{uj}/\partial p_{.s}] = g_{uc} - g_{us}$. Thus, the kr by (c - 1) matrix $I_{\theta p}$ satisfies

$$I_{\theta p} = (\sqrt{n_{a.}}\, [g_{us} - g_{uc}]).$$

To evaluate $I_{\theta p} I_{pp}^{-1} I_{p\theta}$ we first need I_{pp}^{-1}, and by the Lemma 4.1, $n_{..} I_{pp}^{-1} = \text{diag}(p_{.s}) - (p_{.s} p_{.t})$. Now on using

$$\sum_{j=1}^{c-1} (g_{uc} - g_{uj}) p_{.j} = \sum_{j=1}^{c} (g_{uc} - g_{uj}) p_{.j} = g_{uc} - \sum_{j=1}^{c} g_{uj}\, p_{.j} = g_{uc} \text{ and}$$

$$\sum_{j=1}^{c-1} (g_{uj} - g_{uc}) p_{.j}\, (g_{ws} - g_{wc})$$

$$= \sum_{j=1}^{c} \{g_{uj}\, g_{ws} - g_{uc}\, g_{ws} - g_{wc}\, g_{uj} + g_{uc} g_{wc}\} p_{.s}$$

$$= \sum_{j=1}^{c} g_{uj}\, g_{ws}\, p_{.j} + g_{uc} g_{wc} = \delta_{uw} + g_{uc} g_{wc},$$

we find that $I_{\theta p}\, I_{pp}^{-1} I_{p\theta}$ is the direct sum of k equal matrices $I_r - ff^T/n_{..}$. The stated information matrix now follows. It is singular because f is a

eigenvector with zero eigenvalue.

The score statistic involves the inverse of the information matrix. One way to overcome the information matrix being singular is to omit θ_{1r}, ... , θ_{kr} from the model. In modelling terms, the reason for doing this is that the θ's model differences between the row distributions and the average (fitted) distribution $\{N_{.j}/n_{..}\}$. From the average and any r - 1 rows, the other row can be deduced.

Theorem 4.2. For u = 1, ... , k and i = 1, ... , r define $\hat{V}_{ui} = \sum_j N_{ij} \hat{g}_{uj} / \sqrt{n_{i.}}$. The score statistic for the model

$$p_{ij} = \{1 + \sum_{u=1}^{k} \theta_{ui} g_{uj} / \sqrt{n_{i.}}\} p_{.j},$$

for i = 1, ... , r - 1 (**not** r as in (4.1)) and j = 1, ... , c - 1, with $p_{rj} = p_{.j} - p_{1j} - ... - p_{(r-1)j}$ for j = 1, ... , c - 1, and $p_{ic} = 1 - p_{i1} - ... - p_{i(c-1)}$, i = 1, ... , r, is

$$\hat{S}_k = \hat{V}_1^T \hat{V}_1 + ... + \hat{V}_k^T \hat{V}_k$$

in which $\hat{V}_u = (\hat{V}_{u1}, ... , \hat{V}_{ur})^T$. The $\hat{V}_u$ are asymptotically independent.

Proof. After omitting θ_{1r}, ... , θ_{kr} from the model, the information matrix becomes M*, the direct sum of k matrices $I_{r-1} - f^* f^{*T}/n_{..}$, where f* is the (r - 1) by 1 vector formed from f by omitting $\sqrt{n_{r.}}$. From the lemma this matrix has inverse $I_{r-1} + f^* f^{*T}/n_{r.}$, and the inverse of the information matrix is the direct sum of k such matrices. The efficient score is

$$\hat{V}^* = (\hat{V}_{11}, ... , \hat{V}_{1(r-1)}, ... , \hat{V}_{k1}, ... , \hat{V}_{k(r-1)})^T,$$

where the hats indicate that the maximum likelihood estimators of the nuisance parameters are required, namely $\hat{p}_{.j} = N_{.j}/n_{..}$, j = 1, ... , c. If we put $\hat{V}_w^* = (\hat{V}_{w1}, \ldots , \hat{V}_{w(r-1)})^T$, w = 1, ... , k, then $\hat{V}^*$ can also be expressed as $\hat{V}^* = (\hat{V}_1^{*T} , \ldots , \hat{V}_k^{*T})^T$. Substitution gives the score statistic as

$$\hat{S}_k = \hat{V}^{*T} \hat{M}^{*-1} \hat{V}^* = \sum_{w=1}^{k} \hat{V}_w^{*T} \{I_{r-1} + f^* f^{*T}/n_{r.}\} \hat{V}_w^*.$$

The contribution of the w*th* order terms is

$$\hat{V}_w^{*T} \{I_{r-1} + f^* f^{*T}/n_{r.}\} \hat{V}_w^* = \sum_{i=1}^{r-1} \hat{V}_{wi}^2 + \{\sum_{i=1}^{r-1} \sum_{j=1}^{c} N_{ij} \hat{g}_{wj}\}^2/n_{r.}.$$

This simplifies if we first notice that

$$\sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \hat{g}_{wj} = \sum_{j=1}^{c} N_{.j} \hat{g}_{wj} = n_{..} \sum_{j=1}^{c} \hat{g}_{wj} \hat{p}_{.j} = 0,$$

using $N_{.j} = n_{..} \hat{p}_{.j}$ and the zero mean conditions. In terms of the $\hat{V}_w^*$, this is

$$\sqrt{n_{1.}} \hat{V}_{w1} + \ldots + \sqrt{n_{r.}} \hat{V}_{wr} = 0, \; w = 1, \ldots , k.$$

Hence

$$\sum_{i=1}^{r-1} \sum_{j=1}^{c} N_{ij} \hat{g}_{wj} = \sum_{i=1}^{r-1} \sqrt{n_{i.}} \hat{V}_{wi} = - \sqrt{n_{r.}} \hat{V}_{wr} \text{ and}$$

$$\hat{V}_w^{*T} \{I_{r-1} + f^* f^{*T}/n_{r.}\} \hat{V}_w^* = \hat{V}_{w1}^2 + \ldots + \hat{V}_{wr}^2, \text{ for } w = 1, \ldots , k.$$

Note that $\hat{V}_u$ gives information about the deviations of order u from the fitted distribution $\{N_{.j}/n_{..}\}$; there are contributions to this information from all r rows. The asymptotic covariance matrix of $(\hat{V}_{u1}, \ldots , \hat{V}_{u(r-1)})$ is

derived incidentally in the proof of Theorem 4.2 to be $I_{r-1} + f^*f^{*T}/n_{r.}$. It follows that the $\widehat{V}_{ui}$ are correlated and hence are not components. Also, since $\widehat{V}_u$ is asymptotically r - 1 variate normal with mean zero and covariance matrix $I_{r-1} + f^*f^{*T}/n_{r.}$, the contribution to $\widehat{S}_k$ from the u*th* order terms, $\widehat{V}^2_{u1} + \ldots + \widehat{V}^2_{ur}$, asymptotically has the χ^2_{r-1} distribution. This uses a result for multivariate normal random variables, given for example, in Stuart and Ord (1994, 15.14). Given the asymptotic independence of the $\widehat{V}_u$, $\widehat{S}_k$ has asymptotic distribution $\chi^2_{k(r-1)}$.

Theorem 4.3. $\widehat{S}_{(c-1)} = X^2_P$. Although dependent, the $\widehat{V}_{ui}$ partition X^2_P arithmetically in that the sum of the squares of all r(c - 1) $\widehat{V}_{ui}$ add to give X^2_P.

Proof. The proof is similar to that of Rayner and Best (1989a, Theorem 5.1.2). Write $H = (\widehat{g}_{wj})$, and for i = 1, ... , r, $\widehat{U}_i = (\widehat{V}_{1i}, \ldots, \widehat{V}_{(c-1)i})^T$ and $N_i = (N_{i1}, \ldots, N_{ic})^T$. Then by definition $\widehat{U}_i = HN_i/\sqrt{n_{i.}}$. If we now put

$$\widehat{V}^2_{1i} + \ldots + \widehat{V}^2_{(c-1)i} = X^2_i,$$

the sum of the squares of the $\widehat{V}_{ui}$ corresponding to each row, then

$$X^2_i = \widehat{U}^T_i\widehat{U}_i = N^T_i H^T H N_i/n_{i.}.$$

Putting $\widehat{p} = (\widehat{p}_{.1}, \ldots, \widehat{p}_{.c})^T$, the zero mean condition implies $H\widehat{p} = 0$. The orthonormality condition may be expressed as $H^*\mathrm{diag}(\widehat{p}_{.s})H^{*T} = I_c$, where H* is H augmented by a c*th* row of ones. This implies that $\mathrm{diag}(\widehat{p}^{-1}_{.s}) = H^{*T}H^* = H^TH + 1_c1^T_c$, where 1_c is c by 1 vector of ones. This gives

$$X^2_i = \widehat{U}^T_i\widehat{U}_i = (N_i - n_{i.}\,\widehat{p})^T H^T H(N_i - n_{i.}\,\widehat{p})/n_{i.}$$

$$= (N_i - n_{i.}\,\hat{p})^T\{\mathrm{diag}(\hat{p}_{.s}^{-1}) - 1_c 1_c^T\}(N_i - n_{i.}\,\hat{p})/n_{i.}$$
$$= \sum_j (N_{ij} - n_{i.}\,\hat{p}_{.j})^2/(n_{i.}\,\hat{p}_{.j}).$$

This is of the form of Pearson's X_P^2, and is clearly the contribution to X_P^2 from the i*th* row. Summing over rows gives

$$X_P^2 = \sum_i \sum_j (N_{ij} - n_{i.}\,\hat{p}_{.j})^2/(n_{i.}\,\hat{p}_{.j}) = \sum_i X_i^2.$$

The non-diagonal covariance matrix establishes the dependence.

Notice that this result is different from that obtained in Chapter 3, as here only one margin is fixed whereas there both margins were fixed.

As remarked before the statement of Theorem 4.2, there are only (r - 1)(c - 1) functionally independent θ's, for one row may be determined from the average multinomial distribution by the other r - 1 row distributions. Each $\hat{V}_{ua}$ corresponds to a θ_{ua}, and assesses the deviation of the u*th* moment of the i*th* row distribution $\{p_{i1}, \ldots, p_{ic}\}$ from the u*th* moment of the distribution defined by the $\{N_{.1}/n_{..}, \ldots, N_{.c}/n_{..}\}$. In fact $\hat{V}_{ua}^2$ could be derived as the score statistic for the model: $p_{aj} = \{1 + \theta_{ua}\, g_{uj}/\sqrt{n_{a.}}\}p_{.j}$, and $p_{ij} = p_{.j}$ for all (r - 2)(c - 1) other p_{ij} in the first r - 1 rows. Therefore each $\hat{V}_{ua}$ is the basis of a strongly directional test, with one dimensional parameter space $\{\theta_{ua}\}$. In the same vein we confirm that $\hat{V}_{u1}^2 + \ldots + \hat{V}_{ur}^2$ has the χ_{r-1}^2 distribution by observing that it is the score statistic for the model $p_{ij} = \{1 + \theta_{ui}\, g_{uj}/\sqrt{n_{i.}}\}p_{.j}$, i = 1, ... , r - 1, j = 1, ... , c - 1. It has r - 1 dimensional parameter space $\{\theta_{u1}, \ldots, \theta_{u(r-1)}\}$. It thus plays a useful intermediate role, being *more directional* than the (r - 1)(c - 1) dimensional X_P^2, and *more omnibus* than each of the $\hat{V}_{u1}^2, \ldots, \hat{V}_{ur}^2$ singly.

Although we have not done a thorough analysis, we suspect that orthogonal polynomials can also be used as part of a log-linear model approach. In that case, the log-likelihood ratio statistic would be partitioned rather than Pearson's X_P^2. We would expect such test statistics

to perform very similarly to those we have just introduced. Everitt (1992, section 7.3), for example, indicates how to proceed.

4.5 Other Methods for Ordered Data

4.5.1 Nair's Method

Nair (1986) defined location and dispersion statistics for our situation. Nair's location statistic is just the Kruskal-Wallis statistic adjusted for ties which is often applied to ranked one-way analysis of variance data. Eubank et al. (1987) showed that a number of commonly used statistics, like the one on which the Kruskal-Wallis test is based, are components of Pearson's X_P^2.

Nair (1986) defined location scores

$$\ell_k = (t_k - 0.5)/\sqrt{\{\sum_{j=1}^{c} N_{.j}(t_j - 0.5)^2/n_{..}\}}$$

in which $t_k = (N_{.1} + ... + N_{.(k-1)} + N_{.k}/2)/n_{..}$, for $1 \leq k \leq c$, and also dispersion scores

$$d_k = e_k/\sqrt{\{\sum_{j=1}^{c} N_{.j}e_j^2/n_{..}\}}$$

in which $e_k = \ell_k \{\ell_k - (N_{.1}\ell_1^3 + ... + N_{.c}\ell_c^3)/n_{..}\} - 1$, for $1 \leq k \leq c$. If we define location and dispersion effects

$$\tau_i = N_{i1}\ell_1 + ... + N_{ic}\ell_c \text{ and } \omega_i = N_{i1}d_1 + ... + N_{ic}d_c,$$

then $\tau_i/\sqrt{n_{i.}}$ and $\omega_i/\sqrt{n_{i.}}$ are analogous to $\hat{V}_{1i}$ and $\hat{V}_{2i}$.

Nair's location scores are proportional to the midrank for category k. Graubard and Korn (1987) criticized the use of rank scores for contingency table analysis on the grounds that they may not give enough weight to extreme categories. The same sort of criticism may also apply to

other data-dependent or estimated scores such as those given in the next section. Nair's statistics can also be derived as in section 4.3 if we use mid-rank scores rather than the *natural* scores 1, 2, ... , c. Thus his statistics are special cases of our partition of X^2_P statistics.

4.5.2 Logistic Models

The partition of X^2_P given in Theorem 4.3 is relevant when either the rows (or columns) have ordered categories and where columns (or rows) have nominal categories. Another model suggested for use in this situation is the logistic model. McCullagh (1980) suggested the model:

$$\log\{(N_{i(j+1)} + \ldots + N_{ic})/(N_{i1} + \ldots + N_{ij})\} = (\alpha_j + \tau_i')/\omega_i'$$
$$\text{in which } 1 \le i \le r,\ 1 \le j \le c - 1 \text{ and } \tau_1' + \ldots + \tau_r' = 0.$$

Iterative methods are needed for maximum likelihood estimation of the parameters. Agresti (1984, Appendix B.3) gave details. Notice that the parameters α_j, $1 \le j \le c - 1$, estimate the scores which are not arbitrarily assigned while the τ_i' and ω_i' are location and dispersion parameters with $1 \le i \le r$. However it should be noted that there is some evidence, for example Agresti (1984, p. 225), Newell (1986) and Box and Jones (1986), that in some cases there is little difference in both location and dispersion effects for assigned scores and estimated scores.

The logistic method just described requires iteration. However our X^2_P method, which includes Nair's midrank scores method, does not. Further, if we take $\tau_i^* = \widehat{V}_{1i}\sqrt{n_{i.}}$ with $1 \le i \le r$, then

$$\sum_{i=1}^{r} \tau_i^* = \sum_{i=1}^{r} \sqrt{n_{i.}}\,\{\sum_{j=1}^{c} N_{ij}g_1(j)/\sqrt{n_{i.}}\} = \sum_{j=1}^{c} N_{.j}g_1(j) = 0.$$

We now have

$$\sum_{i=1}^{r} \tau_i = \sum_{i=1}^{r} \tau_i' = \sum_{i=1}^{r} \tau_i^* = 0.$$

In this sense, the τ_i, τ_i', and τ_i^* are all contrasts. For completeness we also define $\omega_i^* = \hat{V}_{2i}\sqrt{n_{i.}}$, i = 1, ... , r.

4.5.3 ANOVA Analysis

Box and Jones (1986) and Nair (1990) suggested the use of analysis of variance (ANOVA) methods, and user defined assigned scores, to analyse ordered categorical data. However such an analysis relies on more assumptions to justify its use, and these additional assumptions may be difficult to justify. Sometimes the ANOVA method can give a non-orthogonal analysis, which is less convenient from many standpoints. Further, Brown (1988) has done a small simulation study which indicates that, when compared to X_P^2 tests, the ANOVA tests have actual sizes further from the nominal sizes. For these reasons we do not consider ANOVA methods further, although we have often used them in the past for the analysis of ordered categorical data. It may be worth extending the simulation study given by Brown (1988).

4.6 Small Sample Size and Power Comparisons

The X_P^2 method, that includes Nair's method, and McCullagh's method both partition the total X_P^2 value into location, dispersion and residual effects. The location and dispersion test statistics are each associated with r - 1 degrees of freedom, and have asymptotic χ_{r-1}^2 distributions. We now embark upon a limited simulation study of the Nair and McCullagh methods.

The standard multinomial model for each row of the contingency table will be assumed where row totals are fixed and the multinomial probabilities are the same for each row. For simplicity we take row totals to be equal at (N/r), where N is the total of all the counts. Similar results are obtained when these totals are not equal. Random multinomial samples were generated as suggested by Devroye (1986, p. 558).

Table 4.1 examines the corresponding χ^2_{r-1} location test statistics proposed by Nair (1986) and McCullagh (1980). It shows the actual test size for a nominal 5% size using the χ^2 critical value, and sample moments $\bar{x}$, s^2, g_1 and g_2 of the location test statistics. The moments and sizes are based on Monte Carlo samples of 1,000 using the r, N and p_i values shown in the table. The expected values of $\bar{x}$, s^2, g_1 (skewness) and g_2 (kurtosis) if the χ^2_ν approximation holds are ν, 2ν, $\sqrt{8}/\nu$ and $12/\nu$ respectively.

Calculation of the McCullagh statistics was about 40 times slower than the Nair statistics using a Sun 3/50 workstation. There were also occasional problems of non-convergence with McCullagh's method. These observations

Table 4.1 Size and moment estimates (Nair, McCullagh) based on 1,000 simulations for the location test statistic with r, p_i, i = 1(1)4 and N as shown

r	(p_1, p_2, p_3, p_4)	N	Size (%)	$\bar{x}$	s^2	g_1	g_2
			(Target 5%)	1.0	2.0	2.8	12.0
2	(4*.25)	40	(4.7, 5.3)	(1.02, 1.05)	(1.96, 2.14)	(2.46, 2.58)	(7.75, 8.61)
2	(4*.25)	40	(5.8, 6.1)	(1.08, 1.09)	(2.09, 2.16)	(2.18, 2.22)	(6.23.6.56)
2	(3*.2, .4)	40	(4.9, 5.2)	(1.02, 1.04)	(1.98, 2.15)	(2.56, 2.67)	(8.47, 9.32)
2	(3*.2, .4)	100	(5.1, 5.2)	(1.05, 1.06)	(1.89, 1.95)	(2.19, 2.23)	(6.27, 6.58)
2	(.1, 3*.3)	100	(4.7, 5.1)	(1.06, 1.07)	(1.94, 2.01)	(2.44, 2.50)	(9.65, 10.33)
			(Target 5%)	2.0	4.0	1.4	6.0
3	(4*.25)	60	(5.4, 6.0)	(2.04, 2.11)	(4.24, 4.71)	(1.82, 1.96)	(4.28, 5.34)
3	(4*.25)	150	(4.6, 4.9)	(2.01, 2.03)	(3.44, 3.57)	(1.61, 1.64)	(3.11, 3.34)
3	(3*.2, .4)	60	(6.1, 6.8)	(2.05, 2.11)	(4.28, 4.7)	(1.84, 1.94)	(4.02, 4.64)
3	(3*.2, .4)	150	(4.1, 4.3)	(1.98, 2.00)	(3.46, 3.58)	(1.60, 1.62)	(3.01, 3.12)
3	(.1, 3*.3)	150	(5.0, 5.1)	(2.05, 2.07)	(3.62, 3.75)	(1.64, 1.70)	(3.57, 4.00)

Table 4.2 Size and moment estimates (Nair, McCullagh) based on 1,000 simulations for the (location + dispersion) test statistic with r, p_i, i = 1(1)4 and N as shown

r	(p_1, p_2, p_3, p_4)	N	Size (%)	$\bar{x}$	s^2	g_1	g_2
			(Target 5%)	2.0	4.0	1.4	6.0
2	(4*.25)	40	(5.0, 6.0)	(1.97, 2.04)	(3.70, 4.25)	(1.80, 1.93)	(4.47, 5.17)
2	(4*.25)	100	(5.6, 4.6)	(2.10, 2.12)	(4.04, 4.02)	(1.59, 1.64)	(3.32, 3.32)
2	(3*.2, .4)	40	(4.6, 6.9)	(1.96, 2.05)	(3.55, 4.22)	(1.59, 1.77)	(3.04, 4.04)
2	(3*.2, .4)	100	(4.9, 5.2)	(2.12, 2.14)	(4.06, 4.29)	(1.74, 1.81)	(4.49, 5.21)
2	(.1, 3*.3)	100	(4.1, 5.4)	(2.10, 2.14)	(3.84, 4.15)	(1.60, 1.67)	(3.89, 4.09)
			(Target 5%)	4.0	8.0	0.7	3.0
3	(4*.25)	60	(4.5, 4.6)	(3.91, 4.19)	(8.43, 9.55)	(1.62, 1.50)	(4.23, 3.69)
3	(4*.25)	150	(4.4, 4.9)	(3.92, 4.11)	(7.35, 8.40)	(1.24, 1.59)	(1.88, 3.32)
3	(3*.2, .4)	60	(7.6, 7.1)	(4.62, 4.21)	(9.55, 9.40)	(1.19, 1.38)	(1.57, 2.30)
3	(3*.2, .4)	150	(8.2, 4.9)	(5.11, 4.14)	(9.85, 8.41)	(1.09, 1.27)	(1.36, 1.74)
3	(.1, 3*.3)	150	(4.4, 4.9)	(3.78, 4.08)	(7.16, 8.47)	(1.42, 1.50)	(2.87, 3.22)

Table 4.3 Power estimate (Nair, McCullagh) based on 1,000 simulations for the location test statistic with r, p_i, N as shown. The first r - 1 rows had p_i as shown and the r*th* row had $p_i = 0.25$

r	(p_1, p_2, p_3, p_4)	N	power
2	(0.14, 0.21, 0.28, 0.37)	40	(0.20, 0.21)
2	(0.14, 0.21, 0.28, 0.37)	100	(0.39, 0.38)
2	(0.35, 0.15, 0.15, 0.35)	40	(0.04, 0.04)
2	(0.35, 0.15, 0.15, 0.35)	100	(0.05, 0.05)
3	(0.14, 0.21, 0.28, 0.37)	60	(0.16, 0.19)
3	(0.14, 0.21, 0.28, 0.37)	150	(0.41, 0.44)
3	(0.35, 0.15, 0.15, 0.35)	60	(0.06, 0.07)
3	(0.35, 0.15, 0.15, 0.35)	150	(0.05, 0.06)

Table 4.4 Power estimate (Nair, McCullagh) based on 1,000 simulations for the (location + dispersion) statistic with r, p_i, N as shown. The first r - 1 rows had p_i as shown and the r*th* row had $p_i = 0.25$

r	(p_1, p_2, p_3, p_4)	N	power
2	(0.14, 0.21, 0.28, 0.37)	40	(0.13, 0.14)
2	(0.14, 0.21, 0.28, 0.37)	100	(0.32, 0.32)
2	(0.35, 0.15, 0.15, 0.35)	40	(0.19, 0.20)
2	(0.35, 0.15, 0.15, 0.35)	100	(0.46, 0.45)
3	(0.14, 0.21, 0.28, 0.37)	60	(0.15, 0.15)
3	(0.14, 0.21, 0.28, 0.37)	150	(0.35, 0.33)
3	(0.35, 0.15, 0.15, 0.35)	60	(0.15, 0.19)
3	(0.35, 0.15, 0.15, 0.35)	150	(0.37, 0.45)

are important if Monte Carlo p-values were required for a given data set. It appears from Table 4.1 that in the present instance, the χ^2_{r-1} approximation is reasonable for both location statistics. For McCullagh's method this is in agreement with a conclusion in McCullagh and Nelder (1989). Table 4.2 extends the Table 4.1 calculations to look at test sizes for the location plus dispersion statistic, and the $\chi^2_{2(r-1)}$ approximation is again seen to be reasonable. In neither Table 4.1 nor Table 4.2 is one method seen to be preferable to the other. We could also have looked at sizes and powers (i) of the location test adjusted for dispersion and (ii) the dispersion test adjusted for location for the McCullagh method. However, these are not easily compared with the Nair Method. Similar results were obtained for other numbers of rows and columns and for other marginal probabilities.

Tables 4.3 and 4.4 give some power comparisons for the two methods of analysis assuming that χ^2 critical values are appropriate. Tables 4.1 and 4.2 indicate that this is reasonable. The alternatives chosen should be susceptible to either a location or a dispersion test. It appears that, for the limited range of alternatives considered, the location and location plus dispersion tests for both methods considered have power almost exactly the same. This is hardly surprising, since the X^2_P location and dispersion tests are score tests and so will have the usual asymptotic optimality properties, and the corresponding Nair tests coincide with these tests. The McCullagh method gives likelihood ratio tests and so will also have asymptotic optimality properties. Thus asymptotically the X^2_P, Nair, and McCullagh tests will be almost identical. This present simulation study complements the asymptotics by considering small sample comparisons using Monte Carlo simulations.

Which method might be recommended? The size and power results give little evidence of differences between the methods. The iterative McCullagh method can have numerical problems with convergence. For data sets with large cell counts this is not a concern but for sparse tables when Monte Carlo p-values are needed it is an important consideration. Also there are possible problems in explaining to a client the two analyses, as illustrated in the Taste-Test Example following, for the McCullagh method.

Further, another point in favour of the X^2_P and Nair methods is that

their test statistics can be regarded as nonparametric, relying on fewer assumptions. Given the similarity of size and power performance and taking into account these last points, we suggest use of the orthogonal X_P^2 and Nair analyses.

4.7 Examples

The first example below demonstrates a difficulty with the use of the logistic model/McCullagh analysis.

Taste-Test Example. A taste-test experiment from Bradley et al. (1962) gave the response frequencies shown in Table 4.5. In the following the X_P^2 analysis uses the natural scores 1, 2, ... , 5. The Nair analysis is equivalent to an X_P^2 analysis using midrank scores.

This is a somewhat unusual taste test as it appears the judges did not taste each product and so cannot be eliminated as blocks. However, it could be claimed that presenting one product per judge gives a more realistic consumer assessment of the products (McBride, 1986). For this contingency table, recalling that $\tau_i^* = \hat{V}_{1i}\sqrt{n_{i.}}$ and $\omega_i^* = \hat{V}_{2i}\sqrt{n_{i.}}$, we obtain the summaries shown in Table 4.6.

Table 4.5 Response frequencies from a taste-test experiment

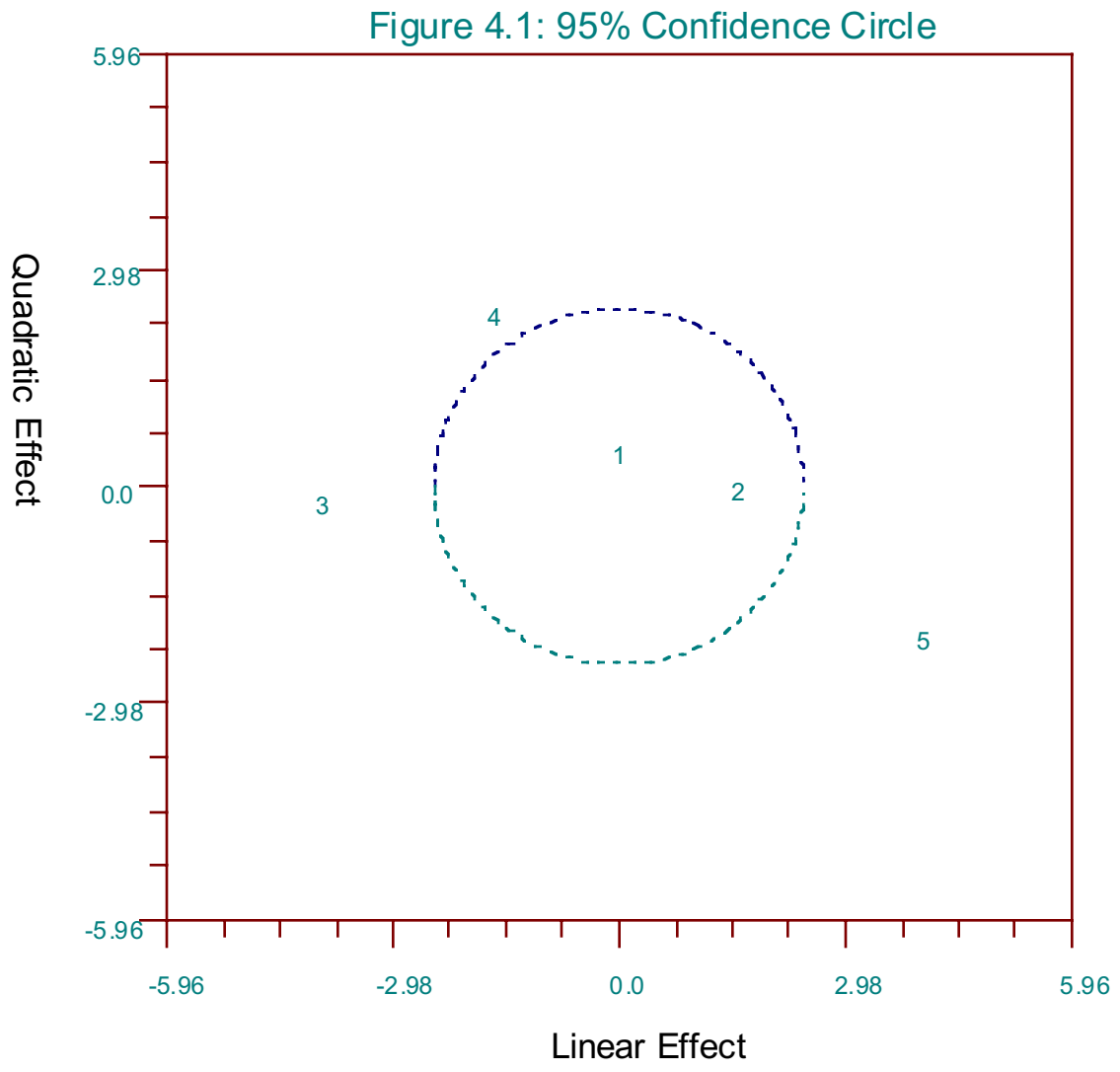
Product	Response Category				
	--	-	ϕ	+	++
1	9	5	9	13	4
2	7	3	10	20	4
3	14	13	6	7	0
4	11	15	3	5	8
5	0	2	10	30	2

Table 4.6 Taste-test data summarized by location and dispersion parameters and three methods of analysis

	Analysis					
	X_P^2 components		McCullagh		Nair	
Product	τ_i^*	ω_i^*	τ_i'	ω_i'	τ_i	ω_i
1	-0.22	2.40	0.07	1.25	-0.35	2.11
2	10.00	-0.87	0.56	1.05	9.85	-2.14
3	-25.06	-1.90	-1.11	0.90	-24.97	-0.04
4	-11.02	14.60	-0.50	1.66	-10.41	16.74
5	26.30	-14.23	0.98	0.51	25.89	-16.68

The analyses are very similar. All indicate that treatment 5 is most liked, as τ_5, τ_5' and τ_5^* are the highest τ values, and that the judges agree on this, as ω_5, ω_5' and ω_5^* are the smallest ω values.

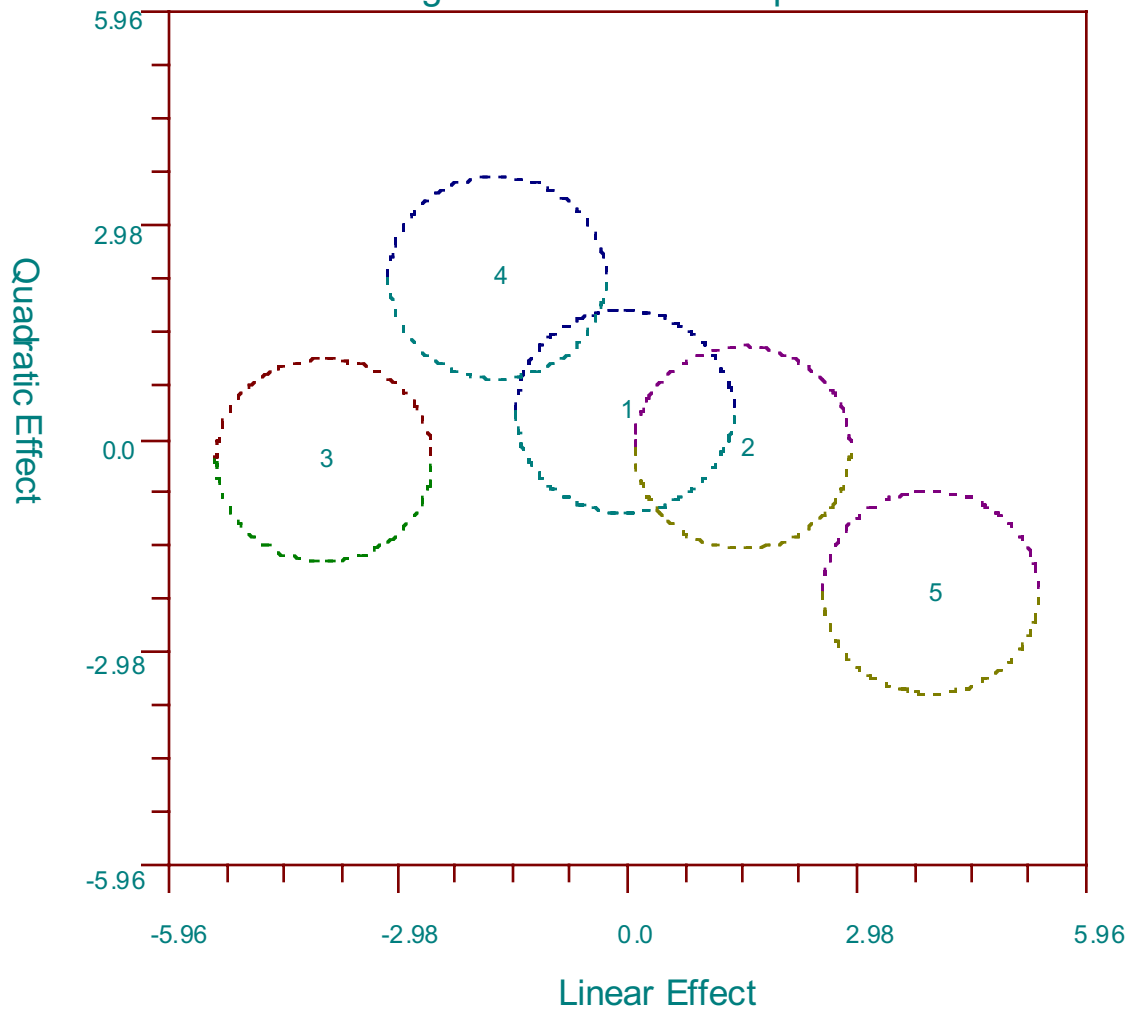
The high ω_4, ω_4' and ω_4^* values indicate the judges' responses were most spread for treatment 4. Reference to the data indicates this spread is real, and not a spurious dispersion effect due to a large location effect, as discussed in Hamada and Wu (1990). Treatment 3 was least liked, as τ_3, τ_3' and τ_3^* are the smallest τ values, and the judges were in some agreement about this, as ω_3, ω_3' and ω_3^* are the second or third smallest of the ω values. Notice the agreement between the analyses and, in particular, the closeness of the X_P^2 and Nair results.



These results are most easily seen from product maps. Figure 4.1 shows a 95% confidence circle. If there are no product differences then it is expected that all products will lie within this circle 95% of the time. The equation of the circle is $x^2 + y^2 = 6$ where 6.0 is the 95% point of the χ^2_2 distribution. Clustering the products can be done using confidence circles and Figure 4.2 gives an example using a radius based on an approximate 95% least significant difference (LSD) value, 1.4142. This value is the approximate standard deviation of the difference between any two of $\hat{V}_{1i}$, $\hat{V}_{2i}$ and $\hat{V}_{3i}$, all of which are well approximated by the standard normal distribution. In one dimension the 95% LSD will be 1.96 times this

standard deviation and so if circles with this radius overlap, the one dimensional LSD will not have been exceeded. Products 1 and 2 are clustered. This clustering given by Figure 4.2 is a two-dimensional analogue of a one-dimensional LSD multiple comparisons procedure. If two dimensions are needed to separate the data then it is less conservative to use confidence circles rather than squares based on two sets of one-dimensional multiple comparisons.

Figure 4.2: Product Map



We can also partition the value of the usual X_P^2 homogeneity statistic. We obtain the analysis shown in Table 4.7.

However, because the logistic analysis is not orthogonal, different

analyses result depending on whether location or dispersion effects are removed first. In fact we have either of the partitions shown in Table 4.8. In this case the conclusions from either logistic analysis are the same, but it is not clear that this would always be so.

Finally we note that the residual has p-value less than 1% in the analyses given above. We recommend continuing the partition until the residual is not significantly small if there is value, in terms of interpreting the data, in so doing. In this case the residual on 8 degrees of freedom and taking the value 27.33 splits into skewness or cubic and kurtosis or quartic terms, each on 4 degrees of freedom, taking the values 23.02 and 4.31 respectively. The corresponding χ^2 p-values are less than 0.001 and 0.376 respectively. There is evidence of skewness but not kurtosis effects.

Table 4.7 Partition of X_P^2 for the taste-test data

Effect	df	SS	χ^2 p-value
Location	4	36.58	0.000
Dispersion	4	9.93	0.042
Residual	8	27.33	0.001
Total	16	73.84	

Table 4.8 Alternative partitions of the log-likelihood statistic

Effect	df	SS_1	χ^2 p-value	SS_2	χ^2 p-value
Location	4	36.11	0.000	40.89	0.000
Dispersion	4	27.76	0.000	22.98	0.000
Residual	8	21.34	0.006	21.34	0.006
Total	16	85.21		85.21	

Ulcer Example. The overall X_P^2 value of 73.84 for the Taste-Test Example was highly significant. However, it can be the case that the overall X_P^2 is not significant but that the examination of location and/or dispersion statistics will indicate a significant effect. To illustrate this point consider the data in Table 4.9 which are taken from Armitage (1955) and which are concerned with two treatments for ulcers and subsequent categorization of ulcer severity. Take the category scores to be 1, 2, 3 and 4.

For these ulcer data $X_P^2 = 5.91$ on three degrees of freedom implies a p-value in excess of 10% if the usual χ^2 approximation for X_P^2 is assumed. However the location statistic $\hat{V}_{11}^2 + \hat{V}_{21}^2 = 5.26$, with a p-value between 1% and 5%. Treatment B is superior to treatment A, tending to have more responses in the number healed/mostly healed categories. The residual is 0.65 on two degrees of freedom, indicating there are no dispersion and no skewness effects. This example emphasizes the need to look at the $\hat{V}_{ui}$ and not just X_P^2, as in X_P^2 the insignificant dispersion and skewness effects have masked a significant location effect.

Table 4.9 Responses for two ulcer treatments

Treatment	# larger	# slightly healed	# mostly healed	# healed
A	12	10	4	6
B	5	8	8	11

Suppose now we do not wish to use the 'natural' scores 1, 2, 3 and 4. Because 'larger' is such an undesirable outcome, and 'healed' is such a desirable outcome, we use the scores -10, -1, 1, 10. These scores give the results in Table 4.10. In this case the different scores do not change the conclusions.

One possible data dependent scoring method uses the ranks. Suppose we take the scores to be the 'mid-ranks', or 'average-ranks'. The mid-rank for 'larger' is (17 + 1)/2 = 9, for 'slightly healed' is 17 + (18 + 1)/2 = 26.5, for 'mostly healed' is 17 + 18 + (12 + 1)/2 = 41.5, and for 'healed' is 17 + 18 + 12 + (17 + 1)/2 = 56. Use of the scores 9, 26.5, 41.5 and 56 gives the results in Table 4.11.

Table 4.10 Partition of X_P^2 for new scores

Source	SS	df	p-value
Linear	4.629	1	0.03
Quadratic	0.281	1	0.60
Residual	0.998	1	0.32
X_P^2	5.908	3	0.12

Table 4.11 Partition of X^2_P for mid-rank scores

Source	SS	df	p-value
Linear	5.339	1	0.02
Quadratic	0.218	1	0.64
Residual	0.351	1	0.55
X^2_P	5.908	3	0.12

Again, for these data the change in scores has not altered the conclusions.

As a matter of interest, the parametric row effects log-linear model has an associated X^2 test statistic value of 5.91 when scores of 1, 2, 3 and 4 are used for the categories. This is the same value as $V_1^T V_1$ for these data. We have noted this similarity previously, in Chapter 3, section 3. Agresti (1984, section 5.2) discusses row effects log-linear models.

5
Further Tests Based on a Product Multinomial Model: Order in the Sign Test and Ordinal Categorical Data with a Factorial Response

---

5.1 Introduction

We now consider two topics that draw directly on the product multinomial model. The first is how order affects the sign test. The successful model for the sign test when order is important is a product binomial - a simpler version of the model used in the previous section to derive Yates' test and extensions of it. The second topic relates to data analysis in which the structure of the model allows us to take advantage of the results of the previous chapter. Examples are given in sections 5.6, 5.9 and 5.10.

5.2 How Order Affects the Sign Test

Gart's (1969) tests are used when the appropriate model is the sign test, but the order in which the treatments are presented to subjects may be important. We subsequently derive score tests for treatment and order effects for this situation. They are equivalent to the tests obtained by Gart using a logistic ancillary framework, which was also the basis of the power study of Nam (1971). We reassess size and power from our simpler perspective.

Taste-test standards such as AS2542.2.1 Standards Australia (1982)

and ISO5495.2 International Standards Organisation (1979) only suggest analysis using the sign test. As a consequence of our results here, we suggest the statistical methods sections in these standards and in sensory evaluation textbooks should discuss Gart's test, as should other discussions of sign tests where order effects are possible.

5.3 The Sign Test and Gart's Tests

Subjects are asked which of two treatments, A and B, they prefer. A fixed number, $n_{1.}$, are given A before B, while $n_{2.}$, also fixed number, are given B before A. Of the $n_{1.}$ who receive A first, N_{11} prefer A, while N_{12} prefer B. Of the $n_{2.}$ who receive B first, N_{21} prefer A, while N_{22} prefer B. The sign test would decide if A and B were equally preferred using $N_{.1}$, the total number of subjects who preferred A. The total number of subjects is $n_{..} = n_{1.} + n_{2.}$, from which the total number who preferred B is $n_{..} - N_{.1} = N_{.2}$. In practice experimenters usually take $n_{1.} = n_{2.}$.

	Prefer A	Prefer B	Total
A first	N_{11}	N_{12}	$n_{1.}$
B first	N_{21}	N_{22}	$n_{2.}$
	$N_{.1}$	$N_{.2}$	$n_{..}$

Under the null hypothesis of no difference between the treatments A and B, the number of subjects who prefer A follows a binomial distribution with parameters $n_{..}$ and 0.5. This leads to the sign test in which the exact probability of $N_{.1} = n_{.1}$ or more subjects preferring A is

$$P_S = \sum_{i = n_{.1}}^{n_{..}} {}^{n_{..}}C_i \, (0.5)^{n_{..}}.$$

Often an approximate test is made by referring the continuity corrected statistic

$$Z_S = \frac{|N_{.1} - N_{.2}| - 1}{\sqrt{n_{..}}}$$

to tables of the standard normal distribution. Clearly the sign test ignores the order in which the treatments are presented.

To give a test for a treatment effect which does account for the order effect, Gart (1969) used a logistic model and an argument involving ancillary statistics, and obtained an exact test of correlated, ordered proportions formally equivalent to Fisher's exact test for 2 by 2 tables. In large samples the continuity corrected statistic

$$Z_{GT} = \frac{\left|N_{11} - \frac{n_{1.}(N_{11} + N_{22})}{n_{..}}\right| - \frac{1}{2}}{\sqrt{\frac{n_{1.}n_{2.}(N_{11} + N_{22})(N_{12} + N_{21})}{n_{..}^2(n_{..} - 1)}}}$$

$$= \frac{|N_{11}N_{21} - N_{12}N_{22}| - \frac{n_{..}}{2}}{\sqrt{n_{1.}\, n_{2.}\, (N_{11} + N_{22})\, (N_{12} + N_{21})/(n_{..} - 1)}}$$

may be referred to tables of the standard normal distribution. Occasionally either $N_{11} + N_{22} = 0$ or $N_{12} + N_{21} = 0$, in which case the test statistic is undefined. Such outcomes should be assigned to either an acceptance or a rejection region before sighting the data. Often, but not always, these outcomes have low probability, and the assignment is inconsequential. Here we assign $Z_{GT} = 0$ when it would otherwise be undefined.

In small samples the p-value for a one-tailed exact test of A being equally preferred to B against a one-sided alternative is

$$P_{GT} = \sum_{tail} {}^{n_{1.}}C_i \; {}^{n_{2.}}C_{N_{11} + N_{22} - i} / {}^{n_{..}}C_{N_{11} + N_{22}},$$

where the tail is determined by ordering probabilities less than or equal to the observed p-value.

Gart's one-tailed exact test of no order effect against a one-sided alternative is based on the continuity corrected statistic

$$Z_{GO} = \frac{\left|N_{11} - \frac{n_{1.}N_{.1}}{n_{..}}\right| - \frac{1}{2}}{\sqrt{\frac{n_{1.}n_{2.}N_{.1}N_{.2}}{n_{..}^2(n_{..} - 1)}}} = \frac{|N_{11}N_{22} - N_{12}N_{21}| - \frac{n_{..}}{2}}{\sqrt{n_{1.}\, n_{2.}\, N_{.1}\, N_{.2}/(n_{..} - 1)}},$$

which, in large samples, may be referred to tables of the standard normal distribution. If either $n_{.1} = 0$ or $n_{.2} = 0$, the test statistic is undefined. For these outcomes we assign $z_{GO} = 0$. The p-value for Gart's one-tailed exact test of no order effect against a one-sided alternative is

$$P_{GO} = \sum_{tail} {}^{n_{1.}}C_i \, {}^{n_{2.}}C_{N.1 - i} / {}^{n_{..}}C_{N.1},$$

where the tail is determined by ordering probabilities less than or equal to the observed p-value.

Nam (1971) compared the power functions of Gart's treatment test and the sign test. He found that if there was no order effect then the sign test was more powerful. If there is an order effect, the sign test is biased, and Gart's test was recommended. As the statistician does not know *a priori* whether there is or is not an order effect, a preliminary test for an order effect, for example, based on Z_{GO}, was needed. We now suggest a new model and, based on this model, that Gart's test can be applied without use of a preliminary test.

5.4 A New Model and Score Test

According to the notation of section two, the row totals are fixed and the column totals are random. This is reasonable, since the number of subjects who receive treatment A before B, $n_{1.}$, is determined before collecting the data. The number who prefer A, $N_{.1}$, can only be determined after collecting the data. A reasonable model here is the product binomial

model.

Let p_{ij} denote the probability of an observation in the (i, j)*th* cell. The row sums are both one: $p_{i1} + p_{i2} = p_{i.} = 1$ for i = 1, 2. The likelihood is

$$L = \frac{n_{1.}!}{n_{11}!\, n_{12}!}\, p_{11}^{n_{11}} p_{12}^{n_{12}} \frac{n_{2.}!}{n_{21}!\, n_{22}!}\, p_{21}^{n_{21}} p_{22}^{n_{22}}.$$

There are only two parameters at our disposal. When they are specified the model is saturated. We now introduce the parameter θ; a positive value reflects a preference for treatment A. A second parameter ϕ is used to model an order effect: a positive value reflects greater preference for the first used treatment. This is achieved if we put

$$p_{11} = 0.5 + \theta + \phi,\ p_{12} = 0.5 - \theta - \phi,\ p_{21} = 0.5 + \theta - \phi$$
$$\text{and } p_{22} = 0.5 - \theta + \phi.$$

The logarithm of the likelihood L has first order derivatives of its logarithm

$$\frac{\partial \log L}{\partial \theta} = \frac{n_{11}}{0.5 + \theta + \phi} - \frac{n_{12}}{0.5 - \theta - \phi} + \frac{n_{21}}{0.5 + \theta - \phi} - \frac{n_{22}}{0.5 - \theta + \phi} \text{ and}$$
$$\frac{\partial \log L}{\partial \phi} = \frac{n_{11}}{0.5 + \theta + \phi} - \frac{n_{12}}{0.5 - \theta - \phi} - \frac{n_{21}}{0.5 + \theta - \phi} + \frac{n_{22}}{0.5 - \theta + \phi}.$$

Now put $p_O = 0.5 + \phi$ and $q_O = 0.5 - \phi$; p_O is $p_{11} = p_{22}$ when $\theta = 0$, and q_O is $p_{12} = p_{21}$ when $\theta = 0$. Also put $p_T = 0.5 + \theta$ and $q_T = 0.5 - \theta$; p_T is $p_{11} = p_{21}$ when $\phi = 0$, and q_T is $p_{12} = p_{22}$ when $\phi = 0$. The quantity p_O is the probability that A is preferred in the absence of a treatment effect and the presence of an order effect, and similarly p_T is the probability that A is preferred in the presence of a treatment effect and the absence of an order effect. These will be used in deriving the score tests.

5.4.1 Score Test for Treatment Effect

To test H_0: $\theta = 0$ against K: $\theta \neq 0$ with ϕ an unspecified nuisance parameter, the score test is based on $\frac{\partial \log L}{\partial \theta}$ evaluated when $\theta = 0$, U say. We easily find

$$U = N_{11}/p_O - N_{12}/q_O + N_{21}/q_O - N_{22}/p_O,$$

which has exact variance $n_{..}/(p_O q_O)$. In U the unknown parameter p_O must be replaced by its maximum likelihood estimator $(N_{11} + N_{22})/n_{..}$. The score test is based on

$$Z_T = \frac{2\left|N_{11}N_{21} - N_{12}N_{22}\right| - \frac{n_{..}}{2}}{\sqrt{n_{..}\left(N_{11} + N_{22}\right)\left(N_{12} + N_{21}\right)}},$$

which includes continuity correction and may be referred to tables of the standard normal distribution. Note the similarity with Z_{GT}.

In small samples tables of possible outcomes should be ordered by values of the test statistics given, but without the continuity corrections. Then p-values are obtained by summing products of binomial probabilities over values of Z_T (without the continuity correction) at least as extreme as the observed. For testing for a treatment effect this is

$$P_T = \sum_{tail} {}^{n_{1.}}C_{n_{11}} \hat{p}_O^{\,n_{11}} \hat{q}_O^{\,n_{1.} - n_{11}}\, {}^{n_{2.}}C_{n_{21}} \hat{q}_O^{\,n_{21}} \hat{p}_O^{\,n_{2.} - n_{21}}$$

where $\hat{p}_O = (N_{11} + N_{22})/n_{..}$.

If we assume $p_O = 0.5$, or, equivalently, $\phi = 0$, then a modification of the derivation above produces the sign test. The statistics Z_S and Z_T differ because Z_S assumes $p_O = 0.5$ whereas Z_T estimates from the data. If $p_O = 0.5$ the sign test should be more powerful than that based on Z_T because it makes that additional assumption.

5.4.2 Score Test for Order Effect

To test H_0: $\phi = 0$ against K: $\phi \neq 0$ with θ an unspecified nuisance parameter, the score test is based on $\frac{\partial \log L}{\partial \phi}$ evaluated when $\phi = 0$, V say. We easily find

$$V = N_{11}/p_T - N_{12}/q_T - N_{21}/p_T + N_{22}/q_T,$$

which has exact variance $n_{..}/(p_T q_T)$. In V the unknown parameter p_T must be replaced by its maximum likelihood estimator $(N_{11} + N_{21})/n_{..} = N_{.1}/n_{..}$. The score test is based on

$$Z_O = \frac{2\,|N_{11}N_{22} - N_{12}N_{21}| - \frac{n_{..}}{2}}{\sqrt{n_{..}\, N_{.1}\, N_{.2}}},$$

which again includes continuity correction, and may be referred to tables of the standard normal distribution. Note the similarity with z_{GO}.

In small samples tables of possible outcomes should be ordered by values of the test statistic given, but without the continuity correction. Then p-values are obtained by summing products of binomial probabilities. For testing for an order effect the p-value is

$$P_O = \sum_{tail} {}^{n_{1.}}C_{n_{11}} \hat{p}_T^{\,n_{11}} \hat{q}_T^{\,n_{1.} - n_{11}}\; {}^{n_{2.}}C_{n_{21}} \hat{p}_T^{\,n_{21}} \hat{q}_T^{\,n_{2.} - n_{21}}$$

in which $\hat{p}_T = N_{.1}/n_{..}$.

Note that if the continuity corrections are ignored, the only difference between the new test statistics and the corresponding Gart statistics is, in both cases, the factor $2\sqrt{\frac{n_{1.}\, n_{2.}}{n_{..}\,(n_{..} - 1)}}$. (For $n_{..}$ fixed, this factor is greater than one for $n_{1.}$ and $n_{2.}$ approximately equal, and less than one otherwise. Thus for $n_{..} = 20$, the factor is greater than one for $n_{1.} = 8$, 9, 10, 11 and 12.) Since the factor does not depend on the data, the Gart tests

derived in the logistic ancillary framework are equivalent to the score tests derived using the product binomial model. However the normal approximations may be different, and an assessment of the power properties within the product binomial rather than logistic ancillary framework is of interest.

Both new statistics are sometimes undefined. As before, we then take their values to be zero.

We also derived and examined the Wald tests. The normal approximation to the null distribution of the Wald tests was markedly inferior to that for the score tests, and after making allowances for size differences, for the small sample sizes we were considering, the powers of the score and Wald tests were comparable. Indeed, as the sample size increases, this must be so, as they are asymptotically equivalent. We concluded that there was little practical difference in the score and Wald tests, and that as users familiar with the methodology were likely to continue to use the entrenched Gart/score tests, we did not consider the Wald tests further.

5.5 Comparison of the Sign and Score Tests

Nam (1971) compared the sign and Gart tests. He found that the sign test was more powerful when there was no order effect. However his comparison was hampered by having to calculate expected powers, by averaging over possible values of $N_{11} + N_{22}$. We find powers using a product binomial model. A permutation test approach would use fixed margins, which is not consistent with our model.

Table 5.1 gives some powers when $n_{1.} = n_{2.} = 30$ for a grid of θ, ϕ values. A critical value of 4.26667 was used for Z_S^2, Z_T^2 and Z_O^2. All tests were two-sided. We see that the power of the test based on Z_T can be comparable with that of the sign test based on Z_S. The powers for the tests based on Z_{GT} and Z_{GO} are equal to those for Z_T and Z_O respectively, if critical values of $\sqrt{4.26667} * 2\sqrt{\frac{n_{1.}\, n_{2.}}{n_{..}\,(n_{..} - 1)}}$ are used for the former. Notice in Table 5.1 that, in agreement with Nam (1971), the test based on Z_S does indeed appear to be biased. Also note that, contradicting the spirit of Nam (1971, Table 4), when $\phi = 0$ the powers of the tests based on Z_S and Z_{GT} are approximately equal. The point is that here, unlike Nam (1971, Table 4), we have $n_{1.} = n_{2.}$.

We have checked the calculations in Nam (1971, Table 4) and find them to be essentially correct. It seems that for fixed $n_{..}$, the powers of the tests based on Z_T are greatest for $n_{1.} = n_{2.}$, and decrease as these quantities become increasingly unequal. The power of the sign test is comparable with the power of the Z_T tests when $n_{1.} = n_{2.}$.

In practice it is usually possible to choose $n_{1.} \approx n_{2.}$, and if we do so then the score test based on Z_T, and equivalently Gart's treatment test based on Z_{GT}, are, when $\phi = 0$ and under the product binomial model, as powerful as the sign test. Again noting our different framework, the slight advantage of the sign test in Nam (1971, Figure 4) for $n_{1.} = n_{2.} = 7$ was not evident in our calculations.

Table 5.1 Powers of nominal $\alpha = 0.05$ tests based on Z_S, Z_T and Z_O when $n_{1.} = n_{2.} = 30$

θ	ϕ	Z_S	Z_T	Z_O
0	0	0.052	0.052	0.052
0	0.3	0.015	0.043	1.000
0	0.4	0.001	0.041	1.000
0.1	0	0.349	0.349	0.048
0.15	0	0.662	0.662	0.044
0.2	0	0.896	0.896	0.042
0.15	0.3	0.698	0.867	1.000
0.1	0.2	0.334	0.367	0.902
0.1	0.3	0.307	0.480	1.000
0.1	0.35	0.282	0.583	1.000

Many users would apply the normal approximation to obtain a p-value or to test for significance. How good is that approximation? Table 5.2 looks at sizes for the 5% critical value for a two-sided test, namely 3.841, for Z_S^2 and Z_T^2, and for the same statistics with continuity corrections, Z_{SC}^2 and Z_{TC}^2. Sizes for Z_O^2, Z_{GT}^2 and Z_{GO}^2 were identical to those for Z_T^2, and hence are not shown.

It appears that in all cases use of the continuity correction does not improve the approximations. For further references and discussion, see Stuart and Ord (1991, section 30.32, especially p. 1184). In view of the factor relating the new and the Gart statistics being approximately one for $n_{1.} = n_{2.} = n_{..}/2$, agreement between the Z_{GT}^2 and Z_T^2 sizes and the Z_{GO}^2 and Z_O^2 sizes was expected.

We also calculated some nominally 5% sizes for $\theta = 0$ and $\phi = 0.2$ when $n_{1.} = n_{2.} = n_{..}/2$. For $n_{..} = 10$, 20, 30 and 40, the sizes for the test based on Z_{GT}^2 are 0.056, 0.036, 0.042 and 0.055 respectively, while for the test based on Z_T^2 the sizes were 0.056, 0.036, 0.057 and 0.055 respectively.

In this situation the sign test is biased and has poor sizes, and the tests based on Z^2_{GO} and Z^2_O give powers, not sizes.

Table 5.2 Sizes for nominal significance level $\alpha = 0.05$ when $n_{1.} = n_{2.} = n_{..}/2$ for various $n_{..}$ and two-sided tests

$n_{..}$	10	20	30	40	50	100
Z^2_S	0.021	0.041	0.043	0.039	0.065	0.057
Z^2_T	0.061	0.042	0.046	0.043	0.065	0.057
Z^2_{SC}	0.021	0.041	0.043	0.039	0.033	0.035
Z^2_{TC}	0.021	0.041	0.043	0.039	0.033	0.035

For $n_{1.} = n_{2.} = n_{..}/2$ it appears that the distributions of all five statistics assessed are reasonably approximated by the standard normal distribution. For unequal $n_{1.}$, $n_{2.}$ it appears that the sign test statistic Z_S and the score statistics Z_T and Z_O are better approximated by the standard normal distribution. In all cases the approximation is not improved if the continuity correction is applied.

If the normal approximation is not being applied, it is worth noting that the distribution of Z_S has fewer points of increase than that of Z_T, and so there are fewer achievable sizes. Hence it will be more difficult to obtain a size close to the nominal α with the sign test than the test based on Z_T.

5.6 Sports Drink Example

Suppose that 60 athletes at a track and field meeting test two sports drinks A and B after they have competed. It is very likely that after competing the athletes will be thirsty, and that the first drink will largely

quench that thirst more than the second. Thus an order effect would be anticipated. Suppose $n_{1.} = n_{2.} = 30$ and that the results are as given below. Which is the preferred sports drink?

Table 5.3 Sports drink preference data

	Prefer A	Prefer B	Total
A first	27	3	30
B first	10	20	30
	37	23	60

The sign test gives $z_S = 1.8074$ and the score treatment test gives $z_T = 2.1936$, with two-sided standard normal p-values without continuity corrections 0.071 and 0.028 respectively. We find $z_O = 4.5140$ with p-value 0.000, confirming the order effect. In spite of the order effect, the score test identifies that A is clearly preferred. The sign test is unable to identify this preference.

5.7 Recommendations

Our calculations indicate that whenever $n_{1.} \approx n_{2.}$, which is often easy to ensure in practice, the sign test is not superior to either the new score test or to Gart's test. As $n_{1.}$ and $n_{2.}$ become unequal, the score and Gart tests become inferior to the sign test if there is no order effect, and the normal approximation to the score statistic is better than the normal approximation to Gart's statistic. If there is an order effect the sign test is biased and should not be used. We therefore recommend the new score test. If a normal approximation is to be used it should be *without* the continuity correction. The suggested tests based on Z_T and Z_O are equivalent to score tests for a product binomial model, and so will have the asymptotic optimality properties of such tests.

5.8 Nonparametric Analysis of Ordinal Categorical Data with Factorial Response

In many sensory evaluation experiments, clinical trials and market research surveys, data are recorded on an ordinal scale. An example is the data quoted in Agresti (1990) and shown in Table 5.4 below. The data, gathered from defence force staff, are concerned with preferences for black olives and are presented as counts on a categorised liking scale. There is thus an ordinal response variable, liking, which is categorised as dislike extremely, dislike moderately, neutral, like slightly, like moderately and like extremely. The statistic Q proposed by Yates (1948, p. 179) gives a simple method of analysis for this sort of data. The data in Table 5.4 have the additional complication that there are two explanatory variables or classifications, namely location and urbanisation.

Table 5.4 Black olive preference counts and linear/quadratic effects

		Preference								
Urbanisation	Location	--	-	ϕ	+	++	+++	i	V_{1i}	V_{2i}
Urban	MW	20	15	12	17	16	28	1	1.1	1.6
	NE	18	17	18	18	6	25	2	-0.2	0.6
	SW	12	9	23	21	19	30	3	2.9	-0.7
Rural	MW	30	22	21	17	8	12	4	-3.9	0.3
	NE	23	18	20	18	10	15	5	-2.1	-0.2
	SW	11	9	26	19	17	24	6	2.1	-1.5

Suppose we have ordinal responses observed within factorial-

structured groups. We thus have data that could be analysed by a standard ANOVA model, but instead of a typical observation x we have a large number of categorized ordinal responses. The ordinal responses must be the same for each observation. We may treat the data as a two-way contingency table with columns the ordered classification. We now have counts N_{ij}, arranged in r rows (the groups) and c columns (the ordinal response categories) and can calculate v_{11}, ... , v_{1r} as defined in section 4.3.

Under these circumstances, for the i*th* row we can calculate a single v_{1i} from the n_i observations. The V_{1i} are approximately independent normal with mean θ_i and constant variance. These are assumptions required for a fixed effects ANOVA model. If we make further standard assumptions about the θ_i, these can be assessed in the usual way. For the olives data given in Table 5.4, we assume

$$\theta_i = \text{grand mean} + \text{urbanisation effect} + \text{location effect.}$$

Standard ANOVA techniques, such as F tests and residual checks, could be applied.

Another difference from the analysis of variance is that Q is a sum of squares about zero rather than about the average V_{1i} value. If all the row totals are the same then this average will be zero, but in general this will not be the case. Thus if an analysis of variance computer routine is used to calculate the partition of Q then $(v_{11} + v_{12} + ... + v_{1r})^2/r$ should be added to the analysis of variance total sum of squares, and the sum of squares associated with each factor.

We can also calculate *row dispersion* effects using the V_{2i}, or indeed any other order V_{ui} of interest, as defined in section 4.3. These could then be analysed in the same way that the V_{1i} were. See the cross cultural study example in section 5.10 where the V_{2i} are assessed.

Probabilities to assess significance can be obtained using the usual χ^2 distributions, or, when counts are small, by using Monte Carlo methods. If we carried out a permutation test on the data, the column totals, $N_{.j}$, would remain constant. Thus Monte Carlo p-values may be obtained using the algorithm of Patefield (1981) which generates random

contingency tables with fixed margins. On the other hand, obtaining Monte Carlo p-values for log-linear models is not always routine, as some random tables may be sparse and lead to numerical convergence problems.

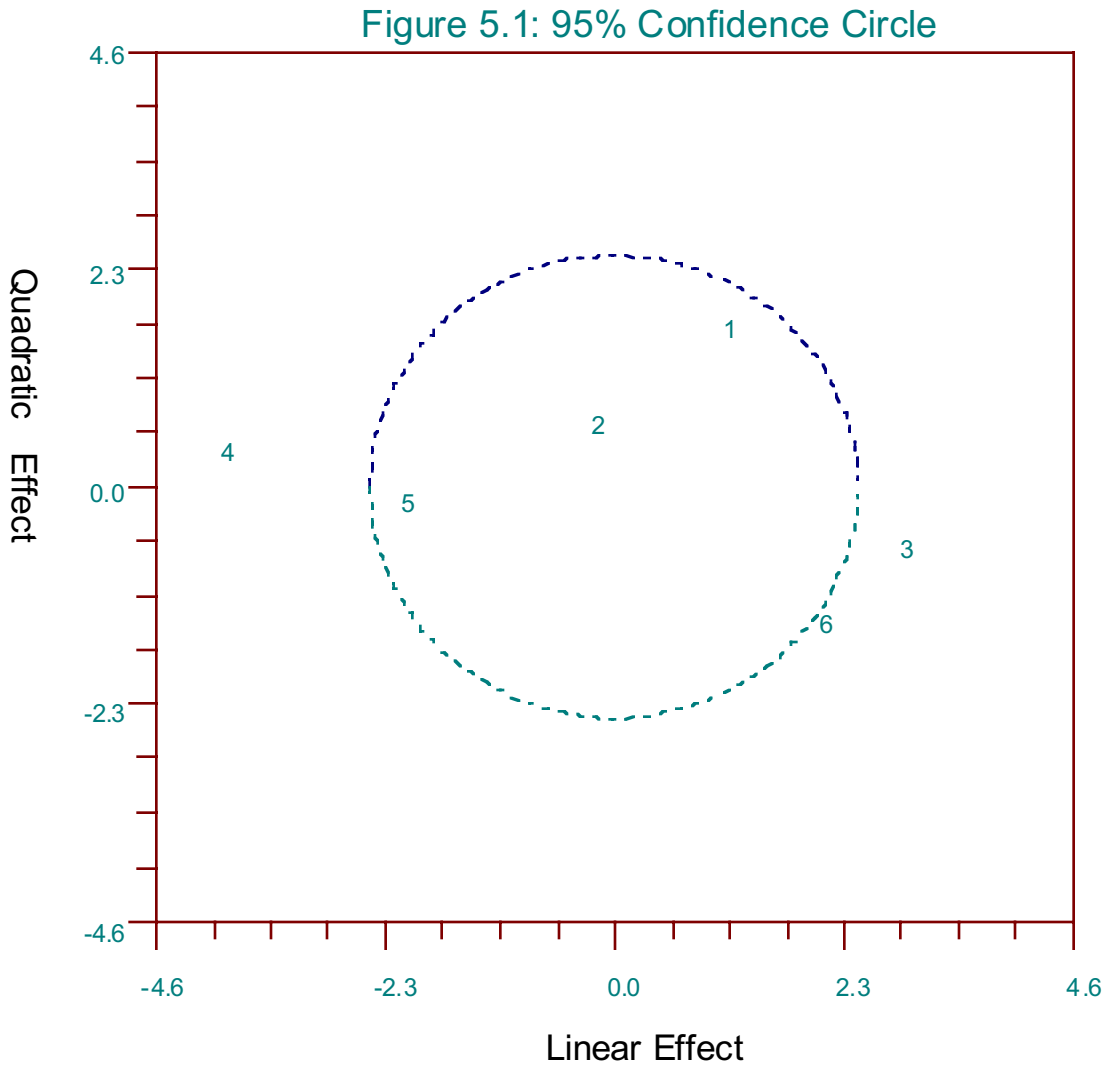
5.9 Olives Data Example

For each of the six rows of Table 5.4 values v_{1i} and v_{2i} were calculated using integer scores: the scores 1, 2, ... , 6 were assigned to the columns. For this data the usual homogeneity chi-squared statistic, X^2_P, takes the value 50.05 on 25 degrees of freedom, with χ^2 p-value 0.0021, so the distributions of the responses for the six regions are not homogeneous.

The sum of the squares of the v_{1i} given in Table 5.4 is Q = 33.61 with χ^2 p-value less than 0.1%. An LSD analysis could be performed on the v_{1i}, but we have another alternative in mind. First note that in addition we have X^2_P - Q = 16.44 on 20 degrees of freedom, with χ^2 p-value 0.689. It appears that the linear effects are the important ones. Nevertheless, examination of the v_{2i} can still be illuminating. A low v_{2i} often indicates a concentration of counts in the middle categories: see v_{26}. A high v_{2i} often indicates counts concentrated at one or both ends of the ordered categories. If the counts are concentrated at both ends or there are two peaks there is market segmentation: see v_{21}.

Table 5.5 Analysis of linear effects for olive data

Source	Chi-squared	Degrees of Freedom	p-value
Urbanisation	10.12	1	0.001
Location	18.83	2	0.001
Remainder	4.66	2	0.097
Q	33.61	5	< 0.001



The statistic Q was partitioned using a computer routine for a two-way ANOVA without replication, with factors being urbanisation and location. This led to the analysis summarised in Table 5.5. Although quadratic effects are not important, Figure 5.1 gives a plot of (v_{1i}, v_{2i}) values and shows a 95% confidence circle. Such figures are discussed in Appendix A.8. There are both urbanisation and location effects as judged by our V_{1i} or linear effects. Olives are liked more in urban than in rural areas (compare 1, 2 and 3 in Figure 5.1 with 4, 5 and 6) and more in the SW than the other locations (compare 1 and 4 and 2 and 5 with 3 and 6).

As a matter of interest, the X^2 value for a parametric row-effects

log-linear model of the Table 5.4 data is 17.7 on 20 degrees of freedom, which is close to the (X^2_P - Q) value of 16.4 given above. Agresti (1984, section 5.2) discusses row-effects log-linear models.

5.10 Cross Cultural Study Example

Japanese and Australian consumers were asked to rate various sweet foods on a seven point category scale with anchors *dislike extremely* and *like extremely*. Table 5.6 shows responses similar to those actually obtained for Japanese chocolate where each category has been assigned an integer score.

Table 5.6 Japanese chocolate responses

Country	City	Sweetness Liking						
		1	2	3	4	5	6	7
Australia	Sydney	2	1	6	1	8	9	6
	Melbourne	1	6	2	2	10	5	5
Japan	Tokyo	0	1	3	4	15	7	1
	Osaka	1	1	2	3	16	6	2

Table 5.7 confirms the *by eye* impression that there are differences between countries in dispersion or quadratic effects, but not in linear or location effects. There is no significant difference between Australian and Japanese consumers in their average liking of the sweetness of Japanese chocolate. However there is significantly more spread in the opinions of Australian consumers than in the opinions of Japanese consumers. This conclusion would not have been reached if the only location tests had been employed. The spread of opinions of Australian consumers indicates market segmentation which, of course, has commercial implications. The partitions of Q and the total sum of squares statistic, $V^2_{21} + V^2_{22} + V^2_{23} + V^2_{24}$, were calculated using a one-way ANOVA routine and the v_{1i} and v_{2i}

values as data. The p-values in Table 5.7 are based on χ^2 approximations. As all the row totals are equal in this example, no changes are needed in the ANOVA output.

Table 5.7 Analysis of linear and quadratic effects for Japan data

Source	df	Linear	p-value	Quadratic	p-value
Countries	1	0.2122	NS	11.1042	< 0.01
Cities within countries	2	0.7524	NS	0.4544	NS
Total	3	Q = 0.9646	NS	11.5586	< 0.05

6
Tests on Complete Randomised Blocks: Extensions to the Friedman and Cochran Tests

6.1 Peach Example

To demonstrate the sort of data analysed by the approach discussed in this chapter, and the power of our methods, we begin with an example that implements our approach.

Table 6.1 Firmness ranking of five canned peach products

	Canning Treatment				
Judge	A	B	C	D	E
1	1	2	3	4	5
2	1	2	3	5	4
3	1	4	5	2	3
4	1	3	4	5	2
5	1	2	3	5	4
6	1	3	5	2	4
7	1	5	4	3	2
8	1	2	3	5	4
9	1	2	4	3	5
10	1	5	2	4	3

Peaches Example. O'Mahony (1986, p. 340) gave an example where ten expert judges evaluated the firmness of sliced peaches canned under five different conditions. The results are given in Table 6.1, along with, in Table 6.2, the same data in terms of the number of times each rank is assigned to each product. In Table 6.3 we give Friedman's statistic, a new dispersion statistic and a residual, with their asymptotic chi-squared p-values and the corresponding p-values found by simulation.

Table 6.2 Ranks for peach data

	Rank				
Treatment	1	2	3	4	5
A	10	0	0	0	0
B	0	5	2	1	2
C	0	1	4	3	2
D	0	2	2	2	4
E	0	2	2	4	2

Table 6.3 Partition of Anderson's statistic for peach data

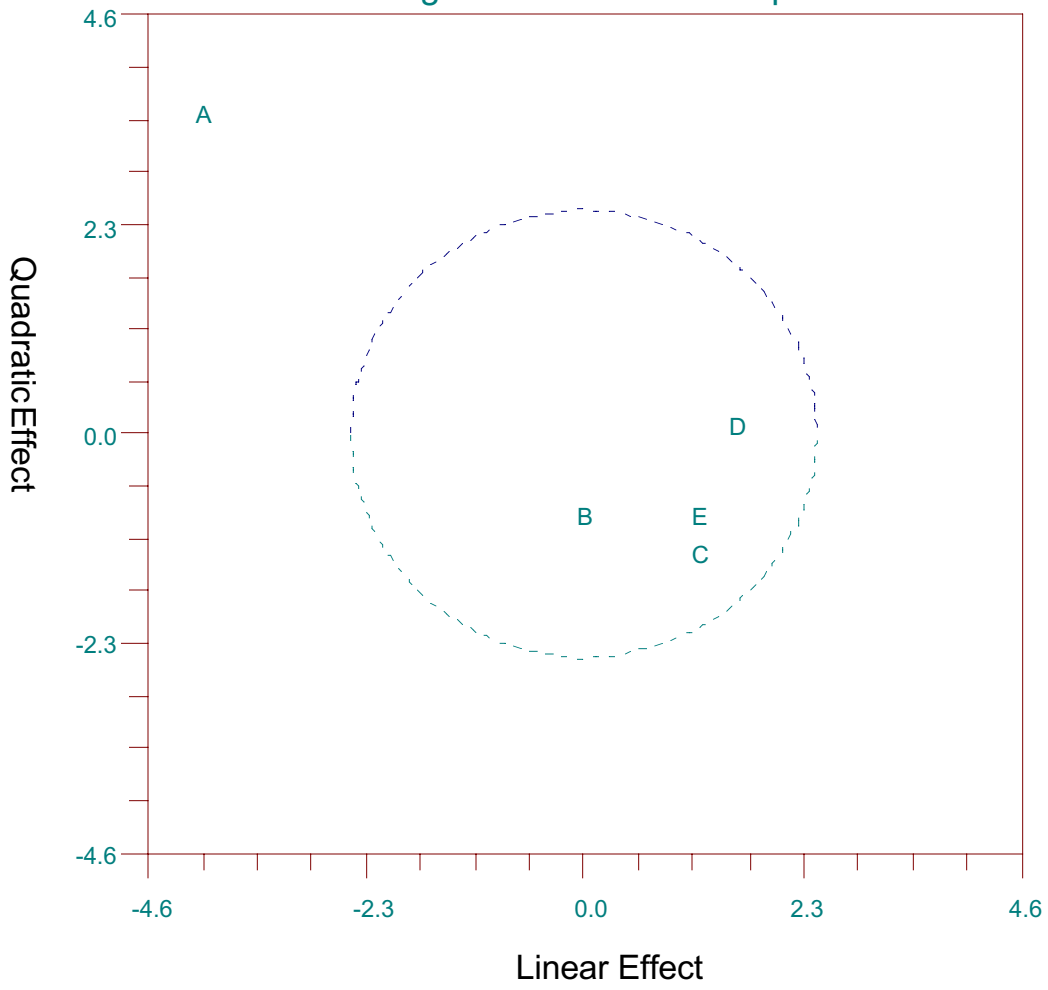
Statistic	df	Value	χ^2 p-value	Simulated p-value
Friedman	4	21.44	< 0.001	< 0.001
Dispersion	4	15.31	0.004	0.002
Residual	8	11.25	0.188	0.181

As we shall discuss in the following section, Anderson (1959) showed that $(4/5)\,X_P^2$, Anderson's statistic, is an appropriate statistic for testing the homogeneity of the row distributions for the data in Table 6.2. The analysis in Table 6.3 shows that there are significant location and dispersion differences, but that the residual is not significant. Using

formulae given in the next section, we find that the component of the Friedman statistic corresponding to A is -4.0, while those corresponding to B to E take values between zero and 1.6. Since these components are approximately standard normal, we conclude that treatments B, C, D and E do not differ from each other, all being inferior to treatment A. This separation explains both the significant location and significant dispersion effects. The nonsignificant residual indicates that we have identified all of the information contained in $(4/5)\,X_P^2$. The adequacy of the χ^2 p-values is confirmed by the simulated p-values.

It is possible to graphically complement the LSD analysis just above. Figure 6.1 gives a product map for the canning treatments and emphasises the obvious differences between treatment A and the other treatments. Notice that the scales for both axes have been chosen to be the same: this helps interpretation.

Figure 6.1: Peaches Map



This analysis is more robust than the usual analysis of variance, in that neither homogeneity of judge variance nor normality are assumed. It is also more informative than simply applying Friedman's test, as this may not be significant, while dispersion and higher order effects might be significant. This was the case in some of the one-way layout examples; see Tables 3.2 and 4.9 and the nearby discussions.

More examples are given at the end of sections 6.4, 6.5, 6.6 and 6.7.

6.2 Friedman's Test and its Extensions

Suppose we have observations x_{ij}, being the i*th* of s treatments on the j*th* of n blocks. These observations are ranked within each block and R_i, the sum of the ranks for each treatment, is calculated. The Friedman statistic is

$$S = \frac{12}{ns(s+1)} \sum_{i=1}^{s} R_i^2/n - 3n(s+1).$$

Table 6.4 Data for Friedman's test and its extensions

	Block 1	Block 2	…	Block n	Total
Treatment 1	x_{11}	x_{12}	…	x_{1n}	n
	…	…	…	…	…
Treatment s	x_{s1}	x_{s2}	…	x_{sn}	n
Total	s	s	…	s	ns

Table 6.5 Presentation of Friedman data in an Anderson table

	# Rank 1	# Rank 2	…	# Rank s	Total
Treatment 1	N_{11}	N_{12}	…	N_{1s}	n
	…	…	…	…	…
Treatment s	N_{s1}	N_{s2}	…	N_{ss}	n
Total	n	n	…	n	ns

To parallel the analysis we gave for the one-way layout, we suggest constructing an s by s contingency table of counts $\{N_{ij}\}$ with N_{ij} being the number of times treatment i receives rank j over the n blocks. We call such a table an *Anderson table*, after Anderson (1959). From this table calculate $\sum_{j=1}^{s} jN_{ij} = R_i$, the rank sum for treatment i, from which S is readily calculated.

Extensions of Friedman's S will now be described. We have s (> 2) treatments ranked by each of n judges. This s by s Anderson table (N_{ij}) records the number of times treatment i receives rank j. Every row and column sums to n, as each treatment and each rank is assigned n times. This is not the usual s by s contingency table with row and column totals fixed. As Anderson (1959) said,

> "repeated sampling is not a random rearrangement of rn items, subject to border restrictions ... ."

Anderson showed that the *usual* X^2 statistic,

$$Q^2 = \frac{s}{n}\sum_{i=1}^{s}\sum_{j=1}^{s}\left(N_{ij} - \frac{n}{s}\right)^2,$$

is such that $A = (s - 1)Q^2/s$, which we call Anderson's statistic, is asymptotically distributed as $\chi^2_{(s-1)^2}$.

If we define

$$V_r = \sqrt{\frac{s - 1}{ns}}\left(\sum_{j=1}^{s} g_r(j)\, N_{ij}\right),\ r = 1, \ldots, s - 1,$$

then we will show in section 6.3 that

$$(s - 1)Q^2/s = V_1^T V_1 + \ldots + V_{s-1}^T V_{s-1},$$

in which $\{g_r(j)\}$ are the polynomials orthonormal on the discrete uniform distribution on s points. The V_r are asymptotically mutually independent and asymptotically have the $N_s(0, I_s - \frac{1}{s} 1^s 1_s^T)$ distribution, so that the $V_r^T V_r$ are asymptotically mutually independent χ^2_{r-1}. In 6.3.4 the first component, $V_1^T V_1$, will be shown to be Friedman's statistic, that detects differences in the treatment means. The r*th* component $V_r^T V_r$ assesses r*th* moment differences between the treatments. If

$$V_{ir} = \sqrt{\frac{s-1}{ns}} \sum_{j=1}^{s} g_r(j) \left(N_{ij} - \frac{n}{s}\right), \text{ for } i = 1, \ldots, s \text{ and } r = 1, \ldots, s-1,$$

then

$$V_r^T V_r = \sum_{i=1}^{s} V_{ir}^2$$

and a useful graphical presentation of the data is a plot of (v_{i1}, v_{i2}), i = 1, ... , s. The V_{i2} can be useful for examining market segmentation in market research data similar to the one-way layout case discussed near Figure 5.1.

We have linked dispersion or quadratic effects with $V_2^T V_2$ and large values of this statistic can be related to different types of agreement or concordance between judges or consumers in, say, market research data. These types of agreement may not be identified by Kendall's coefficient of concordance, CC, which is related to $V_1^T V_1$ by

$$CC = V_1^T V_1 / [s(n-1)].$$

6.3 Derivations

6.3.1 Introduction

Suppose s (> 2) treatments are ranked by each of n judges. An s by s Anderson table (N_{ij}) records the number of times treatment i receives rank j. Anderson (1959) showed that the usual X^2 statistic,

$$Q^2 = \frac{s}{n} \sum_{i=1}^{s} \sum_{j=1}^{s} \left(N_{ij} - \frac{n}{s}\right)^2,$$

is such that $(s - 1)Q^2/s$, which we call Anderson's statistic, is asymptotically distributed as $\chi^2_{(s-1)^2}$. We review his derivation, and, in addition, partition $(s - 1)Q^2/s$ into *components* to make the analysis more informative, showing that $(s - 1)Q^2/s$ is the sum of asymptotically mutually independent random variables, each asymptotically having the χ^2_{s-1} distribution. The first component is Friedman's statistic. Higher order components reflect higher moment discrepancies between the treatments. In particular, the second moment discrepancies can often be as important as the first moment discrepancies. For example, in market research applications, the second moment discrepancies can point to important market segmentation effects.

6.3.2 Distribution of Anderson's Statistic

To derive the distribution of $(s - 1)Q^2/s$, Anderson noted that

$$X_{ij} = \left(N_{ij} - n/s\right) s/\sqrt{n\,(s - 1)}, \text{ with } i, j = 1, \ldots, s,$$

are individually asymptotically normally distributed. Writing

$$X = (X_{11}, \ldots, X_{1s}, X_{21}, \ldots, X_{2s}, \ldots, X_{s1}, \ldots, X_{ss})^T,$$

it follows that

$$[s/(s - 1)]\, Q^2 = X^T X,$$

and that X is asymptotically s^2-variate normal with zero mean and covariance matrix Σ say, written $N_{s^2}(0, \Sigma)$. If

$$R = [s/(s - 1)]\, I_s - [1/(s - 1)]\, 1_s 1_s^T,$$

then Σ is the direct product of R with itself,

$$\Sigma = R \otimes R.$$

It is routine to find that the eigenvalues of Σ are $[s/(s - 1)]^2 = d$ say, $(s - 1)^2$ times, and zero $(2s - 1)$ times.

Suppose H is orthogonal and diagonalises Σ. We then have

$$H^T \Sigma H = d\left(I_{(s - 1)^2} \oplus 0_{(2s - 1)}\right),$$

where "$\oplus$" means direct sum. Define $Y = [(s - 1)/s]\, H^T X$. Then

$$X^T X = d\, Y^T Y$$

in which Y is asymptotically $N_{s^2}(0, \left(I_{(s - 1)^2} \oplus 0_{(2s - 1)}\right))$. It follows that $X^T X/d = Y^T Y = (s - 1)Q^2/s$ asymptotically has the $\chi^2_{(s - 1)^2}$ distribution, as Anderson showed.

6.3.3 Partitioning the Anderson Statistic

The elements Y_i of Y are such that the Y_i^2 partition $(s - 1)Q^2/s$. There is some choice in defining the Y_i, in that H is not yet fully specified. In doing so, our aim is to find Y_i that can be easily and usefully

interpreted.

To achieve such a partition, first suppose that $\{g_r(j)\}$ are the polynomials orthonormal on the discrete uniform distribution on s points. Write g_r for the s by 1 vector with elements $g_r(j)$. Define G by

$$G = \left[\; G_1 \; | \quad \cdots \quad | \; G_s \; \right] / \sqrt{s}$$

in which G_r is the s^2 by s matrix

$$G_r = \begin{bmatrix} g_r & 0 & & \\ 0 & g_r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_r \end{bmatrix}, \; r = 1, \ldots, s - 1, \text{ and}$$

$$G_s = \begin{bmatrix} 1_s & 0 & & \\ 0 & 1_s & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_s \end{bmatrix} \text{ is also } s^2 \text{ by s.}$$

Define $Z = [(s - 1)/s] \; G^T X$. The elements of Z may be considered in blocks of s, the r*th* block corresponding to the polynomial of order r. These blocks are asymptotically mutually independent. Write

$$Z^T = (V_1^T, \ldots, V_s^T), \text{ in which}$$
$$V_1 = (Z_1, \ldots, Z_s)^T, \ldots, V_{s-1} = (Z_{(s-1)^2}, \ldots, Z_{s^2 - s})^T, \text{ and } V_s = 0.$$

Note that all the V_r are s by 1, and that

$$(s - 1)Q^2/s = X^TX/d = Z^TZ = V_1^TV_1 + \ldots + V_{s-1}^TV_{s-1}.$$

This is the partition of Anderson's statistic into components $V_r^TV_r$, r = 1, ... , s - 1. The V_r are asymptotically mutually independent and asymptotically $N_s(0, I_s - (1/s)1_s1_s^T)$, so that the $V_r^TV_r$ are asymptotically mutually independent χ^2_{s-1}. Explicitly, we have, for r = 1, ... , s - 1,

$$\begin{aligned} V_r &= \frac{s - 1}{s} G_r^TX = \frac{s - 1}{s} \frac{1}{\sqrt{s}} \left(\sum_{j=1}^{s} g_r(j) X_{ij} \right) \\ &= \sqrt{\frac{s - 1}{ns}} \left(\sum_{j=1}^{s} g_r(j) \left(N_{ij} - \frac{n}{s} \right) \right) \\ &= \sqrt{\frac{s - 1}{ns}} \left(\sum_{j=1}^{s} g_r(j) N_{ij} \right), \end{aligned}$$

and

$$V_r^TV_r = \frac{(s - 1)^2}{s^3} \sum_{i=1}^{s} \left\{ \sum_{j=1}^{s} g_r(j) X_{ij} \right\}^2.$$

Because V_r involves, through g_r, a polynomial of order r, the elements of V_r are polynomials of order r in the elements of X. Under the null hypothesis E[X] = 0, but when this is not true E[V_r] involves moments of X up to order r. Thus for r = 1, ... , s - 1, $V_r^TV_r$ detects r*th* moment departures from the null hypothesis of similarly distributed rows (treatments). We now show that $V_1^TV_1$ is Friedman's statistic, so that the $V_r^TV_r$ provide extensions to Friedman's test.

6.3.4 The First Component: Friedman's Statistic

We now identify $V_1^TV_1$. Note that $g_1(j) = aj + b$, j = 1, ... , s in which

$$a = \sqrt{\{12/(s^2 - 1)\}} \text{ and } b = -\sqrt{\{3(s + 1)/(s - 1)\}} = -\{(s + 1)/2\}a.$$

The rank sum for treatment i, R_i, is $\sum_{j=1}^{s} jN_{ij}$, i = 1, ... , s.

The i*th* element of V_1 is

$$\sqrt{\frac{s-1}{ns}}\sum_{j=1}^{s} g_1(j)\,N_{ij} = \sqrt{\frac{s-1}{ns}}\sum_{j=1}^{s} a\left(j - \frac{s+1}{2}\right)N_{ij}$$

$$= a\sqrt{\frac{s-1}{ns}}\left\{\sum_{j=1}^{s} j\,N_{ij} - \frac{s+1}{2}\sum_{j=1}^{s} N_{ij}\right\}$$

$$= \sqrt{\frac{12}{ns(s+1)}}\left\{R_i - \frac{n(s+1)}{2}\right\}$$

after recalling that $n_{i.} = n$ (= $n_{.j}$ also). It follows that

$$V_1^T V_1 = \frac{12}{ns(s+1)}\sum_{i=1}^{s}\left\{R_i - \frac{n(s+1)}{2}\right\}^2.$$

This is Friedman's statistic, S say, well known to be sensitive to location departures from the null hypothesis. Since V_r assesses r*th* moment departures, we have partitioned Anderson's $(s-1)Q^2/s$ statistic into asymptotically mutually independent components, $V_r^T V_r$, r = 1, ... , s - 1, so that the r*th* detects r*th* moment departures from the hypothesis of similarly distributed rows (treatments). The first of these is Friedman's statistic. The subsequent components provide extensions of Friedman's test. This decomposition also explains why Anderson's test cannot be very powerful. It is an omnibus test detecting moment departures up to order s. For tables of even moderate size, it is not sufficiently *directional*.

6.4 Page's Test and its Relationship to Friedman's, Anderson's, and Pearson's Tests

We now derive Page's test, which tests for a particular ordering of the treatment means. Here we take the null hypothesis to be that the treatment means are identical, $\mu_1 = \mu_2 = \ldots = \mu_{s-1} = \mu_s$, and the alternative is that $\mu_1 < \mu_2 < \ldots < \mu_{s-1} < \mu_s$. We then relate Page's test to Friedman's, Anderson's, and Pearson's tests.

In 6.3.4 above it was shown that the i*th* element of V_1 is

$$\sqrt{\frac{s-1}{ns}} \sum_{j=1}^{s} g_1(j)\, N_{ij} = \sqrt{\frac{12}{ns(s+1)}} \left| R_i - \frac{n(s+1)}{2} \right|.$$

Now define $U = HV$ where H is partitioned into $H_1, \ldots, H_s$ so that $U_i = H_i V_i$, $i = 1, \ldots, s$. The H_i are taken to be orthogonal. This ensures that the only term of interest to us here, $(U_1)_1$, is asymptotically standard normally distributed. Further, take H_1 to have last row $1/\sqrt{s}$ and first row $(i - \mu)/\sigma$, $i = 1, \ldots, s$, where $\mu = (s+1)/2$ and $\sigma^2 = s(s^2 - 1)/12$. Then

$$(U_1)_1 = \sqrt{\frac{12}{ns(s+1)}} \sum_{i=1}^{s} \left(\frac{i - \mu}{\sigma} \right) \left(R_i - \frac{n(s+1)}{2} \right) = \frac{L - \mu_L}{\sigma_L},$$

after some manipulation and provided we first define

$$L = \sum_{i=1}^{s} i\, R_i,\ \mu_L = \frac{ns(s+1)^2}{4} \text{ and } \sigma_L^2 = \frac{n(s-1)s^2(s+1)^2}{144}.$$

The statistic L is the well known Page test statistic. Another derivation is given in section 6.5.

The derivation here shows that Page's statistic is a component of the Anderson statistic, but not of Friedman's statistic. Nevertheless, we can informatively relate these statistics, albeit crudely. Observe that the Page, Friedman, Anderson, and Pearson statistics all have asymptotic χ^2 distributions, with degrees of freedom respectively 1, s - 1, $(s-1)^2$, and s!

- 1. For $s \geq 3$ note that this is a strict ordering: $1 < s - 1 < (s - 1)^2 < s! - 1$. The degrees of freedom are the dimension of the parameter space specified by the alternative hypothesis. Because the Friedman statistic is a component of the Anderson statistic, the Friedman alternative hypothesis parameter space is a subset of the Anderson alternative hypothesis parameter space. Similar conclusions follow because the Page statistic is a component of the Anderson statistic, and, as was shown by Rayner and Best (1989b), the Page and Friedman statistics are components of Pearson's statistic. We can crudely think of the alternative hypothesis parameter spaces being nested, Page within Friedman within Anderson within Pearson. This means the Page test is very directional, the Friedman less directional and more omnibus, with the tests becoming decreasingly directional and increasingly omnibus.

A small simulation study is now performed to support the notions above. It is related to a study done by Kepner and Robinson (1984). Powers were computed by first assuming that X_{ij} is the measurement on the j*th* of k products by the i*th* of n judges. Further

Table 6.6 Power of Page's (L), Friedman's (S), Anderson's (A), and Pearson's (X_P^2) tests, based on 10,000 simulations, three treatments, an exact size of 0.05, and N(0, 1) treatment errors

(a) $(T_1, T_2, T_3) = (\delta, 0, -\delta)$

		$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$
(i) n = 5	L	0.050	0.317	0.761	0.968	0.999
	S	0.050	0.146	0.491	0.827	0.970
	A	0.050	0.127	0.406	0.718	0.900
	X_P^2	0.050	0.063	0.250	0.565	0.818
(i) n = 10	L	0.050	0.562	0.974	1.000	1.000
	S	0.050	0.327	0.884	0.997	1.000
	A	0.050	0.247	0.788	0.987	1.000
	X_P^2	0.050	0.206	0.710	0.972	0.999

(b) $(T_1, T_2, T_3) = (\delta/2, -\delta, \delta/2)$

		$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$
(i) n = 5	L	0.050	0.036	0.021	0.011	0.004
	S	0.050	0.119	0.372	0.689	0.898
	A	0.050	0.104	0.327	0.641	0.878
	X_P^2	0.050	0.052	0.153	0.268	0.346
(i) n = 10	L	0.050	0.038	0.027	0.016	0.012
	S	0.050	0.251	0.760	0.979	0.999
	A	0.050	0.187	0.656	0.955	0.998
	X_P^2	0.050	0.159	0.577	0.915	0.992

$$X_{ij} = \mu + B_i + T_j + E_{ij}$$

in which the B_is are judges effects, the T_js are product effects, and the E_{ij}

are independent and identically distributed errors from a continuous population with common variance σ^2. The restrictions $B_1 + ... + B_n = 0$ and $T_1 + ... + T_k = 0$ are imposed, and without loss of generality we can take $B_i = 0$ for $i = 1, ... , n$, $\mu = 0$ and $\sigma^2 = 1$. The results in Tables 6.6 and 6.7 are for a normal error distribution but change little if uniform or double exponential distributions are used. Similarly, the same relative performance occurs for other choices of size α and number of judges n. After the X_{ij} were generated, ranks for each judge were found and the statistics calculated. A significance level of 0.05 was used and critical values were randomised so that a size of exactly 0.05 was possible.

Table 6.7 Power of Page's (L), Friedman's (S), Anderson's (A), and Pearson's (X_P^2) tests, based on 10,000 simulations, four treatments, an exact size of 0.05, and N(0, 1) treatment errors

(a) $(T_1, T_2, T_3, T_4) = (\delta, \delta/3, -\delta/3, -\delta)$

		$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$
(i) n = 5	L	0.050	0.348	0.829	0.989	1.000
	S	0.050	0.163	0.546	0.881	0.985
	A	0.050	0.091	0.291	0.582	0.832
	X_P^2	0.050	0.076	0.234	0.500	0.765
(i) n = 10	L	0.050	0.621	0.989	1.000	1.000
	S	0.050	0.328	0.905	0.999	1.000
	A	0.050	0.210	0.732	0.981	1.000
	X_P^2	0.050	0.136	0.523	0.905	0.995

(b) $(T_1, T_2, T_3, T_4) = (\delta, -\delta, -\delta, \delta)$

		$\delta = 0.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$
(i) n = 5	L	0.050	0.023	0.005	0.001	0.000
	S	0.050	0.285	0.820	0.989	1.000
	A	0.050	0.127	0.338	0.523	0.594
	X_P^2	0.050	0.116	0.392	0.651	0.747
(i) n = 10	L	0.050	0.028	0.008	0.001	0.000
	S	0.050	0.561	0.993	1.000	1.000
	A	0.050	0.342	0.927	1.000	1.000
	X_P^2	0.050	0.238	0.852	0.955	1.000

For the mean shift alternatives $(\delta, 0, -\delta)$ and $(\delta, \delta/3, -\delta/3, \delta)$, Page's test is clearly most powerful, followed by Friedman's, then Anderson's and then Pearson's tests. Since the labelling of the judges is unlikely to be

relevant in the applications we envisage, detection of these alternatives to the exclusion of any others is not desirable. For the other alternatives, Page's test has poor power. Thus it appears from this simulation study that Friedman's test provides a better overall test of product differences than Page's, Anderson's or Pearson's tests.

Anderson's Bean Example. Anderson (1959) gave the data in Table 6.8 for the consumer rankings of three varieties of snap beans.

Table 6.8 Anderson's snap bean data

	Rank			
Variety	1	2	3	Total
Variety 1	42	64	17	123
Variety 2	31	16	76	123
Variety 3	50	43	30	123

For these data we find S = 24.797 with χ^2_2 p-value 0.000, L = 1481, with normal distribution p-value 0.375 and A = 53.041, with χ^2_4 p-value 0.000. If the labels on varieties two and three are exchanged, the standard (normal) score for Page's test is now 4.463, with p-value 0.000. There is a clear difference in the rankings of the beans, from variety 1 to variety 3 to variety 2.

6.5 An Alternative Partition of the Anderson Statistic: an Umbrella Test

We now review work from Best and Rayner (1999b) in which an alternative partition of the Anderson statistic that also leads to Page's statistic, and to a new umbrella statistic, is given. First define an s by s orthogonal matrix L, say, by

$$L = \begin{pmatrix} g_1(1) & \cdots & g_{s-1}(1) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ g_1(s) & \cdots & g_{s-1}(s) & 1 \end{pmatrix} / \sqrt{s},$$

and then put $G = L \otimes L$ and $Z = \frac{s-1}{s} G^T X$. Then Z has typical element

$$\frac{s-1}{s^2} \sum_{i=1}^{s} \sum_{j=1}^{s} X_{ij}\, g_\ell(i)\, g_m(j) = \sqrt{\frac{s-1}{s}}\, \frac{1}{\sqrt{ns}} \sum_{i=1}^{s} \sum_{j=1}^{s} \left(N_{ij} - \frac{n}{s}\right) g_\ell(i)\, g_m(j)$$
$$= U_{\ell m} \text{ say.}$$

From the orthogonality $\sum_{i=1}^{s} g_\ell(i) = \sum_{j=1}^{s} g_m(j) = 0$. It follows that

$$U_{\ell m} = \sqrt{\frac{s-1}{s}}\, \frac{1}{\sqrt{ns}} \sum_{i=1}^{s} \sum_{j=1}^{s} N_{ij}\, g_\ell(i)\, g_m(j) \text{ and}$$
$U_{\ell m} = 0$ if either ℓ or $m = s$.

The latter follows if we put $g_s(j) = 1$ for all j, giving

$$U_{\ell s} = \sqrt{\frac{s-1}{s}}\, \frac{1}{\sqrt{ns}} \sum_{i=1}^{s} \left\{ g_\ell(i) \sum_{j=1}^{s} \left(N_{ij} - \frac{n}{s}\right) \right\} = 0 \text{ since}$$
$$\sum_{j=1}^{s} \left(N_{ij} - \frac{n}{s}\right) = 0 \text{ for all i.}$$

It follows that Z has $(s-1)^2$ nondegenerate elements $U_{\ell m}$. As before

$$(s-1)Q^2/s = Z^T Z = \sum_{\ell=1}^{s-1} \sum_{m=1}^{s-1} U_{\ell m}^2.$$

As in 6.3.2, Z is asymptotically $N_{s^2}(0, (I_{(s-1)^2} \oplus 0_{(2s-1)}))$. It follows

that the nondegenerate elements are asymptotically independent and asymptotically have the standard normal distribution.

Recall from 6.3.4 that $g_1(j) = aj + b$, $j = 1, \ldots, s$, in which

$$a = \sqrt{\{12/(s^2 - 1)\}} \text{ and } b = -\sqrt{\{3(s + 1)/(s - 1)\}} = -\{(s + 1)/2\}a.$$

The rank sum for treatment i, R_i, is $\sum_{j=1}^{s} jN_{ij}$, $i = 1, \ldots, s$, and $L = \sum_{i=1}^{s} iR_i$, with mean and variance respectively

$$\mu_L = \frac{ns(s + 1)^2}{4} \text{ and } \sigma_L^2 = \frac{n(s - 1)s^2(s + 1)^2}{144}.$$

By definition

$$\begin{aligned} U_{11} &= \sqrt{\frac{s - 1}{s}} \frac{1}{\sqrt{ns}} \sum_{i=1}^{s} \sum_{j=1}^{s} N_{ij}\, g_1(i)\, g_1(j) \\ &= a^2 \sqrt{\frac{s - 1}{s}} \frac{1}{\sqrt{ns}} \sum_{i=1}^{s} \sum_{j=1}^{s} N_{ij} \left(i - \frac{s + 1}{2}\right) \left(j - \frac{s + 1}{2}\right) \\ &= \frac{12}{s^2 - 1} \sqrt{\frac{s - 1}{ns^2}} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{s} ijN_{ij} - ns\left(\frac{s + 1}{2}\right)^2 \right\} \\ &= (L - \mu_L)/\sigma_L. \end{aligned}$$

An the alternative derivation of this relation was given in section 6.4.

It follows that the $U_{\ell m}$ are extensions of Page's statistic. Of particular interest is U_{21} that can be used to test the null hypothesis of no treatment differences against the umbrella alternative that as we pass through the rows (treatments) the columns either increase and then decrease (one tail) or decrease and then increase (the other tail).

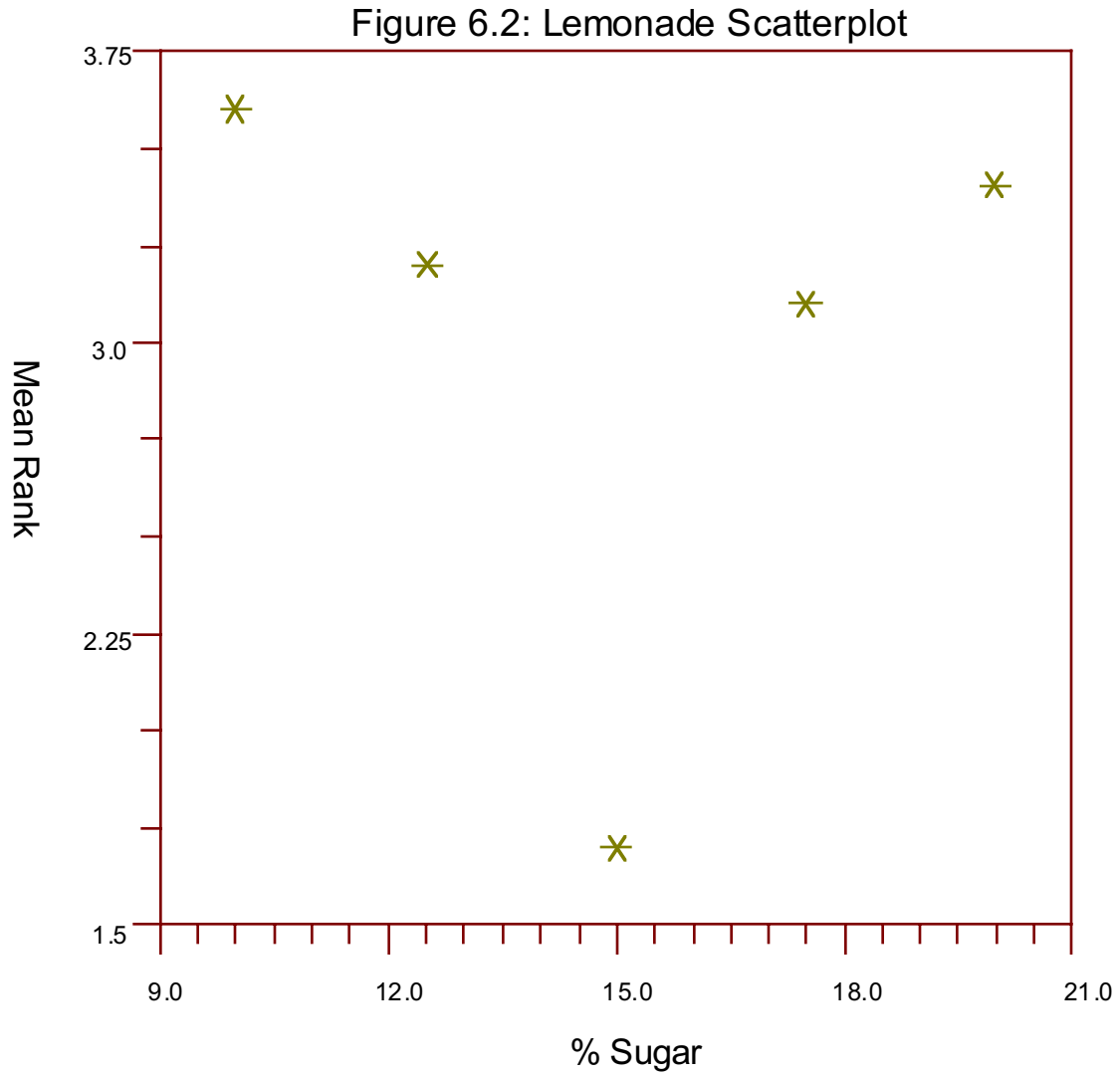
Lemonade Example. Products A to E are lemonade drinks with increasing sugar levels where the 'just right' level of sugar is that in lemonade drink C. The data are given in Table 6.9 and converted into counts given in Table 6.10, from which we calculate our test statistics.

Table 6.9 Lemonade data; 5 products ranked by 10 judges

	Judge									
Product	1	2	3	4	5	6	7	8	9	10
A	5	3	4	5	3	4	5	3	1	3
B	2	5	3	2	5	3	2	5	4	1
C	1	2	2	1	2	2	1	2	2	2
D	3	1	5	3	1	5	3	1	5	4
E	4	4	1	4	4	1	4	4	3	5

Table 6.10 Counts $\{N_{ij}\}$ for lemonade data

	Rank				
Product	1	2	3	4	5
A	1	0	4	2	3
B	1	3	2	1	3
C	3	7	0	0	0
D	3	0	3	1	3
E	2	0	1	6	1



The Friedman and Page Monte Carlo p-values, found using *StatXact* (1995), are 0.55 and 0.63, respectively. They give no evidence of differences between treatments. The umbrella statistic U_{21} has one-sided p-value 0.01 using the standard normal distribution. Thus as we go from A to E there are more 3, 4 and 5 rankings for A and E and more 1 and 2 rankings for C.

6.6 Ties

If tied rankings are permitted there are various ways in which we could proceed. We will illustrate with reference to Page's test. With tied ranks it is still possible to construct a table of counts $\{N_{ij}\}$ where, as before, i = 1, 2, ... , K but where now the number of columns depends on which mid-ranks occur. Schach (1979, p. 549) noted that the Anderson-type statistic calculated on these counts for the tied data no longer has an asymptotic χ^2 distribution. Further, textbooks and computer software provide different approaches to calculating Page's test. We illustrate by considering an example.

Ties Example. Daniel (1990, p. 281) gave an example where there are three ordered conditions and 12 subjects with the ranks within subjects. A table of counts for these data is given in Table 6.11.

For these data L = 148.5. Daniel (1990) stated that p > 0.05, but he used critical values based on the assumption of no ties. Accounting for ties can make considerable differences in p-values. For example, if we use the approximate standard normal statistic Z_L as Daniel did, we find z_L = 0.919 for the Table 6.11 data, with p-value 0.18. The IMSL (1989) package also gives this result. Conover (1998, p. 381) uses the same approximations as Daniel (1990).

A different value of Z_L, 1.125, can be obtained if only the eight rows in Table 6.11 that are not completely tied (that is, consist of 2, 2, 2) are used. *StatXact* (1995) gives this value of Z_L. Pirie (1985) gave a formula for σ_L^2 that accounts for ties. If this is used then z_L = 1.248 - a third value! The p-value for the Pirie statistic is 0.11.

Yet another nonparametric approach to the analysis of these data is to use a Cochran-Mantel-Haenszel (CMH) statistic, the so-called correlation CMH statistic. This is applied with each judge being a stratum, using midranks as scores, and having product totals within each stratum all equal to unity. For the present data the value of the correlation CMH is equal to the square of the Pirie z_L value. Details of the correlation CMH statistic are given, for example, in Landis et al. (1979).

Table 6.11 Ties example data

	Conditions		
Subject	1	2	3
1	2	2	2
2	1	3	2
3	1.5	1.5	3
4	1.5	1.5	3
5	2	3	1
6	2	2	2
7	2.5	1	2.5
8	2.5	1	2.5
9	1	2.5	2.5
10	2	2	2
11	2.5	1	2.5
12	2	2	2

Table 6.12 Counts $\{N_{ij}\}$ for the ties example data

	Mid-Rank				
Condition	1.0	1.5	2	2.5	3.0
1	2	2	5	3	0
2	3	2	4	1	2
3	1	0	5	4	2

A permutation test on the Table 6.11 data is another nonparametric approach, and gives, for the present data, a p-value of 0.14: close to the Pirie and correlation CMH p-value. Further work needs to be done to confirm that the Pirie Z_L (or correlation CMH statistic) is a preferred approach.

The table of counts for these data are shown in Table 6.12. In any such example the number of columns cannot be known before sighting the data. Moreover while, typically, the column totals in Table 6.10 are all 10, and known before sighting the data, the same cannot be said of the column totals of Table 6.12.

6.7 Cochran's Test

Cochran's test assesses treatment effects in a randomised block design in which the data are zeros and ones. We derive a Wald-type test statistic which is a simple multiple of the Cochran (1950) test statistic. In section 6.8 this result is extended to permit more complex outcomes than are zeros and ones.

Suppose we have 10 doctors each examining 50 patients to assess whether or not each patient has a particular complaint. The outcome is a bald *yes* or *no* for each patient. If the doctors are considered to be treatments and the patients blocks, this is a randomised block design. Since the data are only zeros and ones, the analysis of Cochran (1950), described for example in Conover (1998), may be used. Other interesting situations in which this design arises are when food products (treatments) are being assessed by consumers (blocks), and when the portfolios of various companies (treatments) are being assessed by accountants (blocks) for potential bankruptcy.

We now give a model for this problem and derive a Wald-type test statistic. Suppose there are r (> 1) treatments, c (> 1) blocks and the outcomes are ones, a positive response, and zeros, a negative response. For u = 1, ... , r and v = 1, ... , c, let N_{uv} = 1 if treatment u on block v is assessed positively, and zero otherwise. By (rough) analogy with the two factor analysis of variance, assume that

$$N_{uv} = \theta_u + p_v + E_{uv}$$

where θ_u reflects the inclination for treatment u to be assessed positively over and above the consensus of the other treatments, where p_v is the probability, averaging over treatments, that block v is assessed positively, and where the E_{uv} are mutually independent random errors, each with mean zero.

Now

$$\{P(N_{1v} = 1) + ... + P(N_{rv} = 1)\}/r = p_v \text{ for } v = 1, ... , c.$$

Thus since $r\, p_v = \sum_u (\theta_u + p_v)$, $\theta_1 + ... + \theta_r = \theta_.$, say, satisfies $\theta_. = 0$, one of the θ_u must be a dependent variable. We take this to be θ_r.

Since $E[N_{uv}] = E[\theta_u + p_v + E_{uv}] = \theta_u + p_v = \pi_{uv}$, say, $var(N_{uv}) = \pi_{uv}(1 - \pi_{uv}) = p_v(1 - p_v)$ under the null hypothesis of no treatment differences. The model here is a product of rc binomial $(1, \pi_{uv})$ random variables, so that

$$N_{uv} = \hat{\pi}_{uv} = \hat{\theta}_u + \hat{p}_v, \text{ for } u = 1, ... , r \text{ and } v = 1, ... , c.$$

Summing N_{uv} over u gives, since $\hat{\theta}_. = 0$,

$$\hat{p}_v = N_{.v}/r, \; v = 1, ... , c,$$

while summing N_{uv} over v gives, after rearrangement,

$$c\hat{\theta}_u = N_{u.} - \frac{N_{..}}{r} = N_{u.}\,(1 - \frac{1}{r}) - \frac{1}{r} \sum_{u' \neq u} N_{u'.}, \; u = 1, ... , r.$$

Writing $\theta = (\theta_1, ... , \theta_{r-1})^T$, we wish to test the null hypothesis H: $\theta = 0$ against K: $\theta \neq 0$. In general the Wald test is based on the statistic $\hat{\theta}^T \left\{cov^{-1}(\hat{\theta})\right\} \hat{\theta}$ in which the asymptotic covariance matrix $cov(\hat{\theta})$ depends on

θ, and in its usable form, in $\text{cov}(\hat{\theta})$ we replace θ by $\hat{\theta}$; see Appendix A.5. A Wald-type test may be based on $\hat{\theta}^T \left\{\text{cov}^{-1}(\hat{\theta})\right\} \hat{\theta}$ in which the exact rather than the asymptotic covariance matrix is used. Here we will use the covariance matrix evaluated under the null hypothesis. This results in a much simpler form of the test statistic than the usual Wald test. Now write, for any positive integer n, I_n for the n by n unit matrix, 1_n for the n by 1 vector of ones, and J_n for the n by n matrix given by

$$J_n = I_n + 1_n 1_n^T, \text{ so that } J_n^{-1} = I_n - \frac{1}{n+1} 1_n 1_n^T.$$

Direct evaluation, using the expression for $c\hat{\theta}_u$, gives for u = 1, ... , r

$$\begin{aligned} c^2 \text{ var}_0(\hat{\theta}_u) &= (1 - \tfrac{1}{r})^2 \sum_v p_v (1 - p_v) + (\tfrac{1}{r})^2 (r - 1) \sum_v p_v (1 - p_v) \\ &= \left(\frac{r - 1}{r}\right) \sum_v p_v (1 - p_v), \end{aligned}$$

and similarly, for $u \neq u' = 1, \ldots, r$,

$$c^2 \text{ cov}_0(\hat{\theta}_u, \hat{\theta}_{u'}) = -\frac{1}{r} \sum_v p_v (1 - p_v),$$

together yielding

$$\text{cov}_0(\hat{\theta}) = \{\sum_v p_v (1 - p_v)\} J_{r-1}^{-1} / c^2.$$

This leads to the test statistic

$$W_E = \hat{\theta}^T \text{cov}_0^{-1}(\hat{\theta}) \hat{\theta} = \frac{c^2 \sum_{u=1}^{r} \hat{\theta}_u^2}{\sum_{v=1}^{c} \hat{p}_v(1 - \hat{p}_v)} = \frac{\sum_{u=1}^{r} \left(N_{u.} - \frac{N_{..}}{r}\right)^2}{\sum_{v=1}^{c} \hat{p}_v(1 - \hat{p}_v)}.$$

The subscript E signifies the exact covariance matrix has been used rather

than an asymptotic approximation. Under the null hypothesis of no treatment differences and for c large, the Wald-type statistic W_E approximately follows the χ^2_{r-1} distribution. When all $\hat{p}_v$ are either zero or one, W_E is undefined. Then the responses on any given block are identical, and there is no information about differences between treatments; W_E cannot be informative.

We recognise that $(r - 1)W_E/r = Q$, the test statistic derived by Cochran (1950). Cochran claimed that when c is large, Q will have the χ^2_{r-1} distribution, but does not claim his proof to be rigorous. In this situation Shah and Claypool (1985) derive W_E, and argue that its asymptotic distribution is χ^2_{r-1}. However they claim inserting the factor $(r - 1)/r$ improves the asymptotic distribution, and attribute this to Conover (1980). However Conover gives no justification for the claim. When all the marginals are fixed, Landis et al. (1979) derive $(r - 1)W_E/r$ as a Mantel-Haenszel statistic, again with the χ^2_{r-1} distribution.

Clearly if r, the number of treatments, is small, there will be a considerable difference in the distributions of W_E and Q. Nevertheless, whether the margins are fixed or not, it appears from preliminary studies of our own that Q is better approximated by χ^2_{r-1} than W_E. However the number of blocks, c, is not often large enough that the asymptotic distribution can be assumed with confidence. Then a resampling test should be performed, and this would be conditional on the row and column totals being known, in which case W_E and Q are equivalent statistics. Our inclination would be to use a parametric bootstrap, using the maximum likelihood estimators $\hat{p}_v$ and $\hat{\theta}_u$.

Milk Example. Suppose milk has been refrigerated for three days in four different types of containers: opaque plastic (A), clear plastic (B), cardboard (C) and glass (D). A question of interest is whether the containers are equal in their ability to keep the milk fresh. Six *in-house* or expert judges are asked "is the milk fresh?" The judges do not know from which container each sample of milk comes. Each judge tastes a milk sample from each of A, B, C and D. Samples are presented in random order. A "yes" response is shown as "1" in the following table and a "no" response as "0".

Table 6.13 Responses by six expert judges for a milk storage trial

Judge	A	B	C	D
1	1	1	1	0
2	1	1	1	0
3	1	0	1	1
4	1	0	1	1
5	1	0	1	1
6	1	0	1	0

Suppose we carry out some conventional statistical tests on these data. The ANOVA F test for containers gives p-value 0.027, the Friedman test adjusted for ties, which is just Cochran's Q for binary data, gives p-value 0.046 and a logistic regression for binary data (which is a Generalized Linear Model with binomial errors and logit link) doesn't have a solution!

The reason there is no logistic regression for binary data is that the maximum likelihood estimators don't exist. This is a known problem; see, for example, Agresti (1996, p. 134). However, we can still get a logistic regression solution if we use a score test rather than the likelihood ratio test used in most of the statistics packages (such as MINITAB, GENSTAT and SPLUS). The logistic regression score test is available in the LOGXACT statistics package. However, would we believe p-values based on large sample theory for such a small data set of binary data? LOGXACT calculates an *exact* p-value as well as giving a large sample p-value, but it is a conditional *exact* p-value. Conditional p-values sometimes seem very conservative. The LOGXACT *exact* p-value is obtained by conditioning on the rows sums for each judge. Now, if we repeated the taste test with the same expert judges there is no guarantee these judges would respond with the same row sums as they did previously. Hence it is not clear that the LOGXACT *exact* p-value is appropriate. Further discussion of the LOGXACT approach is given, for example, in Mehta and Patel (1995).

The *exact* conditional p-value is a permutation test p-value. How can we get unconditional *exact* p-values? Applying the bootstrap to each

judge's results would seem a way of generating many data sets under the null hypothesis of no difference between container types. This is equivalent to estimating the probability of a fresh assessment for each judge and then simulating more taste tests with the same judges by generating random Bernoulli trials (binomials with n = 1) with the estimated probabilities. If we simulate many new taste tests in this way, calculate test statistics for each of the simulated taste tests, and compare the test statistic for the actual data with these we get an unconditional *exact* p-value. Notice that both the permutation and bootstrap p-values can be called *exact* p-values.

It is easy enough to find bootstrap p-values for the ANOVA F test and for the Cochran's Q test. The usual logistic regression for binary data requires an iterative calculation for each simulated taste test. This requires a lot of computing.

To overcome this difficulty we note that the binary logistic score test for differences between milk containers is just Cochran's Q test. The LOGXACT Users Manual (1996, section 8.6) demonstrates this for r = 2. We find a bootstrap p-value 0.030 for Cochran's Q, while a permutation test has a p-value of 0.053: a value some would consider nonsignificant. Further

$$Q = b(t - 1)F/(F + b - 1)$$

where b = number of judges; see Shah and Claypool (1985, p. 1178). Thus the bootstrap p-value for F is also the same as for Q. The good agreement between the p-value for the F test found from the F distribution as in the usual ANOVA and the bootstrap p-value demonstrated above has been repeated in other small data sets we have looked at.

To summarize, the p-values obtained using the different methods are:

Test Statistic	*p-value*	*Method*
Anova F	0.027	F distribution
Cochran's Q	0.046	chi-squared
Cochran's Q	0.053	permutation
F and Q	0.030	bootstrap

6.8 Stuart's Test and its Extensions

Suppose now that instead of just two outcomes, as in section 6.7, k are possible. Thus a doctor may permit uncertainty in his or her diagnosis: a patient definitely has the disease, probably has it, probably doesn't have the disease, or definitely doesn't. For u = 1, ... , r, v = 1, ... , c and w = 1, ... , k, let N_{uvw} = 1 if treatment u on block v is assessed in the w*th* category, and zero otherwise. Put

$$N_{uvw} = \theta_{uw} + p_{vw} + E_{uvw},$$

where θ_{uw} reflects the inclination for treatment u to be categorised in the w*th* category over and above the consensus of the other treatments, where p_{vw} is the probability, averaging over treatments, that block v is categorised into the w*th* category, and where the E_{uvw} are mutually independent errors each with mean zero.

Now $\sum_w P(N_{uvw} = 1) = 1$, since treatment u on block v is categorised into one of the k categories. Thus

$$\theta_{u.} = \theta_{u1} + ... + \theta_{uk} = 0 \text{ for all u and}$$
$$p_{v.} = p_{v1} + ... + p_{vk} = 1 \text{ for all v.}$$

Since p_{vw} is the average indicated, $\theta_{.w} = \theta_{1w} + ... + \theta_{rw} = 0$ for all w. Noting that $\theta_{u.} = 0$, $\theta_{.w} = 0$ and $p_{v.} = 1$ for all u, v and w respectively, there are (r - 1)(k - 1) θ_{uw}s, and c(k - 1) p_{vw}s.

This model is a product of rc multinomials. We find

$$N_{uvw} = \hat{\theta}_{uw} + \hat{p}_{vw} \text{ for all u, v and w,}$$

and summing over u gives

$$\hat{p}_{vw} = \frac{N_{.vw}}{r} \text{ for all v and w,}$$

and substituting back, summing over v and dividing by c gives

$$\hat{\theta}_{uw} = \frac{N_{u.w}}{c} - \frac{N_{..w}}{rc} \text{ for all u and w.}$$

Write $\theta = (\theta_{11}, \ldots, \theta_{1(k-1)}, \ldots, \theta_{(r-1)1}, \ldots, \theta_{(r-1)(k-1)})^T$. The covariance matrix for $c\hat{\theta}$ is the sum of the covariance matrices corresponding to each block, since blocks are independent. We now consider only the v*th* block.

If $p_{vw} = 0$ the w*th* category should be eliminated from the model, as none of the treatments are categorised into this category. Thus we may assume that $p_{vw} \neq 0$ for all v and w. If

$$D_v = \text{diag}\left(\frac{1}{p_{v1}}, \ldots, \frac{1}{p_{v(k-1)}}\right) + \frac{1}{p_{vk}} 1_{k-1} 1_{k-1}^T$$

then $D_v^{-1} = \text{diag}(p_{vw}) - (p_{vw}p_{vz})$. Analysis similar to that in section 6.7 shows that when $\theta = 0$, $\hat{\theta}$ has covariance matrix

$$\mathrm{cov}_0(\hat{\theta}) = \begin{pmatrix} (1 - \frac{1}{r})D_v^{-1} & \frac{1}{r}D_v^{-1} & \cdots & \frac{1}{r}D_v^{-1} \\ \frac{1}{r}D_v^{-1} & (1 - \frac{1}{r})D_v^{-1} & \cdots & \frac{1}{r}D_v^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{r}D_v^{-1} & \frac{1}{r}D_v^{-1} & \cdots & (1 - \frac{1}{r})D_v^{-1} \end{pmatrix} = J_{r-1}^{-1} \otimes D_v^{-1},$$

in which $\otimes$ is the direct or Kronecker product.

Now consider all blocks. The covariance matrix of $\hat{\theta}$ is given by

$$c^2 \, \mathrm{cov}_0(\hat{\theta}) = J_{r-1}^{-1} \otimes (D_1^{-1} + \ldots + D_c^{-1}),$$

from which

$$\mathrm{cov}_0^{-1}(\hat{\theta}) = J_{r-1} \otimes A$$

in which A is defined by

$$c^2 A^{-1} = D_1^{-1} + \ldots + D_c^{-1} = \mathrm{diag}\left(\sum_{v=1}^{c} p_{vw}\right) - \left(\sum_{v=1}^{c} p_{vw}p_{vz}\right).$$

Write $\theta_u = (\theta_{u1}, \ldots, \theta_{u(k-1)})^T$ for $u = 1, \ldots, r$ so that $\theta = (\theta_1^T, \ldots, \theta_{r-1}^T)^T$. The Wald-type statistic for testing H: $\theta = 0$ against K: $\theta \neq 0$ for this model, W_{Ek}, say, is given by

$$W_{Ek} = \hat{\theta}^T \, \mathrm{cov}_0^{-1}(\hat{\theta}) \, \hat{\theta}$$

in which p_{vw} is replaced by $\hat{p}_{vw}$. If $\hat{A}$ denotes A in which p_{vw} is replaced by $\hat{p}_{vw}$, then

$$W_{Ek} = \sum_{u=1}^{r-1} \hat{\theta}_u^T \hat{A} \hat{\theta}_u + \sum_{u=1}^{r-1} \sum_{u'=1}^{r-1} \hat{\theta}_u^T \hat{A} \hat{\theta}_{u'}.$$

Using $\hat{\theta}_{.w} = 0$ it can be shown that

$$\sum_{u=1}^{r-1} \sum_{u'=1}^{r-1} \hat{\theta}_u^T \hat{A} \hat{\theta}_{u'} = \hat{\theta}_r^T \hat{A} \hat{\theta}_r,$$

so that

$$W_{Ek} = \sum_{u=1}^{r} \hat{\theta}_u^T \hat{A} \hat{\theta}_u.$$

The test statistic W_{Ek} is calculated from this expression, noting that

$$c\,\hat{\theta}_u = \left(N_{u.1} - \frac{N_{..1}}{r}, \ldots, N_{u.(k-1)} - \frac{N_{..(k-1)}}{r}\right)^T \text{ and}$$

$$\hat{A} = c^2 \left\{ \frac{1}{r} \operatorname{diag}\left(N_{..w}\right) - \frac{1}{r^2} \left(\sum_{v=1}^{c} N_{.vw} N_{.vz} \right) \right\}^{-1}.$$

This new W_{Ek} statistic can be derived from the generalized Cochran-Mantel-Haenszel statistic, QCMH, given in Landis et al. (1979). This can be done by observing that the scores made by k*th* consumer can be given as an r by s contingency table with a one in the i*th* row and j*th* column if product i is rated by consumer j into category k, and zeroes elsewhere. The statistic W_{Ek} has an asymptotic $\chi^2_{(r-1)(k-1)}$ distribution. The situation here is similar to that with Cochran's Q: $[(r - 1)/r]W_{Ek}$ is better approximated by this chi-squared distribution than is W_{Ek}. The Anderson statistic, A of section 6.2, is a special case of $[(r - 1)/r]W_{Ek}$.

Observe that the QCMH statistic and hence our W_{Ek} statistic is different from a statistic of Bhapkar (1970), who used a population average approach; see the discussion in Agresti (1990, p. 405). Also observe that Agresti (1990, Chapter 11) discussed modelling approaches that provide alternative tests to the W_{Ek} test. These alternative tests involve iterative calculations, that may not converge, while the test based on W_{Ek} does not.

'Just Right' Example for Two Products. Suppose 100 consumers rate two food products X and Y for sweetness on a three point *just right* scale with categories *too strong*, *just right* and *too weak*. The data for the 100 pairs of ratings can be conveniently summarized in a 3 by 3 table of counts as shown in Table 6.14.

The response frequencies can also be summarized as a 2 by 3 table of marginal counts as shown in Table 6.15. Notice that it is not appropriate to calculate the usual chi-squared statistic for homogeneity for Table 6.15 and test its significance using a χ^2_2 distribution. For that to be appropriate, different consumers would need to have tasted the different products.

To see if the ratings from products X and Y are significantly different a common statistical procedure is to assign scores, say 1, 2, 3, to the three categories and to carry out a randomized blocks analysis of variance. If this is done for the Table 6.14 data then the means of X and Y are both 2.0.

Table 6.14 Response frequencies for a just right rating of two products by 100 consumers

	Product X		
Product Y	Too strong	Just right	Too weak
Too strong	8	24	8
Just right	7	7	6
Too weak	10	19	11

Table 6.15 Marginal frequencies from a just right rating of two products by 100 consumers

Product	Rating		
	Too strong	Just right	Too weak
X	25	50	25
Y	40	20	40

The between products test statistic F takes the value 0, so that it may be concluded that there is no difference between products: each is *just right*. However note that the use of ANOVA here is a little suspect as the data are too discrete to be considered normally distributed and the choice of scores is rather arbitrary.

If we calculate $[(r - 1)/r]W_{Ek}$, which for r = 2 products is equivalent to the Stuart (1955) test, we find $[(r - 1)/r]W_{Ek} = 16.2$. As this has an approximate χ^2_2 distribution, it is very highly significant. This conflicts with the conclusion based on the ANOVA because the test based on W_{Ek} compares the two *distributions* and not just the two means of the distributions. For consumer data it is often, as here, important to compare both distributions and means. For the Table 6.15 data it appears there is market segmentation for product Y and this is important information identified by the W_{Ek} test and not by the ANOVA F test. Observe that W_{Ek} does not use assigned scores as the ANOVA does.

Agresti (1990, section 10.3.1) discusses a test for marginal homogeneity due to Caussinus. This test uses parametric log-linear models and here gives a value of 16.2, the same value as the Stuart statistic.

'Just Right' Example for Three Products. Suppose it is wished to compare three products X, Y and Z on the three point just right scale and that the categories are labelled 1, 2, 3. The data from eight consumers are shown in Table 6.16.

Table 6.16 Just right ratings for three products

Consumer	Product X	Y	Z
1	2	1	2
2	2	3	2
3	2	1	2
4	1	2	1
5	1	2	1
6	1	3	2
7	1	3	2
8	1	3	1

The randomized blocks ANOVA F statistic takes the value 3.37 with a p-value, based on the $F_{2,14}$ distribution, of 0.06. Friedman's ranking statistic for randomized blocks takes the value 4.00 with p-value 0.14. Actual sensory consumer tests we have analysed often show that F is more sensitive than Friedman's ranking statistic as is the case with this example. However, Friedman's ranking statistic has the advantage of not using arbitrary scores or of assuming normality as F does. We also find $[(r - 1)/r]W_{Ek} = 9$ on 4 degrees of freedom with p-value 0.06.

Notice that Friedman's ranking statistic can be expressed as a sum of squares and so can be partitioned using linear contrasts in a similar manner to the treatment sum of squares in the ANOVA. The linear contrast comparing one product or treatment with the average of the others gives a generalisation of the statistic of Miettinen (1969). Further, if the products have an ordering associated with them, then Page's test corrected for ties can be used to assess the significance of this ordering. As in section 6.6, see Pirie (1985) for an appropriate formula.

7
Further Tests on Randomised Blocks: Extensions to Durbin's Test

7.1 Introduction

In this chapter we begin by considering balanced incomplete block designs for ranked data, traditionally analysed using Durbin's test to assess differences between treatments that we subsequently identify as location differences. We begin, as in Chapter 6, with an example to demonstrate the rather complete analysis we are able to give with our approach to data following a balanced incomplete block design.

Dried Egg Example. In taste-testing there is evidence to suggest that tasting more than four or five samples at one sitting results in reduced acuity arising from confusion or fatigue associated with making a large number of comparative judgements. For this reason incomplete block designs are employed. Hanson et al. (1951) gave results from a taste test on ten dried egg samples, A, B, ... , J, where the design used was a balanced incomplete block design. In Table 7.1 below we give the design for 15 sittings, and in Table 7.2 the corresponding mean scores for *off-flavour* from seven judges. These mean scores may be ranked 1, 2, 3, 4 and counts made of how many times each sample received each rank. These are given in Table 7.3. Finally we partition Anderson's statistic into components $V_1^T V_1$, which is Durbin's statistic, $V_2^T V_2$, a new dispersion statistic, and $V_3^T V_3$, a residual, all with associated p-values. See Table 7.4.

A treatment map is shown in Figure 7.1. The experimenter expected the A to J location ordering. The map also shows that ranks for treatments D, F, G, H are more dispersed than A or J, and the ranks for E are dispersed like those of C and D.

The partition of Anderson's statistic shows strong evidence of both location and dispersion differences between treatments, and that all of the variability in the data has been identified.

Table 7.1 Balanced incomplete block design for dried egg taste test

Sitting	Samples	Sitting	Samples	Sitting	Samples
1	A B D E	6	B G H I	11	A F G J
2	B C F J	7	B E H J	12	C D I J
3	B D F G	8	E G I J	13	A F H I
4	A C E G	9	A B C I	14	C D G H
5	A D H J	10	D E F I	15	C E F H

Table 7.2 Off-flavour scores for samples as specified by BIB design above

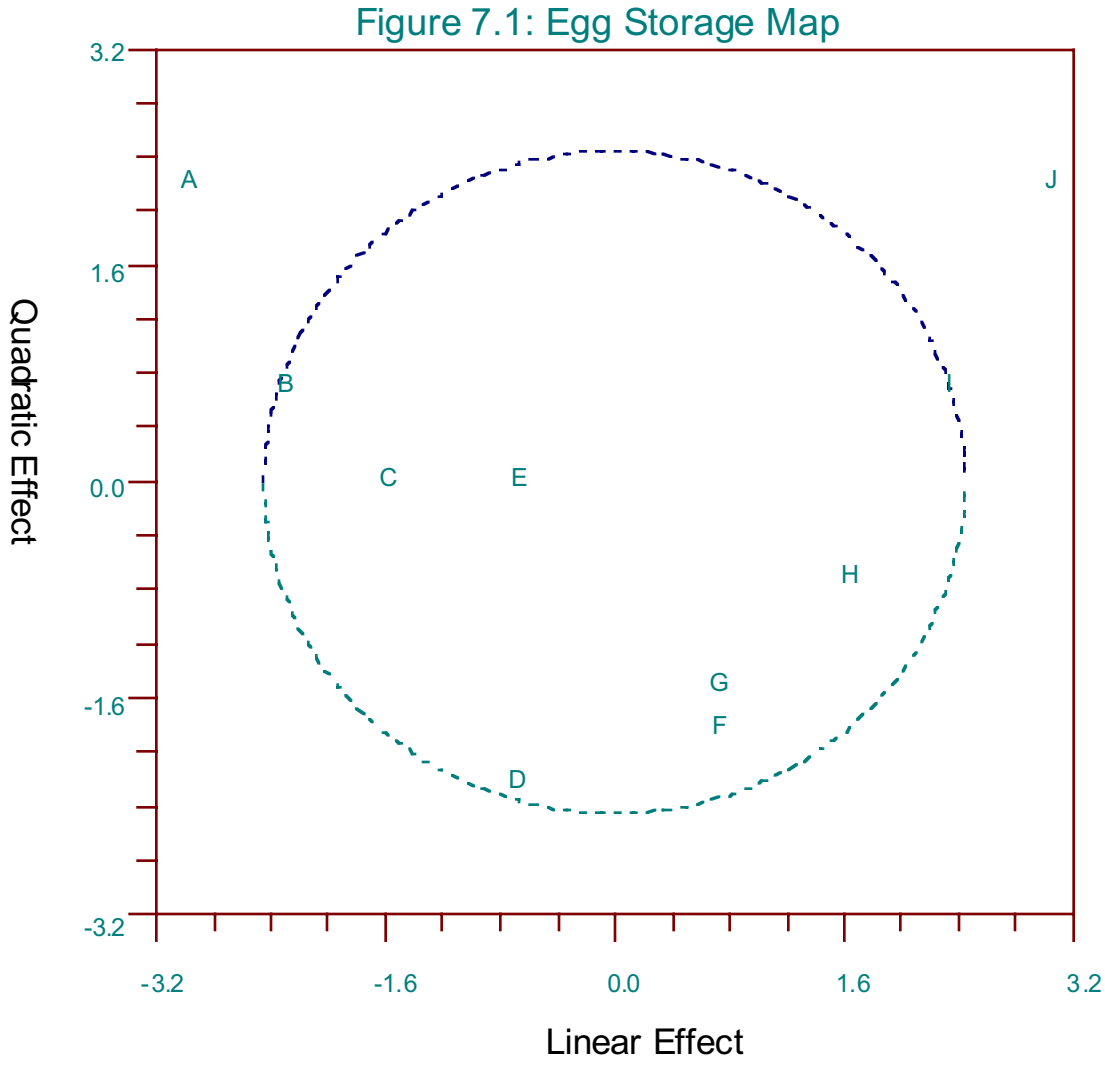
Sitting	Scores	Sitting	Scores	Sitting	Scores
1	9.7, 8.7, 5.4, 5.0	6	9.6, 5.1, 4.6, 3.6	11	9.7, 6.7, 6.6, 2.8
2	9.6, 8.8, 5.6, 3.6	7	9.8, 7.4, 4.4, 3.8	12	9.3, 8.1, 3.7, 2.6
3	9.0, 7.3, 3.8, 4.3	8	9.4, 6.3, 5.1, 2.0	13	9.8, 7.3, 5.4, 4.0
4	9.3, 8.7, 6.8, 3.8	9	9.4, 9.3, 8.2, 3.3	14	9.0, 8.3, 4.8, 3.8
5	10.0, 7.5, 4.2, 2.8	10	8.7, 9.0, 6.0, 3.3	15	9.3, 8.3, 6.3, 3.8

Table 7.3 Counts of ranks for dried egg data

Product	Rank			
	1	2	3	4
A	6	0	0	0
B	4	2	0	0
C	3	2	1	0
D	0	5	1	0
E	2	2	1	1
F	0	2	3	1
G	0	2	3	1
H	0	0	4	2
I	0	0	2	4
J	0	0	0	6

Table 7.4 Partition of Anderson's statistic for the dried egg data

Statistic	df	Value	χ^2 p-value	Monte Carlo p-value
$V_1^T V_1$	9	39.12	< 0.001	< 0.001
$V_2^T V_2$	9	22.80	0.007	0.002
Residual	9	10.08	0.344	0.337



The plot in Figure 7.1 is typical of data where there is a product ordering. It would be useful to examine this ordering using the analysis introduced in section 7.4 below. Of course these data can be analysed using ANOVA, but unlike our ranks analysis, ANOVA assumes variance homogeneity - which clearly does not hold here - and normality.

It is known that the approximate χ^2 probabilities for Durbin's test are not particularly good: see Van der Laan (1988). Thus the use of Monte Carlo p-values is recommended.

Further examples are given at the end of sections 7.3, 7.4, and 7.5.

7.2 Durbin's Test and its Extensions

Data from a balanced incomplete block experiment may be presented in a contingency table with both marginal totals fixed. This table, like Table 6.5, is not the usual contingency table, and X_P^2 does not follow the χ^2 distribution, although a simple multiple of it does. As in the Friedman case, the appropriate test statistic may be partitioned into components. The first component is Durbin's (1951) test statistic, and subsequent components give moment-type extensions.

To give the first two components explicitly note that 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 $k < t$, $r < b$, $bk = rt$, and $\lambda(t - 1) = r(k - 1)$. Durbin's statistic, D_1, and the new statistic, D_2, that detects dispersion differences between treatments, are given by

$$D_1 = \frac{12\,(t - 1)}{rt\,(k^2 - 1)} \sum_{i=1}^{t} R_i^2 - 3r(t - 1)(k + 1)/(k - 1) \text{ and}$$

$$D_2 = \frac{180\,(t - 1)}{rt\,(k^2 - 1)\,(k^2 - 4)} \sum_{i=1}^{t} \{S_i - (k + 1)R_i + r(k + 1)(k + 2)/6\}^2,$$

where $R_i = \sum_{j=1}^{k} jN_{ij}$ and $S_i = \sum_{j=1}^{k} j^2N_{ij}$, in which N_{ij} is the number of times product i receives rank j. We now show how to derive D_1 and D_2 and further information useful for data analysis.

7.3 Derivations

In Table 7.3 all the row totals are 6 and all the column totals are 15. This is typical in that data from a balanced incomplete block experiment may be always presented in a contingency table with both marginal totals fixed.

As noted above, this is not the usual contingency table, and the 'usual' X^2 does not follow the χ^2 distribution, although we show that a simple multiple of X^2 does. We partition the appropriate test statistic into components. The first component is simply related to Durbin's (1951) test, and subsequent components give moment-type extensions. The data are given in a vector of counts, for which the covariance and correlation matrices are derived. The correlation matrix has the same form as that derived by Anderson (1959), and a derivation parallel to that in previous sections of this paper establishes the desired extensions.

We use the description and notation from Conover (1998, section 5.9), given in the previous section. On each of the b blocks, k treatments are compared. This produces k! possible preference sequences for each block; a typical preference sequence $s_1 s_2 \ldots s_k$ indicates that treatment 1 is given rank s_1, treatment 2 is given rank s_2, ... , treatment k is given rank s_k. Under the null hypothesis that all treatments are equally preferred, each of the (k! b)preference sequences is equally likely, with probability $p = (k!\ b)^{-1}$. If Y_h is the number of times the h*th* preference sequence is selected, then the vector (Y_h) follows a multinomial distribution with parameters $b = \sum Y_h$ = the number of judgements made (the summation is over all sequences), and probability p for each of the (k! b) preferences sequences. The Y_h have means bp, variances bpq, and covariances - bp^2.

As before, define N_{ij} to be the number of times treatment i is given rank j. The set $\{N_{ij}\}$ forms a t by k contingency table with all row totals r and all column totals b. The random variable N_{ij} is the sum over (k - 1)! r of the Y_h. It follows that, on using $p = (k!\ b)^{-1}$,

$$E[N_{ij}] = [(k - 1)!\ r]\ bp = r/k,$$

and, ultimately,

$$\begin{aligned} \mathrm{var}(N_{ij}) &= [(k - 1)!\ r]\ bpq - bp^2\ [(k - 1)!\ r][(k - 1)!\ r - 1] \\ &= b(t - 1)/t^2. \end{aligned}$$

This also uses the result that since $N_{ij} = \sum_{c \text{ cells}} Y_h$,

$$\text{var}(N_{ij}) = \sum_{c \text{ terms}} \text{var}(Y_h) + \sum_{c(c-1) \text{ terms}} \text{cov}(Y_h, Y_k),$$

where here $c = (k - 1)!\, r$. For $j \neq t$,

$$\text{cov}\ (N_{ij}, N_{it}) = - bp^2\, [(k - 1)!\, r]^2 = - b/t^2$$

ultimately, because N_{ij} and N_{it} count the number of times variety i is given rank j and rank t respectively, and these events cannot occur together; so N_{ij} and N_{it} are sums over $[(k - 1)!\, r]$ Y_h corresponding to non-overlapping cells. It now follows that

$$\text{correlation}\ (N_{ij}, N_{it}) = - bp^2\, [(k - 1)!\, r]^2 = - 1/(t - 1) \text{ for } j \neq t.$$

Similarly, for $i \neq s$, correlation $(N_{ij}, N_{sj}) = - bp^2\, [(k - 1)!\, r]^2 = - 1/(t - 1)$.

We now consider cov (N_{ij}, N_{st}) for $i \neq s$ and $j \neq t$. The random variables N_{ij} and N_{st} involve sums over $[(k - 2)!\, \lambda]$ Y_h corresponding to non-overlapping cells. It follows that for $i \neq s$ and $j \neq t$

$$\text{cov}\ (N_{ij}, N_{st}) = (k - 2)!\, \lambda\, bpq - bp^2\, [(k - 1)!\, r][(k - 1)!\, r - (k - 2)!\, \lambda]$$

$$= b/[t^2(t - 1)].$$

This uses the result that $N_{ij} = U + V$, $N_{st} = V + W$, where $V = \sum_{c \text{ cells}} Y_h$ in which $c = (k - 2)!\, \lambda$, and U and W each involves summing Y_h's corresponding to $[(k - 1)!\, r - (k - 2)!\, \lambda]$ non-overlapping cells. This leads to $[(k - 1)!\, r][(k - 1)!\, r - (k - 2)!\, \lambda]$ equal covariance terms.

It now follows that

$$\text{correlation}\ (N_{ij}, N_{st}) = 1/(t - 1)^2 \text{ for } i \neq s \text{ and } j \neq t.$$

If we put

$$X_{ij} = (N_{ij} - r/k)t/\sqrt{[b(t-1)]} \text{ for } i = 1, \ldots, t \text{ and } j = 1, \ldots, k, \text{ and}$$
$$X = (X_{11}, \ldots, X_{1k}, X_{21}, \ldots, X_{2k}, \ldots, X_{t1}, \ldots, X_{tk})^T,$$

then X will be asymptotically tk variate normal with zero mean and covariance matrix

$$\Sigma = R_t \otimes R_k, \text{ in which } R_n = [t/(t-1)]\, I_n - [1/(t-1)]\, 1_n 1_n^T.$$

The eigenvalues of R_n are $t/(t-1)$ a total of $(n-1)$ times, and zero once. The eigenvalues of Σ are thus $[t/(t-1)]^2$ a total of $(t-1)(k-1)$ times, and zero $(t+k-1)$ times. If we note that the expected number of the (i, j)*th* cell is $r/k = b/t$, then

$$X_P^2 = \frac{k}{r} \sum_{i=1}^{t} \sum_{j=1}^{k} \left(N_{ij} - \frac{r}{k}\right)^2.$$

Following the steps in 6.3.2, we see that $(t-1)X_P^2/t$ has the $\chi^2_{(t-1)(k-1)}$ distribution.

If we follow the derivation of 6.3.3, we obtain

$$(t-1)\, X_P^2/t = V_1^T V_1 + \ldots + V_{k-1}^T V_{k-1},$$

in which the V_u are asymptotically mutually independent and asymptotically $N_t(0, I_t - \frac{1}{t} 1_t 1_t^T)$, so that the $V_u^T V_u$ are asymptotically mutually independent χ^2_{t-1}. Explicitly, for $u = 1, \ldots, k-1$, we have

$$V_u = \frac{t-1}{t} \frac{1}{\sqrt{k}} G_u^T X = \frac{t-1}{t} \frac{1}{\sqrt{k}} \left(\sum_{j=1}^{k} g_u(j)\, X_{ij} \right) = \sqrt{\frac{t-1}{bk}} \left(\sum_{j=1}^{k} g_u(j)\, N_{ij} \right).$$

The first component is of the form of the sum of the squares of

$$a\sqrt{\frac{t-1}{rt}}\sum_{j=1}^{k}\left(j-\frac{k+1}{2}\right)N_{ij}=\sqrt{\frac{12}{(k^2-1)}}\sqrt{\frac{(t-1)}{rt}}\left(R_i-\frac{k+1}{2}r\right).$$

Here a is analogous to a similar quantity defined in section 6.3.4, and we use rt = bk. This gives Durbin's statistic,

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

As before, the subsequent components, $V_u^TV_u$, u = 2, ... , k - 1, assess quadratic (loosely dispersion) and higher order differences between the treatments that we loosely label moment differences. In the example of section 7.1 we have calculated $V_1^TV_1 = D_1$, Durbin's statistic, $V_2^TV_2 = D_2$, a dispersion detecting statistic having, like D_1, the χ^2_{t-1} distribution, and

$$(t-1)X_P^2/t - V_1^TV_1 - V_2^TV_2 = V_3^TV_3 + \ldots + V_{k-1}^TV_{k-1},$$

a residual having the $\chi^2_{(t-1)(k-3)}$ distribution. Note that if k = t then D_1 becomes Friedman's statistic and $(t-1)X_P^2/t$ becomes Anderson's statistic discussed in section 6.2.

Ice Cream Example. Conover (1998, 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 7.5 shows, for each judge, the rank given for each variety; Table 7.6 gives the corresponding counts $\{N_{ij}\}$.

Table 7.5 Ranks of seven judges of seven ice cream varieties

	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

We see that $\{R_1, R_2, R_3, R_4, R_5, R_6, R_7\} = \{8, 9, 4, 3, 5, 6, 7\}$ and $\{S_1, S_2, S_3, S_4, S_5, S_6, S_7\} = \{22, 27, 6, 3, 9, 14, 17\}$. Thus for these data $D_1 = 12$, $D_2 = 5.1$, $(t - 1)X_P^2/t = 17.1$ and $U_{11}^2 = 0.98$. Using the approximating chi-squared distributions the corresponding p-values are 0.062, 0.526, 0.125 and 0.325. Thus, if a 5% significance level is used, none of these would be declared significant and this is the conclusion of Conover (1998, p. 390). However, using the more exact Monte Carlo method described subsequently, the p-value for D_1 is 0.020, and so at least D_1 is significant at the 5% level. There is some evidence of differences in mean product rankings. The insignificant D_2 means we fail to find evidence of market segmentation in ice cream preferences.

The Monte Carlo p-values for D_2, $(t - 1)X_P^2/t$ and U_{11}^2 are 0.138, 0.801 and 0.370 respectively. This example identifies the fact that the chi-squared approximations to the statistics D_1, D_2, $(t - 1)X_P^2/t$ and U_{11}^2 may not be accurate. In relation to the D_1 statistic, Daniel (1990, p. 286) said

> "The chi-square approximation is good only when r is large, and it should be realized that the results are probably very crude when r is small."

Van der Laan (1988) gave examples of errors obtained when using the chi-squared approximation. However, with modern computing facilities there is no need to use the chi-squared approximation. Monte Carlo p-values can easily be produced and these should be used when at all feasible.

Schach (1979) refers without details to an omnibus chi-squared type statistic, our $(t - 1)X_P^2/t$, and a statistic which tests for a predetermined ordering of product rank sums. We call this a Page-type statistic.

Table 7.6 Counts $\{N_{ij}\}$ for ice cream data

	Rank		
Ice Cream	1	2	3
S	0	1	2
U	0	0	3
V	2	1	0
W	3	0	0
X	1	2	0
Y	1	1	1
Z	0	2	1

7.4 A Page-Type Test

We now give an alternative partition of the $(t - 1)X_P^2/t$ that also leads to Page-type statistic, and to a new umbrella statistic. First define an s by s orthogonal matrix L_s, say, by

$$L_s = \begin{pmatrix} g_1(1) & \cdots & g_{s-1}(1) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ g_1(s) & \cdots & g_{s-1}(s) & 1 \end{pmatrix} / \sqrt{s},$$

and then put $G = L_t \otimes L_k$. Note that L_t involves $\{g_\ell(j)\}$, the polynomials orthonormal on the discrete uniform distribution taking t values, whereas L_k involves $\{h_m(j)\}$, the polynomials orthonormal on the discrete uniform distribution taking k values. Further, put $Z = \frac{t - 1}{t} G^T X$. Then Z has typical element, $U_{\ell m}$ say, given by

$$U_{\ell m} = \frac{t - 1}{t} \frac{1}{\sqrt{tk}} \sum_{i=1}^{t} \sum_{j=1}^{k} X_{ij}\, g_\ell(i)\, h_m(j)$$
$$= \sqrt{\frac{t - 1}{bkt}} \sum_{i=1}^{t} \sum_{j=1}^{k} \left(N_{ij} - \frac{r}{k}\right) g_\ell(i)\, h_m(j).$$

From the orthogonality $\sum_{i=1}^{t} g_\ell(i) = \sum_{j=1}^{k} h_m(j) = 0$. It follows that

$$U_{\ell m} = \sqrt{\frac{t - 1}{bkt}} \sum_{i=1}^{t} \sum_{j=1}^{k} N_{ij}\, g_\ell(i)\, h_m(j) \text{ and}$$
$U_{\ell m} = 0$ if either $\ell = t$ or $m = r$.

The latter follows by an argument parallel to that in section 6.5, if we put $g_t(j) = 1$ for all j, and $h_k(j) = 1$ for all j. It follows that Z has $(k - 1)(t - 1)$ nondegenerate elements $U_{\ell m}$. As before

$$(t - 1)Q^2/t = Z^T Z = \sum_{\ell=1}^{t-1} \sum_{m=1}^{k-1} U_{\ell m}^2.$$

By an argument similar to that in 6.3.2, Z is asymptotically distributed as $N_{tk}(0, (I_{(k - 1)(t - 1)} \oplus 0_{(k + t - 1)}))$. It follows that the nondegenerate elements are asymptotically independent and asymptotically have the standard normal distribution.

First recall from 6.3.4 that for j = 1, ... , s, $g_1(j) = \sqrt{\frac{12}{s^2 - 1}}(j - \frac{s + 1}{2})$. Now note that the rank sum for treatment i is $R_i = \sum_{j=1}^{k} jN_{ij}$, and $L = \sum_{i=1}^{t} iR_i$. Finally put

$$\mu_L = \frac{rt(k + 1)(t + 1)}{4} \text{ and } \sigma_L^2 = \frac{rt^2(k^2 - 1)(t + 1)}{144}.$$

By definition

$$\begin{aligned} U_{11} &= \sqrt{\frac{t - 1}{bkt}} \sum_{i=1}^{t} \sum_{j=1}^{k} N_{ij}\, g_1(i)\, h_1(j) \\ &= \sqrt{\frac{t - 1}{bkt}} \sqrt{\frac{12}{t^2 - 1}} \sqrt{\frac{12}{k^2 - 1}} \sum_{i=1}^{t} \sum_{j=1}^{k} N_{ij} \left(i - \frac{t + 1}{2}\right) \left(j - \frac{k + 1}{2}\right) \\ &= \sqrt{\frac{144}{rt^2(t + 1)(k^2 - 1)}} \left\{ \sum_{i=1}^{t} \sum_{j=1}^{k} ijN_{ij} - \frac{bk(t + 1)(k + 1)}{4} \right\} \\ &= (L - \mu_L)/\sigma_L, \end{aligned}$$

using bk = rt again. When k = t the statistic U_{11} reduces to the well known Page statistic. Schach (1979) mentioned the existence of a statistic like this but did not give a formula for its calculation.

It follows that the $U_{\ell m}$ are extensions of this Page statistic U_{11}. Again U_{21} is of interest to test the null hypothesis of no treatment differences against the umbrella alternative that as we pass through the rows (treatments) the columns either increase and then decrease (one tail) or decrease and then increase (the other tail).

Dried Egg Example. Consider again this example from section 7.1. Suppose that *a priori* product A is ranked first, product B second, and so on. Then, for these data, L = 988.0, μ_L = 825.0 and σ_L^2 = 687.5, giving U_{11}

= 6.22 with a one-sided p-value of 0.000. Clearly there is a significant ordering of treatments A, B, ... , J.

Mouthfeel Thickness Example. Four unflavoured corn syrup blends, W, X, Y, Z say, were compared for *mouthfeel thickness*. Twelve judges evaluated each of the six pairs WX, WY, WZ, XY, XZ, YZ. Table 7.7 shows the number of times (out of 12) each 'row' blend was chosen as being thicker than each 'column' blend.

If a rank of one is assigned to the thicker blend and a rank of two to the thinner, the table of counts $\{N_{ij}\}$ is as given by Table 7.8. To obtain this table note that the number of ranks of 1 for X is the sum of the X row of Table 7.7, while the number of ranks of 2 for X is the sum of the X column. From Table 7.8, $R_1 = R_W = 71$, $R_2 = R_X = 52$, $R_3 = R_Y = 48$ and $R_4 = R_Z = 45$. As $k = 2$, $(t - 1)X_P^2/t = D_1$. We find $D_1 = 34.17$ and $U_{11}^2 = 28.02$, both with p-values less than 0.001. Meilgaard et al. (1999) also gave this D_1 value but not the U_{11}^2 value. Clearly there are very highly significant differences in mean ranks for the products and there is a very highly significant ordering W > X > Y > Z.

Table 7.7 Number of times, out of 12, that one of a pair of blends was chosen as being thicker than the other blend in the pair (data from Meilgaard et al., 1999)

	Thinner			
Thicker	W	X	Y	Z
W	-	0	1	0
X	12	-	6	2
Y	11	6	-	7
Z	12	10	5	-

Table 7.8 Counts of $\{N_{ij}\}$ for Table 7.7 data

	Blend			
Rank	W	X	Y	Z
1	1	20	24	27
2	35	16	12	9

The analysis given here is as useful and much simpler than the Bradley-Terry model approach described, for example, in Gacula and Singh (1984, Chapter 11), or in Agresti (1990, Chapter 10). Tests derived from the Bradley-Terry model give very similar p-values for the above and other examples we have considered.

7.5 Paired Comparisons with a 2^n Factorial Structure

Paired comparisons are an important taste-test application involving ranked data and balanced incomplete blocks. They are suitable for consumer work as they require almost no training or experience. In this section we develop general ideas through the particular example of the Meilgaard et al. (1999) mouthfeel thickness example. Suppose in that example the products W and X had a high concentration of corn syrup while products Y and Z had a low concentration. Further, suppose products Y and W had a high level of thickening agent while products X and Z had a low level. It could be said that a 2^2 factorial structure exists for products. Does corn syrup concentration or thickening agent alter the mouthfeel thickness? To examine this question we need to calculate product effects v_i, say, given by

$$v_i = (N_{i2} - N_{i1})\sqrt{\frac{t - 1}{N}},\ i = 1, \ldots, t$$

in which $N = \sum_{i=1}^{t} (N_{i1} + N_{i2})$, the sum of the counts in the table of N_{ij} values.

It can be shown that

$$(t - 1)X_P^2/t = D_1 = \sum_{i=1}^{t} v_i^2 \text{ and } \sum_{i=1}^{t} v_i = 0.$$

The sum of squares, $(t - 1)X_P^2/t$, can be partitioned into effects due to each factor. It can also be shown that each term in the partition has an approximate χ_1^2 distribution. For the corn syrup example we find

$v_1 = 34/c$, $v_2 = -4/c$, $v_3 = -12/c$ and $v_4 = -18/c$, in which $c = 4\sqrt{3}$.

Thus $(t - 1)X_P^2/t = 34.17$ and this value can be partitioned as in Table 7.9 using $(v_1 + v_2 - v_3 - v_4)^2/4$ as the corn sum of squares, $(v_1 - v_2 + v_3 - v_4)^2/4$ as the agent sum of squares, with the interaction sum of squares being found by difference.

To demonstrate the latter note that the V_i correspond to the d_i of David (1988, p. 34), who showed that, in the general case, $V = (V_i)$ is asymptotically $N_t(0, \Sigma)$, with $\Sigma = I_t - \frac{1}{t}1_t1_t^T$, where I_t is the t by t identity matrix and 1_t is the t by 1 vector of ones. It is routine to show that 0 is a
eigenvalue of Σ with corresponding eigenvector 1_t, and that 1 is a eigenvalue with multiplicity t - 1; the eigenvectors corresponding to the eigenvalues 1 are orthonormal to 1_t.

Now construct H to be orthogonal with first row $1_t^T/\sqrt{t}$. The remaining rows are orthonormal to this row. It follows that

$$HH^T = I_t \text{ and } H\Sigma H^T = 0 \oplus I_{t-1}.$$

Put $Y = HV$. Then asymptotically Y is $N_t(0, H\Sigma H^T = 0 \oplus I_{t-1})$. This means

that Y_1 is degenerate, but the remaining Y_i are asymptotically independent and asymptotically standard normal.

With t = 4 a possible choice of H is

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} /2.$$

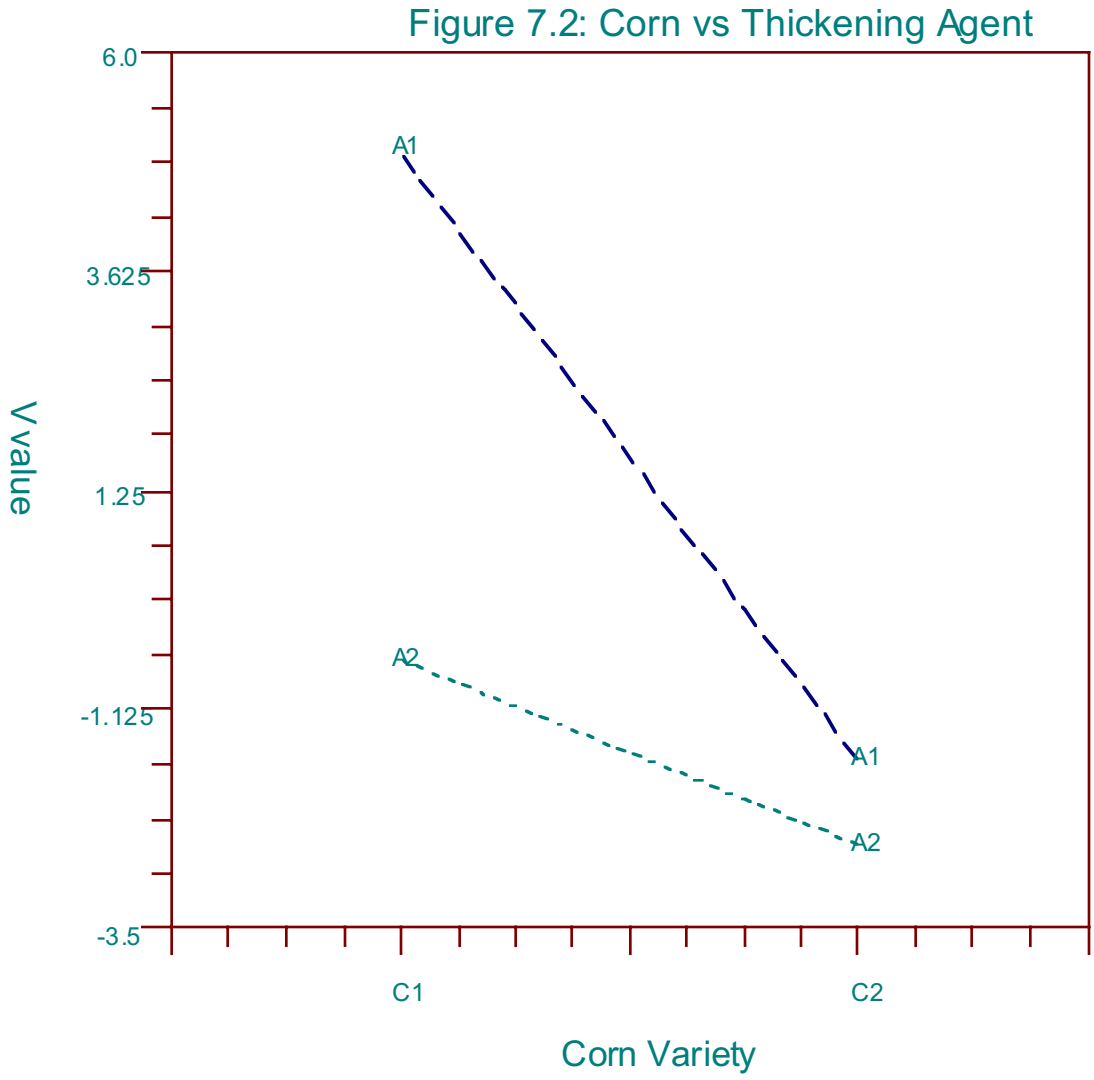
Then $Y_2 = (V_1 + V_2 - V_3 - V_4)/2$ and $Y_3 = (V_1 - V_2 + V_3 - V_4)/2$ are asymptotically independent and asymptotically standard normal. Their squares are hence approximately independent χ^2_1 random variables. Moreover

$$(t - 1)X^2_P/t = V_1^2 + V_2^2 + V_3^2 + V_4^2 = V^TV = Y^TY = Y_2^2 + Y_3^2 + Y_4^2$$

as suggested. In fact the interaction is simply $Y_4^2 = (V_1 - V_2 - V_3 + V_4)^2/4$.

Table 7.9 Partitioning $(t - 1)X^2_P/t$ for the mouthfeel thickness data with factorial structure

Source	Sums of Squares	df	χ^2 p-value
Corn	18.75	1	< 0.001
Agent	10.08	1	< 0.01
Interaction	5.33	1	< 0.05
$(t - 1)X^2_P/t$	34.17	3	< 0.001



For the mouthfeel thickness data there are differences due to both factors, corn concentration and thickening agent, and also a significant interaction between these factors. Details of the analysis are given in Table 7.9. The interaction is manifest by the difference between high and low levels of corn concentration being greater at the higher level of thickening agent than at the lower level of thickening agent. See Figure 7.2. The 2^2 factorial analysis here is very much simpler than that outlined by Bradley (1984).

8
Extensions to a Nonparametric Correlation Test: Spearman's Test

8.1 Introduction

The bulk of this chapter is based on Rayner and Best (1996a) and Best and Rayner (1996), and considers doubly ordered two-way contingency tables of counts N_{ij}, for i = 1, ... , r and j = 1, ... , c. Initially no row or column totals are assumed to be fixed. In this setting we are interested in Pearson's product moment correlation for grouped data r_P. This correlation is generally thought of as a measure of both independence and linearity between these variables; but see section 10.3. Ultimately we want to consider at least the possibility of more than two random variables. It is then sensible to think of r_P as just one of many possible measures of association. In the two-way case, r_P reflects the lowest order bivariate moment, assessing how the data differ from what might be expected under the null, independence model in the $(1, 1)th$ moment. Other bivariate moment-based measures of association - or generalised correlations - are available, are readily and practically interpreted, and these may be extended to multi-way tables.

To develop these moment-based measures of association, we first develop smooth models similar to those introduced in section 4.4 for one-way layouts, and similar to other models that have proved useful for assessing goodness of fit. The smooth omnibus score test for assessing independence is based on Pearson's X^2_P statistic. If the sample size is n, and if natural scores are used, then it will subsequently be shown that the $(1, 1)th$ component of X^2_P is $r_P\sqrt{n}$. The $(1, 2)th$ and $(2, 1)th$ components can be considered to assess bivariate skewness: assessing if these third order moments are consistent what might be expected under the independence model. Similarly the $(3, 1)th$, $(2, 2)th$ and $(1, 3)th$

components can be considered to assess bivariate kurtosis. If the categorical variable values are replaced by their ranks, the omnibus test is still based on X_P^2. If Spearman's rho is written r_S, and if there are no ties, the (1, 1)*th* component of X_P^2 is $r_S\sqrt{n}$. Again the (r, s)*th* component may be interpreted as assessing deviations up to the (r, s)*th* bivariate moment in the data from what might be expected under the independence model.

Intelligence Example. To demonstrate the efficacy of our approach consider the following synthetic data of Mack and Wolfe (1981), given in Table 8.1. Three males in each of five age groups were given a standard intelligence test. From these data we can construct the 5 by 15 table of zeros and ones relating age groups and intelligence ranks, given in Table 8.2.

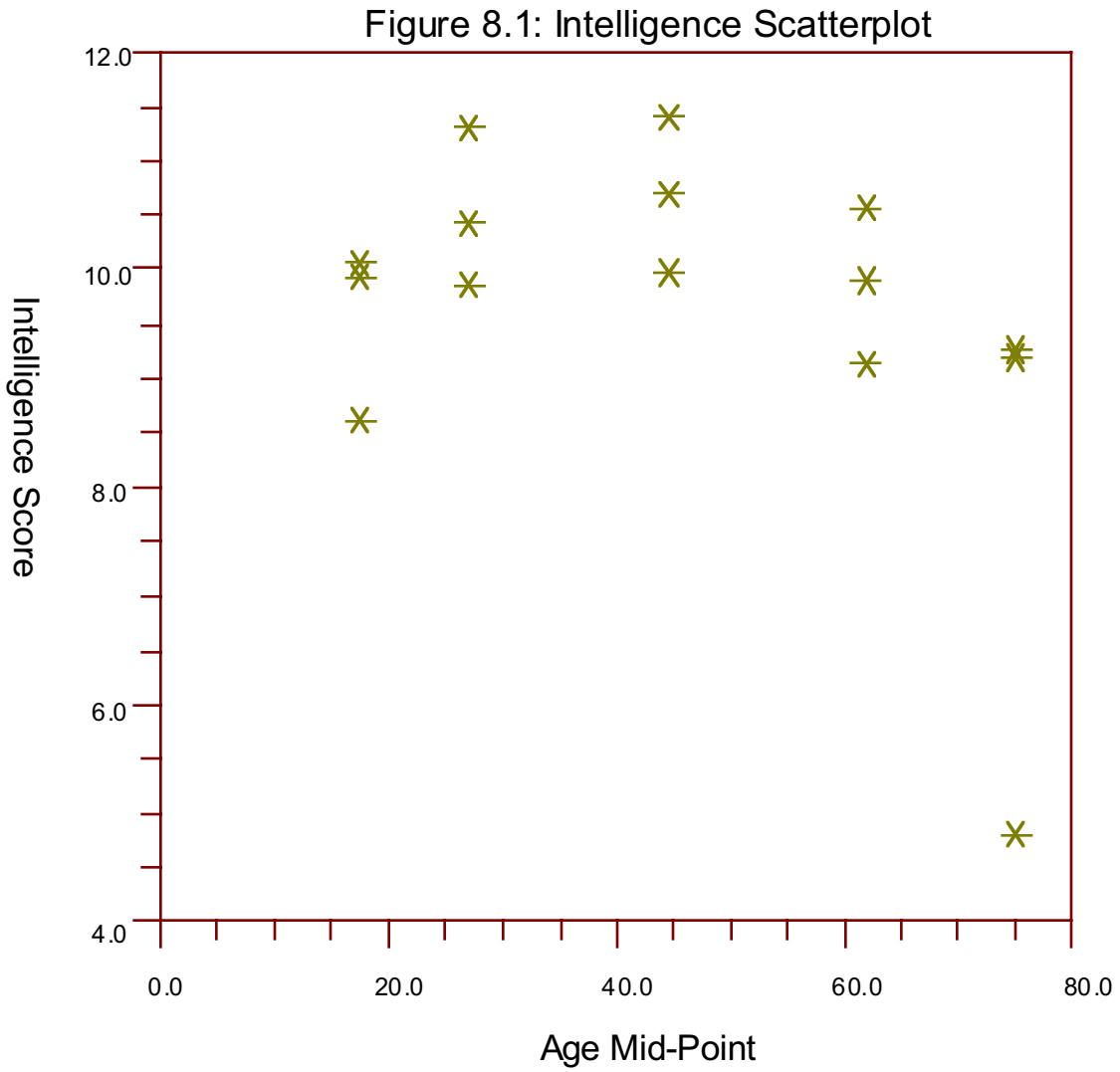
Table 8.1 Wechsler adult intelligence scale scores

Age Group				
16-19	20-34	35-54	55-69	69 plus
8.62	9.85	9.98	9.12	4.80
9.94	10.43	10.69	9.89	9.18
10.06	11.31	11.40	10.57	9.27

Table 8.2 Age groups and intelligence ranks for intelligence scale data

Age Group	Intelligence ranks 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16-19	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0
20-34	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0
35-54	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1
55-69	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
69 plus	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0

A Kruskal-Wallis test, which does not take into account the doubly ordered nature of the table of zeros and ones, gives for these data a Monte Carlo p-value, based on 10,000 simulations, of 0.104. The χ^2 p-values cannot be relied upon because of the sparseness of the table; so throughout this example we will calculate permutation test Monte Carlo p-values based on 10,000 simulations. We calculate r_S to be -0.3164, with Monte Carlo p-value 0.257; independence would seem to be an acceptable model. We could also have checked independence using Jonckheree's nonparametric test described, for example, by Sprent (1998, pp. 191-193).



It may be expected that intelligence will increase and then decrease with age. This is reflected in the data, for there are certainly higher scores for the middle age groups. A scatterplot that captures this effect is given in Figure 8.1. The expected relationship is not monotonic; so a directional test based on r_S will not have good power in detecting it. The skewness or umbrella test components of X_P^2 are approximately standard normally distributed, and yield values 0.672 and 2.393. The corresponding Monte Carlo p-values are 0.538 and 0.019 respectively. The second statistic here confirms the alternative hypothesis to independence suggested above. It is similar to Mack and Wolfe's (1981) umbrella statistic, which reached the

same conclusion. We discussed umbrella tests for two-way layout data in section 6.5.

More examples are given in sections 8.4 and 8.6.

8.2 A Smooth Model and Tests for Independence

We test for independence of the cell probabilities p_{ij} by first setting, for i = 1, ... , r and j = 1, ... , c,

$$p_{ij} = \left\{ 1 + \sum_{u=1}^{k_1} \sum_{v=1}^{k_2} \theta_{uv}\, g_u(i)\, h_v(j) \right\} p_{i.}\, p_{.j}. \qquad (8.1)$$

Here the cell probabilities sum to one: $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{c} p_{.j} = 1$, and the θ_{uv} are real valued parameters. The $g_u(i)$ are orthonormal functions on the marginal row probabilities $p_{i.} = p_{i1} + ... + p_{ic}$ and the $h_v(j)$ are orthonormal functions on the marginal column probabilities $p_{.j} = p_{1j} + ... + p_{rj}$. The subsequent derivations do not rely on the choice of orthonormal function. Usually k_1 and k_2 would each be chosen to be at most four, and more usually two. In the goodness of fit context (see Rayner and Best, 1989a, p. 7) we call the k's the *order* of the model. We have $k_1 \leq r - 1$ and $k_2 \leq c - 1$. When $k_1 = r - 1$ and $k_2 = c - 1$ the model becomes saturated, and an identity results. Further discussion of the model (8.1) is given in section 8.5.

One possible choice for the $\{g_u(i)\}$ is the set of orthogonal polynomials given in Appendix A.3. Subsequent $g_u(i)$ can be defined using the recurrence relations in Emerson (1968).

Whatever the choice of $\{g_u(i)\}$ and $\{h_v(j)\}$, in Theorems 8.1 and 8.2 the $p_{i.}$ are estimated by $\hat{p}_{i.} = N_{i.}/n_{..}$, giving the set $\{\hat{g}_u(i)\}$, and the $p_{.j}$ are estimated by $\hat{p}_{.j} = N_{.j}/n_{..}$, giving the set $\{\hat{h}_v(j)\}$.

Two theorems are now given. For convenience put $t = (u - 1)k_2 + v$, so that θ_{uv} becomes θ_t. Now write $\theta = (\theta_t)$. To test for independence, we test H_0: $\theta = 0$ against K: $\theta \neq 0$ with $p_{1.}, ... , p_{(r-1).}, p_{.1}, ... , p_{.(c-1)}$ as

nuisance parameters; $p_{r.}$ and $p_{.c}$ are omitted from the set of nuisance parameters because the constraints $p_{1.} + ... + p_{r.} = 1$ and $p_{.1} + ... + p_{.c} = 1$ show that $p_{r.}$ and $p_{.c}$ can be considered as algebraically dependent variables. First write

$$\widehat{V}_t = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \, \widehat{g}_u(i) \, \widehat{h}_v(j)/\sqrt{n_{..}} \text{ for } u = 1, ... , k_1, \text{ and } v = 1, ... , k_2.$$

Theorem 8.1. The score statistic for the model (8.1) is $\widehat{V}_1^2 + ... + \widehat{V}_{k_1k_2}^2$. Under the null hypothesis the $\widehat{V}_t$ are asymptotically independent and asymptotically have the standard normal distribution.

Proof. Suppose θ_a and p_b are typical elements of θ and $p = (p_{1.}, ... , p_{(r-1).}, p_{.1}, ... , p_{.(c-1)})^T$ respectively, and that the logarithm of the likelihood for our model is ℓ. As usual, the efficient score is defined by $U = (\partial\ell/\partial\theta_a)$, and the asymptotic covariance matrix by

$$\Sigma = I_{\theta\theta} - I_{\theta p} I_{pp}^{-1} I_{p\theta},$$

in which

$$I_{\theta\theta} = (- E[\partial^2\ell/\partial\theta_a\partial\theta_b]), \; I_{\theta p} = (- E[\partial^2\ell/\partial\theta_a\partial p_b]),$$
$$I_{p\theta} = I_{\theta p}^T, \text{ and } I_{pp} = (- E[\partial^2\ell/\partial p_a\partial p_b]).$$

See A.5.2. The score statistic is of the form $U_0^T\Sigma_0^{-1}U_0$, in which the subscript zero indicates evaluation under the null hypothesis.

If sample values of the N_{ij} are written n_{ij}, the logarithm of the likelihood for our model is

$$\ell = \text{constant} + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log p_{ij}.$$

Routine differentiation leads to $U_t = \sum_i \sum_j N_{ij} \, \widehat{g}_u(i) \, \widehat{h}_v(j)$, where the hats indicate that the orthonormal polynomials use the maximum likelihood estimates of the marginal probabilities, namely $\{N_{i.}/n_{..}\}$ and $\{N_{.j}/n_{..}\}$.

Then $U_0 = (U_t)$.

Elements of the information matrix are obtained by taking minus the expectations of the second order derivatives and evaluating under the null hypothesis. We find

$$
\begin{aligned}
&- E[\partial^2 \ell / \partial\theta_t \partial\theta_{t'}] \mid_{\theta=0} = n_{..}\, \delta_{tt'},\\
&- E[\partial^2 \ell / \partial\theta_t \partial p_{d.}] \mid_{\theta=0} = E[\partial^2 \ell / \partial\theta_t \partial p_{.a}] \mid_{\theta=0} = 0,\\
&- E[\partial^2 \ell / \partial p_{a.} \partial p_{b.}] \mid_{\theta=0} = n_{..}\delta_{ab}/p_{a.} + n_{..}/p_{r.},\\
&- E[\partial^2 \ell / \partial p_{.d} \partial p_{.e}] \mid_{\theta=0} = n_{..}\delta_{de}/p_{.d} + n_{..}/p_{.c}, \text{ and}\\
&- E[\partial^2 \ell / \partial p_{a.} \partial p_{.d}] \mid_{\theta=0} = 0.
\end{aligned}
$$

It now follows that the information matrix is the direct sum of $n_{..}$ times the k_1k_2 unit matrix $I_{k_1k_2}$, $n_{..}\mathrm{diag}(p_{r.}^{-1}) + n_{..}/p_{r.}\, J_{r-1}$ and $n_{..}\mathrm{diag}(p_{.a}^{-1}) + n_{..}/p_{.c}\, J_{c-1}$, in which J_n is the n by n matrix with every element 1. Hence Σ_0^{-1} is $I_{k_1k_2}/n_{..}$. The elements of U_0, being uncorrelated and asymptotically normal, are asymptotically independent. The results are more conveniently expressed in terms of $\widehat{V}_t = U_t/\sqrt{n_{..}}$. This gives Theorem 8.1.

Theorem 8.2. The score statistic, based on the saturated model (8.1) with $k_1 = r - 1$ and $k_2 = c - 1$, is Pearson's statistic,

$$X_P^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (N_{ij} - E_{ij})^2 / E_{ij},$$

in which, as usual,

$$E_{ij} = n_{..}\, \widehat{p}_{i.}\, \widehat{p}_{.j} = N_{i.}\, N_{.j}\, / n_{..}\, .$$

Proof. If we consider the table $\{N_{ij}/n_{..}\}$, then it follows from the definition of ϕ^2 (Lancaster 1969, p. 91), after a little arithmetic, that $n_{..}\, \phi^2 = X_P^2$. From the definition of the coefficients of correlation, ρ_t in (Lancaster

1969, Theorem 2.1), we have $\rho_t = \hat{V}_t / \sqrt{n_{..}}$ for t = 1, ... , (r - 1)(c - 1). Now from Parseval's equality (Lancaster 1969, Theorem 2.2)

$$\phi^2 = \sum_{t=1}^{(r-1)(c-1)} \rho_t^2 = \sum_{t=1}^{(r-1)(c-1)} \hat{V}_t^2 / n_{..} = X_P^2 / n_{..}.$$

This gives Theorem 8.2.

The components $\hat{V}_t$ partition X_P^2 in that the sum of their squares is X_P^2, and these components are asymptotically independent and asymptotically standard normal variables under the independence hypothesis. Each $\hat{V}_t^2$ could be derived as the score statistic for the following model: for j = 1, ... , c and i = 1, ... , r,

$$p_{ij} = \{1 + \theta_{uv}\, g_u(i)\, h_v(j)\}\, p_{i.}\, p_{.j}\ .$$

It follows that each of the (r - 1)(c - 1) components in X_P^2 forms the basis for a strongly directional test. This accounts for every degree of freedom. One practical option is to assess lack of independence by scrutinising all the components in a data analytic fashion.

Suppose now we have data with no ties, as, for example, in Table 8.2. Our contingency table consists of r rows with row totals $n_{1.}$, ... , $n_{r.}$, and $c = n_{..} = n_{1.} + ... + n_{r.}$ columns each with column total one. Now $E_{ij} = N_{i.} / n_{..}$ and

$$X_P^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (N_{ij} - E_{ij})^2 / E_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}^2 / E_{ij} - n_{..}.$$

Since each column has precisely one observation one with the rest being zero,

$$X_P^2 = \sum_{i=1}^{r} N_{ij}^2 / E_{ij} - n_{..} = \sum_{i=1}^{r} n_{..} / n_{i.} - n_{..} = n_{..} \{ n_{1.}^{-1} + ... + n_{r.}^{-1} - 1 \}.$$

As with the Kruskal-Wallis case in section 3.4, X_P^2 is not a suitable test

statistic since its value is independent of the data. However Theorem 8.2 remains true: the components partition X^2_P. However the model on which our derivations were based is no longer true. Resampling p-values may be calculated for the test statistics, but we may no longer be assured they will have χ^2 distributions asymptotically, nor that they will be asymptotically independent. This situation will arise when we decide, before sighting the data, that there can be no tied ranks.

8.3 Smooth Extensions

To demonstrate the relationship between the first component of X^2_P with Pearson's product moment correlation for grouped data, r_P, take $\{g_u(i)\}$ to be the orthogonal polynomials on $\{p_{i.}\}$, and $\{h_v(j)\}$ to be the orthogonal polynomials on $\{p_{.j}\}$. Now $g_0(i) = 1$ for all i, and $g_1(i) = (i - \mu_X)/\sigma_X$, where $\mu_X = \sum_{i=1}^{r} i p_{i.}$ and $\sigma^2_X = \sum_{i=1}^{r} i^2 p_{i.} - \mu^2_X$. Similarly $h_1(j) = (j - \mu_Y)/\sigma_Y$, and the first component is

$$V_1 = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} (i - \mu_X)(j - \mu_Y) / (\sigma_X \sigma_Y \sqrt{n_{..}}).$$

If we replace $p_{i.}$ by $N_{i.}/n_{..}$ and $p_{.j}$ by $N_{.j}/n_{..}$ we obtain $\widehat{V}_1$:

$$\widehat{V}_1 = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \widehat{g}_1(i) \widehat{h}_1(j)/\sqrt{n_{..}},$$

where the hats indicate that in g_1 the $p_{i.}$ have been replaced by $N_{i.}/n_{..}$, and in h_1 the $p_{.j}$ have been replaced by $N_{.j}/n_{..}$. We recognise $\widehat{V}_1$ as $r_P \sqrt{n_{..}}$. As we have previously indicated, the higher order components are extensions of those of low order, and hence of Pearson's product moment correlation.

A parallel development is possible for the corresponding situation in

which ranks replace the scores. Suppose we have n bivariate data points $(x_1, y_1), \ldots, (x_n, y_n)$, and we wish to test if the X and Y classifications are independent. A non-parametric measure of this independence is Spearman's rho. The X's are replaced by their ranks, $R_1, \ldots, R_n$, and similarly the Y's are replaced by their ranks, $S_1, \ldots, S_n$. Spearman's rho is the correlation between the R_i and the S_j:

$$r_S = 12 \sum_{a=1}^{n} \left[R_a - \frac{n+1}{2}\right]\left[S_a - \frac{n+1}{2}\right] / \left[n(n^2 - 1)\right].$$

We assume there are no ties. Form an n by n table of counts $\{N_{ij}\}$ with $N_{ij} = 1$ if (X_a, Y_a) has rank $(X_a) = i$ and rank $(Y_a) = j$ and $N_{ij} = 0$ otherwise. We now show that $\hat{V}_1 = \sqrt{n_{..}}\, r_S$. In $\hat{g}_1$ and $\hat{h}_1$ we need the mean and variance of the Xs and Ys. Now

$$\hat{\mu}_X = \sum_{i=1}^{r} iN_{i.}/n_{..} = \sum_{a=1}^{n} R_a/n_{..} = (n_{..} + 1)/2 = \hat{\mu}_Y, \text{ and}$$

$$\hat{\sigma}_X^2 = \sum_{i=1}^{r} i^2 N_{i.}/n_{..} - \hat{\mu}_X^2 = \sum_{a=1}^{n} R_a^2/n_{..} - \hat{\mu}_X^2 = (n_{..}^2 - 1)/12 = \hat{\sigma}_Y^2.$$

Similarly $\sum ij\ N_{ij} = \sum R_a S_a$, so that

$$\begin{aligned} n_{..}\,[(n_{..}^2 - 1)/12]\, r_S &= \sum_{a=1}^{n} [R_a - (n_{..} + 1)/2][S_a - (n_{..} + 1)/2] \\ &= \sum_{a=1}^{n} R_a\ S_a - (n_{..} + 1)^2/4 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} ij\ N_{ij} - \hat{\mu}_X\ \hat{\mu}_Y \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\,(i - \hat{\mu}_X)\ \ (j - \hat{\mu}_Y). \end{aligned}$$

Thus

$$\sqrt{n_{..}}\, r_S = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}\, \hat{g}_1(i)\, \hat{h}_1(j) / \sqrt{n_{..}} = \hat{V}_1, \text{ as promised.}$$

If there are ties, then $\hat{V}_1$ gives a generalisation of Spearman's rho. If, on the other hand, there are no ties, then the comments at the conclusion of the previous section apply.

8.4 Interpretation of the Components

For the remainder of this chapter we will write $\hat{V}_{uv}$ instead of $\hat{V}_t$, to more clearly identify the bivariate moment being assessed. As in Rayner and Best (1989a), it may be shown that for $\theta_{uv} = O(n^{-0.5})$, $\hat{V}_{uv}$ is asymptotically $N(\theta_{uv}, 1)$. Since g_u and h_v are polynomials, $E[\hat{V}_{uv}]$ involves bivariate moments up to the (u, v)*th*. Thus in large samples, $\hat{V}_{uv}$ reflects how the (u, v)*th* bivariate moment of the distribution defined on $\{p_{i.}p_{.j}\}$ differs from that defined on (8.1). Beh and Davy (1998, 1999) would describe $\hat{V}_{12}$ as reflecting row location and column dispersion effects. See section 8.6.

This *diagnostic* use of the components has recently been cast in doubt. See Horswell and Looney (1993) and Rayner, Best and Mathews (1995), where the focus is on goodness of fit testing, but the conclusions apply equally here. A diagnostic interpretation would occur when a particular bivariate component, $\hat{V}_{uv}$ say, was found to be significantly large, and was interpreted to mean that the independence model failed because the (u, v)*th* bivariate moment differed from what could be expected under independence. The reason this may not be the correct interpretation is that variance of $\hat{V}_{uv}$ involves moments up to the (2u, 2v)*th*, and in small samples may be quite different from 1, especially if the assumption that $\theta_{uv} = O(n^{-0.5})$ is dubious. It is then possible for $E[\hat{V}_{uv}]$ to be zero and yet for $\hat{V}_{uv}$ to be significantly large. If $\hat{V}_{uv}$ is significantly large, it is certainly reasonable to claim that bivariate moments up to the (2u, 2v)*th* may not be consistent with what may be expected under independence. If looking at several components in a data analytic fashion,

it may be possible to isolate particular moments and identify them as the cause of the model failure.

In the goodness of fit context, Henze, in a series of articles, has converted certain component test statistics so that they can be interpreted diagnostically. There is sometimes a cost of loss of power relative to the unconverted components. This will be discussed further in the chapter on goodness of fit. We simply note here that a parallel conversion could be considered for the components considered in this chapter. It would then be sensible to conduct a power study to assess the cost of having this interpretability.

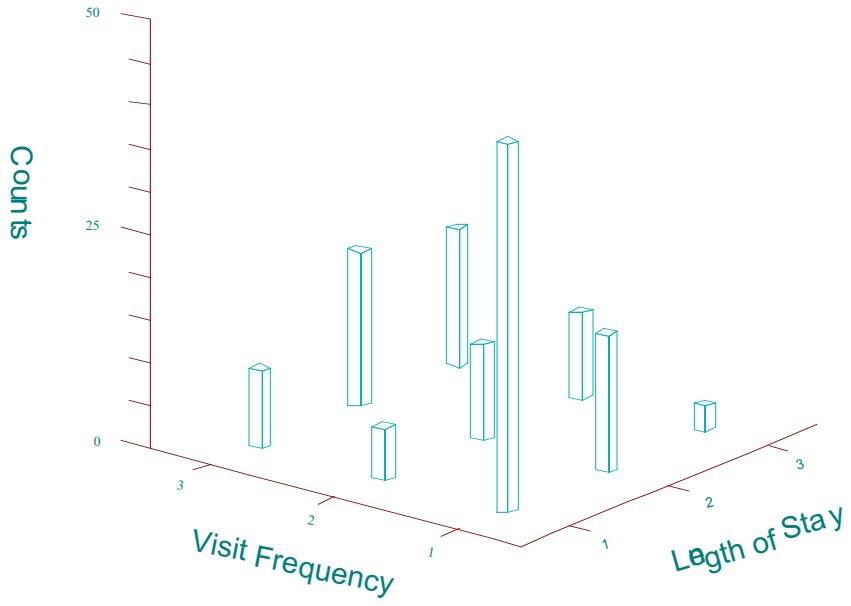
It is certainly true that X_P^2 is an omnibus test, assessing (r - 1)(c - 1) dimensions of the parameter space for a failure of the null hypothesis. Each component $\hat{V}_{uv}$ assesses a single dimension, and thus produces a more directional and more powerful test. Strictly the (u, v)*th* component should be interpreted as assessing deviations up to the (2u, 2v)*th* bivariate moment in the data from what might be expected under the independence model, but most data analysts should be, in the first instance, most suspicious of the (u, v)*th* bivariate moment.

Hospital Example. The example in section 8.1 deals with ranks and calculates components of X_P^2. The following example calculates components of X_P^2 when "natural" scores are used instead of ranks. This requires a routine generalisation of the derivation given here. The formulae are essentially the same but with the substitution of the generalised polynomials given in Appendix A.3. Lancaster (1953, p. 7) first gave a component analysis of this sort. The data in Table 8.3 below, given and analysed in Haberman (1974), relate the length of stay in hospital and the visit frequency for mental patients. See Figure 8.2. Since the cell counts are substantial we have used χ^2 rather than Monte Carlo p-values.

Table 8.3 Length of stay in hospital vs. visit frequency for mental patients

	Length of Stay (Days)		
Visit Frequency	≥ 2, < 10	≥ 10, < 20	≥ 20
Regular	43	16	3
Less than monthly	6	11	10
Never	9	18	16

Figure 8.2: Mental Patients' Blockplot



For these data X_P^2 = 35.2, with p-value less than 0.001. With four

degrees of freedom we can calculate up to four components, but which four is up to the statistician. We assign scores of 1, 2 and 3 for the classes in both the visit frequency and length of stay categorisations. Components $\hat{v}_{rs}$ of X^2_P with r, s = 1, 2 and their associated χ^2 p-values are: $\hat{v}_{11}$ = 5.4186 (< 0.001), $\hat{v}_{12}$ = -0.6872 (0.246), $\hat{v}_{21}$ = -2.301 (0.021) and $\hat{v}_{22}$ = 0.2122 (0.832). The highly significant (1, 1)*th* component suggests that the longer the length of stay in hospital, the less frequent the visits. The component $\hat{v}_{21}$ relates to the quadratic by linear association, or bivariate skewness. This suggests that as the length of stay increases, the variability of the visit frequency increases and then decreases. Referring to the data, for length of stay at least two but less than 10 days, the standard deviation is 0.750; for length of stay at least 10 and less than 20 days the standard deviation increases to 0.878; and for at least 20 days the standard deviation now reduces to 0.686. Observe that $\hat{V}_{11}$ is a linear by linear association statistic and that exact probabilities for $\hat{v}_{11}$ conditional on fixed margins are given in *StatXact* (1995).

As a matter of interest, the X^2 statistic for a parametric linear by linear log-linear association model has a value of 29.36, which agrees reasonably with $\hat{v}^2_{11}$ = 27.86. We have observed such good agreement in a number of similar data sets. Agresti (1984, section 5.1) discusses linear by linear log-linear association models. Beh and Davy (2000) show that if σ^2_1 is μ_2 in Appendix A.3 defined on the columns of Table 8.3 and if σ^2_2 is μ_2 in Appendix A.3 defined on the rows, then $\hat{V}_{11}/(\sigma_1\sigma_2)$ approximates the log-linear model parameter for linear by linear association. We suspect this approximation is better when $\hat{v}_{11}$ is not too large.

The next example shows the use of monotonic scorings and how our methods work with sparse tables.

Whiskey Grading Example. O'Mahony (1986, p. 363) had an imaginary whiskey expert grade eight whiskies and obtained the table below.

Using mid-rank scores we have $\{x_1, x_2, x_3\} = \{2, 4.5, 7\}$ and $\{y_1, y_2, y_3\} = \{2, 5, 7.5\}$. This gives r_S = 0.73 and an associated two-sided Monte Carlo p-value of 0.09, again based on 10,000 simulations. There appears to be some evidence of a correlation between maturity and grade. With

small counts as in this table it is important to use Monte Carlo p-values. The asymptotic p-value of 0.04 falsely indicates a more significant correlation.

A bootstrap standard error of r_S, based on 200 bootstrap samples, was calculated as in Efron and Tibshirani (1993, algorithm 6.1) and found to be 0.20. Thus r_S is large in comparison with its standard error and this again is some evidence of a correlation between maturity and grade.

Table 8.4 Cross-classification of whiskey for age and grade

	Grade		
Years/Maturity	First	Second	Third
7	2	1	0
5	1	1	1
1	0	0	2

As a matter of interest, the X^2 test statistic for a parametric linear by linear log-linear association model has a value of 4.35, which agrees reasonably with $\hat{v}_{11}^2 = 4.30$. Similar agreement was also seen in the previous example. Of course, with data as sparse as these, the usual asymptotic standard errors associated with log-linear models are likely to be incorrect. Use of r_S and a bootstrap standard error is likely to be a better option for describing association.

8.5 Discussion

The theory here is based on what, in sections 8.2 and 8.3, we called a smooth model. The model may be thought of as distribution free, but is certainly parametric. Related models have proved very successful in constructing tests of goodness of fit, for which see Chapter 9, and extensions of Yates' test, for which see Chapter 4. The low order tests

based on these smooth models agree with well known and established tests, while the extensions to higher order tests are useful for data analysis.

The same tests can be derived by looking at the vector of counts given in lexicographic order, using asymptotic normality, diagonalising the covariance matrix and constructing components. Nair (1987) used a similar technique. However the score tests have general optimality properties, and it was for this reason that we chose this approach.

Colleagues have commented that although models of the form (8.1) are well known to sometimes give excellent results, they can also produce negative probabilities. However, assuming the θ_{uv} are $O(n^{-0.5})$, (8.1) is asymptotically equivalent to

$$p_{ij} = C(\theta)\, p_{i.}\, p_{.j} \exp\left\{\sum_u \sum_v \theta_{uv}\, g_u(i)\, h_v(j)\right\}$$

where the summations are for u = 0, 1, ... , k_1, v = 0, 1, ... , k_2, but not (u, v) = (0, 0). This model of course cannot produce negative probabilities. Score tests could have been derived for both models; (8.1) was chosen for its simplicity. Whatever the asymptotic optimality probabilities of the extensions, they should be judged on their practicality, convenience, and small sample properties. Thus, although our tests were derived using a model that can result in negative probabilities, this does not imply that poor tests have resulted.

Our experience in goodness of fit testing suggests that only the first few components will be important. We therefore do not recommend calculation of the extensions beyond those assessing skewness and kurtosis.

8.6 Multi-Way Tables

Beh and Davy (1998) extended some of the considerations of the previous sections to higher multi-way tables, mainly through consideration of three-way tables. Their concern was more with data analysis rather than generalising and extending nonparametric analysis. We include a brief

outline of their approach to three-way tables. We first indicate that although our discussion has been restricted to two-way tables it generalises readily to higher order tables, and second, that although we have not always focused on data analysis with our decomposition of X_P^2 approach, it can be profitable to do so.

Suppose all margins are completely ordered and only the grand total of the observations is known before collecting the data. If we have counts N_{ijk} and corresponding cell expectations E_{ijk} on IJK cells, then Pearson's independence X_P^2 is defined, as usual, by

$$X_P^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (N_{ijk} - E_{ijk})^2 / E_{ijk}.$$

The expectations E_{ijk} are given, again as usual, by the product of the row, column and tube totals divided by the square of the grand total of the observations, n. There are IJK - 1 - (I - 1) - (J - 1) - (K - 1) = IJK - I - J - K + 2 degrees of freedom, since I - 1 row marginal probabilities $p_{1..}, \ldots, p_{(I-1)..}$ must be estimated, with p_I found by difference, and similarly for columns and tubes. Beh and Davy (1998) showed that

$$X_P^2 = \sum_{u=0}^{I-1} \sum_{v=0}^{J-1} \sum_{w=0}^{K-1} Z_{uvw}^2$$

where

$$Z_{uvw} = \sqrt{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_u(i)\, b_v(j)\, c_w(k)\, p_{ijk}$$

in which p_{ijk} is estimated by N_{ijk}/n for all i, j and k, and $\{a_u(i)\}$, $\{b_v(j)\}$ and $\{c_w(k)\}$ are normalised orthogonal polynomials on $\{N_{i..}/n\}$, $\{N_{.j.}/n\}$ and $\{N_{..k}/n\}$ respectively.

There are apparently IJK components in all, but Z_{000} is identically $\sqrt{n}$, and $Z_{u00} = Z_{0v0} = Z_{00w} = 0$, for all u, v and w. The Z_{uv0} are the components of the X_P^2 obtained by summing over tubes in the original

table, and calculating components as in earlier sections of this chapter; similarly for Z_{u0w} and Z_{0vw}. The components Z_{uvw} with no u, v or w zero assess genuine trivariate association.

If we define

$$X^2_{UVW} = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} Z^2_{uvw}$$

$$X^2_{UV} = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} Z^2_{uv0}$$

$$X^2_{UW} = \sum_{u=1}^{I-1}\sum_{w=1}^{K-1} Z^2_{u0w} \text{ and}$$

$$X^2_{VW} = \sum_{v=1}^{J-1}\sum_{w=1}^{K-1} Z^2_{0vw},$$

then

$$X^2_P = X^2_{UVW} + X^2_{UV} + X^2_{UW} + X^2_{VW}.$$

Here X^2_{UV}, X^2_{UW} and X^2_{VW} are two-way X^2_P statistics corresponding to collapsing tubes, columns and rows respectively, and X^2_{UVW} assesses genuine trivariate association between rows, columns and tubes.

Some collections of components may have relevance to the analysis of a particular hypothesis. Thus $\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} Z^2_{1vw}$ assesses the effect of the row location component on the three-way association, and $\sum_{v=1}^{J-1} Z^2_{2v0}$ assesses the effect of the dispersion component on the two-way association between rows and columns.

To see the practical application of these ideas, an example from Beh and Davy (1998) is given.

Happiness Example. The data in Table 8.5 (reproduced, with permission, from Beh and Davy, 1998) relate the way 1517 people assess their

happiness (not too happy to very happy), number of completed years of schooling (four categories) and their number of siblings (five categories).

Using consecutive integer scores for all categories, Beh and Davy (1998) reported the permutation test p-values for X^2_{UV}, X^2_{UW} and X^2_{VW}, based on 10,000 simulations, of 0, 0 and 0.0008 respectively, while that associated with X^2_{UVW} is 0.3426. The overall X^2_P p-value is also zero. Three-way association does not contribute significantly to the overall significant association, but all the two-way associations do. Since X^2_{UV} was 72% of X^2_P, but only 24% of the degrees of freedom, the association between years of schooling and number of siblings is the most significant contributor to the overall association.

Table 8.5 Cross-classification of 1517 people according to happiness, schooling, and number of siblings

Years of Schooling	Number of Siblings 0-1	2-3	4-5	6-7	8+
			Not too happy		
< 12	15	34	36	22	61
12	31	60	46	25	26
13 - 16	35	45	30	13	8
17+	18	14	3	3	4
			Pretty happy		
< 12	17	53	70	67	79
12	60	96	45	40	31
13 - 16	63	74	39	24	7
17+	15	15	9	2	1
			Very happy		
< 12	7	20	23	16	36
12	5	12	11	12	7
13 - 16	5	10	4	4	3
17+	2	1	2	0	1

In X^2_{UV}, the column location component $\sum_{u=1}^{3} Z^2_{u10}$ has p-value zero. This suggests that, ignoring happiness, there are location differences in different rows (years of schooling), or the mean number of siblings is different for different years of schooling. As Beh and Davy (1998) say,

> "the difference in the Number of siblings is due to the difference at each level across Years of completed schooling when the happiness of people is not of interest."

Table 8.6 Partition of X^2_P statistic into component values

Term	Component	Value	df	p-value
X^2_{UV}	Column components			
	Location	222.2234	3	0
	Dispersion	7.7034	3	0.0528
	Residual	5.3720	6	0.5035
	Row components			
	Location	209.9878	4	0
	Dispersion	24.2943	4	0.0001
	Residual	1.0168	4	0.9156
		235.2988	12	0
X^2_{UW}	Tube components			
	Location	28.6582	3	0
	Dispersion = Residual	12.4805	3	0.0061
	Row components			
	Location	31.6954	2	0
	Dispersion	6.2973	2	0.0466
	Residual	3.1460	2	0.2127
		41.1387	6	0
X^2_{VW}	Tube components			
	Location	8.7582	4	0.0770
	Dispersion = Residual	17.0634	4	0.0009
	Column components			
	Location	18.0973	2	0.0001
	Dispersion	0.9724	2	0.6172
	Residual	6.7517	4	0.1492
		25.8215	8	0.0008
X^2_{UVW}	Tube components			
	Location	10.9158	12	0.5507
	Dispersion = Residual	15.3925	12	0.2250
	Column components			
	Location	4.0458	6	0.6683
	Dispersion	12.4921	6	0.0574
	Residual	9.7705	12	0.6389
	Row components			
	Location	4.2737	8	0.8360
	Dispersion	18.9646	8	0.2000
	Residual	3.0701	8	0.9245
		26.3084	24	0.3426
X^2_P		328.5674	50	0

The corresponding column dispersion component $\sum_{u=1}^{3} Z^2_{u20}$ has p-value 0.05, and the residual, $\sum_{u=1}^{3}\sum_{v=3}^{5} Z^2_{uv0}$, is not significant: there are no further column effects. In considering the row components, both location and dispersion have zero p-values, and the residual is again large. This suggests that, ignoring happiness, the mean and variance of years of schooling are different for different numbers of siblings. There are other conclusions that can be drawn from these data, but the above is sufficient to demonstrate the approach.

The Z_{uvw} are asymptotically standard normal, and can be readily interpreted. Thus $Z_{110} = -14.42$, suggesting that "those with many siblings tend to finish school earlier than those with few siblings", and $Z_{012} = 3.47$, suggesting that "as the number of siblings increases, happiness tends to decrease, then increase". A complete analysis of the data involves careful interpretation of all the components. These are given in Table 8.6, which is reproduced, with permission, from Beh and Davy (1998, Table 2).

Beh and Davy (1998) relate their model with the corresponding log-linear model. This enables them to find excellent direct approximations to the maximum likelihood estimates of the parameters of the log-linear model, direct in that no iteration is required to calculate the approximations. In addition they observe that their models have no difficulty with selection of the optimum model.

More discussion of three way tables is given in section 10.2.

9
One and S-Sample Smooth Tests of Goodness of Fit

9.1 Introduction

The following account incorporates and updates the review paper of Rayner and Best (1990a) on one-sample smooth goodness of fit testing. More detail on the earlier material is given in Rayner and Best (1989a). Recent work indicates that the Anderson-Darling test provides a good omnibus test. It can be used in conjunction with the components of the smooth tests. In this chapter we also extend the smooth approach to S-sample goodness of fit testing and demonstrate the power and flexibility of the smooth tests.

We have developed tests for many distributions, such as the discrete uniform, binomial, univariate and bivariate Poisson, univariate and multivariate normal. For each distribution studied small sample assessments of size and power are necessary. We usually find score tests for testing for each distribution. These are essentially omnibus tests, but their components provide useful and powerful directional tests. In general we recommend a data-analytic approach to testing: use the omnibus test to test for the distribution, and calculate p-values of the components to scrutinise, in a data-analytic fashion, how the data may deviate from the specified distribution in particular moment effects. A relatively recent development is this interpretation of the components, and we make some comments on this. The formulation of categorised smooth models leads to X^2 tests and their components. A generalisation of the smooth categorised model, when allied with Hall's (1985) idea of overlapping, leads to focused tests, and to an alternative to pooling.

The S-sample problem asks if S random samples may be regarded as coming from the same population. Our smooth formulation requires that a target distribution be specified. This is the distribution that, if the S

samples were found to be consistent, we would be most inclined to hypothesise as the common distribution. Having made this specification, it is possible to assess if the samples agree in regard to their location, dispersion, skewness and higher moments. If the samples are not consistent, an LSD analysis is available to detail the differences between the samples.

9.2 One-Sample Testing for Uncategorised Distributions

Suppose we wish to test for a probability density function $f_X(x;\beta)$, where $\beta = (\beta_1, \ldots, \beta_q)^T$ is a q by 1 vector of parameters (such as μ and σ in the normal case). This null probability density function is nested in the *order k* probability density function

$$C(\theta,\beta)\exp\{\sum_{i=1}^{k} \theta_i h_i(x;\beta)\} f_X(x;\beta), \qquad (9.1)$$

where $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of k real parameters, $C(\theta,\beta)$ is a normalising constant and $\{h_i(x;\beta)\}$ is a set of orthonormal functions on $f_X(x;\beta)$, so that for $f_X(x;\beta)$ continuous

$$\int_{-\infty}^{\infty} h_i(x;\beta)\, h_j(x;\beta)\, f_X(x;\beta)\, dx = \delta_{ij}.$$

The Legendre polynomials are orthonormal on the continuous uniform (0, 1) distribution, and the Hermite-Chebyshev polynomials (sometimes just called the Hermite polynomials) are orthonormal on the standard normal distribution. In each case the first few polynomials are given in Appendix A.3. See also Abramowitz and Stegun (1972).

Sometimes $C(\theta,\beta)$ will not exist but the statistic we are about to define will; see Kallenberg et al. (1997). Using a random sample $X_1, \ldots, X_n$, we test for $f_X(x;\beta)$ parametrically, by testing H_0: $\theta = 0$ against K: $\theta \neq 0$ with β a vector of nuisance parameters. For k small the alternatives vary

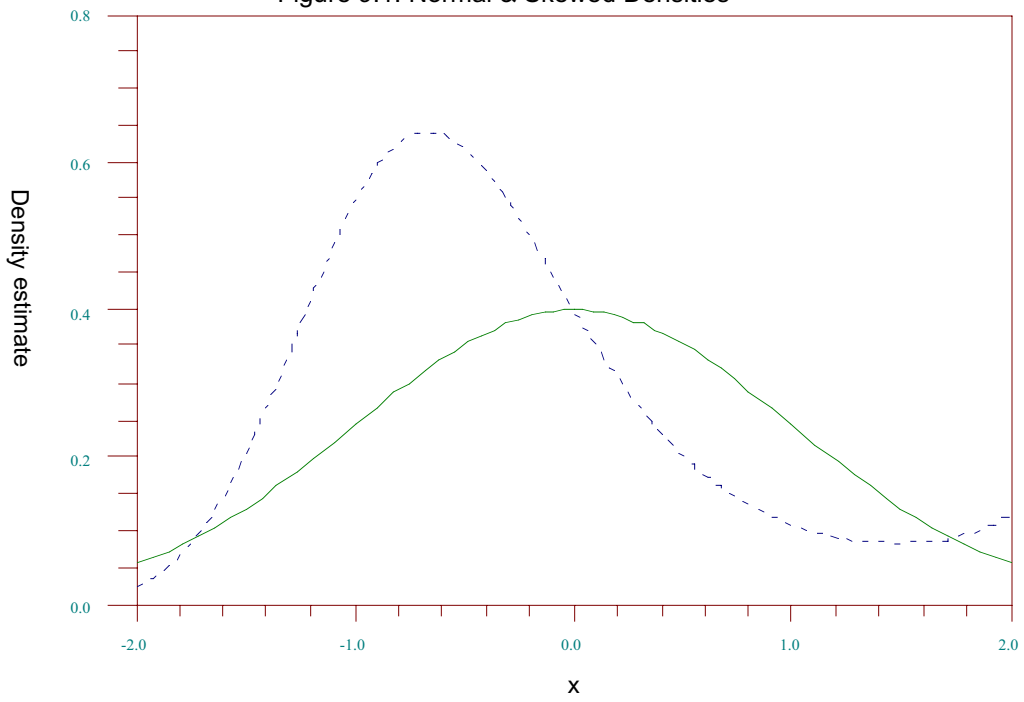
'smoothly' from the null, in the sense of Neyman (1937). As an example of this consider testing for the standard normal distribution, so that the mean and variance are specified, and need not be estimated. The order one alternative is a normal distribution with a mean shift, the order two alternative is a normal distribution with a simultaneous mean and variance shift, and the third order alternative is more involved, with skewness differences to what might be expected were the distribution normal. See Figure 9.1.

The score test is based on the statistic

$$\hat{S}_k = \sum_{i=1}^{k} \hat{V}_i^2 \text{ in which } \hat{V}_i = \sum_{j=1}^{n} h_i(X_j,\hat{\beta})/\sqrt{n},$$

where $\hat{\beta}$ is the maximum likelihood estimate of β. The $\hat{V}_i$ are components of the $\hat{S}_k$, and provide directional tests with high power in one direction and poor power elsewhere, complementing the omnibus tests based on the $\hat{S}_k$ which has moderate power in all directions. Since q elements of β must be estimated, we find in some important cases (but not always!) that q of the $\hat{V}_i$ are zero, and $\hat{S}_k$ has the χ^2_{k-q} distribution under both the null and contiguous alternatives. We usually redefine the test statistic, calling it $\hat{S}_{k-q}$, to emphasise the number of useful components. It can be shown the score test for testing H_{0i}: $\theta_i = 0$ against K_i: $\theta_i \neq 0$ is based on the statistic $\hat{V}_i$. Being score tests, these tests are weakly optimal. If there are no nuisance parameters, no estimation is required and the results indicated hold with q = 0 and the ^'s removed from the test statistics.

Figure 9.1: Normal & Skewed Densities



We emphasise that different $\{h_i(x;\beta)\}$ detect different alternatives using different tests. Which $\{h_i(x;\beta)\}$ should be chosen depends on which alternatives one hopes to most powerfully detect. If the distribution were uniform, as in Neyman's case, we could take the series $\{x^i\}$. The i^{th} component would still compare the data and hypothesised distribution moments up to the i^{th}. However we prefer, like Neyman, to use orthonormal functions because the components are simply defined and are, at least asymptotically, independent. If it were desirable to test for periodic alternatives we would consider using the series $\{\sqrt{2}\sin(i\pi x)\}$. For the "bump" alternatives to uniformity of Cressie (1978), we suggest the orthonormal series $\{\xi_i(x)\}$ given in Hamdan (1974). The aim is to find an orthonormal series that represents the alternatives of interest in as few terms as is possible. Greater power results from doing this.

Each family of distributions to be detected requires an independent study. Given the choice of orthonormal family, the small sample distributions of the test statistics and their components should be

investigated. Usually we use regression techniques to calculate corrections to the asymptotic null critical points. The score statistics for testing for some distributions, like the continuous uniform, approach their asymptotic distributions quite quickly, while those for other distributions, like the normal, approach their asymptotic distributions relatively slowly. See Carolan and Rayner (2000d). In general we recommend using bootstrap p-values. Those users for whom this is not an option should use the asymptotic corrections whenever practicable.

Again for each family of distributions we may compare, via simulation studies, the powers of the smooth tests we propose and the commonly used competitor tests. We give the general advice that in a testing context a powerful goodness of fit test is that based on the sum of the squares of the first two, three or four components. Another powerful omnibus test is that based on the Anderson-Darling statistic. However in particular cases there are variations to this advice. Thus in the Poisson case we find that the component based on the quadratic orthonormal function provides a powerful test against skewed alternatives. See Best and Rayner (1999a).

Over a period of many years we have studied the continuous and discrete uniform, univariate, bivariate and multivariate normal, exponential, geometric, binomial, univariate and bivariate Poisson. Details are in Rayner and Best (1989a) and papers detailed in the references. In all these cases the smooth tests are generally at least as powerful as the competitors studied, and the smooth tests have the added flexibility of several meaningful components being available. Thus if the null hypothesis is rejected, its failure can be attributed to particular θ_i being non-zero, and an alternative model is suggested, namely

$$C(\hat{\theta}, \hat{\beta}) \exp\{\sum_{i=1}^{k} \hat{\theta}_i h_i(x;\hat{\beta})\} f_X(x;\hat{\beta}),$$

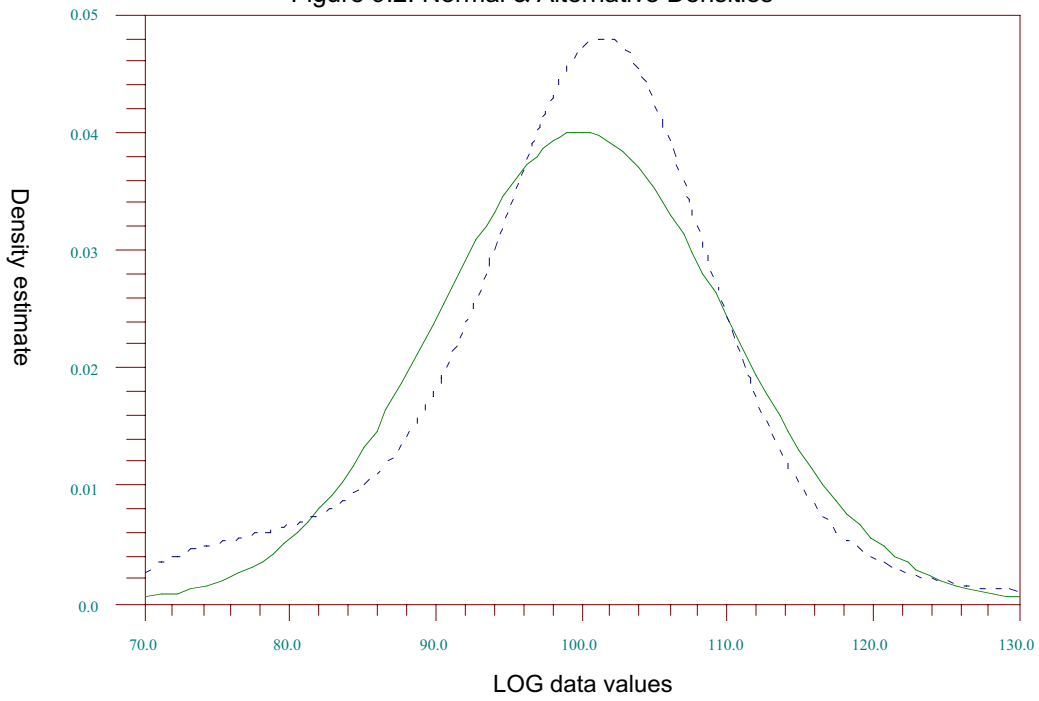
where the $\hat{\theta}_i = \hat{V}_i / \sqrt{n}$ are unbiased, consistent and efficient estimators of the θ_i. This density estimate is a fortuitous by-product of the testing, and may be used to graph the data. It is also related to the partially parametric approach described in section 10.4. In passing note that the application of smooth tests to distributions other than those listed above

may be less straightforward. For example see Boulerice and Ducharme (1995).

We began this section by assuming the need to test for a given probability density function. Of course not all statistical investigations involve hypothesis testing. In the assessment of any statistical model we recommend the calculation of the omnibus score statistic and as many components as is convenient and meaningful. The associated p-values will give a detailed and informative data-analytic insight into the behaviour of the data. However hypothesis testing is the usual context in which the methods described here arise.

LOG Example. D'Agostino and Stephens (1986, p. 532) gave the LOG data set. One analysis of it (D'Agostino and Stephens 1986, p. 75) used a Pearson type test to test for normality. With 25 approximately equiprobable classes using data dependent boundaries and estimating the parameters using the original uncategorised data, they find a p-value between 8.6% and 14.0%. The ambiguity is because the asymptotic distribution of the test statistic can only be bounded. This could be resolved by using resampling methods. In this case we find using a categorised test unnecessary, as the data are given to two decimals. Using the Hermite-Chebyshev polynomials (see A.3) we test normality using $\hat{S}_4 = \hat{V}_3^2 + \hat{V}_4^2 + \hat{V}_5^2 + \hat{V}_6^2$, because the normal mean and variance are to be estimated from the data, using $\hat{V}_1 = \hat{V}_2 = 0$.

Figure 9.2: Normal & Alternative Densities



This test will assess the data in relation to the first six moments of the normal distribution, surely an intensive scrutiny. We calculate $\hat{S}_4 = 6.152$, with a corresponding p-value (based on 10,000 simulations) of 8.4%. However we also have $\hat{V}_3 = -0.870$, $\hat{V}_4 = 1.271$, $\hat{V}_5 = 1.216$ and $\hat{V}_6 = -1.517$. These components are asymptotically independent and asymptotically have the standard normal distribution.

The p-value casts some doubt on the normality hypothesis. Although none of the components is significantly large, none is small. Figure 9.2 is the density estimate using $\hat{V}_3, \ldots, \hat{V}_6$ discussed previously. It suggests negative skewness and positive kurtosis. Notice that the omnibus statistic, components and density estimate complement each other nicely. As noted above, the Anderson-Darling statistic (A^2) is often the basis of a powerful test of fit. For the present data $A^2 = 0.852$ with a p-value of 0.025. This p-value was determined from D'Agostino and Stephens (1986, Table 4.7).

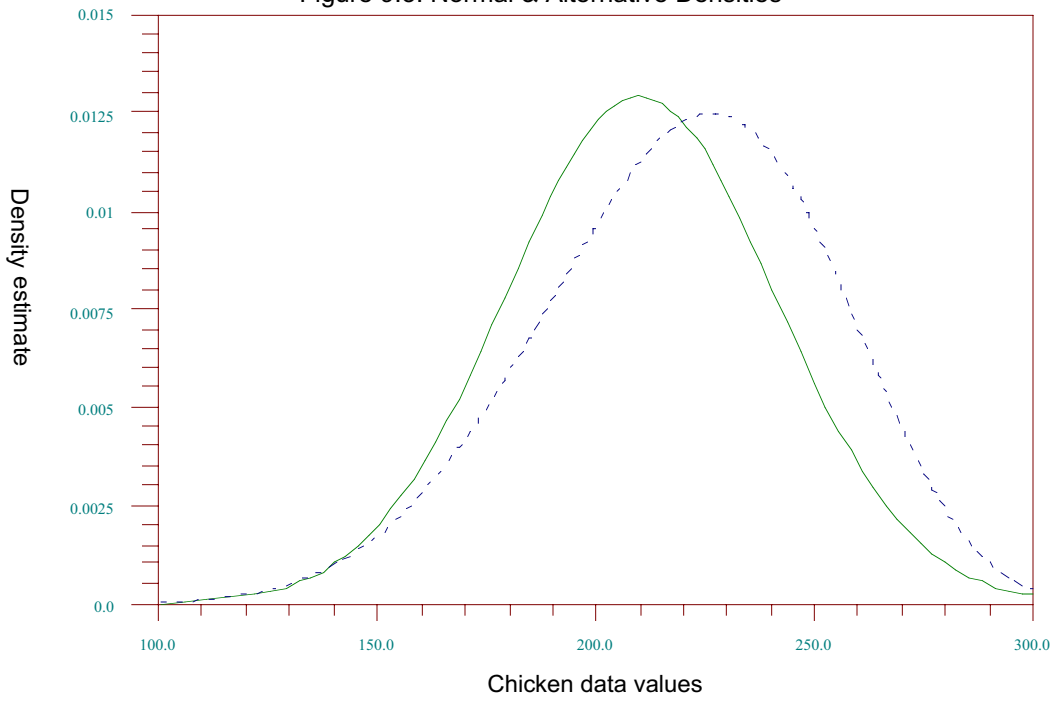
Aside: We claim it is inconsistent to use the uncategorised data for

estimation, and then categorise to form an inconvenient Pearson-type test statistic based on the sum (observed - expected)/expected. This is what is done in forming the so-called Chernoff-Lehmann X^2 test (see Rayner and Best 1989a, p. 27), and by D'Agostino and Stephens in this example.

Leghorn Example. D'Agostino and Stephens (1986, p. 98) gave the Leghorn Chick data due to Bliss (1967). In their Example 4.4.1, they applied several tests, none of which are significant at levels less than 25%. We test for the normal distribution with mean 200 and variance 1225 using $S_4 = V_1^2 + V_2^2 + V_3^2 + V_4^2$. Although the data have mean 209.6 and variance 939.2, we find $S_4 = 2.168$ with a p-value (based on 10,000 simulations) of 53.6%. Moreover $V_1 = 1.227$, $V_2 = -0.621$, $V_3 = -0.422$ and $V_4 = -0.316$. Figure 9.3 is the density estimate using these components, and confirms the observation that the mean of the data is larger and the variance of the data is slightly smaller than hypothesised. However there is insufficient evidence to doubt normality.

Hypothesis tests can be performed using the corrected χ^2 critical points in Rayner and Best (1989a). However the correction factors vary with the significance level, and have only been calculated for a few significance levels. Obtaining p-values by simulation is often more convenient.

Figure 9.3: Normal & Alternative Densities



We conclude this section by noting and commending research in two quite different directions. First, Kallenberg and Ledwina (1997) gave a procedure whereby the data drive the choice of the number of components used in the omnibus score test. Research continues into this important area.

Second, Horswell and Looney (1993) noted that symmetric data may be rejected by the skewness test for normality. Rayner, Best and Mathews (1995) identified the cause. The variance of the r*th* component is asymptotically one under contiguous alternatives, but, in general and certainly in small samples, may involve moments up to the 2r*th*. Therefore significance of the r*th* component suggests the cause is the discrepancy of the data with the r*th* moment of the distribution of interest, but moments up to the 2r*th* may in fact be the cause. Henze has published a series of papers showing how smooth tests may be converted to be diagnostic, that is, so that significance of the r*th* component may be interpreted as the r*th* moment of the data differ from that of the null distribution. See, for example, Henze and Klar (1996). However the Henze tests do not appear to always have good small sample properties. See also Carolan and

Rayner (2000e), who show that even when only one θ_r is non-zero, several V_r may be significant. Not withstanding these facts, throughout this book we have consistently *suggested* that significant components may be interpreted diagnostically. This is not strictly true, but in practice we expect it will only occasionally be false.

9.3 One-Sample Testing for Categorised Distributions

The X^2 tests of Pearson, Pearson and Fisher and of Rao and Robson may be shown to be smooth tests. Complete definitions are given in Rayner and Best (1989a), but the Pearson tests are based on the familiar statistic $\sum$(observed - expected)^2/expected, and we emphasise that the expected cell frequencies are calculated from the *categorised* data. By using the smooth formulation, asymptotically independent components may be derived. Subsequently the test statistics corresponding to the Pearson, Pearson and Fisher and Rao and Robson tests are denoted by X^2_P, X^2_{PF} and X^2_{RR} respectively.

In the categorised smooth model, cell probabilities under the alternative hypothesis are specified by

$$\pi_j(\beta) = C(\theta,\beta) \exp\{\sum_{i=1}^{k} \theta_i h_{ij}(\beta)\} p_j(\beta), j = 1, \ldots, m, \qquad (9.2)$$

where β is a q by 1 vector of nuisance parameters, $p_j(\beta)$ are the null cell probabilities, and $C(\theta,\beta)$ is a normalising constant. The $h_{ij}(\beta)$ are yet to be specified. If the cell counts are $N = (N_1, \ldots, N_m)$, and if $p = (p_j(\beta))$, $H = (h_{ij}(\beta))$ and Σ is the asymptotic covariance matrix of $H(N - np)/\sqrt{n}$, then the score test is based on the statistic

$$\hat{S}_k = (N - n\hat{p})^T \hat{H}^T \hat{\Sigma}^{-1} \hat{H}(n - n\hat{p}),$$

where the ^'s indicate that β has been replaced by the maximum likelihood estimator using the *grouped* data. Specifically Σ is given by

$$\Sigma = H\{D - pp^T - W^T(WD^{-1}W^T)^{-1}W\}H^T,$$

in which $D = \text{diag}(p_1, \ldots, p_m)$ and $W = (\partial p_j/\partial\beta_u)$.

An analogous model and statistic hold if there are no nuisance parameters present. The Pearson X^2 goodness of fit test results by choosing H appropriately in the no nuisance parameters model, and the Pearson-Fisher test may be obtained from the model with nuisance parameters present, again provided H is chosen appropriately. These H are essentially orthogonal. The Rao-Robson X^2 goodness of fit test may be obtained from the model (9.1) using indicator functions for the $h_i(x;\beta)$.

By modifying the models (9.1) and (9.2) the components of $\hat{S}_k$ may be found. For (9.2) this requires defining ϕ so that $\theta = B\phi$, where $\theta = (\theta_1, \ldots, \theta_q)$, B is a k by q matrix of elements $b_{ij}(\beta)$, and working with the model

$$\pi_j(\beta) = C(B\phi,\beta)\exp\{\sum_{i=1}^{k}(B\phi)_i\, h_{ij}(\beta)\}\, p_j(\beta),\ j = 1, \ldots, m.$$

The score statistic is

$$\hat{S}_k = (N - n\hat{p})^T \hat{H}^T \hat{B}\,(\hat{B}^T\hat{\Sigma}\,\hat{B})^{-1}\,\hat{B}^T\hat{H}(n - n\hat{p}),$$

and $\hat{B}$ may be chosen to diagonalise $\hat{\Sigma}$. In this way components of both the Pearson and Pearson-Fisher tests can be obtained.

As in the uncategorised case, different $H = (h_{ij}(\beta))$ result in different tests and in the detection of different alternatives. Not all choices of H result in X^2-type tests. However in the NOR Example, H is chosen to give X_P^2 and moment-type components. Other choices permit us to compare the r*th* cell with the mean of its predecessors, cell focusing and cell overlapping, as we now outline.

Overlapping is an alternative to pooling cells which avoids the loss of information implicit in pooling. Typically the χ^2 approximations to the

distributions of the Pearson and Pearson-Fisher X^2 statistics improve with increasing cell expectation. The i*th* element of H(N - np) is a linear combination in the variables observed - expected for the i*th* cell: $\sum_j h_{ij}(N_j - np_j)$. Thus if the cell expectations are small, the h_{ij} can be taken as indicator functions to amalgamate cells so that the amalgamation of cells have moderate to large expectation. This can be done without losing the information in particular cells. For example, all cells beyond the r*th* can be combined in one contrast, all beyond the (r + 1)*th* in the next, and so on. Details for the no nuisance parameter case are in Rayner (1986), and for the composite case in Rayner and McAlevey (1990). We first met this idea in Hall (1985), where it is called overlapping. S_k (or $\hat{S}_k$) is the quadratic form in this contrast that is weakly optimal for detecting a specifiable smooth alternative. The test statistic has an asymptotic χ^2 distribution. Another option is to focus on particular cells (for example by taking $h_{1i} = 1$, $h_{1j} = 0$ for $j \neq i$) or on particular sets or contrasts between cells (for example by taking $h_{1i} = 1$, $h_{1,i+1} = -1$, $h_{1j} = 0$ for $j \neq i$).

In choosing the $\{h_i(x;\beta)\}$ in the uncategorised case or $\{h_{ij}(\beta)\}$ in the categorised case, it is not necessary that the functions be orthogonal, or powers of the cumulative distribution function. Sometimes other choices are appropriate, and inventive choices may result in powerful and useful tests.

NOR Example. D'Agostino and Stephens (1986, p. 529) gave the NOR data set. One analysis of it (D'Agostino and Stephens 1986, p. 73) used a Pearson type test to test for a completely specified normal distribution. With 25 equiprobable classes under the null hypothesis, they find Pearson's X^2 test statistic, X_P^2, takes the value 28.0, with a corresponding p-value of 26%.

The distributions most commonly tested for are the common distributions with orthogonal functions well investigated. Most have recurrence formulae enabling the higher order orthogonal functions to be quickly and conveniently calculated from their predecessors. In this case where no parameters need to be estimated we may choose equiprobable cell counts. The appropriate orthogonal functions are then the discrete Chebyshev (see Rayner and Best, 1989a, p. 141) polynomials. We find

$$V_1 = -0.596,\ V_2 = 1.369,\ V_3 = -0.616,\ V_4 = -0.779,\ V_5 = 0.790,$$
$$V_6 = -1.999,\ \ldots\ ,\ V_{10} = 2.099,\ \ldots\ ,\ V_{14} = 1.764,\ \ldots\ ,\ V_{24} = 0.325.$$

If we calculate $S_k = V_1^2 + V_2^2 + \ldots + V_k^2$ and the corresponding p-values for k = 1, 2, ... , 24, we find the p-value drops below 10% around k = 15.

Now obviously it is not valid to apply 24 significance tests to a data set and focus on only the most critical of them. We recommend that when testing a distributional hypothesis, the statistician construct a test based on S_k or $\hat{S}_k$, using the first two, three or four components. Fewer components give a more directional test, more components a more omnibus test. A residual such as $X_P^2 - S_k$ may be assessed to see if the remaining components are relevant. However much data analysis is not hypothesis testing, and the calculation of several components enables the analyst to scrutinise the data minutely.

The expression of Pearson's test in terms of its components demonstrates why it often has weak power. In this example, it was assessing deviations from the hypothesised distribution with equal weight in each of 24 dimensions. It is unnecessary to check, say, 20 of these dimensions; and doing so reduces the effectiveness of the test in assessing the first four dimensions of the parameter space.

EMEA Example. D'Agostino and Stephens (1986, p. 548) gave the EMEA data set of heights of maise plants, and investigated whether or not the data were normally distributed with unspecified mean and variance. One analysis of the data (D'Agostino and Stephens 1986, p. 75) supposedly used an X^2 test with cells estimated to be equiprobable. The test statistic was found to be 554, which would be highly significant. However it is not clear to us which X^2 test was being used.

They claimed that "a more intelligent procedure is to use fixed cells with unit width centred at the integers". Using the grouped estimators and a Sheppard's correction, they found $(\hat{\mu}, \hat{\sigma})$. Unfortunately these are not the maximum likelihood estimators. Using a bisection method on the likelihood, we find $(\hat{\mu}, \hat{\sigma}) = (14.5396, 2.2138)$ leading to $X_{PF}^2 = 6.54$,

compared to D'Agostino and Stephens' (14.5396, 2.2159) and $X^2_{PF} = 7.56$. The p-values for either analysis are quite large, implying acceptability of the normality hypothesis. For most purposes the grouped estimates adjusted by Sheppard's correction is an acceptable approximation.

Rayner and McAlevey (1990) described how to calculate the components of X^2_{PF}. Their r*th* component $\widehat{V}_r$ is a contrast in the differences observed minus expected for each cell, standardised to be asymptotically independent and standard normal. Such constructions are not unique. Rayner and McAlevey (1990) constructed $\widehat{V}_r$ to involve the first r + q + 1 cells only, r = 1, ... , m - q - 1. If such a statistic is judged to be significantly large, the interpretation would be that the r*th* cell differs from its predecessors in the comparison of the cell expectations and observations. In the sense that other cells could have been focused on, the construction is not unique. However with no preconceived ideas about this data, we will use these $\widehat{V}_r$. For this data set we find these $\widehat{V}_r$ and the corresponding p-values (from the χ^2 distribution) to be

$\widehat{V}_1 = 0.5773$ (0.56), $\widehat{V}_2 = -0.6371$ (0.52), $\widehat{V}_3 = -0.5933$ (0.55),
$\widehat{V}_4 = -1.0711$ (0.28), $\widehat{V}_5 = -0.1634$ (0.87), $\widehat{V}_6 = -0.7950$ (0.42),
$\widehat{V}_7 = -0.8775$ (0.38), $\widehat{V}_8 = -0.6125$ (0.54), $\widehat{V}_9 = -0.9056$ (0.37),
$\widehat{V}_{10} = 0.5289$ (0.60), $\widehat{V}_{11} = -1.0716$ (0.28), $\widehat{V}_{12} = -0.4534$ (0.65).

Since all the p-values are relatively large, the data have stood up to an extremely close scrutiny of the normality hypothesis and resoundingly confirmed it. The equiprobable analysis that rejected the normality model was clearly questionable. This is not to say that a different choice of H and therefore a different scrutiny may have failed to confirm the hypothesis. After all, the statistician's conclusion is only that there is evidence supplied by the tests used either for or against a certain hypothesis.

With modern computing power it seems to us that if there is a choice, we would not group the data for the purposes of goodness of fit testing. This would then avoid most of the problems demonstrated in the examples above. If the data are not grouped, as above, then a useful omnibus goodness of fit statistic is the Anderson-Darling statistic

discussed, for example, in D'Agostino and Stephens (1986).

9.4 S-Sample Testing

In the S-sample goodness of fit problem we wish to test if S random samples may be considered to come from the same population. Typically that population is not initially specified, although ultimately it may be. For other approaches to this problem, see Kiefer (1958), Eplett (1982), Conover (1998), Scholz and Stephens (1987), Boos (1986) and Eubank and LaRiccia (1990).

The development in this section extends to the S-sample case, the advantages of the one sample smooth goodness of fit tests. These advantages include the following:

(i) a wide class of tests is available;
(ii) these tests are score tests and are thus weakly optimal;
(iii) omnibus tests with directional component tests are available;
(iv) the test statistics have convenient null and alternative distributions and are easily interpreted.

The approach adopted here has two stages. Initially we test if the data are consistent with coming from the same unspecified population: the S-sample problem. If so, subsequent testing for a specified population, a one-sample goodness of fit problem, may be appropriate. A typical situation is a two-sample location problem. A two-sample goodness of fit test assesses if both samples can be regarded as coming from the same population. If so, we ask if the combined samples are consistent with the normal distribution, a one-sample goodness of fit problem. A positive answer leads to the pooled t-test or Welch's test; a negative answer suggests use of the Wilcoxon test. However if both samples cannot be regarded as coming from the same population, this information may be more informative than the answer to the original location question. A Wilcoxon test may still be applied, but the location question may be too simplistic. Perhaps density estimates should be investigated. The question "why has the model (distribution) changed?" needs to be addressed.

As with the one-sample smooth tests, the use of components

permits a close scrutiny of the data. Thus for example, we could optimally test whether the samples agree in their location, dispersion, skewness and kurtosis measures. As a corollary we may then test whether these samples, individually or collectively, are consistent with a specified distribution such as the exponential or the normal. An omnibus assessment bases the decision on moments up to specified order r; the directional components give information on each moment. If the samples are not consistent with each other, an LSD-type analysis will determine the nature of the differences between the populations. Again this analysis gives information on the differences between the samples for each moment up to order r. That we can give much more than a single p-value in assessing the S-sample goodness of fit problem is a considerable strength of this approach.

We first consider the two-sample problem. The hypothesized common distribution is unknown, and we model it as an order k alternative to a convenient "target" distribution, assumed to have no nuisance parameters. Subsequently we extend to the S-sample problem with the target having no nuisance parameters. Some readers may have reservations about the subjective choice of the target. An immediate response would be to say that the choice of target is akin to the assumption of normality in the analysis of variance. Something has to be assumed, and the hope is that the technique is robust to the assumptions. In fact we can do better. From our experience with one-sample problems, if a target can be found that enables all the samples to be represented by a low-order alternative to the target distribution, then the test that results will have good power. The cost of a poor choice of target is loss of power. Whether or not the S samples are assessed to be consistent with each other, if the k parameter alternative chosen is not a good model for all the samples, then further parameters are required to model them, as differences between the samples in the higher parameters may exist.

9.4.1 Two-Sample Smooth Goodness of Fit Testing

We wish to test if two random samples $X_{s1}, \ldots, X_{sn_s}$, s = 1 and 2, may be considered to come from the same population. There is no immediate specification of a probability density function for this common

population. However, we seek to model the probability density functions for each population by functions of the form (9.1) with the same $f_X(x;\beta)$. Then we test if the corresponding parameters are the same, in which case we can conclude that the populations are consistent with each other.

If the current testing problem is preliminary in some sense to assessing if the two populations have the same *specified* probability density function, then that specified probability density function will be used in our model. Otherwise we seek a *target* probability density function that models our populations by order k probability density functions of the form (9.1) with k not large: hopefully at most four. Our expectation from the one-sample case is that this limitation on k will lead to more powerful procedures.

With k large enough, the data can be modelled exactly. In practice we find that few densities project substantially into more than four dimensions: see Rayner, Best and Dodds (1985). However, if the target is poorly chosen, then more than order four differences will need to be assessed. Reduced power will result, and this is the cost of poor targeting.

We model the s*th* probability density function, s = 1, 2, by

$$C_s(\xi_{s1}, \ldots, \xi_{sk}, \beta) \exp\{\sum_{i=1}^{k} \xi_{si} h_i(x;\beta)\} f_X(x;\beta).$$

Although the $\{h_i(x;\beta)\}$ need not be orthonormal functions in general, here we will make that assumption. Were $f_X(x;\beta)$ assumed to be the continuous uniform distribution on (0, 1), the $\{h_i(x;\beta)\}$ could be taken to be the Legendre polynomials, or $\{\sqrt{2}\cos(r\pi j)\}$, or some other set of functions that will enable representation of the data using relatively few terms. If the populations are not uniform, then some ξ_{si} will be non-zero.

In one sample goodness of fit testing, we wish to test if all the parameters in the order k alternative are zero. If the problem was specifying a parsimonious model, we would seek an $f_X(x;\beta)$ with as few ξ_{si} as are needed to describe the data. For the two-sample goodness of fit problem, we want to test if $\xi_1 = (\xi_{1i}) = \xi_2 = (\xi_{2i})$. In many practical

situations, if the populations are consistent with each other, it will also be of interest to know if they are consistent with the target probability density function. Having found that $\xi_1 = \xi_2$, we would then ask if $\xi_1 = \xi_2 = 0$, corresponding to a one sample goodness of fit testing problem.

9.4.2 S-Sample Smooth Goodness of Fit Testing

Suppose we have S independent random samples, the s*th* of which is $X_{s1}, \ldots, X_{sn_s}$. We wish to test if these samples are consistent with coming from the same population. The s*th* sample is assumed to come from distribution with probability density function

$$C_s(\xi_{s1}, \ldots, \xi_{sk};\beta) \exp\left\{\sum_{i=1}^{k} \xi_{si} h_i(x;\beta) / \sqrt{(n_1+\ldots+n_s)}\right\} f_X(x;\beta), \; s = 1, \ldots, S.$$

In the two-sample case we transform from $\{\xi_{si}\}$ to $\{\theta, \beta\}$; see section 9.5.1. There is considerable choice in how we transform the parameters, in both the two and the S-sample cases. Here we use a transformation based on the Helmert matrices discussed, for example, in Lancaster (1965). We reparametrize by putting

$$(\theta_1^T, \ldots, \theta_{S-1}^T, \beta)^T = (H_S \otimes I_S)\,(\xi_1^T, \ldots, \xi_S^T)^T,$$

in which $\otimes$ is the Kronecker product, I_n is the n by n identity matrix, and H_n is the n by n Helmert matrix with last row $(1, \ldots, 1)/\sqrt{n}$ and $(r + 1)$*th* row $(1, \ldots, 1)/\sqrt{[r(r + 1)]}$, $r = 1, \ldots, n - 1$. Writing $G_s = (\sum_j h_r(X_{sj};\gamma_s)$, $s = 1, \ldots, S$, leads to a score vector proportional to

$$((G_1 - G_2 - E_\beta[G_1 - G_2])^T, (G_1 + G_2 - 2G_3 - E_\beta[G_1 + G_2 - 2G_3])^T, \ldots,$$
$$(G_1 + \ldots + G_{S-1} - (S - 1)G_S - E_\beta[G_1 + \ldots + G_{S-1} - (S - 1)G_S])^T)^T.$$

At this point we propose a hierarchical procedure. The difficulty is that the maximum likelihood equations are equivalent to

$$G_1 + ... + G_S = E_\beta[G_1 + ... + G_S],$$

and substituting this into the score vector does not yield appealing components. On the other hand, under the null hypothesis that the first two samples are consistent with having come from the same population, we may use $G_1 + G_2 = E_\beta[G_1 + G_2]$ for the maximum likelihood equations, and substitute this into $(G_1 - G_2 - E_\beta[G_1 - G_2])$. We are led to the vector of components

$$(n_2G_1 - n_1G_2)/\sqrt{[n_1n_2(n_1 + n_2)]}.$$

If the null hypothesis is accepted, we may proceed to the next step of the hierarchical testing. Under the null hypothesis that the first three samples are consistent with coming from the same population, we use $G_1 + G_2 + G_3 = E_\beta[G_1 + G_2 + G_3]$ for the maximum likelihood equations in $G_1 + G_2 - 2G_3 - E_\beta[G_1 + G_2 - 2G_3]$, leading to the vector of components

$$[n_3(G_1 + G_2) - (n_1 + n_2)G_3]\sqrt{[(n_1 + n_2)n_3(n_1 + n_2 + n_3)]}.$$

At the s*th* step (s = 2, ... , S - 1), we assess the consistency of the (s + 1)*th* sample with the preceding s samples by using

$$\frac{n_{s+1}(G_1 + ... + G_s) - (n_1 + ... + n_s)G_{s+1}}{\sqrt{[(n_1 + ... + n_s)n_{s+1}(n_1 + ... + n_{s+1})]}}.$$

The covariance matrix of this vector is asymptotically the k by k unit matrix. Write $V_s = (\sum_j h_r(X_j;\hat{\beta})/\sqrt{n_s})$. In terms of the sample components, at the s*th* step we assess each element of

$$\frac{n_{s+1}(V_1\sqrt{n_1} + ... + V_s\sqrt{n_s}) - (n_1 + ... + n_s)V_{s+1}\sqrt{n_{s+1}}}{\sqrt{[(n_1 + ... + n_s)n_{s+1}(n_1 + ... + n_{s+1})]}}$$

relative to the standard normal distribution. The sum of the squares of these k elements may be compared with the χ^2_k distribution to give an omnibus assessment of the null hypothesis.

It is not necessary to test hierarchically. Doing so assumes that there is an appropriate order for the testing, and that may not be so. Substitution of the maximum likelihood estimators used in this section can be justified on the grounds that they are root n consistent. The omnibus tests mentioned in the previous paragraph are then seen to be C(α)-tests. These omnibus tests can be justified in other ways; see Rayner and Rayner (1998). Moreover the contrasts used in the components may be chosen depending on the situation.

It is also worth noting that we would usually recommend using a completely specified target distribution. Estimating the mean and variance when assuming a normal target means that we cannot assess location and dispersion differences between the hypothesised populations.

9.4.3 Discussion and Example

As always with omnibus tests and their directional component tests, we have the option of informal data analytic assessment, and formal tests of significance. Our preference is a combination of both: using the omnibus test as the basis for a formal test of significance, and using the components informally to assess the nature of any differences should there be some.

To test for a common distribution, we assess the first two samples by calculating the components $(V_{r1}\sqrt{n_2} - V_{r2}\sqrt{n_1})/\sqrt{(n_1 + n_2)}$, and their sum of squares for an omnibus comparison. At the s*th* step, s = 2, ... , S - 1, we use

$$\frac{n_{s+1}\left(V_{r1}\sqrt{n_1} + \ldots + V_{rs}\sqrt{n_s}\right) - \left(n_1 + \ldots + n_s\right)V_{r(s+1)}\sqrt{n_{s+1}}}{\sqrt{\left[\left(n_1 + \ldots + n_s\right)n_{s+1}\left(n_1 + \ldots + n_{s+1}\right)\right]}}$$

to assess the consistency of the order r term of the (s + 1)*th* sample with the order r terms of the preceding consistent samples. The sum of the squares of these statistics up to order k gives an omnibus assessment based on a statistic with the χ^2_k distribution. If at any step the samples are considered to be inconsistent, testing stops. Then the V_{rs} may be used to assess the nature of the inconsistency. With uniform target and $\{h_r(x)\}$ the

Legendre polynomials, the V_{rs} reflect moments of the s*th* population up to the r*th*. This assists in giving an informative explanation of the differences.

If the null hypothesis of a common distribution is accepted, we may use the components to see if the target distribution is that common distribution. The statistic $V_{r1}^2 + \ldots + V_{rS}^2$ assesses order r deviations from the target distribution, and a reasonable omnibus assessment would be given by using the statistic

$$\sum_{s=1}^{S} \sum_{r=1}^{4} V_{rs}^2 .$$

Even if we decide the data are not consistent with the samples being from a common distribution, we can still assess if the s*th* sample is consistent with the target distribution by using $V_{1s}^2 + V_{2s}^2 + V_{3s}^2 + V_{4s}^2$.

Smoothness Example. Lehmann (1975) gave four sets of eight measurements each of the smoothness of a certain type of paper, obtained from four laboratories. The Kruskal-Wallis test gave evidence of significant differences with a p-value of 0.005. The same data were analysed by Scholz and Stephens (1987) using various versions of their Anderson-Darling test. They reported p-values of around 0.002. We could also apply the Yates type analysis of Chapter 4 using the measurements as scores and obtain a sparse 4 by 29 contingency table. However, our interest here is to look at goodness of fit for certain "target" distributions. The data are reproduced in Table 9.1.

With no preconceptions about the data it seems reasonable to assume a target distribution that is uniform between 30 and 60. The value 58.0 looks suspiciously large. We used the Legendre polynomials so that the components could be given a moment interpretation. The V_{rs} are given in Table 9.2. Testing hierarchically, we find samples one, two and three are consistent, but sample four is inconsistent with the other three. The vector test statistic for the final comparison is $(V_{r1} + V_{r2} + V_{r3} - 3V_{r4})/\sqrt{12}$. The elements of this vector take the values 1.51, -1.54, -0.59 and 2.68. Their sum of the squares is 12.16, which yields a p-value of 0.016

on using the χ^2_4 distribution.

Table 9.1 Four sets of eight measurements each of the smoothness of a certain type of paper, obtained from four laboratories

Laboratory	Smoothness							
A	38.7	41.5	43.8	44.5	45.5	46.0	47.7	58.0
B	39.2	39.3	39.7	41.4	41.8	42.9	43.3	45.8
C	34.0	35.0	39.0	40.0	43.0	43.0	44.0	45.0
D	34.0	34.8	34.8	35.4	37.2	37.8	41.2	42.8

Table 9.2 Components V_{rs} for the data in Table 9.1 using a uniform (30, 60) target and normalised Legendre polynomials

Population (s)	Order (r) 1	2	3	4
1	0.2327	-1.9439	0.7992	1.8767
2	-1.0859	-2.5016	2.0570	1.2663
3	-1.5105	-1.6286	1.5978	0.4755
4	-2.5311	-0.2467	2.1608	-1.8855

Had we chosen initially to compare the first and fourth samples, the component values are 1.95, -1.20, -0.96 and 2.66, with a sum of the squares of 13.26, and a p-value of 0.01. An LSD-type comparison of the $V_{rs}/\sqrt{n_s}$ shows clearly in what respects the samples differ. The details are in Rayner and Rayner (1997). Define

$$\text{LSD} = z_{(\alpha / {}^{I}C_2)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

in which $z_{(\alpha / {}^{I}C_2)}$ is the point giving equal $50\alpha/{}^{I}C_2\%$ tails with the standard

normal distribution.

For an overall 5% level LSD = 1.32. Using the uniform (30, 60) target, the only effect is an order 4 difference between the first and fourth samples. The LSD method is known for its conservatism, and so it is of interest to weaken the overall level to 10%. This identifies in addition a first order difference between the first and fourth samples and a fourth order difference between the second and fourth samples.

We repeated the analysis using a $N(\mu, \sigma^2)$ target with the Hermite polynomials. The components are given in Table 9.3 when using $\mu = 40$, $\sigma = 10$. While the data are consistent with this target (p-value 0.199), none of the omnibus comparison statistics are significant at either the 5% or 10% levels. With a N(40, 5) target, differences are of order one between the first and fourth samples and order two between samples one and two. Moreover the first sample has greater order three effects than both the second and third samples, and the fourth sample has greater order four effects than the remaining populations.

Table 9.3 Components V_{rs} for the data in Table 9.1 using a N(40, 10) target and normalised Hermite polynomials

Population (s)	Order (r) 1	2	3	4
1	1.6157	-0.7796	-0.9948	0.4217
2	0.4738	-1.8516	-0.5423	1.4847
3	0.1061	-1.6975	-0.1442	1.2295
4	-0.7778	-1.6670	0.8655	1.1795

The ordering of the samples by the uniform (30, 60) and N(40, 5) targets is identical for orders one and four, but permuted somewhat for orders two and three.

We repeated the analysis for uniform (0, 100) and uniform (34, 58) targets. It seems that the different target distributions are defining different scales, and thus give slightly different foci on the various effects.

Varying the overall significance level achieves a similar result. The consensus from the different perspectives appears to confirm that there are location and kurtosis differences between the first and fourth samples: laboratories A and D.

A uniform (30, 60) distribution is not a good description of the data (p-value 0.0002). The latter is reflected in Table 9.2 in that the components of orders two and three all have the same sign, and that this is almost true for the components of orders one and four. With a uniform target further ξ_{si} are needed to model the probability density functions corresponding to the different samples. We could assess this by calculating further V_{rs}. It may well be that there are differences in moments beyond the four considered here, although whether we really want to know this is a moot point. Also, with all samples being of size eight, the asymptotic null distributions are questionable. This could be overcome by calculating permutation test p-values for the V_i and S_i.

9.5 Derivations and Simulation Study

9.5.1 Derivation of the Two-Sample Test Statistic

We now derive the score test of H: $\xi_1 = \xi_2$ against K: $\xi_1 \neq \xi_2$ given two random samples, $X_{s1}, \ldots, X_{sn_s}$, s = 1 and 2. Abbreviating meaningfully, the likelihood is

$$C_1^{n_1}(\xi_1) \exp\left\{\sum_{i=1}^{k} \xi_{1i} \sum_{j=1}^{n_1} h_i(x_{1j})\right\} \prod_{j=1}^{n_1} f(x_{1j}) C_2^{n_2}(\xi_2) \exp\left\{\sum_{i=1}^{k} \xi_{2i} \sum_{j=1}^{n_2} h_i(x_{2j})\right\} \prod_{j=1}^{n_2} f(x_{2j}).$$

Write $G_{sr} = \sum_j h_r(x_{sj})$, for s = 1, 2 and r = 1, ... , k ≤ n - 1, $f_{X_1} = \prod_j f(x_{1j})$ and $f_{X_2} = \prod_j f(x_{2j})$. The logarithm of the likelihood is then

$$\ell = n_1 \ell n\, C_1 + \sum_i \xi_{1i} G_{1i} + \ell n\, f_{X_1} + n_2 \ell n\, C_2 + \sum_i \xi_{2i} G_{2i} + \ell n\, f_{X_2}.$$

Now reparametrize by putting $\theta = \xi_1 - \xi_2 = (\theta_i)$ and $\beta = \xi_1 + \xi_2 = (\beta_i)$. Then $\xi_1 = (\theta + \beta)/2$ and $\xi_2 = (\beta - \theta)/2$. We wish to test H: $\theta = 0$ against K: $\theta \neq 0$ in the presence of the vector of nuisance parameters β.

Differentiation yields

$$\frac{\partial \ell}{\partial \theta_r} = \frac{n_1}{2}\frac{\partial\, \ell n\, C_1}{\partial \xi_{1r}} + \frac{G_{1r}}{2} - \frac{n_2}{2}\frac{\partial\, \ell n\, C_2}{\partial \xi_{2r}} - \frac{G_{2r}}{2},$$

$$\frac{\partial \ell}{\partial \beta_r} = \frac{n_1}{2}\frac{\partial\, \ell n\, C_1}{\partial \xi_{1r}} + \frac{G_{1r}}{2} + \frac{n_2}{2}\frac{\partial\, \ell n\, C_2}{\partial \xi_{2r}} + \frac{G_{2r}}{2},$$

$$\frac{\partial^2 \ell}{\partial \theta_r \partial \theta_s} = \frac{n_1}{4}\frac{\partial^2\, \ell n\, C_1}{\partial \xi_{1r}\partial \xi_{1s}} + \frac{n_2}{4}\frac{\partial^2\, \ell n\, C_2}{\partial \xi_{2r}\partial \xi_{2s}} = (I_{\theta\theta})_{rs},$$

$$\frac{\partial^2 \ell}{\partial \theta_r \partial \beta_s} = \frac{n_1}{4}\frac{\partial^2\, \ell n\, C_1}{\partial \xi_{1r}\partial \xi_{1s}} - \frac{n_2}{4}\frac{\partial^2\, \ell n\, C_2}{\partial \xi_{2r}\partial \xi_{2s}} = (I_{\theta\beta})_{rs},$$

$$\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = \frac{n_1}{4}\frac{\partial^2 \ell n\, C_1}{\partial \xi_{1r}\partial \xi_{1s}} + \frac{n_2}{4}\frac{\partial^2\, \ell n\, C_2}{\partial \xi_{2r}\partial \xi_{2s}} = (I_{\beta\beta})_{rs}.$$

From Lehmann (1959, p. 58), $\partial \ell n\, C_i / \partial \xi_{ir} = - E_{\xi_i}[h_r]$ and $\partial^2 \ell n C_i / \partial \xi_{ir}\partial \xi_{is} = - \mathrm{cov}_{\xi_i}[h_r, h_s]$, for i = 1, 2. It follows that the elements of the efficient score, U_r, are given by

$$2U_r = G_{1r} - G_{2r} - \{n_1 E_{\xi_1}[h_r] - n_2 E_{\xi_2}[h_r]\},\ r = 1, \ldots, k.$$

The expectations in U_r are unknown, but in the usable form of the score test are estimated by maximum likelihood under the null hypothesis. These equations are $\partial \ell / \partial \beta_r = 0$, r = 1, ... , q, and yield, if we write E_β for the expectation under the null hypothesis,

$$G_{1r} + G_{2r} = (n_1 + n_2) E_\beta[h_r].$$

Substituting in the efficient score and simplifying gives, under the null

hypothesis,

$$U_r = (n_2 G_{1r} - n_1 G_{2r})/(n_1 + n_2).$$

Write $\Omega = (\mathrm{cov}_\xi(h_r, h_s))$. The partitioned information matrix is given by

$$\begin{pmatrix} I_{\theta\theta} I_{\theta\beta} \\ I_{\beta\theta} I_{\beta\beta} \end{pmatrix} = \begin{pmatrix} (n_1+n_2)\Omega & (n_1-n_2)\Omega \\ (n_1-n_2)\Omega & (n_1+n_2)\Omega \end{pmatrix} /4 \ .$$

The asymptotic covariance matrix of U_θ is $I_{\theta\theta} - I_{\theta\beta} I_{\beta\beta}^{-1} I_{\beta\theta} = n_1 n_2 \Omega/(n_1 + n_2)$. Denote Ω when evaluated under the null hypothesis by Ω_β. The score statistic is then

$$(n_1 + n_2)(U_r)^T \Omega_\beta^{-1} (U_r)/(n_1 n_2).$$

To make this more accessible, we replace the ξ_i in the original model by $\xi_i/\sqrt{(n_1 + n_2)}$, i = 1 and 2. The s*th* probability density function is thus assumed to be

$$C_s(\xi_{s1}, \ldots, \xi_{sk}) \exp\{\sum_{i=1}^{k} \xi_{si} h_i(x)/\sqrt{(n_1+n_2)}\} f_\chi(x), \ s = 1, 2. \qquad (9.3)$$

Such distributions are said to be *contiguous*, and tend to the target distributions with probability density functions $f_\chi(x)$ as the sample sizes tend to infinity. Each probability density function is now asymptotically equivalent to that used in Rayner and Best (1986, Theorem 1). From the proof of that theorem, under the alternative hypothesis there, it follows that *asymptotically* $E_\beta[G_{sr}] = n_s \theta_r/\sqrt{(n_1 + n_2)}$, s = 1, 2, and $\Omega = I_k$, the k by k identity matrix. Under this revised model the score statistic is

$$S_k = V_1^2 + \ldots + V_k^2, \text{ in which}$$
$$V_r = \{n_2 \textstyle\sum_j h_r(X_j) - n_1 \sum_j h_r(Y_j)\}/\sqrt{\{n_1 n_2 (n_1 + n_2)\}}.$$

These V_r are asymptotically independent and asymptotically have the standard normal distribution. Moreover they may be derived as score statistics in their own right, and hence are weakly optimal for detecting specifiable alternatives.

Define, as in Rayner and Best (1989a, Chapter 4), $V_{r1} = \sum_j h_r(X_j)/\sqrt{n_1}$ and $V_{r2} = \sum_j h_r(Y_j)/\sqrt{n_2}$, the usual components from the one sample goodness of fit problem. Then

$$V_r = \{\sqrt{n_2}V_{r1} - \sqrt{n_1}\,V_{r2}\}/\sqrt{(n_1 + n_2)},\ r = 1, \ldots, k.$$

Thus we can calculate the one sample components and combine them as indicated to assess whether the two populations can be considered to be the same. As a corollary, the one sample components could then be used for subsequent testing for the so-called target distribution. Incidentally, $E[V_{rs}] = \sqrt{n_s}\ E[h_r(X)]$, and if $h_r(x)$ is the orthonormal polynomial of order r, $E[h_r(X)]$ involves central moments up to order r. For this reason we typically interpret a significant V_r as indicating differences in the populations in the moments up to order r.

9.5.2. Simulation Study

In order to assess the distributional conclusions of the preceding section, a small Monte Carlo study was undertaken. The target distribution was taken to be the uniform, and the $\{h_r(x)\}$ was taken to be the set of Legendre polynomials. We calculated the proportion of exceedances for the asymptotic 99%, 95%, 90%, 80%, 20%, 10%, 5% and 1% points for each of the first four components. These components all asymptotically follow the standard normal distribution. If the common null distribution of the two samples is uniform, the simulated exceedances are within sampling error of the nominal exceedances for sample sizes as low as $n_1 = n_2 = 5$. Taking into account the possibility of using the probability integral transformation to uniformity, it appears that if the target is correctly specified, the V_r appear to follow the anticipated normal distribution for surprisingly small sample sizes.

We next considered the family of truncated exponential-type distributions with probability density functions

$$f_X(x; \theta) = \lambda\exp(\lambda x/\sqrt{n})/\{\sqrt{n}[\exp(\lambda/\sqrt{n}) - 1]\},\ x \in (0, 1),$$
zero otherwise, in which $\lambda > 0$. (9.4)

With $\lambda = 5$ the nominal and simulated exceedances were close, even for $n_1 = n_2 = 10$. Table 9.4 gives results for V_1 and V_4 for $\lambda = 10$. In all cases the number of simulations is 1,000. Results for V_2 and V_3 suggest the agreement between the actual and nominal sizes is best for V_1, then V_2, then V_3, and then V_4. Greater detail is given in Rayner and Rayner (1998).

The reason for the good agreement for small λ is that (9.4) is well approximated by probability density functions of the form (9.3) with k small, or, equivalently, most of the ξ_{si} small; that is, if the target distribution is well chosen. Also the asymptotic distribution of the V_i depends on the Central Limit Theorem. For unimodal $h_r(x)$ the Central Limit Theorem will converge rapidly; for higher order polynomials - larger r - convergence will be slower.

It is interesting to note that for non-contiguous null hypotheses, the exceedances can be poor and do *not* improve with increasing sample size. The experiment of the preceding paragraph was repeated, but with the common null distribution having cumulative distribution function $F_X(x) = x^{0.25}$. The nominal 5% exceedances for V_1 with $n_1 = n_2 = 10$, 50, 100 and 200 were 3.9%, 3.3%, 4.4% and 3.6% respectively. The corresponding exceedances can be poor and do *not* improve with increasing sample size. The experiment of the preceding paragraph was repeated, but with the common null distribution having cumulative distribution function $F\chi(x) = x^{0.25}$. The nominal 5% exceedances for V_1 with $n_1 = n_2 = 10$, 50, 100 and 200 were 3.9%, 3.3%, 4.4% and 3.6% respectively. The corresponding exceedances for V_4 were 16.8%, 14.5%, 15.2% and 14.6%. In view of this poor approximation for V_4, some users may prefer to calculate permutation test p-values when there is any doubt that the asymptotics may not yield good approximations.

Table 9.4 Proportions exceeding nominal percentage points for a contiguous truncated exponential distribution with $\lambda = 10$ and for components based on the first four Legendre polynomials. Based on 1,000 simulations for equal sample sizes $n_1 = n_2 = 5, 10, 20, 50$

	Nominal Exceedances							
	99%	95%	90%	80%	20%	10%	5%	1%
V_1								
$n_1 = n_2 = 5$	0.999	0.991	0.970	0.894	0.121	0.036	0.010	0.000
$n_1 = n_2 = 10$	0.999	0.979	0.943	0.851	0.167	0.060	0.018	0.003
$n_1 = n_2 = 20$	0.997	0.970	0.930	0.830	0.168	0.077	0.034	0.006
$n_1 = n_2 = 50$	0.993	0.959	0.908	0.805	0.147	0.075	0.037	0.010
V_4								
$n_1 = n_2 = 5$	0.979	0.908	0.863	0.778	0.246	0.133	0.076	0.018
$n_1 = n_2 = 10$	0.982	0.944	0.890	0.796	0.215	0.106	0.059	0.014
$n_1 = n_2 = 20$	0.978	0.931	0.879	0.773	0.193	0.104	0.057	0.016
$n_1 = n_2 = 50$	0.985	0.947	0.902	0.790	0.205	0.113	0.057	0.010

A substantive power study was also undertaken and is reported in greater detail in Rayner and Rayner (1998). That study compared the omnibus tests based on S_2, S_3 and S_4, their components V_2, V_3 and V_4, and A^2_{2n}, the Anderson-Darling statistic of Scholz and Stephens (1987). For alternatives of the form (9.4) with different λ, it was found that there appears to be a non-uniform ordering of the tests: V_1^2, A^2_{2n}, S_2, V_2^2, S_3, S_4, V_3^2, V_4^2 in decreasing effectiveness. We also examined trigonometric-type alternatives to uniformity. These confirmed our initial conclusion that the A^2_{2n} test appears to be predominantly what we would call a first order test, being sensitive especially to location shifts. Although its distribution theory is most adequate for applications, it is reasonable to ask if the Anderson-Darling test provides significantly more information than that

supplied by the Wilcoxon test.

The smooth tests provide a wealth of information. If there is some preliminary knowledge about the alternatives it is best to protect against, this information can be built into the choice of target distribution and the orthonormal polynomials. If the target distribution is well chosen, the asymptotic null distributions of the smooth tests should be adequate for moderate sample sizes. For small sample sizes, or if there is any doubt about the adequacy of the target, p-values should be based on a permutation test.

10
Conclusion

10.1 Introduction

We conclude this discussion of our approach to nonparametric methods with some comments on three related areas: the analysis of multi-way contingency tables, the notion of correlation in higher-way tables and testing when there is some parametric structure.

A consistent assumption throughout this book has been that the data may be presented in a contingency table of counts. Instead of then analysing the data by standard methods, such as log-linear models, we have used a partition of X^2 approach. That this approach ultimately leads to many of the fundamental and powerful tests of nonparametric analysis is, we believe, a good reason for deeper consideration of these methods. They can be used instead of the deeply entrenched parametric methods, to some advantage. We consider how in section 10.2.

In two-way tables correlation is used as both a measure of independence and a measure of linearity. In the latter role it has a very appealing property: that linearity occurs if and only if the correlation takes either of the values 1 or -1. Here we consider correlation in three and higher dimensions. We suggest that for independence to hold there are many parameters that must be zero, and all could be called correlations. There are corresponding sample statistics that can be thought of as sample correlations, and these can be used to make inferences about the population correlations. In more than two dimensions correlations do not seem to be successful in approaching questions about linearity. These matters are considered in section 10.3.

A goodness of fit test may reveal that a set of data do not follow a particular distribution, perhaps the normal, failing only in regard to, say, skewness. It may be that this failure is sufficiently serious in the circumstances to exclude the use of the normal theory based method of

analysis that was originally envisaged. However, instead of falling back on a totally distribution-free analysis, we can proceed using a parametric analysis that adjusts for the failure of the normal model. In 10.4 we demonstrate that this approach may lead to substantial gains.

Finally, some brief concluding remarks are given in 10.5.

10.2 Partitioning Pearson's Chi-Squared Statistic for at Least Partially Ordered Three-Way Contingency Tables

10.2.1 Introduction

In this section we extend the discussion of section 8.6 on completely ordered tables to at least partially ordered tables, mainly by reporting on papers by Beh and Davy (1998 and 1999) who partitioned Pearson's X^2 statistic for three-way contingency tables. Their approach generalises immediately to multi-way tables, but the three-way description is sufficient to give the flavour of what is intended. This material is included to demonstrate that our approach to two-way tables, that produced both known and new nonparametric tests, can also be used for data analysis, and extends to higher-way tables.

Beh and Davy (1998) relate their model with the corresponding log-linear model. Their parameters approximate corresponding parameters in the log-linear model, but closed expressions are possible for the Beh-Davy parameters. It follows that these parameter values could be regarded as initial values in the iterative schemes for finding the maximum likelihood estimates in the log-linear models, and as default values when the iterative schemes fail to converge. In addition Beh and Davy (1998) observe that their models have no difficulty with selection of the optimum model, give more information regarding the structure of the model, and can give a better fit of the data to a model.

10.2.2 Notation

Suppose we have a three way contingency table of counts $\{N_{ijk}\}$ with i = 1, ... , I, j = 1, ... , J and k = 1, ... , K. We use the standard dot notation so that, for example, $N_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} N_{ijk}$. Define $p_{i..} = N_{i..}/n$, $p_{.j.} = N_{.j.}/n$, and $p_{..k} = N_{..k}/n$, where $n = N_{...}$. Further, define normalised orthogonal polynomials $\{a_u(i)\}$ on $\{p_{i..}\}$, $\{b_v(j)\}$ on $\{p_{.j.}\}$ and $\{c_w(k)\}$ on $\{p_{..k}\}$.

Three sets of "standardised sums" will be required. Put

$$Z_{ujk} = \sqrt{\frac{n}{p_{.j.}p_{..k}}} \sum_{i=1}^{I} a_u(i)\, p_{ijk},$$

$$Z_{uvk} = \sqrt{\frac{n}{p_{..k}}} \sum_{i=1}^{I} \sum_{j=1}^{J} a_u(i)\, b_v(j)\, p_{ijk}, \text{ and}$$

$$Z_{uvw} = \sqrt{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} a_u(i)\, b_v(j)\, c_w(k) p_{ijk}.$$

The transition from Z_{uvw} to Z_{uvk} involves replacing the summation over k and $c_w(k)$ with $\frac{1}{\sqrt{p_{..k}}}$, and the transition from Z_{uvk} to Z_{ujk} involves replacing the summation over j and $b_v(j)$ with $\frac{1}{\sqrt{p_{.j.}}}$. The subscripts u, v and w are used for ordered categories, while the subscripts i, j and k are used for unordered categories.

As usual, the Pearson X^2 statistic is defined as the sum over all cells of $(\text{obs} - \text{exp})^2/\text{exp}$, where obs is the observed cell count and exp is the expected cell count under the null hypothesis.

For three-way tables we now look at the partition of the Pearson X^2 statistic when there are one, two and three ordered sets of categories. We assume the ordered categories come first; so, for example, when there are two ordered sets of categories they are on the rows and columns.

10.2.3 Partitioning Pearson's X^2 statistic

For one ordered set of categories Pearson's X^2 statistic X^2_P is given by

$$X^2_P = \sum_{u=0}^{I-1}\sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{ujk} = \sum_{u=1}^{I-1}\sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{ujk} + \sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{0jk}.$$

The term $\sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{0jk}$ is equal to X^2_{JK}, Pearson's X^2 statistic defined for a two-way table with JK cells and neither category ordered. The term $\sum_{u=1}^{I-1}\sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{ujk}$ includes summands that assess genuine trivariate association, and is denoted by X^2_{UJK}. Thus

$$X^2_P = X^2_{UJK} + X^2_{JK}.$$

Here and henceforth, when a summation starts at zero, that summation is broken into two parts: the zero term and the remainder. Ordered summations start from zero while unordered summations start from one.

For two ordered sets of categories Pearson's X^2 statistic X^2_P satisfies

$$X^2_P = \sum_{u=0}^{I-1}\sum_{v=0}^{J-1}\sum_{k=1}^{K} Z^2_{uvk}$$

$$= \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{uvk} + \sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{0vk} + \sum_{u=1}^{I-1}\sum_{k=1}^{K} Z^2_{u0k}$$

$$= X^2_{UVK} + X^2_{VK} + X^2_{UK} \text{ say.}$$

In a similar notation to previously, here X^2_{VK} denotes Pearson's X^2 statistic defined for a two-way table with the first category ordered and the second not. The term X^2_{UVK} assesses genuine trivariate association.

For three ordered sets of categories the Pearson independence X^2 statistic X^2_P satisfies

$$X^2_P = \sum_{u=0}^{I-1} \sum_{v=0}^{J-1} \sum_{w=0}^{K-1} Z^2_{uvw}$$

$$= \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} \sum_{w=1}^{K-1} Z^2_{uvw} + \sum_{v=0}^{J-1} \sum_{w=0}^{K-1} Z^2_{0vw} + \sum_{u=0}^{I-1} \sum_{w=0}^{K-1} Z^2_{u0w} + \sum_{u=0}^{I-1} \sum_{v=0}^{J-1} Z^2_{uv0}$$

$$= X^2_{UVW} + X^2_{VW} + X^2_{UW} + X^2_{UV},$$

in obvious notation, and this agrees with section 8.6.

10.2.4 Degrees of Freedom

We now look at the degrees of freedom for these cases. In all cases Pearson's X^2 statistic has degrees of freedom

the number of cells - 1 - the number of linear constraints.

Thus for the singly ordered case, X^2_P has degrees of freedom $df(X^2_P)$ given by

$$df(X^2_P) = IJK - 1 - (I - 1) - (J - 1) - (K - 1) = IJK - I - J - K + 2$$

since X^2_P involves $p_{i..}$, $p_{.j.}$ and $p_{..k}$, and, for example, $p_{i..} = N_{i..}/n$ imposes I - 1 linear constraints.

An alternative view gives the degrees of freedom of $X^2_{UJK} = \sum_{u=1}^{I-1} \sum_{j=1}^{J} \sum_{k=1}^{K} Z^2_{ujk}$ as (I - 1)(JK - 1) since, first, for each of I - 1 values of u we have the Z_{ujk} subject to the constraint

$$\sum_{j=1}^{J}\sum_{k=1}^{K} z_{ujk}\sqrt{p_{.j.}\,p_{..k}} = \sum_{i=1}^{I} a_u(i)\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk} = \sqrt{n}\sum_{i=1}^{I} a_u(i)\,p_{i..} = 0$$

using the orthonormality, thus verifying that there are only JK - 1 such z_{ujk}. Second, $df(X^2_{JK}) = (J - 1)(K - 1)$, as is usual for a two-way contingency table. This confirms that for singly ordered three-way tables

$$df(X^2_P) = df(X^2_{UJK}) + df(X^2_{JK}).$$

For two ordered sets of categories we have the degrees of freedom of $X^2_{UVK} = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} z^2_{uvk}$ are (I - 1)(J - 1)K, since there are this number of asymptotically independent z_{ujk}, and we have $df(X^2_{VK}) = (J - 1)(K - 1)$ and $df(X^2_{UK}) = (I - 1)(K - 1)$, as usual for two-way contingency tables, confirming

$$df(X^2_P) = df(X^2_{UVK}) + df(X^2_{VK}) + df(X^2_{UK}).$$

For three ordered sets of categories we have the degrees of freedom of $X^2_{UVW} = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} z^2_{uvw}$ are (I - 1)(J - 1)(K - 1), since there are this number of asymptotically independent z_{ujk}. We also have $df(X^2_{VW}) = (J - 1)(K - 1)$, $df(X^2_{UW}) = (I - 1)(K - 1)$ and $df(X^2_{UV}) = (I - 1)(J - 1)$ as usual for two-way contingency tables, confirming that for triply ordered three-way tables

$$df(X^2_P) = df(X^2_{UVW}) + df(X^2_{VW}) + df(X^2_{UW}) + df(X^2_{UV}),$$

as given in section 8.6 above.

10.2.5 Interpretation

The null hypothesis of interest may loosely be described as independence, which includes homogeneity across specified tubes, or columns and tubes.

For singly ordered tables Z_{ujk} describes the u*th* moment effect on the (j, k)*th* pair of categories, and $\sum_{j=1}^{J}\sum_{k=1}^{K} Z^2_{2jk}$ describes the quadratic or dispersion effect over the whole table.

For doubly ordered tables Z_{uvk} describes the (u, v)*th* bivariate moment effect on the k*th* tube, and $\sum_{k=1}^{K} Z^2_{12k}$ describes how, over the whole table, the data differ from what might be expected under independence in the (1, 2)*th* bivariate moment.

For any given table we typically have alternative ways to use the partition of X^2. Thus the term $\sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{uvk}$ may be further decomposed via

$$\sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{uvk} = \sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{1vk} + \sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{2vk} + \sum_{u=3}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Z^2_{uvk}.$$

The first term describes the overall row location effect, the second the overall row dispersion effect, and the third term is a residual. An alternative approach would partition the same quantity into overall column location, dispersion and residual effects. Clearly we have to be careful about performing several tests of significance simultaneously on the same data.

Happiness Example Continued. In the light of this development, the reader may wish to now refer back to the Happiness Example in section 8.6 and reconsider those data assuming happiness (tubes) is an unordered response variable, so the overall table has only a partial ordering. The partition of X^2_P statistic into component values is given in Table 10.1, and

this should be compared with Table 8.6.

What is distinctive is that for the triply ordered table we had $X^2_P = X^2_{UVW} + X^2_{VW} + X^2_{UW} + X^2_{UV}$, whereas for a doubly ordered table $X^2_P = X^2_{UVK} + X^2_{VK} + X^2_{UK}$. Numerically $X^2_{UV} = X^2_{UK}$ and $X^2_{VW} = X^2_{VK}$, and this can be shown algebraically. Interestingly in going from the triply ordered case to the doubly ordered, there are more genuine trivariate degrees of freedom, and with the loss of one of the bivariate X^2 terms, fewer bivariate degrees

Table 10.1 Partition of X^2_P statistic into component values

Term	Component	Value	df	p-value
X^2_{UK}	Row components			
	Location	31.6954	2	0
	Dispersion	6.2973	2	0.0466
	Residual	3.1460	2	0.2127
		41.1387	6	0
X^2_{VK}	Column components			
	Location	18.0973	2	0.0001
	Dispersion	0.9724	2	0.6172
	Residual	6.7517	4	0.1492
		25.8215	8	0.0008
X^2_{UVK}	Row components			
	Location	214.2615	12	0
	Dispersion	43.2588	12	0.0001
	Residual	4.0869	12	0.9819
	Column components			
	Location	226.2692	9	0
	Dispersion	20.1955	9	0.0159
	Residual	15.1425	18	0.6551
		261.6072	36	0
X^2_P		328.5674	50	0

of freedom. This may affect the data analysis. Hence Beh and Davy (1999) point out that when the level of happiness is not considered to be of an ordinal nature, the trivariate association is statistically significant, contrary to when it is considered to be ordinal. Detailed conclusions may be gleaned from Beh and Davy (1999), or by reference to Table 10.1.

Although we have not pursued it here, the ability to analyse data such as these suggests that multivariate analogues can be developed for

Yates' test, and the extensions we have developed in Chapters 4 and 5. In the next section we look at such a development in the case of Spearman's rho and its extensions.

10.3 Generalised Correlations

10.3.1 Generalised Bivariate Correlations

If the aim of an analysis of bivariate contingency table data is to assess independence, and if bivariate normality can be assumed, then it is known that independence occurs if and only if the correlation is zero. Independence of a bivariate (X, Y) distribution can be assessed by testing if the sample Pearson correlation is consistent with zero. In this case, the parameter space is one dimensional. On the other hand, suppose that no distributional assumptions can be made other than the bivariate distribution is discrete, with I rows and J columns. The distribution is specified by the cell probabilities p_{ij}, i = 1, ... , I and j = 1, ... , J. Independence is equivalent to $p_{ij} = p_{i.}\, p_{.j}$ for i = 1, ... , I and j = 1, ... , J, with $p_{i.}$ being a marginal row probability, and $p_{.j}$ a marginal column probability. Assuming category scores of $\{x_i\}$ and $\{y_j\}$, a saturated model for p_{ij}, for i = 1, ... , I and j = 1, ... , J, is

$$P(X = x_i, Y = y_j) = p_{ij} = \left\{ 1 + \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} \theta_{uv}\, g_u(x_i)\, h_v(y_j) \right\} p_{i.}\, p_{.j},$$

in which the θ_{uv} are real-valued parameters and $\{g_u(x_i)\}$ are orthonormal functions on the $\{p_{i.}\}$ and $\{h_v(y_j)\}$ are orthonormal functions on the $\{p_{.j}\}$. With natural scores this is (8.1) with k_1 replaced by its maximum value I - 1, and k_2 replaced by its maximum value J - 1. Testing for independence is equivalent to testing if every θ_{uv} is consistent with zero. This is a (I - 1)(J - 1) dimensional parameter space. This highlights just how strong a distributional assumption can be: in this case, equivalent to a (I - 1)(J - 1) - 1 dimensional parameter space! Note that for u = 1, ... , I - 1, and v = 1, ... , J - 1

$$\theta_{uv} = \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}\, g_u(x_i) h_v(y_j).$$

Suppose $E[X] = \mu_X$, $E[Y] = \mu_Y$, $var(X) = \sigma_X^2$ and $var(Y) = \sigma_Y^2$. If $g_1(x_i) = (x_i - \mu_X)/\sigma_X$ and $h_1(y_j) = (y_j - \mu_Y)/\sigma_Y$ then $\theta_{11} = \rho_P$, the Pearson population correlation.

From section 8.2 recall that we have defined, again for u = 1, ... , I - 1, and v = 1, ... , J - 1,

$$\hat{V}_{uv} = \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}\, g_u(x_i)\, h_v(y_j)/\sqrt{n},$$

in which n is the grand total of the observations. It is straightforward to show that if the orthonormal functions use weights the marginal distributions from the contingency table, $\hat{V}_{uv}^2$ is a score statistic for testing H_0: $\theta_{uv} = 0$ against K: $\theta_{uv} \neq 0$, and the sample Pearson correlation is $r_{11} = \hat{V}_{11}/\sqrt{n}$. In large samples the $\hat{V}_{uv}$ are independent and have the standard normal distribution under the null hypothesis. Under the alternative hypothesis, and assuming contiguous alternatives, $\hat{V}_{uv}$ has the $N(\theta_{uv}, 1)$ distribution. Since $\theta_{11} = \rho_P$, the population Pearson correlation, inference about ρ_P can be based on the $N(\rho_P, 1)$ distribution of $\hat{V}_{11}$.

Using the null N(0, 1) distribution of $\hat{V}_{uv}$, a value $\hat{v}_{uv}$ of $\hat{V}_{uv}$ may be assigned a position $2P\left(\hat{V}_{uv} > \left|\hat{v}_{uv}\right|\right)$ on the (0, 1) scale, and this may be used as a weight of evidence indicator for the null hypothesis of independence. If the asymptotic normal is in doubt, because the sample size is small, or the contiguity assumption unlikely to be satisfied, then resampling methods should be used.

Now $\hat{V}_{11}$ is related to the Pearson correlation, but all the $\hat{V}_{uv}$ are relevant to the independence hypothesis, and we suggest that all

$$r_{uv} = \hat{V}_{uv}/\sqrt{n} \text{ for } u = 1, \ldots, I - 1, \text{ and } v = 1, \ldots, J - 1$$

be thought of as generalised correlations. It is shown in Davy et al. (2000) that $-1 \leq r_{uv} \leq 1$, with $r_{uv} = \pm 1$ if and only if $g_u(i) = \pm h_v(j)$. In data analysis, the conclusions that may be made are "the hypothesised independence fails with regard to deviations of order (1, 1), which suggest linearity between the variables", or, perhaps, fails with regard to deviations of order (3, 1), which may loosely be regarded as a kurtosis deviation of the data from what might be expected under independence".

Although the sharp bounds for the correlation are of theoretical interest, they do not seem to be in direct use for inference. If the sample Spearman correlation is 0.75, we may ask, "is this value consistent with a correlation of zero?" (when independence is of interest) but we would far less often ask "is this value consistent with a correlation of one?" (when linearity is the question). It would not be unreasonable to use $\hat{V}_{11}$ to test for a particular value of the corresponding population correlation, ρ_0 say. However in testing for $\rho_0 = 1$, a test for linearity, the asymptotic $N(\rho, 1)$ distribution is clearly inappropriate. Resampling p-values could be obtained, but we suspect this would not be done in practice. The main use of the sharp bounds seems to be in *suggesting* an alternative to independence: if r_{uv} is significantly different from zero, there is a suggestion of a tendency for the data to have a relationship of the form $g_u(i) = \pm h_v(j)$.

10.3.2 Triply Ordered Trivariate Distributions

Suppose now we have a discrete trivariate distribution (X, Y, Z) with joint probabilities $p_{ijk} = P(X = x_i, Y = y_j, Z = z_k)$ for i = 1, ... , I, j = 1, ... , J, and k = 1, ... , K. The strong assumption is made that all three margins are ordered. The following results may be proved as in the bivariate case, or are well known.

- A saturated model for p_{ijk} is, for i = 1, ... , I, j = 1, ... , J, and k = 1, ... , K,

$$p_{ijk} = \left\{ 1 + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} \theta_{uv0}\, a_u(x_i)\, b_v(y_j) + \sum_{u=1}^{I-1}\sum_{w=1}^{K-1} \theta_{u0w}\, a_u(x_i)\, c_w(z_k) + \sum_{v=1}^{J-1}\sum_{w=1}^{K-1} \theta_{0vw}\, b_v(y_j)\, c_w(z_k) + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} \theta_{uvw}\, a_u(x_i)\, b_v(y_j)\, c_w(z_k) \right\} p_{i..}\, p_{.j.} p_{..k},$$

in which the θ_{uvw} are real-valued parameters and, as in section 10.2, $\{a_u(x_i)\}$, $\{b_v(y_j)\}$ and $\{c_w(z_k)\}$ are orthonormal functions on the $\{p_{i..}\}$, $\{p_{.j.}\}$ and $\{p_{..k}\}$ respectively. If all the subscripts u, v and w in θ_{uvw} are nonzero, then the corresponding θ_{uvw} reflect genuine trivariate association between the margins corresponding to these subscripts. If precisely one of the subscripts is zero, then the corresponding θ_{uvw} reflect bivariate association between the margins corresponding to the nonzero subscripts. If precisely two of the subscripts are zero, then so are the corresponding θ_{uvw}. Of course it is possible to identify distributions with characteristics such as pairwise but not mutual independence, when all of the θ_{uvw} with precisely one subscript zero are zero and at least one of the θ_{uvw} with precisely three subscripts nonzero is nonzero.

- For u = 1, ... , I - 1, v = 1, ... , J - 1 and w = 1, ... , K - 1,

$$\theta_{uvw} = E[a_u(X)b_v(Y)c_w(Z)].$$

- A score test statistic for testing H_0: $\theta_{uvw} = 0$ against K: $\theta_{uvw} \neq 0$ is $\hat{V}_{uvw}^2$, where

$$\hat{V}_{uvw} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} N_{ijk}\, a_u(x_i)\, b_v(y_j)\, c_w(z_k)/\sqrt{n},$$

in which the weights in the orthonormal functions are the marginal distributions from the contingency table. In large samples the $\hat{V}_{uvw}$ are independent and each has the N(0, 1) distribution under the null

independence hypothesis. Under the alternative hypothesis, and assuming contiguous alternatives, $\widehat{V}_{uvw}$ has the $N(\theta_{uvw}, 1)$ distribution.

Our previous argument, that the $\widehat{V}_{uv}/\sqrt{n}$ in the bivariate scenario can be regarded as generalised sample correlations, extends immediately to the trivariate scenario and the $r_{uvw} = \widehat{V}_{uvw}/\sqrt{n}$ may also be regarded as generalised sample correlations. However care must be taken with the interpretation of these correlations, for a reason now demonstrated.

Consider the special case X = Y = Z, reflecting linearity. Then

$$\theta_{111} = E[(X - \mu_X)^3]/\sigma_X^3,$$

the standardised skewness coefficient. In spite of the strong linear relationship, θ_{111} can take almost any value, including zero. In this situation we have both linearity and this form of trilinear independence. The trivariate structure permits both. Thus unlike θ_{11}, θ_{111} is not a measure of linearity.

Our preferred interpretation of $\widehat{V}_{uvw}$ is that this statistic reflects the order (u, v, w) variation of the data from what might be expected under independence. An order (u, v, w) variation will occur if the trivariate (u, v, w)*th* moment is different from what might be expected under the independence hypothesis, but may also occur if some higher trivariate moments are different from what might be expected under the independence hypothesis. What is lost relative to the bivariate case is the interpretation that an extreme value of the correlation reflects a relationship like $a_u(x_i) = \pm b_v(y_j)$.

10.3.3 Doubly Ordered Trivariate Distributions

Suppose that the previous models apply, with the exception that the Z margin cannot be ordered. Define, for u = 1, ... , r - 1, v = 1, ... , c - 1 and k = 1, ... , K,

$$\widehat{V}_{uv(k)} = \sum_{i=1}^{I} \sum_{j=1}^{J} N_{ijk}\, a_u(x_i)\, b_v(y_j)/\sqrt{(N_{..k})}.$$

It could be considered that the (u, v)*th* term for the whole table is

$$W_{uv} = \sum_{k=1}^{K} \widehat{V}^2_{uv(k)}.$$

Since under the null independence hypothesis the $\widehat{V}_{uv(k)}$ are asymptotically independent and asymptotically N(0, 1), the W_{uv} asymptotically have the χ^2_K distribution. Provided n is large enough for the asymptotic normality of the $\widehat{V}_{uv(k)}$, $P(W_{uv} > w_{uv} \mid W_{uv}$ is $\chi^2_K)$ is a weight of evidence for accumulated order (u, v) independence for the whole table.

Happiness Example Continued. For the Happiness data previously discussed in sections 8.6 and 10.2, and using natural scores, we calculated the 50 sample correlations r_{uvw} given in Table 10.2.

When the data are considered to be completely ordered (as all three variables are ordinal), the three bilinear correlations r_{110}, r_{101} and r_{011} are all significantly large. Interpreting these as suggesting linearity between the appropriate bivariate marginals it appears that:

- those who finish school early tend to be those with many siblings;
- those with only a few years of schooling are happier than those who have been schooled a long time; and
- those with a few siblings tend to be not as happy as those with many siblings.

There is other structure too, and this refutes the notion that the data are independent. The highly significant r_{210} suggests that as the number of siblings increases, the number of years of completed schooling increases and then decreases. This information, and that from r_{110}, suggests the relationship between years of completed schooling and number of siblings is more complicated than simple linearity.

There is only one significantly large genuine trivariate correlation. With so many correlations being computed, it would not be surprising for this to be the case and for the data to have genuine trivariate independence. It appears that most of the dependence structure is in the marginal distributions.

Suppose now that the years of schooling and the number of siblings are regarded as ordinal variables, while the happiness variable is considered nominal. We then assume apply the doubly ordered analysis. The value of W_{11}, the overall linear-linear term for the row and column categories, is 208.0, which, when compared with the χ^2_3 distribution, is highly significant. There is strong evidence of accumulated bilinear dependence between the row and column categories.

Table 10.2 Correlations r_{uvw} for the happiness data with two-sided p-values in parentheses; correlations with p-values < 0.05 are in bold

(a)	r_{uv0}	u	1	2	3
	v	1	**-0.3701** (0.000)	**0.0961** (0.000)	-0.0163 (0.525)
		2	0.0161 (0.531)	**0.0688** (0.007)	0.0093 (0.717)
		3	-0.0251 (0.330)	0.0011 (0.966)	0.0089 (0.728)
		4	-0.0234 (0.363)	0.0453 (0.078)	-0.0154 (0.545)
(b)	r_{u0w}	u	1	2	3
	w	1	**0.1267** (0.000)	0.0325 (0.206)	-0.0424 (0.099)
		2	**-0.0696** (0.007)	**0.0556** (0.030)	0.0167 (0.515)
(c)	r_{0vw}	w	1	2	
	v	1	**0.0632** (0.014)	**0.0891** (0.001)	
		2	-0.0179 (0.487)	0.0179 (0.485)	
		3	-0.0249 (0.333)	0.0384 (0.135)	
		4	-0.0290 (0.259)	0.0390 (0.129)	
(d)	r_{uv1}	u	1	2	3
		1	-0.0015 (0.954)	0.0463 (0.089)	0.0034 (0.895)
		2	-0.0068 (0.791)	-0.0477 (0.063)	-0.0173 (0.501)
		3	0.0150 (0.559)	0.0096 (0.709)	-0.0106 (0.681)
		4	0.0063 (0.807)	0.0448 (0.081)	0.0137 (0.594)
	r_{uv2}	u	1	2	3
		1	0.0115 (0.654)	-0.0241 (0.3480	-0.0062 (0.811)
		2	-0.0274 (0.286)	**0.0687** (0.007)	-0.0119 (0.645)
		3	-0.0340 (0.186)	0.0304 (0.237)	-0.0348 (0.175)
		4	-0.0216 (0.399)	-0.0003 (0.989)	0.0049 (0.850)

10.4 Partially Parametric Testing

The fundamental premise underlying Carolan (2000) is that a goodness of

fit test has been done as a prelude to a parametric analysis. Suppose for now that the parametric analysis is a one-way analysis of variance, aiming to test if the means of the S samples collected are consistent. If a smooth test for normality is conducted (see section 9.2), a close scrutiny of the data is possible. One model is to assume the s*th* probability density function, s = 1, ... , S, is given by

$$C_s(\xi_{s3}, \ldots, \xi_{sk}) \exp\{\sum_{i=3}^{k_s} \xi_{si} h_i(x;\beta_s)\} f_X(x;\beta_s). \quad (10.1)$$

Here $f_X(x;\beta)$ is the $N(\mu, \sigma^2)$ probability density function, and $\{h_i(x;\beta)\}$ are the standardised Hermite-Chebyshev polynomials (see A.3)

$$h_i(x;\beta) = H_i((x - \mu)/\sigma),$$

where the $\{H_i(z)\}$ are orthonormal using the standard normal distribution as weight function. In this formulation ξ_{si} reflects the i*th* moment of the s*th* sample, so there are no ξ_{s1} and ξ_{s2} terms in (10.1), as means and variances are modelled by the β_s. A common variance can be modelled by taking $\beta_s = (\mu_s, \sigma^2)^T$. If this proves too restrictive, we can take $\beta_s = (\mu_s, \sigma_s^2)^T$, with the intention of using Welch's test, which does not assume a common variance, to assess a common mean. In this formulation $\hat{V}_3$ is a standardised version of the sample skewness and $\hat{V}_4$ is a standardised version of the sample kurtosis.

Analysis of the data may find, for example, that for some of the samples, their $\hat{V}_3$ and $\hat{V}_4$ are significantly large. Now the usual alternatives are, first, to assume that the analysis of variance F test is robust and proceed with the parametric test. The difficulty here is that if the failure of the model is extreme, even a robust procedure cannot be relied upon. The second option is to use a nonparametric test such as the Kruskal-Wallis. Here one problem is that this analysis is less powerful than the parametric analysis when both are appropriate, and the nonparametric analysis may be unable to detect a genuine failure of the model.

One of the early criticisms of Pearson's X_P^2 test was that if the null hypothesis was rejected, no alternative model is available. That is not the case with the smooth tests: (10.1) specifies possible probability density functions. The parametric hypothesis of interest may be tested assuming the richer family of probability density functions. This is thus a third option: to base the analysis on a less restrictive parametric family of probability density functions.

But first, let us take a step backwards. If $\hat{V}_3$ is significant there is evidence of skewness. In this case, the usual measures of location, the mean, median and mode, are all different. The analysis of variance F and the Kruskal-Wallis tests are testing different hypotheses. Before retreating to a nonparametric analysis as an alternative to a parametric one, it is important to consider carefully if the nonparametric analysis is appropriate. It may be that the information provided by the smooth goodness of fit test means the analysis to be undertaken needs reconsideration. Most other goodness of fit tests do not give the detailed information needed to do this. As an example, consider a two-sample problem in which one sample appears symmetric and the other is significantly skewed. Because the mean and median of the skewed sample are well removed from each other, the means of the two samples may be consistent while the medians are not. Blindly applying just the parametric test, "because it is robust," or the nonparametric test, "because it is always applicable," fails to identify an important feature of the data. The original question needs to be revisited in light of this information.

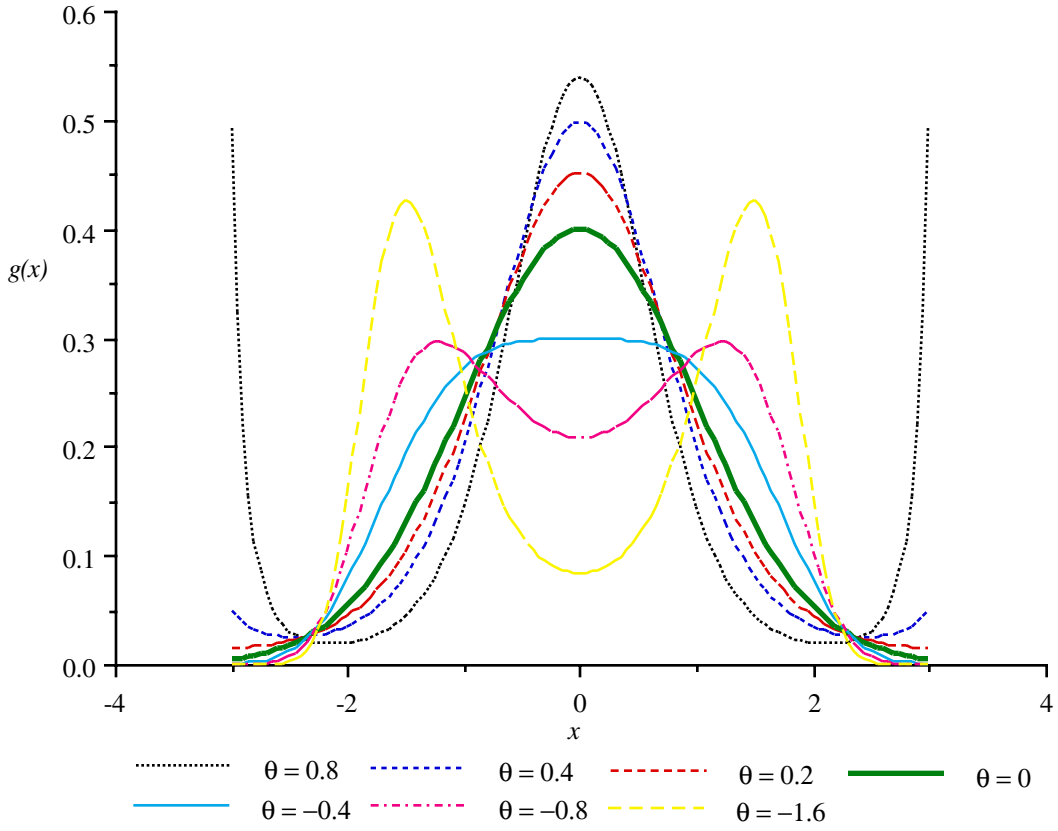
One of the features of basing the analysis on (10.1) is that the larger k_s, the more parameters are involved in the analysis. This ultimately makes that analysis computer-intensive, but if the software is available, the benefit is that much more powerful procedures are available. As a particular k_s increases, the family of densities becomes larger and larger, until for maximal k_s there is no restriction on the underlying parametric family, and the analysis is distribution-free. The term semiparametric is usually used in the context of models that are part parametric and part nonparametric. That is not the case here, where we visualise a continuum between parametric and distribution-free models, with that continuum being indexed by increasing k_s. These models may be reasonably called

partially parametric.

For problems in which symmetry is not a reasonable assumption, Carolan and Rayner (2000a) has focused on a choice of polynomials in place of the $\{h_i(x;\beta)\}$; this permits modes to be compared. Means and medians may still be compared by the standard tests, even if there is a loss of power in doing so. There are tests for the number of modes in the literature, but little about tests concerning the location of the modes.

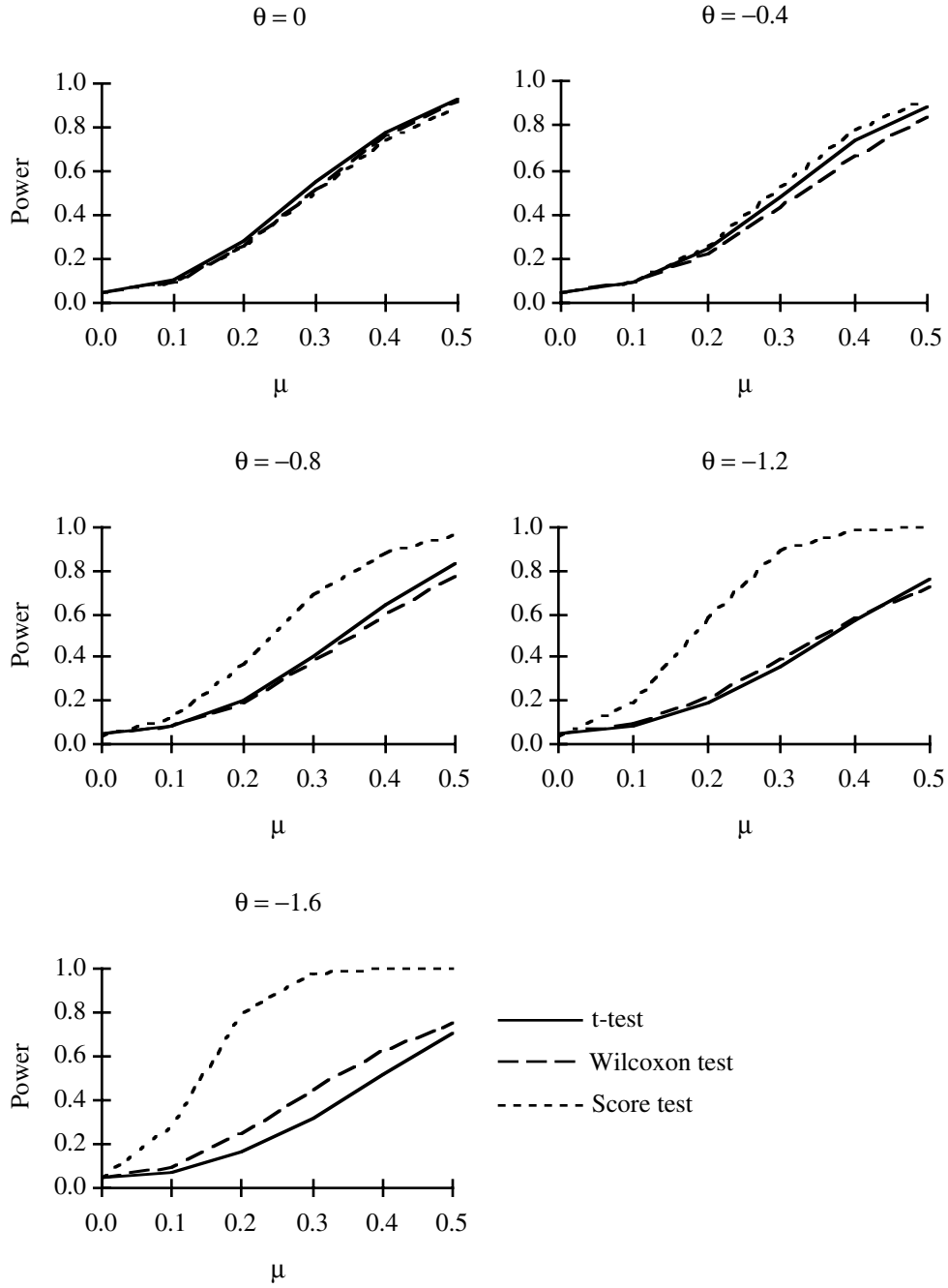
For problems in which symmetry is a reasonable assumption, there will be no terms ξ_{si} in (10.1) with i odd. There may be few or many terms ξ_{si} with i even included in the model, and this provides a rich family of distributions. Compared to being constrained to just the normal, this family gives far more flexibility in obtaining a distribution consistent with the data. Figure 10.1 shows some simple density functions in the family(10.1).

Figure 10.1 Probability density functions $g_4(x; 0, 1, \theta)$ for various values of θ



If a model has been determined, the 'obvious' approach is to use the asymptotically optimal tests. All are quite involved. The likelihood ratio test requires estimation under both the null and the full models, and this makes assessing power by simulation and the calculation of resampling p-values quite time consuming. In power studies there are economies in using the Wald tests, in that fewer maximum likelihood estimations need to be calculated. In addition it seems that for the scenarios considered here, the Wald tests achieve their asymptotic optimality properties more quickly, that is, for smaller sample sizes, than the score tests. Thus the likelihood ratio test is less preferred than the score test, which is less preferred than the Wald test.

Figure 10.2 Comparison of power curves for testing H_0: $\mu = 0$ against K: $\mu \neq 0$ at the 5% level using the t-test, Wilcoxon test and score test for symmetric data becoming progressively more nonnormal. Powers are based on 5,000 simulations from the $g_4(x; \mu, 1, \theta)$ distribution



The efficacy of the partially parametric tests will now be demonstrated for the case of one sample tests of location for nonnormal symmetric data. First, a random sample of size n is assumed to come from a distribution with probability density function

$$g_4(x; \mu, \sigma, \theta) = C(\theta) \exp\{\theta h_4(x; \mu, \sigma)\} f_X(x; \mu, \sigma), -\infty < x < \infty.$$

Without loss of generality we take $\sigma = 1$ and assume $\theta < 0$, since otherwise $g_4(x; \mu, \sigma, \theta)$ does not define a proper probability density function. Densities of this form are shown in Figure 10.1. Powers of the t-test, Wilcoxon test and the score test based on the correct $g_4(x; \mu, \sigma, \theta)$ model are given in Table 10.3 and power curves, based on additional simulations, are shown in Figure 10.2.

Table 10.3 Powers of the t-test, Wilcoxon test and score test based on the correct $g_4(x; \mu, 1, \theta)$ model for testing H_0: $\mu = 0$ against K:$\mu \neq 0$ with a nominal test size of 5% for 5,000 simulated samples of size 50

θ	μ	t-test	Wilcoxon test	Score test
0	0.0	0.048	0.044	0.046
	0.1	0.102	0.097	0.096
	0.2	0.281	0.265	0.257
	0.3	0.557	0.523	0.504
	0.4	0.783	0.760	0.742
	0.5	0.928	0.917	0.898
-0.8	0.0	0.051	0.047	0.041
	0.1	0.087	0.085	0.117
	0.2	0.208	0.194	0.370
	0.3	0.408	0.377	0.686
	0.4	0.647	0.597	0.885
	0.5	0.838	0.778	0.967
-1.6	0.0	0.051	0.050	0.044
	0.1	0.075	0.094	0.281
	0.2	0.166	0.254	0.789
	0.3	0.320	0.446	0.983
	0.4	0.516	0.624	1.000
	0.5	0.711	0.751	1.000

The greater detail in Carolan and Rayner (2000b and 2000c) shows that the t-test and Wilcoxon test are size robust even for small samples. In terms of size robustness, the score test performs reasonably for sample sizes greater than 20. For normal data the t-test has greatest power, which is consistent with it being the uniformly most powerful unbiased test: see Lehmann (1959). When the true model is nonnormal, the score test is best, and the power advantage increases as the true model becomes increasingly nonnormal. It is telling that when $\theta = -1.6$, the score test

power is, for a large part of the parameter space, three and four times that of the t-test. Also as θ decreases and the probability density functions are increasingly nonnormal, the Wilcoxon test overtakes the t-test.

The choice of k_s in (10.1) is critical. Greater power results if smaller k_s are appropriate, but incorrectly choosing k_s to be too small will also lead to loss of power. In applications the smooth tests of goodness of fit may identify significant θ_4, θ_6 and θ_8, but sometimes the significant θ_6 and θ_8 are merely 'harmonics' of the significant θ_4, and the additional terms in the model are not needed. The powers in Table 10.4 involve simulations from the uniform distribution, where it is not clear what the correct choice of k_s (since this is a one sample problem) is.

Table 10.4 Powers of the t-test, Wilcoxon test, score tests S_4 and S_6 and Bartlett corrected Wald tests W_4 and W_6 for testing H_0: $\mu = 0$ against K:$\mu \neq 0$ with a nominal test size of 5% for 5,000 simulated samples of size 50 from the uniform distribution

μ	t-test	Wilcoxon test	S_4	S_6	W_4	W_6
0.0	0.05	0.05	0.04	0.03	0.05	0.05
0.1	0.08	0.08	0.14	0.13	0.14	0.12
0.2	0.21	0.20	0.38	0.32	0.42	0.35
0.3	0.42	0.39	0.67	0.56	0.74	0.66
0.4	0.63	0.58	0.89	0.76	0.93	0.89
0.5	0.83	0.78	0.97	0.79	0.99	0.98

Score and Wald tests S_4 and W_4 were constructed using the model

$$g_4^*(x; \mu, \theta) = C(\theta) \exp\{\theta_2 h_2(x; \mu, 1) + \theta_4 h_4(x; \mu, 1)\} f_X(x; \mu, 1),$$

where $-\infty < x < \infty$ and $\theta = (\theta_2, \theta_4)^T$ (for technical reasons σ has been replaced by θ_2 in the g_4 model) as well as score and Wald tests S_6 and W_6

using the model $g_6^*(x; \mu, \theta)$ given by

$$g_6^*(x; \mu, \theta) = C(\theta) \exp\{\theta_2 h_2(x; \mu, 1) + \theta_4 h_4(x; \mu, 1) + \theta_6 h_6(x; \mu, 1)\} f_\chi(x; \mu, 1), -\infty < x < \infty.$$

The Wald tests are not size robust, but are tolerably so after implementing a Bartlett correction. Powers given in Table 10.4 show that here, for n = 50, the t-test is superior to the Wilcoxon test, and both are uniformly inferior to the score and Wald tests. Since S_4 is superior to S_6, and W_4 is superior to W_6, it appears that for n = 50 only one parameter, θ_4, is needed to model the uniform distribution rather than both θ_4 and θ_6. Moreover, for almost all values of μ, W_4 has superior power to S_4. However the sample size has an effect on such conclusions. For small sample sizes S_4 and W_4 have almost no advantage over the t-test, while for larger samples sizes S_6 and W_6 may be superior to S_4 and W_4.

Carolan (2000) extends the investigation of the partially parametric tests to the one-way analysis of variance and efficiently demonstrates that these tests can achieve huge power gains over their parametric and nonparametric competitors. The message is that any analysis should be implemented only after careful consideration of the data, the questions to be addressed, and the techniques that are relevant and available.

10.5 Concluding Comments

We thank those readers who have persisted to the end of this book. In a sense the material you have read is not so much *modern nonparametrics*, as *modernised traditional nonparametrics*. However there are many new things to consider.

- Our view about *ties* is perhaps very simple and quite important, but may be overlooked. In our view ties should be modelled; scoring categories for data presented in contingency tables does this.
- We have shown here that many common nonparametric tests can be obtained as low order components of Pearson's X_P^2 statistic. This unifies a

body of heavily used nonparametric testing. The subsequent higher order components give potentially very useful extensions that define new nonparametric tests.

- In this way we have given an extended Stuart test for categorical data. This test is linked to generalised Cochran-Mantel-Haenszel tests.
- Extensions of our partitioning technique to three-way tables have been suggested, as have correlation coefficients for trivariate data. Some investigations into partially parametric tests were also reported.
- Further work needs to be done on reconciling our testing approach with tests associated with log-linear models. In particular, links with row effects models, linear by linear association models, Bradley Terry models and the Caussinus test for categorical data could be examined.

We hope workers in many fields will find our new tests useful and will use our techniques to develop further useful statistical methodology.

APPENDICES

This appendix gives definitions and results of relevance in the body of the book. We suggest readers familiar with the topics covered can skim the material to familiarise themselves with the notation.

A.1 Statistical Prerequisites

A.1.1 Binomial and Multinomial Distributions

Fundamental in the sign test is the *binomial distribution*, b(n, p) say, where n is a known constant and p is typically an unknown parameter. The binomial b(n, p) distribution has probability function

$$f_X(x; p) = P(X = x) = {}^nC_x p^x q^{n-x}, \text{ for } x = 0, 1, \ldots, n,$$
$$\text{in which } p = 1 - q \in (0, 1),$$

and has mean np and variance npq. For n large and p not extreme, the binomial may be approximated by the normal distribution with the same mean np and variance npq as the binomial: N(np, npq). Thus

$$P(X = x) = P\left(\frac{x - np - 0.5}{\sqrt{npq}} < Z < \frac{x - np + 0.5}{\sqrt{npq}} \mid Z \text{ is } N(0, 1)\right).$$

Any discrete distribution that is approximated by a continuous distribution has a *continuity correction*, which is a misnomer: a continuity *adjustment* would be more correct, as it must always be used. If X takes integer values, the probability P(X = x) is treated as the area of a rectangle with base (x - 0.5, x + 0.5) and height P(X = x), and is approximated by the area under the probability density function of the N(np, np(1 - p)) distribution, between x - 0.5 and x + 0.5. The normal approximation to the binomial

above is a typical example.

In the *multinomial distribution*, $m(n, p_1, \ldots, p_m)$ say, n is a known constant and the p_i are probabilities. The multinomial has probability function

$$P(N_1 = n_1, \ldots, N_m = n_m) = \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m}$$
$$\text{for } n_i = 0, 1, \ldots, n,\ i = 1, \ldots, m,\ n_1 + \ldots + n_m = n,$$
$$\text{in which } p_i \in (0, 1),\ i = 1, \ldots, m \text{ and } p_1 + \ldots + p_m = 1.$$

The N_i have means np_i, $i = 1, \ldots, m$, variances $np_i(1 - p_i)$, $i = 1, \ldots, m$, and the covariance between N_i and N_j is $\text{cov}(N_i, N_j) = -np_ip_j$, $i \neq j$. Thus the vector variate $N = (N_1, \ldots, N_m)^T$ has mean $n(p_1, \ldots, p_m)^T$ and covariance matrix

$$(\text{cov}(N_i, N_j)) = n\{\text{diag}(p_1, \ldots, p_m) - (p_ip_j)\}.$$

A.1.2 Hypothesis Testing

Hypothesis testing may be viewed as a choice between two decisions: to accept or to reject the null hypothesis. A multiple decision procedure allows other possibilities, such as to continue sampling in order to obtain more data, or to stop sampling and reserve judgement. In hypothesis testing a significance level α is set, usually at 10%, 5%, 1% or 0.1%. When the null hypothesis is *simple*, specifying a single value of a parameter, this α is the probability of rejection of the null hypothesis when in fact it is true.

A valid but old-fashioned approach to hypothesis testing would be to construct a rejection region that depends on α, and a complementary acceptance region, and to accept/reject the null hypothesis when the value of the test statistic falls in the acceptance/rejection region.

It is more usual nowadays to find the *p-value*, *the probability of observations at least as extreme as the observed*. If the p-value is less than α, the null hypothesis is rejected at the 100α% level. Otherwise it is accepted at this level.

Thus the p-value may be used to assess the consistency of the data with the null hypothesis, bearing in mind the alternative hypothesis (see, for example, Cox and Hinkly, 1974, Chapter 3). Hence a p-value of 30% may be viewed as comfortable acceptance of the null hypothesis at the 5% level, whereas a p-value of 6% accepts the null hypothesis, but suggests to researchers a more detailed followed up study with careful control of more factors. A p-value of 8% indicates rejection at the 10% level, but not the confident resounding rejection a p-value of 0.001% may indicate. Moreover if the statistician reports a p-value of 3%, different users who may use different significance levels can come to their own conclusions; this isn't possible if, for example, only "accept at the 5% level" is reported. This is sometimes called the "data analytic" approach to inference, rather than a more rigid formal hypothesis testing approach.

Competing tests of the same null hypothesis tested against the same alternative when both tests have the same level of significance may be compared using the *powers* of the tests. Power is the probability of rejection for a given alternative. The power function expresses the power as a function of the alternatives, and is often presented graphically. More power is desirable: so in a sense, statisticians should be megalomaniacs! In general the greater the level of significance the greater the power; so it is not possible to compare tests with different levels of significance.

Test efficiency should not be confused with estimation efficiency. Estimation efficiency is the quotient of the variances of unbiased estimators of a given parameter. The idea is that less variability of the estimates around the true value is desirable. There are many different definitions of test efficiency. For tests with the same level of significance, Pitman's *asymptotic relative efficiency* is the limit of the quotient of the sample sizes required for the tests to achieve the same power. Here the idea is that a test that requires a smaller sample size to achieve a given power will save the cost of the additional data and so is more desirable.

Efficiency is an omnibus measure of the desirability of a test. In a sense, it summarises the information in the power function. There are other considerations in choosing between tests. For example, a test statistic that takes on relatively few values will achieve relatively few levels of significance. A test that has more achievable levels will, in general, be able to get closer to nominated levels such as 1% and 5%, and if observations are not expensive, this may be a more important

consideration than power.

A.1.3 Pearson's Goodness of Fit Test

In Pearson's test of goodness of fit we have counts $N_1, \ldots, N_m$ that follow a multinomial $m(n, \pi_1, \ldots, \pi_m)$ distribution. The aim is test the null hypothesis H_0: $(\pi_1, \ldots, \pi_m) = (p_1, \ldots, p_m)$ against K: $(\pi_1, \ldots, \pi_m) \neq (p_1, \ldots, p_m)$. Pearson's test is based on the statistic

$$X_P^2 = \sum_{j=1}^{m} \frac{(N_i - np_i)^2}{np_i}.$$

Under H_0 the test statistic X_P^2 has the χ^2_{m-1} distribution. Large values indicate rejection of H_0.

Many writers refer to this test as Pearson's χ^2 test. We consider it undesirable to identify a test by the asymptotic distribution of its test statistic. After all, many test statistics have the χ^2 distribution. Better identifiers would be Pearson's test, Pearson's X_P^2 test, or Pearson's goodness of fit test.

Birth Times Example. Rayner and Best (1989a, pp. 13-15 and pp. 52-54) give 37 birth times originally given by Mood et al. (1974, p. 509). The hypothesis of interest is that births occur uniformly throughout the day. Grouping the data into three classes, there are 10 births from midnight to 8 am, 18 from 8 am to 4 pm, and 9 from 4 pm to midnight. For these data

$$X_P^2 = \sum_{j=1}^{m} \frac{(N_i - np_i)^2}{np_i} = \sum_{j=1}^{m} \frac{N_i^2}{np_i} - n = \frac{10^2 + 18^2 + 9^2}{\left(\frac{37}{3}\right)} - 37 = 3.9459.$$

The exact p-value is 0.16. There is no evidence, at the 5% level, that birth times are not uniformly distributed.

A.1.4 Components and Contrasts

The parameter space for Pearson's test is m - 1 dimensional. In as much as this test seeks alternatives in all m - 1 dimensions, it is an *omnibus* test, and can be expected to have only moderate power throughout the parameter space. More focused or *directional* tests seek alternatives in fewer dimensions and have superior power for some alternatives, and weaker power for others. We now describe some directional tests that complement Pearson's test.

Suppose $\{g_r(j)\}$ are orthonormal on $\{p_j\}$, so that

$$\sum_{j=1}^{m} g_u(j)g_v(j)p_j = \delta_{uv}.$$

Here δ_{uv} is the Kronecker delta: $\delta_{uv} = 1$ if $u = v$ and $= 0$ if $u \neq v$. More detail on orthonormal functions is given in A.3. Now put $g_0(j) = 1$ for $j = 1, \ldots, m$ and

$$V_r = \sum_{j=1}^{m} N_j g_r(x_j)/\sqrt{n} \quad \text{for } r = 1, \ldots, m - 1.$$

We can show that the *components* V_r are asymptotically mutually independent and asymptotically have the standard normal distribution. With some caveats that we discuss later, if the categories have a natural ordering (for example, height or weight), and the $\{g_r(j)\}$ are the set of orthonormal polynomials on $\{p_j\}$, then the component V_r detects r*th* moment departures of the data from the distribution $\{p_j\}$.

Birth Times Example Continued. In A.3 it will be shown that for $m = 3$ $g_1(j)$ takes the values $-\sqrt{\frac{3}{2}}$, 0 and $\sqrt{\frac{3}{2}}$ for $j = 1, 2$ and 3, while $g_2(j)$ takes the values $\frac{1}{\sqrt{2}}$, $-\frac{2}{\sqrt{2}}$ and $\frac{1}{\sqrt{2}}$ for $j = 1, 2$ and 3. It follows that

$$V_1 = \sqrt{\frac{3}{2}}\,(-10 + 0 + 9)/\sqrt{37} = -0.2013 \text{ and}$$
$$V_2 = \frac{1}{\sqrt{2}}\,(10 - 36 + 9)/\sqrt{37} = -1.9762.$$

The corresponding χ^2_1 p-values are 0.727 and 0.048. There is no evidence of a mean shift from what would be expected under the hypothesis of uniformity, but there is some evidence of a variability shift. There is a greater concentration of births in the daylight hours than uniformity would predict.

In essence the m - 1 components V_r each assess a different dimension of the m - 1 dimensional parameter space for departures from the null, while X^2_P assesses all dimensions simultaneously. Using two to four components in a data analytic fashion will give a close and informative scrutiny of the data. Alternatively the statistic $V_1^2 + V_2^2 + V_3^2 + V_4^2$ will generally result in more power than Pearson's test, as most alternatives that arise in practice tend to project into at most four dimensions. Further consideration of components is given later in section A.6.

Some readers will be familiar with *contrasts* from the analysis of variance. Contrasts have similar properties to our components, in that they split an omnibus statistic up into parts. Both components and contrasts:

- are at least asymptotically mutually independent;
- have sum, or sum of squares, the original omnibus statistic;
- have, at least asymptotically, convenient distributions; and
- have an immediate and useful interpretation.

In addition, contrasts achieve another useful purpose: comparison. Components can be constructed to do this also. This can be achieved if, for example, the orthonormal functions are based on the Helmert matrix given in section A.4. More usually, effects are compared using a method of *paired comparisons*, such as the least significant difference (LSD) technique. If $X_1, \ldots, X_n$, are mutually independent with X_i having the $N(\mu_i, \sigma^2)$ distribution, then X_i and X_j may be judged to be different at the ${}^nC_2\alpha$ level if $|X_i - X_j| > 2t_{df}(\alpha)S = \text{LSD}$, where S^2 is an estimate of σ^2 based on df degrees of freedom. It is not unusual to use the LSD approach

when some of these assumptions hold only approximately. This usually needs an extensive simulation study to confirm that the technique is valid to an acceptable approximation.

A.1.5 Maximum Likelihood Estimation

The joint probability of the observations, regarded as a function of a vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_q)^T$, is called the *likelihood function* of the sample. For n independent observations from the same distribution the likelihood $L(\theta \mid x)$ is given by

$$L(\theta \mid x) = f_X(x_1 \mid \theta) * f_X(x_2 \mid \theta) * \ldots * f_X(x_n \mid \theta),$$

where $f_X(x \mid \theta)$ is the probability density function of the sampled distribution.

The maximum likelihood (ML) principle says choose θ to be that value, $\hat{\theta}$, say, that makes $L(\theta \mid x)$ as large as possible:

$$L(\hat{\theta} \mid x) \geq L(\theta \mid x) \text{ for all } \theta.$$

We call $\hat{\theta}$ the maximum likelihood estimator (MLE) of θ. Suppose Θ is the parameter space, the set of admissible values of θ. If Θ is discrete then different techniques for maximising the likelihood will be required than if Θ is continuous. This will be obvious in the examples that follow.

Suppose now that Θ is one dimensional, an interval, and that $L = L(\theta \mid x)$ is continuous with respect to θ and twice differentiable. Provided L does not have a terminal maximum, the estimator $\hat{\theta}$ occurs at a value of

$$L' = dL/d\theta = 0$$

for which $L'' < 0$. It is often more convenient to work with $\ell(\theta) = \ell = \log L$. This follows because

$$d\ell/d\theta = \{dL/d\theta\}/L,$$

so that $\ell' = 0$ if and only if $L' = 0$. But since

$$\ell'' = \{LL'' - (L')^2\}/L^2,$$

when $\ell' = 0$ (and also $L' = 0$), $\ell'' = L''/L$. At stationary points $\ell' > (<) \; 0$ as $L' > (<) \; 0$. That is, stationary points of L and ℓ occur together.

It is, of course, possible that there are many stationary points. These must then be compared to find the maximum.

Normal Maximum Likelihood Example. Find $\hat{\sigma}$, the maximum likelihood estimator of σ when sampling from a $N(\mu_0, \sigma^2)$ distribution, with μ_0 known.

$$\begin{aligned} L &= \prod_{i=1}^{n} \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x_i - \mu_0)^2}{2\sigma^2}\} \right\} \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2\}, \end{aligned}$$

so that

$$\begin{aligned} &\ell = -(n/2) \log(2\pi) - n \log \sigma - \Sigma_i (x_i - \mu_0)^2/(2\sigma^2), \\ &\ell' = -n/\sigma + \Sigma_i (x_i - \mu_0)^2/\sigma^3 \text{ and} \\ &\ell'' = n/\sigma^2 - 3\Sigma_i (x_i - \mu_0)^2/\sigma^4. \end{aligned}$$

We find ℓ' equals zero if and only if $\sigma = \sqrt{\{\Sigma_i (x_i - \mu_0)^2/n\}} = \hat{\sigma}$. Then

$$\ell'' = n/\hat{\sigma}^2 - 3n\hat{\sigma}^2/\hat{\sigma}^4 = -2n/\hat{\sigma}^2 < 0,$$

so $\hat{\sigma}$ is the maximum likelihood estimator. It is worthwhile to observe, by putting $\theta = \sigma^2$ and repeating the process, that the maximum likelihood

estimator of σ^2 is $(\hat{\sigma})^2$.

Hypergeometric Maximum Likelihood Example. From an unknown number N of fish in a pond, W are caught and marked. After returning the fish to the pond and allowing them to mix, a sample of size n is caught; suppose x of these are marked. This event has probability

$$L_N = {}^{W}C_x \, {}^{N-W}C_{n-x} \, / \, {}^{N}C_n$$

and we wish to find the maximum likelihood estimator $\widehat{N}$ of N. Clearly N takes values 1, 2, 3, ... , so that calculus may not be used to find $\widehat{N}$. Now

$$L_N / L_{N-1} = \{{}^{W}C_x \, {}^{N-W}C_{n-x} \, / \, {}^{N}C_n\} / \{ {}^{W}C_x \, {}^{N-1-W}C_{n-x} \, / \, {}^{N-1}C_n\}$$
$$= \frac{(N - W)\,(N - n)}{(N - W - n + x)\,N}$$

which is > (<) 1 as $nW/x > (<) N$. If N_0 is the largest integer less than nW/x (when this is not an integer), then $\widehat{N} = N_0$.

Double Exponential Maximum Likelihood Example. We seek the maximum likelihood estimator $\widehat{\theta}$ of the location parameter from a double exponential distribution, which has probability density function

$$f_X(x \mid \theta) = \{\exp(-\mid x - \theta \mid)\}/2, \; -\infty < x < \infty,$$

in which $\theta \in (-\infty, \infty)$. Here

$$\ell = -\, n \log 2 - \Sigma_i \mid x_i - \theta \mid .$$

Write the order statistics as $Y_1 < Y_2 < ... < Y_n$. Then $\ell = \ell(\theta)$ is continuous everywhere, and differentiable everywhere except at the values $\theta = y_1, y_2, ... , y_n$. If $\theta < y_1$ then

$$\ell = - n \log 2 + (\theta - y_1) + (\theta - y_2) + ... + (\theta - y_n),$$

and $\ell' = n > 0$. Similarly $\ell' = - n < 0$ for $\theta > y_n$. For $\theta \in (y_k, y_{k+1})$,

$$\ell = - n \log 2 - (\theta - y_1) - (\theta - y_2) - ... - (\theta - y_k) + (\theta - y_{k+1}) + (\theta - y_{k+2}) + ... + (\theta - y_n),$$

so that

$$\ell' = - k + (n - k) = n - 2k.$$

If $n = 2r + 1$, ℓ' takes successive values $2r + 1, ... , 5, 3, 1, -1, -3, ... , -(2r + 1)$. We find $\hat{\theta} = Y_{r+1}$. If $n = 2r$, $\hat{\theta}$ is any value in (Y_r, Y_{r+1}), as a sketch of ℓ against θ will show.

Cauchy Maximum Likelihood Example. For the Cauchy distribution with

$$f_X(x \mid \theta) = \{\pi[1 + (x - \theta)^2]\}^{-1} \text{ for } -\infty < x < \infty,$$
$$L(\theta \mid x) = \pi^{-n} \prod \{1+(x_i - \theta^2)\}^{-1}, \text{ so that}$$

$$\ell = - n \log \pi - \Sigma_i \log \{1 + (x_i - \theta)^2\}, \text{ and}$$
$$\ell' = \Sigma_i \, 2(x_i - \theta)/\{1 + (x_i - \theta)^2\}.$$

$\ell' = 0$ can only be solved numerically. Newton-Raphson is most frequently used, with the sample median as the starting point. In general the equation has multiple roots and the problem is to find the absolute maximum among the relative maxima.

A.1.6 Scales of Measurement

Typically data are available on four scales of measurement.

- *Nominal*. Variables differ in kind rather than amount. The data are

categorized, for example, into colour or gender.

• *Ordinal*. Ordinal scales are based on qualitative rather than quantitative variables; however some ordering is also implied. This frequently involves ranks.

• *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. Probabilities are an example of the ratio scale, and temperatures are an example of the interval scale.

Parametric procedures typically involve interval or ratio scales. They are not available for nominal and ordinal scales, but nonparametric procedures are available for all measurement scales. Nonparametric procedures also provide an alternative analysis for data analysed by traditional parametric - usually normal theory - methods. When the parametric assumptions are valid, there is usually some loss of power and efficiency in using nonparametric methods rather than the corresponding parametric methods. Balancing this, there is considerable security in knowing that the nonparametric procedures are available when the parametric assumptions are in any way dubious.

Some statisticians retreat to nonparametric methods even when there is considerable parametric structure that could be taken advantage of to produce more powerful analyses. We do not recommend use of nonparametric methods when there are more powerful and valid parametric tests possible. By valid we mean that the assumptions have been assessed and found to be appropriate. See A.6 for assessment of univariate normality, and Rayner and Best (1989a) for a broader assessment: one sample smooth tests of goodness of fit.

A.2 Orthogonal Matrices

Linear transformations of vector random variables will be an important tool in developing nonparametric extensions. These transformations often involve orthogonal matrices. The material in this section is usually covered in undergraduate linear algebra courses.

- The n by n *unit matrix*, I_n, has ones on the diagonal and zeros elsewhere:

$$(I_n)_{rr} = 1 \text{ and } (I_n)_{rs} = 0 \text{ for } r \neq s.$$

- The *Kronecker delta* δ_{rs} is such that $\delta_{rs} = 1$ if r = s and zero otherwise. It follows that I_n is the unit matrix with $(I_n)_{rs} = \delta_{rs}$.
- The n by 1 vector with all elements 1 is written 1_n: $(1_n)_r = 1$.
- The *trace* of a matrix is the sum of its diagonal elements. Thus

$$\text{tr } A = \sum_i (A)_{ii}.$$

- *Trace is invariant under cyclic rotation*. Hence

$$\text{tr } AB = \sum_i \sum_j a_{ij} b_{ji} = \sum_i \sum_j b_{ij} a_{ji} = \text{tr } BA$$

and, similarly,

$$\text{tr } ABC = \text{tr } BCA = \text{tr } CAB.$$

- An n by n matrix $H = (h_{rs})$ is *orthogonal* if and only if $H^T H = I_n$.
- It follows from the definitions of orthogonal and a matrix inverse that

$$H^{-1} = H^T,$$

so that, for example, $HH^T = I_n$ also.

- The *Helmert* matrices are useful orthogonal matrices used in statistics to construct contrasts in the analysis of variance. The 4 by 4 standard Helmert matrix is

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{3}{\sqrt{12}} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The general form of the standard Helmert matrix has last row

$$(\frac{1}{\sqrt{n}}, \ldots , \frac{1}{\sqrt{n}})$$

and r*th* row one, r times, followed by -r and the remaining entries zero, all normalised:

$$(\frac{1}{\sqrt{r(r+1)}}, \ldots , \frac{1}{\sqrt{r(r+1)}}, -\frac{r}{\sqrt{r(r+1)}}, 0, \ldots, 0).$$

See Lancaster (1965).

- The *eigenvalues* of the matrix A are the solutions λ to

$$Ax = \lambda x, \text{ where } x \neq 0.$$

The vector x in this equation is the corresponding *eigenvector*. Eigenvalues and eigenvectors are sometimes known as latent roots and latent vectors, and characteristic roots and characteristic vectors.

- The *determinant* of a matrix is the product of its eigenvalues. Clearly if A has eigenvalues $\lambda_1, \ldots , \lambda_n$, then $\det A = \prod_{i=1}^{n} \lambda_i$.

- All real symmetric matrices are *diagonalisable*. It follows that for a symmetric n by n matrix A with eigenvalues $\lambda_1, \ldots , \lambda_n$ say, there exists an orthogonal matrix H such that

$$H^T A H = \text{diag}(\lambda_1, \ldots, \lambda_n).$$

- The columns of H are the eigenvectors of A.
- The *rank* of a matrix is the number of non-zero eigenvalues.
- A matrix is *idempotent* if it is equal to its own square: $A^2 = A$.
- *The eigenvalues of an n by n real symmetric idempotent matrix of rank r are one r times and zero n - r times*. Suppose A is real, symmetric and idempotent. There exists an orthogonal matrix H that diagonalises A, so that

$$H^T A H = \text{diag}(\lambda_1, \ldots, \lambda_n) = D \text{ say. Then}$$

$$D = H^T A H = H^T A^2 H = H^T A H H^T A H = D^2, \text{ and}$$
$$\text{diag}(\lambda_1, \ldots, \lambda_n) = \text{diag}(\lambda_1^2, \ldots, \lambda_n^2),$$

so that $\lambda_i^2 = \lambda_i$, for all i, and all eigenvalues are either zero or one. Since A is of rank r, r eigenvalues are one; the remaining (n - r) eigenvalues must be zero.

- *The rank of a real symmetric idempotent matrix is equal to its trace*. For if A is such a matrix, and diagonalisable by the orthogonal matrix H, then

$$\text{tr } A = \text{tr } AHH^T = \text{tr } H^T A H = \text{tr diag}(1, \ldots 1, 0, \ldots, 0) = \text{rank } A.$$

- Suppose u_n is an n by 1 vector with $u_n^T u_n = 1$. Direct multiplication shows that $I_n - u_n u_n^T$ is idempotent. Its rank is equal to its trace, which is

$$\text{tr } (I_n - u_n u_n^T) = \text{tr } (I_n) - \text{tr } (u_n u_n^T) = n - \text{tr } (u_n^T u_n) = n - 1.$$

A.3 Orthonormal Functions

The main tool we use to exploit the ordering of categories for data presented in contingency tables is orthonormal functions. For discrete random variables, our main interest, there is a close connection between

orthogonal matrices and orthonormal functions.

A set $\{g_r(x)\}$ of functions of x are *orthogonal* with respect to the probability density function $f_X(x)$ if and only if

$$\int_{-\infty}^{\infty} g_r(x)g_s(x)\, f_X(x)\, dx = 0.$$

If each $g_r(x)$ satisfies a *normality* constraint as well, namely that

$$\int_{-\infty}^{\infty} g_r^2(x)\, f_X(x)\, dx = 1 \text{ for all r,}$$

then $\{g_r(x)\}$ is *orthonormal*. The orthonormality conditions may be written compactly as

$$\int_{-\infty}^{\infty} g_r(x)g_s(x)\, f_X(x)\, dx = \delta_{rs},$$

where again, δ_{rs} is the Kronecker delta. If $f_X(x) = 1$ for $0 < x < 1$, zero otherwise, the continuous uniform (0, 1) distribution, the first few polynomials are zero outside of (0, 1), and in (0, 1) are given by

$$g_0(x) = 1,\ g_1(x) = \sqrt{3}\,(2x - 1),\ g_2(x) = \sqrt{5}\,(6x^2 - 6x + 1),$$
$$g_3(x) = \sqrt{7}\,(20x^3 - 30x^2 + 12x - 1),$$
$$g_4(x) = 3(70x^4 - 140x^3 + 90x^2 - 20x + 1).$$

These are the Legendre polynomials.

If X is the standard normal distribution N(0, 1), the first few Hermite or Hermite-Chebyshev polynomials have domain $(-\infty, \infty)$ and are given by

$$g_0(x) = 1,\ g_1(x) = x\ ,\ g_2(x) = (x^2 - 1)/\sqrt{2}\ ,$$

$$g_3(x) = (x^3 - 3x)/\sqrt{6}, \text{ and } g_4(x) = (x^4 - 6x^2 + 3)/\sqrt{24}.$$

For information about orthonormal functions in general, see Abramowitz and Stegun (1972). Rayner and Best (1989a, Appendix 1) gave information about the orthogonal polynomials for the standard normal, Poisson, exponential, binomial, geometric, and the continuous and discrete uniform distributions.

Here we are mainly interested in discrete random variables on finitely many points. Suppose X has probability function $p_X(x_j)$, j = 1, ... , m. In this case *orthonormality* reduces to the conditions

$$\sum_{j=1}^{m} g_r(x_j) g_s(x_j)\, p_X(x_j) = \delta_{rs}.$$

Note that if $H = (h_{ij})$ is an m by m orthogonal matrix, then

$$\sum_{j=1}^{m} h_{jr}\, h_{js} = \delta_{rs}.$$

Now take X to be a discrete uniform random variable with probability function $p_X(x_j) = m^{-1}$, j = 1, ... , m, and $g_r(x_j) = h_{ij}\sqrt{m}$. Then the orthonormality conditions are satisfied. Conversely, if given a set of orthonormal functions $\{g_r(x_j)\}$ with weight $\{p_X(x_j)\}$, then we can construct an orthogonal matrix H by putting $h_{jr} = g_r(x_j)\sqrt{p_X(x_j)}$.

Readers unfamiliar with orthonormal functions should take comfort from this result. In our setting orthonormal functions and orthogonal matrices are essentially equivalent.

Orthonormal polynomials are perhaps the simplest orthonormal functions to construct. Suppose X is the discrete uniform random variable with probability function $p_X(x_j) = m^{-1}$, j = 1, ... , m. The order zero orthonormal polynomial may be taken to be one; the first orthonormal polynomial is aj + b, and must satisfy

$$E[g_0(X)g_1(X)] = E[g_1(X)] = \sum_{j=1}^{m} (aj + b)/m = 0 \text{ and}$$

$$E[g_1^2(X)] = \sum_{j=1}^{m} (aj + b)^2/m = 1$$

Solving these equations gives

$$a = \sqrt{\frac{12}{m^2 - 1}} \text{ and } b = -\sqrt{\frac{3\,(m + 1)}{(m - 1)}} = -\frac{(m + 1)}{2}\,a, \text{ so that}$$

$$g_1(j) = \sqrt{\frac{12}{m^2 - 1}}\left(j - \frac{m + 1}{2}\right).$$

On m points at most m orthogonal functions may be defined; so $g_2(j)$ is only defined for $m > 2$. Solving the orthogonality and normality conditions leads to

$$g_2(j) = \sqrt{\frac{180}{(m^2 - 1)(m^2 - 4)}}\left\{\left[j - \frac{m + 1}{2}\right]^2 - \frac{m^2 - 1}{12}\right\}.$$

Similarly,

$$g_3(j) = \sqrt{\frac{2800}{(m^2 - 1)(m^2 - 4)(m^2 - 9)}}\left\{\left[j - \frac{m + 1}{2}\right]^3 - \frac{3m^2 - 7}{20}\left[j - \frac{m + 1}{2}\right]\right\}.$$

The calculation of subsequent functions in this way is tedious. Tables of orthonormal polynomials may be found in Fisher and Yates (1970, Table XXIII) and in Pearson and Hartley (1970, Table 47). Recurrence formulae particularly suitable for computer generation may be found in Emerson (1968). Lancaster (1969, p. 49) gave a compact formula involving determinants.

It is useful to generalise $g_1(j)$ and $g_2(j)$ above. Suppose X takes values x_j with probabilities p_j, $j = 1, \ldots, m$. Put

$$\mu = \sum_{j=1}^{m} x_j p_j \text{ and } \mu_r = \sum_{j=1}^{m} (x_j - \mu)^r p_j.$$

Then the first three orthonormal polynomials are, for j = 1, ... , m,

$$g_0(x_j) = 1,\ g_1(x_j) = (x_j - \mu)/\sqrt{\mu_2} \text{ and}$$
$$g_2(x_j) = a\{(x_j - \mu)^2 - \mu_3(x_j - \mu)/\mu_2 - \mu_2\}, \text{ in which}$$
$$a = (\mu_4 - \mu_3^2/\mu_2 - \mu_2^2)^{-0.5}.$$

A.4 Direct Sums and Products

Throughout the main body of the text, mainly two techniques are used. One is to set up a model and use the asymptotically optimal tests to arrive at test statistics: see section A.5. The other is to set the data up in a table of counts, and then put those counts into a single vector. Typically this vector is asymptotically multivariate normal (see A.7) with readily accessible mean and covariance matrix that can be expressed in terms of direct sums or, more usually, direct products. Diagonalisation of the covariance matrix using orthogonal matrices (see A.2) then enables us to find the distribution of important statistics for testing some interesting and important hypotheses. Some definitions and properties of direct sums and direct products are now given.

The *direct sum* of two square matrices A and B, of sizes m by m and n by n respectively, is the square matrix of size (m + n) by (m + n), given by

$$A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

It is routine to show that

$$A \oplus (B \oplus C) = (A \oplus B) \oplus C;$$

$$(A \oplus B) + (C \oplus D) = (A + C) \oplus (B + D);$$

$$(A \oplus B)(C \oplus D) = AC \oplus BD;$$

$$(A \oplus B)^T = A^T \oplus B^T;$$

$$I_m \oplus I_m = I_{m+n};$$

$$(A \oplus B)^{-1} = A^{-1} \oplus B^{-1} \text{ and}$$

$$\det(A \oplus B) = \det(A)\det(B).$$

For example, $(A \oplus B)(A^{-1} \oplus B^{-1}) = AA^{-1} \oplus BB^{-1} = I_{m+n}$.

The *direct product* of two square matrices A and B, of sizes m by m and n by n respectively, is the square matrix of size mn by mn, given by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{pmatrix}.$$

If $A = (a_{ij})$ and $B = (b_{rs})$, then the typical elements of each side are $a_{ij}b_{rs}$ and $(A \otimes B)_{(i-1)n+r,\,(j-1)n+s}$.

It may be shown that

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C;$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD;$$

$$(A \otimes B)^T = A^T \otimes B^T;$$

$$I_m \otimes I_m = I_{mn};$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1};$$

$$\mathrm{tr}(A \otimes B) = \mathrm{tr}A\ \mathrm{tr}B \text{ and}$$

$$\det(A \otimes B) = \{\det(A)\}^m\ \{\det(B)\}^n.$$

It is useful to define the direct product of two vectors so that the elements of the product are arranged in dictionary or lexicographic order. Thus if $x = (x_1, \ldots, x_m)^T$ and $y = (y_1, \ldots, y_n)^T$, define

$$(x \otimes y)^T = (x_1y_1, \ldots, x_1y_n, \ldots, x_my_1, \ldots, x_my_n)^T.$$

Now suppose that A has eigenvalue λ with corresponding eigenvector x, and B has eigenvalue μ with corresponding eigenvector y. This means for $x \neq 0$, $Ax = \lambda x$ and for $y \neq 0$, $By = \mu y$. It follows that $A \otimes B$ has eigenvalue $\lambda\mu$ with corresponding eigenvector $x \otimes y$, since $x \otimes y \neq 0$ and

$$(A \otimes B)\,(x \otimes y) = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mm}B \end{pmatrix} \begin{pmatrix} x_1y \\ x_2y \\ \vdots \\ x_my \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}x_1By & a_{12}x_2By & \cdots & a_{1m}x_mBy \\ a_{21}x_1By & a_{22}x_2By & \cdots & a_{2m}x_mBy \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}x_1By & a_{m2}x_2By & \cdots & a_{mm}x_mBy \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}x_1 & a_{12}x_2 & \cdots & a_{1m}x_m \\ a_{21}x_1 & a_{22}x_2 & \cdots & a_{2m}x_m \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}x_1 & a_{m2}x_2 & \cdots & a_{mm}x_m \end{pmatrix} \otimes By$$

$$= Ax \otimes By = \lambda x \otimes \mu y = (\lambda\mu)\,(x \otimes y).$$

Direct sums and products are sometimes known as *Kronecker* sums and products respectively.

A.5 Likelihood Ratio, Score, and Wald Tests

The likelihood ratio, score, and Wald tests are all asymptotically optimal: they provide tests with good properties in large samples. Precisely what these properties are will not be discussed here. See Rayner and Best (1989a, section 3.4), and the references there. Intuitively speaking, an optimal test is one that is as critical of the data as is possible. Best use is made of the available data, so that 20, say, observations are sufficient to give an specified chance of detecting some alternative, whereas an alternative test may require 30 or 40 observations. Hence the use of optimal procedures assists efficient use of information.

The study of "small sample" optimality, when the sample size is fixed, seems to have lost its impetus; see Lehmann (1959). One difficulty is that there are several notions of optimality, with some being convoluted and unnatural. One that could be interpreted this way is seen in Neyman's (1937) quest for tests that are locally uniformly most powerful unbiased and symmetric of size α. Another problem is that many statisticians do not accept the Neyman-Pearson approach inherent in the quest for small sample optimality: fixing the test size and then maximising the power subject to appropriate restrictions. Whatever objections one

may have to this philosophy, it is hard to deny that it has been extremely productive over the years. Our approach is to assess the properties of tests from the Neyman-Pearson perspective, looking at size and power properties. The desirable tests so found can be applied in a data analytic way. No data analyst can afford to lose too much power in the pursuit of simplicity or convenience.

The asymptotically optimal tests all have the same good asymptotic properties and asymptotic χ^2 distributions. However the likelihood ratio test requires estimation under both the null and alternative hypotheses, whereas the score test requires estimation only under the null hypothesis, and the Wald test requires estimation only under the alternative hypothesis. If one of these estimations is difficult, that immediately handicaps the likelihood ratio test, but one of the other tests is available. If all three tests are easily calculated, it makes sense to do the extra work and assess all three. Although equivalent in large samples, these tests are different in small samples, and this difference may be important. It may be that one test requires an inconvenient iteration for its calculation, while another has an actual significance level greater than that nominated when using its asymptotic null distribution.

See Buse (1982) for an elementary exposition of these tests, and Cox and Hinkley (1974) for more detail. Here we are aiming at a level of treatment somewhere in between these.

A.5.1 The Likelihood Ratio, Score, and Wald Tests for a Simple Null Hypothesis

Suppose we are given a random sample $X_1, \ldots, X_n$ from a distribution with probability (density) function $f_X(x; \theta)$ in which $\theta = (\theta_1, \ldots, \theta_k)^T \in \Omega$, the parameter space. We aim to test the simple null hypothesis $H_0: \theta = \theta_0$ against the composite alternative K: $\theta \neq \theta_0$, where, to avoid continuity problems, θ_0 is not on the boundary of Ω.

The likelihood ratio test for testing H_0 against K was proposed by Neyman and Pearson (1928), and is based on the statistic

$$L = 2\ell(\hat{\theta};X) - 2\ell(\theta_0;X),$$

where $\ell(\theta;X)$ is the logarithm of the likelihood,

$$\ell(\theta;x) = \sum_{i=1}^{n} \log f_X(x_i;\theta),$$

and where $\hat{\theta}$ is the maximum likelihood estimator of θ, chosen to maximise the likelihood $f_X(x_1;\theta) * ... * f_X(x_n;\theta)$ for all $\theta \in \Omega$. Expectation with respect to the probability density function $f_X(x;\theta)$ is denoted by E_θ. Now define the efficient score and information matrix respectively by

$$U(\theta) = (U_i(\theta)) = (\partial\ell(\theta;X)/\partial\theta_i), \text{ and}$$
$$I(\theta) = (I_{ij}(\theta)) = (-E_\theta[\partial^2\ell(\theta;X)/\partial\theta_i\partial\theta_j]).$$

To test H_0 against K, Wald (1943) suggested the test statistic

$$W = (\hat{\theta} - \theta_0)^T I(\hat{\theta}) (\hat{\theta} - \theta_0),$$

while Rao (1948) proposed

$$S = \{U(\theta_0)\}^T\{I(\theta_0)\}^{-1}\{U(\theta_0)\}.$$

S does not require the calculation of the maximum likelihood estimator, but does require the existence of the inverse of the information matrix. For more detail on alternative forms and properties of the score statistic, see Bera and McKenzie (1986) or Rayner and Best (1989a).

The null hypothesis H_0 is rejected for large values of L, W and S, determined by the null distributions of the test statistics. Under the null hypothesis L, W and S are all asymptotically distributed as central χ^2_k random variables, where k is the number of elements of θ, or, equivalently, the dimension of Ω. Alternatively, the degrees of freedom k is the difference between the number of parameters estimated under the full model (Ω) and under the null hypothesis. Under certain *contiguous*

alternatives, that tend to θ_0 as the sample size tends to infinity, the distribution is non-central χ^2_k. Note that the development presented here uses the information that the parameter space under the null hypothesis contains just one point.

Exponential Example. Given a random sample $X_1, \ldots, X_n$ from an exponential (λ) population, suppose we wish to test H_0: $\lambda = \lambda_0$ against K: $\lambda \neq \lambda_0$. The likelihood is

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda s},$$

with logarithm $n \log \lambda - \lambda s$. It is routine to show that

$$\frac{d \log L}{d\lambda} = \frac{n}{\lambda} - s \text{ and } \frac{d^2 \log L}{d\lambda^2} = -\frac{n}{\lambda^2},$$

from which the maximum likelihood estimator is

$$\hat{\lambda} = \frac{n}{S}, \text{ and}$$

$$U(\lambda) = \frac{n}{\lambda} - S \text{ and } I(\lambda) = \frac{n}{\lambda^2}.$$

Substituting in the formulae gives

$$L = 2n \log \left(\frac{n}{\lambda_0 S}\right) + 2(\lambda_0 S - n), \text{ and } S = (n - \lambda_0 S)^2 / r = W.$$

A.5.2 The Likelihood Ratio, Score, and Wald Tests for a Composite Null Hypothesis

In applications it is unlikely that all the parameters of the hypothesised distribution are known. When this occurs, the unknown parameters enter the problem as unspecified or "nuisance" parameters.

The theory of A.5.1 can be modified so that $X_1, \ldots, X_n$ is a random sample from a distribution with probability (density) function $f_X(x;\gamma)$, where the parameter vector γ is partitioned into elements that are tested for, and the remaining "nuisance" elements: $\gamma = (\theta^T, \beta^T)^T$. In the partitioning of γ, θ is a k by 1 vector of real parameters, $\theta \in \Omega$, and β is a q by 1 vector of real nuisance parameters, $\beta \in B$. We wish to test H_0: $\theta = \theta_0$ against K: $\theta \neq \theta_0$ without specifying β; again θ_0 should be an interior point of Ω to avoid continuity problems. A typical example is testing for the mean of a $N(\mu, \sigma^2)$ distribution without specifying σ^2; for example, testing H_0: $\mu = 0$ against K: $\mu \neq 0$, with σ^2 unspecified. The logarithm of the likelihood is

$$\ell(\gamma; x) = \sum_{i=1}^{n} \log f_X(x_i;\gamma)$$

and the natural extension of L from section A.5.1 is

$$\hat{L} = 2\ell(\hat{\gamma};X) - 2\ell(\hat{\gamma}_0;X)$$

where $\hat{\gamma} = (\hat{\theta}^T, \hat{\beta}^T)^T$ is the maximum likelihood estimator of γ under the full model, subject only to $\theta \in \Omega$ and $\beta \in B$, and where $\hat{\gamma}_0 = (\theta_0^T, \hat{\beta}_0^T)^T$ is the maximum likelihood estimator of γ under the null hypothesis, in which $\hat{\beta}$ is restricted to $\beta \in B$ subject to $\theta = \theta_0$.

Denoting expectation with respect to the distribution with probability density function $f_X(x;\gamma)$ by E_γ, the efficient score and the information matrix are given respectively by

$$U(\gamma) = (\partial\ell(\gamma; X)/\partial\gamma_i) \text{ and } I(\gamma) = -E_\gamma[(\partial^2\ell(\gamma;X)/\partial\gamma_i\partial\gamma_j)].$$

Now U and I may be partitioned as is γ, so that

$$U = U(\gamma) = \begin{pmatrix} U_\theta(\gamma) \\ U_\beta(\gamma) \end{pmatrix} \text{ and } I(\gamma) = \begin{pmatrix} I_{\theta\theta} I_{\theta\beta} \\ I_{\beta\theta} I_{\beta\beta} \end{pmatrix}.$$

Define $\Sigma(\gamma)$ by

$$\Sigma(\gamma) = I_{\theta\theta}(\gamma) - I_{\theta\beta}(\gamma) I_{\beta\beta}^{-1}(\gamma) I_{\beta\theta}(\gamma).$$

It follows from the discussion in Cox and Hinkley (1974, section 9.3) that $\Sigma(\gamma)$ is the asymptotic covariance matrix of $U_\theta(\gamma)$ and that $\{\Sigma(\gamma)\}^{-1}$ is the asymptotic covariance matrix of $\hat{\theta}$. The generalisation of the score statistic is given by

$$\hat{S} = \{U_\theta(\hat{\gamma}_0)\}^T \Sigma(\hat{\gamma}_0)^{-1} \{U_\theta(\hat{\gamma}_0)\}.$$

This requires the calculation of the restricted maximum likelihood estimator $\hat{\gamma}_0$ but not the unrestricted maximum likelihood estimator $\hat{\gamma}$, and that the inverse is defined. The generalised Wald statistic is given by

$$\hat{W} = (\hat{\theta} - \theta_0)^T \Sigma(\hat{\gamma}) (\hat{\theta} - \theta_0).$$

This requires the calculation of $\hat{\gamma}$ but not $\hat{\gamma}_0$. Of course $\hat{L}$ requires both $\hat{\gamma}$ and $\hat{\gamma}_0$. The statistics $\hat{L}$, $\hat{S}$ and $\hat{W}$ all have asymptotic χ^2_k distributions.

Normal Example. Suppose that a random sample of size n is drawn from a $N(\mu, \sigma^2)$ population. We wish to test H: $\mu = 0$ against K: $\mu \neq 0$. The variance is an unspecified nuisance parameter. The parameter γ is $(\mu, \sigma)^T$. We find

$$\partial \log L/\partial\mu = \sum_{i=1}^{n} (x_i - \mu)/\sigma^2,$$

$$\partial \log L/\partial\sigma = - n/\sigma + \sum_{i=1}^{n} (x_i - \mu)^2/\sigma^3,$$

$$\partial^2 \log L/\partial\mu^2 = - n/\sigma^2,$$

$$\partial^2 \log L/\partial\mu\partial\sigma = -2\sum_{i=1}^{n} (x_i - \mu)/\sigma^3 \text{ and}$$

$$\partial^2 \log L/\partial\sigma^2 = n/\sigma^2 - 3 \sum_{i=1}^{n} (x_i - \mu)^2/\sigma^4.$$

Hence

$$U = U(\gamma) = \left(\frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2} \quad - \frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3} \right)^T,$$

$$I(\gamma) = - E\left[\begin{pmatrix} - \frac{n}{\sigma^2} & \frac{- 2 \sum_{i=1}^{n} (x_i - \mu)}{\sigma^3} \\ \frac{- 2 \sum_{i=1}^{n} (x_i - \mu)}{\sigma^3} & \frac{n}{\sigma^2} - \frac{3 \sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^4} \end{pmatrix} \right]$$

$$= \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}, \text{ and}$$

$$\Sigma(\gamma) = I_{\theta\theta}(\gamma) - I_{\theta\beta}(\gamma) I_{\beta\beta}^{-1}(\gamma) I_{\beta\theta}(\gamma) = \frac{n}{\sigma^2}.$$

It follows that the maximum likelihood estimator of σ under the null hypothesis is

$$\hat{\sigma}_{\omega}^2 = \sum_{i=1}^{n} X_i^2/n,$$

while in the full parameter space is

$$\hat{\mu}_\Omega = \bar{X} \text{ and } \hat{\sigma}^2_\Omega = \sum_{i=1}^{n} (X_i - \bar{X})^2/n.$$

Substituting in the given formulae yields

$$\hat{L} = 2n \log\left(\frac{\hat{\sigma}_\omega}{\hat{\sigma}_\Omega}\right) = n \log\left(\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right),$$

$$\hat{S} = \left(\frac{n(\bar{X} - \mu_0)}{\hat{\sigma}^2_\omega}\right)^2 \frac{\hat{\sigma}^2_\omega}{n} = \frac{n^2 \bar{X}^2}{\sum_{i=1}^{n} X_i^2}, \text{ and}$$

$$\hat{W} = \frac{n\hat{\mu}^2_\Omega}{\hat{\sigma}^2_\Omega} = \frac{n^2 \bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

For further interesting examples see Rayner (1997).

A.6 Assessing Univariate Normality

Nonparametric methods may be applied to non-normal data such as counts and ranks. They make fewer assumptions than parametric methods such as the analysis of variance, and so are good for non-statisticians to use: there is less chance they will be used improperly. The greater applicability of non-parametric methods is usually said to be balanced by their being less powerful or less efficient than the corresponding parametric tests, but this statement isn't quite true. If the assumptions underlying the parametric tests are valid, the parametric tests will be superior. If not, nonparametric tests can be considerably more

powerful than competing parametric tests.

If the necessary parametric assumptions are not satisfied, then the parametric tests are not available, and nonparametric alternatives should be used. Since normality is the most common parametric assumption, we focus here on assessments of normality that eliminate normality-based tests. A wider discussion of statistical model assessment is given in section 9.1.

Assessments of normality are broadly subjective - mainly graphical, and objective - mainly statistical tests, the so-called goodness of fit tests for normality. Graphical techniques include normal Q-Q (quantile-quantile) plots and density plots. Well-known tests include the Shapiro-Wilk and its modification, the Shapiro-Francia, the Cramer-von Mises and Anderson-Darling. Historically these are all preceded by tests based on the sample skewness g_1 and sample kurtosis g_2.

Rayner and Best (1986, 1989a) recommended an omnibus "smooth" test based on the sum of the squares of the components (see sections 2.6, A.1.4 and A.3) $\widehat{V}_3$, $\widehat{V}_4$, $\widehat{V}_5$ and $\widehat{V}_6$. These components are asymptotically mutually independent and asymptotically standard normal; so $\widehat{S}_4 = \widehat{V}_3^2 + \widehat{V}_4^2 + \widehat{V}_5^2 + \widehat{V}_6^2$ asymptotically has the χ^2_4 distribution. If the component $\widehat{V}_r$ is significantly large, it is suggested that the data differ from what is expected under normality in the r*th* moment. However Rayner, Best and Mathews (1995) found that this suggestion may not be true due to higher moment differences between the data and normality. Note that

$$\widehat{V}_3 = g_1\sqrt{\frac{n}{6}} \text{ and } \widehat{V}_4 = g_2\sqrt{\frac{n}{24}},$$

so that the skewness and kurtosis components are standardised versions of the sample skewness and kurtosis respectively. See also the discussion at the end of section 9.2.

PCB Example. Risebrough (1972) gave data on the concentration of polychlorinated biphenyl (PCB) in the yolk lipids of 65 Anacapa birds. Rayner and Best (1989a, Example 1.4.3) gave the data and found $\widehat{V}_3$ = 2.31, $\widehat{V}_4$ = 1.99, $\widehat{V}_5$ = 0.41, $\widehat{V}_6$ = -0.65 and $\widehat{S}_4$ = 9.89. The corresponding p-

values, based on Monte Carlo simulations, are 0.021, 0.047, 0.682, 0.516 and 0.019. The omnibus test based on $\widehat{S}_4$ finds evidence of non-normality. The mean and variance of the fitted normal distribution are taken from the data. The small p-values for $\widehat{V}_3$ and $\widehat{V}_4$ indicate the skewness and kurtosis are not consistent with normality, while the large p-values for $\widehat{V}_5$ and $\widehat{V}_6$ suggest the fifth and sixth central moments are consistent with normality, although note the caution of the previous paragraph. For these data the Anderson-Darling statistic A^2 takes the value 0.742 with p-value about 0.050. This p-value is based on D'Agostino and Stephens (1986, Table 4.7).

A.7 Multivariate Normality

This section introduces two definitions of multivariate normality and explores some of the properties of this distribution. The results here underpin one of the fundamental approaches used in this book. When a transformation is used to diagonalise a covariance matrix, the underlying mechanism is that the random vector transformed has a multivariate normal distribution, at least asymptotically. Then by the results in this section, the diagonal covariance matrix implies that pairs elements of the corresponding random vector are not only uncorrelated, but independent. For those who are interested, somewhat more theory than is strictly needed is developed.

The first definition is in terms of a probability density function, and is frequently not convenient for establishing results. The second definition is a characterization definition, and is equivalent to the first under certain circumstances. Some properties of the multivariate normal distribution are then established.

A p-dimensional random variable is one that takes values in E^p, p-dimensional Euclidean space. Suppose $X^T = (X_1, \ldots, X_p)$ is such a p-dimensional random variable, while $\mu^T = (\mu_1, \ldots, \mu_p)$, and $\Sigma = (\sigma_{ij})$ is a p by p real symmetric matrix. Anderson (1958) defined the p-variate normal as follows.

Definition A.7.1 Provided Σ is non-singular, the random variable X is p-

variate normal with parameters μ and Σ, written $N_p(\mu, \Sigma)$, if and only if X has probability density function

$$f_X(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-(x - \mu)^T\Sigma^{-1}(x - \mu)/2\}.$$

We may confirm that for p = 1 this is the usual univariate normal probability density function; so N_1 denotes univariate normality. By p-fold integration it follows that $E[X] = (E[X_i]) = \mu$ and $V(X) = (\text{cov}(X_i, X_j)) = \Sigma$. The density does not exist if Σ is singular. The approach here is closer to that of Rao (1965) than Anderson (1958).

Definition A.7.2 A p-dimensional random variable X has the p-variate normal distribution if and only if every linear function of X has a univariate normal distribution.

If $a = (a_1, \ldots, a_p)^T$ then $a^TX = a_1X_1 + \ldots + a_pX_p$ is a typical linear function.

Lemma A.7.1. Suppose X is p-variate normal by definition A.7.2. Then E[X] and V(X) exist and may be denoted by μ and Σ. For any p by 1 vector of constants a, a^TX is $N_1(a^T\mu, a^T\Sigma a)$.

Proof. For i = 1, ... , p in turn take $(a)_i = 1$, $(a)_j = 0$ for $j \neq i$. Then $a^TX = X_i$, and by definition this is univariate normal, say with finite mean μ_i and finite variance σ_{ii}. Denote $\text{cov}(X_i, X_j)$ by $\sigma_{ij} = \sigma_{ji}$ for all $i \neq j$. By the Cauchy-Schwarz inequality $\sigma_{ij} \leq \sigma_{ii}\sigma_{jj} < \infty$, so σ_{ij} is well defined. Write $\mu = (\mu_i)$ and $\Sigma = (\sigma_{ij})$. For a arbitrary, a^TX is N_1 by definition. Hence

$$E[a^TX] = E[\sum_{i=1}^{n} a_iX_i] = \sum_{i=1}^{n} a_i\mu_i = a^T\mu \text{ and}$$

$$V(a^TX) = V(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} V(a_iX_i) + 2\sum_{i<j} \text{cov}(a_iX_i, a_jX_j)$$

$$= \sum_{i=1}^{n} a_i^2\sigma_{ii} + 2\sum_{i<j} a_ia_j\sigma_{ij} = a^T\Sigma a.$$

Lemma A.7.2. The moment generating function of X is

$$M_X(\theta) = \exp(\theta^T\mu + \theta^T\Sigma\theta/2).$$

Proof. Recall that if Y is distributed as $N_1(\mu, \sigma^2)$ then the moment generating function of Y is given by

$$E[\exp(tY)] = \exp(\mu t + \sigma^2 t^2/2)$$

and conversely. Since a^TX is $N_1(a^T\mu, a^T\Sigma a)$, this random variable has moment generating function $E[\exp(ta^TX)] = \exp(a^T\mu t + a^T\Sigma a t^2/2)$. Now put $\theta = ta$.

Henceforth the p-variate normal given by definition A.7.2 will be denoted by $N_p(\mu, \Sigma)$, since only μ and Σ are required to specify the moment generating function and hence the distribution. This will be reconciled with definition 1, so there is no ambiguity.

Next follows a lemma that is the converse of Lemma A.7.1.

Lemma A.7.3. If there exist a vector μ and a symmetric matrix Σ such that for every a, a^TX is $N_1(a^T\mu, a^T\Sigma a)$, then X has the $N_p(\mu, \Sigma)$ distribution.

Proof. Normality follows by definition. The parameters follow by putting first $a_i = 1$, $a_j = 0$ for $j \neq i$, so that

$$E[X_i] = E[a^TX] \text{ (this a)} = a^T\mu = \mu_i \text{ and}$$
$$V(X_i) = V(a^TX) \text{ (this a)} = a^T\Sigma a = \sigma_{ii}.$$

Second consider $a_i = a_j = 1$, $a_k = 0$ for $k \neq i$ or j. Now

$$V(X_i + X_j) = V(X_i) + 2\text{cov}(X_i, X_j) + V(X_j) = V(a^TX) \text{ (this a)}$$
$$= a^T\Sigma a = \sigma_{ii} + 2\sigma_{ij} + \sigma_{jj}.$$

This uses the symmetry of Σ. It follows that $\text{cov}(X_i\ X_j) = \sigma_{ij}$.

Lemma A.7.4. If X is p variate normal, then any subset of q elements of X is q variate normal.

Proof. Every linear function of the subset is univariate normal. Alternatively consider $a^T X$ where $a_i = 0$ if X_i is not in the subset. Now partition X, a and μ into their first q elements and the remainder, so that

$$X^T = (X_1^T \mid X_2^T),\ a^T = (a_1^T \mid a_2^T) \text{ and } \mu^T = (\mu_1^T \mid \mu_2^T),$$

in which X_1, a_1 and μ_1 are all q by 1, X_2, a_2 and μ_2 are all (p - q) by 1, and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

in which Σ_{11} is q by q. Since Σ is symmetric, so are Σ_{11} and Σ_{22}, and $\Sigma_{12}^T = \Sigma_{21}$. If $a_2 = 0$, then $a^T X = a_1^T X_1$, which is distributed as $N_1(a_1^T \mu_1, a_1^T \Sigma_{11} a_1)$, and hence by Lemma A.7.3, X_1 is distributed as $N_q(\mu_1, \Sigma_{11})$. Similarly X_2 is distributed as $N_{p-q}(\mu_2, \Sigma_{22})$. For the particular case of the subset being the first q elements of X, we have now completely specified the distribution in Lemma A.7.4.

Lemma A.7.5. If X is distributed as $N_p(\mu, \Sigma)$, then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.

Proof. Put $\theta^T = (\theta_1^T, \theta_2^T)$, with θ_1 being q by 1. Then

$$M_X(\theta) = M_{X_1}(\theta) M_{X_2}(\theta)$$

if and only if

$$\exp(\theta^T \mu + \theta^T \Sigma \theta / 2) = \exp(\theta_1^T \mu_1 + \theta_2^T \mu_2 + \theta_1^T \Sigma_{11} \theta_1 / 2 + \theta_2^T \Sigma_{22} \theta_2 / 2).$$

But $\theta^T\mu = \theta_1^T\mu_1 + \theta_2^T\mu_2$ and

$$\theta^T\Sigma\theta = \begin{pmatrix} \theta_1^T & \theta_2^T \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \theta_1^T\Sigma_{11}\theta_1 + 2\theta_1^T\Sigma_{12}\theta_2 + \theta_2^T\Sigma_{22}\theta_2.$$

Thus $M_X(\theta) = M_{X_1}(\theta)M_{X_2}(\theta)$ if and only if $\Sigma_{12} = 0$.

Lemma A.7.6. If $Y = CX$, in which C is a q by p matrix of constants, then Y has the $N_q(C\mu, C\Sigma C^T)$ distribution.

Proof. If $\ell = C^Ta$ then $a^TY = a^TCX = \ell^TX$, and by Lemma A.7.1 this has the $N_1(\ell^T\mu, \ell^T\Sigma\ell)$ distribution. Substituting for ℓ, a^TY has the $N_1(a^TC^T\mu, a^TC^T\Sigma Ca)$ distribution. The result now follows from Lemma A.7.3.

Lemma A.7.7. The Reproductive Property of N_p. Suppose that for i = 1, ... , n, X_i are mutually independent $N_p(\mu_i, \Sigma_i)$ random variables, and ℓ_1, ... , ℓ_n are constants. Then $Y = \ell_1X_1 + ... + \ell_nX_n$ has the $N_p(\ell_1\mu_1 + ... + \ell_n\mu_n, \ell_1^2\Sigma_1 + ... + \ell_n^2\Sigma_n)$ distribution.

Proof. Note that $a^TY = \ell_1a^TX_1 + ... + \ell_na^TX_n$. Now since $\ell_ia^TX_i$ has the $N_1(\ell_ia^T\mu_i, a^T\ell_i^2\Sigma_ia)$ distribution, $\sum_{i=1}^{n} \ell_ia^TX_i$ has the $N_1(\sum_{i=1}^{n} \ell_ia^T\mu_i, \sum_{i=1}^{n} a^T\ell_i^2\Sigma_ia)$ distribution, and a^TY has the $N_1(a^T\sum_{i=1}^{n} \ell_i\mu_i, a^T\sum_{i=1}^{n} \ell_i^2\Sigma_ia)$ distribution. The result now follows from Lemma A.7.3.

Corollary. $\overline{X}$, the sample mean of a random sample from the $N_p(\mu, \Sigma)$ distribution, has the $N_p(\mu, \Sigma/n)$ distribution.

Suppose that Σ has rank m. We then say that $N_p(\mu, \Sigma)$ also has rank

m.

We next show that our two definitions are equivalent. This requires the spectral decomposition below. This result may be viewed as an algebraic aside. Some readers may wish to skip to the end of the section where the subsequent use made of the results of this section is discussed.

The spectral decomposition of a real symmetric p by p matrix.

Suppose that the eigenvalues and corresponding eigenvectors of the real symmetric p by p matrix Σ are λ_i and h_i, i = 1, ... , p. Assume also that $\lambda_1 > \lambda_2 > ... > \lambda_p$. Note that if the eigenvalues are not all distinct the decomposition is similar but not unique. See Seber (1984, A1.3, p. 517). Now construct $H = (h_1, ... , h_p)$.

By definition the h_i are normalized to be of unit length: so $h_i^T h_i = 1$, i = 1, ... , p and $h_i^T h_j = 0$, $i \neq j$, = 1, ... , p. H is orthogonal, because $(H^T H)_{ij} = h_i^T h_j = \delta_{ij}$, the Kronecker delta. Write $(H)_{uv} = (h_v)_u = h_{uv}$. First

Lemma A.7.8. $h_1 h_1^T + ... + h_p h_p^T = I_p$ and (A1)

$\lambda_1 h_1 h_1^T + ... + \lambda_p h_p h_p^T = \Sigma$. (A2)

Proof. $H^T H = I_p$ implies $H^T = H^{-1}$, which implies $HH^T = I_p$. Now

$$(\sum_r h_r h_r^T)_{ij} = \sum_r (h_r h_r^T)_{ij} = \sum_r (h_r)_i (h_r)_j = \sum_r h_{ir} h_{jr} = (HH^T)_{ij} = \delta_{ij}.$$

This gives (A1). Next, for i = 1, ... , p, $\Sigma h_i = \lambda_i h_i$, so

$$\Sigma = \Sigma(I_p) = \Sigma h_1 h_1^T + ... + \Sigma h_p h_p^T \text{ (by (A1) above)}$$
$$= \lambda_1 h_1 h_1^T + ... + \lambda_p h_p h_p^T \text{ (by the definition of the eigenvalues).}$$

This is the required decomposition.

The following theorem provides a necessary and sufficient condition for p-variate normality.

Theorem A.7.1. X is $N_p(\mu, \Sigma)$ with rank m if and only if $X = \mu + BZ$, in which μ is a p by 1 vector of constants, B is a p by m matrix of rank m with $BB^T = \Sigma$, and $Z_1, \ldots, Z_m$ are mutually independent standard normal random variables.

Proof.
Necessity: If $X = \mu + BZ$ then $a^TX = a^T\mu + a^TBZ$. Since a^TBZ is a linear combination of normal variables, it is normally distributed and so is a^TX. Now

$$E[a^TX] = a^T\mu + a^TBE[Z] = a^T\mu \text{ since } E[Z] = 0, \text{ and}$$
$$V(a^TX) = V(a^T\{X - \mu\}) = E[a^T(X - \mu)(X - \mu)^Ta] = a^T\Sigma a.$$

Since a^TX has the $N_1(a^T\mu, a^T\Sigma a)$ distribution, by Lemma A.7.3, X is $N_p(\mu, \Sigma)$.
Sufficiency: Assume now that X is $N_p(\mu, \Sigma)$ and that $Z_1, \ldots, Z_m$ are mutually independent standard normal random variables. We define $Y_0, Y_1, \ldots, Y_p$ to be p-variate random variables, given in terms of the Z_i, μ and through its spectral decomposition, Σ, so that X and $Y_0 + Y_1 + \ldots + Y_p$ have the same moment generating function. The result then follows from $X = Y_0 + Y_1 + \ldots + Y_p$ and matrix algebra.

Define Y_0 to be a random variable that takes the value μ with probability 1. Then $M_{Y_0}(\theta) = \exp(\theta^T\mu)$. Set $Y_i = \sqrt{\lambda_i}h_iZ_i$, i = 1, ... , m. Then Y_i is $N_p(0, \lambda_ih_ih_i^T)$ distributed, for

- $a^TY_i = \sqrt{\lambda_i}(a^Th_i)Z_i$ is univariate normal for arbitrary a,
- $E[a^TY_i] = \sqrt{\lambda_i}(a^Th_i)E[Z_i] = 0$, and
- $V(a^TY_i) = E[a^TY_iY_i^Ta] = E[\{\sqrt{\lambda_i}(a^Th_i)Z_i\}\{\sqrt{\lambda_i}(a^Th_i)Z_i\}^T]$
  $= E[\sqrt{\lambda_i}a^Th_iZ_iZ_i^Th_i^Ta\sqrt{\lambda_i}] = a^T\lambda_ih_ih_i^Ta$

since $E[Z_iZ_i^T] = E[Z_i^2] = 1$. Incidentally, this construction shows that the

definition of multivariate normality is not vacuous. Now from Lemma A.7.2,

$$M_{Y_i}(\theta) = \exp\{\lambda_i(\theta^T h_i)^2\},\ i = 1, \ldots, m.$$

In addition the Y_i are mutually independent. For from Lemma A.7.5 it is sufficient to show that $\text{cov}(Y_i, Y_j) = 0$. But

$$\begin{aligned}\text{cov}(Y_i, Y_j) &= E[Y_i Y_j^T] = E[\{\sqrt{\lambda_i} h_i Z_i\}\{\sqrt{\lambda_j} h_j Z_j\}^T] \\ &= \sqrt{(\lambda_i \lambda_j)}\, h_i E[Z_i Z_j^T] h_j^T = 0\end{aligned}$$

because the Z_i are mutually independent, and hence have zero covariances. It now follows that the moment generating function of $Y_0 + Y_1 + \ldots + Y_p$ is the product of the moment generating functions of the independent summands, namely

$$\begin{aligned}&\exp(\theta^T\mu)*\exp\{\lambda_1(\theta^T h_1)^2\}* \ldots * \exp\{\lambda_p(\theta^T h_p)^2\} \\ &= \exp\{(\theta^T\mu + \lambda_1(\theta^T h_1)^2 + \ldots + \lambda_p(\theta^T h_p)^2\}.\end{aligned}$$

Now note that $\theta^T\Sigma\theta = \theta^T(\lambda_1 h_1 h_1^T + \ldots + \lambda_p h_p h_p^T)\theta$ (from (A2))

$$= \lambda_1(\theta^T h_1)^2 + \ldots + \lambda_p(\theta^T h_p)^2.$$

It now follows that the moment generating function of $Y_0 + Y_1 + \ldots + Y_p$ is $\exp(\theta^T\mu + \theta^T\Sigma\theta/2)$, which, from Lemma A.7.2, is the moment generating function of X. Thus we have

$$X = Y_0 + Y_1 + \ldots + Y_m, \text{ in which } Y_0 = \mu,\ Y_i = \sqrt{\lambda_i} h_i Z_i,\ i = 1, \ldots, m.$$

Hence

$$\begin{aligned}X &= \mu + \sum_{i=1}^{n} \sqrt{\lambda_i} h_i Z_i = \mu + (h_1\sqrt{\lambda_1}, \ldots, h_m\sqrt{\lambda_m})(Z_1, \ldots, Z_m)^T \\ &= \mu + BZ,\end{aligned}$$

defining B and Z. With these definitions,

$$BB^T = (h_1\sqrt{\lambda_1}, \ldots, h_m\sqrt{\lambda_m})\,(h_1\sqrt{\lambda_1}, \ldots, h_m\sqrt{\lambda_m})^T$$
$$= \sum_{i=1}^{n} h_i h_i^T \lambda_i = \Sigma \text{ (from (A2)), as required.}$$

Aside. $B = (h_1\sqrt{\lambda_1}, \ldots, h_m\sqrt{\lambda_m})$ has rank m because, for any matrix A, $\text{rank}(AA^T) = \text{rank}(A^TA) = \text{rank}(A) = \text{rank}(A^T)$. Thus $\text{rank}(B) = \text{rank}(\Sigma) = m$.

Density for the multivariate normal distribution of full rank

Suppose that X is $N_p(\mu, \Sigma)$ where $\text{rank}(\Sigma) = p$. Then by the theorem above, $X = \mu + BZ$, in which B is a p by p matrix of rank p with $BB^T = \Sigma$, and $Z_1, \ldots, Z_p$ are mutually independent standard normal random variables. Since B is of full rank, B^{-1} exists. Now $Z = B^{-1}(X - \mu)$ has probability density function

$$f_Z(z) = (2\pi)^{-p/2} \exp(-z^Tz/2),\ -\infty < z_i < \infty,\ i = 1, \ldots, p.$$

We change variable from Z to X, applying

$$f_X(x) = f_Z(z(x))\ |\partial z(x)/\partial x|$$

and noting that the Jacobian is $|\partial \{B^{-1}(x - \mu)\}/\partial x| = |B^{-1}|$. Now since

$$BB^T = \Sigma,$$
$$|\Sigma| = |BB^T| = |B|^2$$

since $|B^T| = |B|$, we have $|B| = |\Sigma|^{1/2}$ and the Jacobian is

$$|B^{-1}| = |B|^{-1} = |\Sigma|^{-1/2}.$$

Thus

$$f_X(x) = f_Z(z(x))\ |\partial z(x)/\partial x|$$

$$= (2\pi)^{-p/2} \exp\{ -z(x)^T z(x)/2\} \ |\Sigma|^{-1/2}$$
$$= (2\pi)^{-p/2} \ |\Sigma|^{-1/2} \exp\{- [B^{-1}(x - \mu)]^T [B^{-1}(x - \mu)]/2\}$$
$$\text{(using } Z = B^{-1}(X - \mu) \text{)}$$
$$= (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{ -(x - \mu)^T (B^{-1})^T B^{-1}(x - \mu)/2\}.$$

Now $\Sigma = BB^T$ implies that

$$\Sigma^{-1} = (BB^T)^{-1} = (B^T)^{-1}B^{-1} = (B^{-1})^T B^{-1}.$$

This gives

$$f_X(x) = (2\pi)^{-p/2} \ |\Sigma|^{-1/2} \exp\{ -(x - \mu)^T \Sigma^{-1}(x - \mu)/2\}.$$

If Σ is not of full rank, we may use either definition A.7.2 or the necessary and sufficient condition (the theorem) to establish further properties.

The main use we make of the results of this section is as follows. By the Central Limit Theorem, sums of counts are asymptotically normally distributed. An arbitrary linear combination of the elements of such a vector of sums will be asymptotically normal, so that vector of sums will be asymptotically multivariate normal. Now by Lemma A.7.5, if the elements of a multivariate normal random variable have zero covariance, they are mutually independent. Hence if, for example, a vector is asymptotically p-variate normal with mean vector zero and covariance matrix I_p, the elements of the vector are asymptotically mutually independent and asymptotically have the standard normal distribution. Now $(X_1, \ldots X_p)^T$ is a vector with the p-variate normal distribution and with covariance matrix Σ. If H is orthogonal with $H^T\Sigma H$ diagonal, then by Lemma A.7.6, $Y = H^T X$ has covariance matrix $H^T\Sigma H$ which is diagonal, and hence its elements are mutually independent.

A.8 Confidence Circles and Correspondence Analysis Plots

In this section we consider plots of data that can be presented in two-way contingency tables.

For many of the models we consider in this book t treatments are compared using a statistic such as Pearson's X_P^2, which we may decompose into components L, Q, C and so on, representing respectively linear, quadratic, cubic and higher moment effects:

$$X_P^2 = L^2 + Q^2 + C^2 + \ldots .$$

Each summand may in turn be decomposed into components attributable to the treatments:

$$L^2 = L_1^2 + \ldots + L_t^2$$
$$Q^2 = Q_1^2 + \ldots + Q_t^2$$
$$C^2 = C_1^2 + \ldots + C_t^2.$$

The L_i, Q_i and so on are all asymptotically mutually independent, or effectively so depending on the model, and all are asymptotically standard normally distributed.

Now suppose that $P(W < a(\alpha) \mid W$ has the χ_2^2 distribution$) = \alpha$, and we observe $(L_i, Q_i) = (\ell_i, q_i)$, $i = 1, \ldots , t$. Then the $100\alpha\%$ *confidence circle* for (ℓ_i, q_i) is the circle with equation

$$(x - \ell_i)^2 + (y - q_i)^2 = a(\alpha).$$

Using cubic and higher order effects we can construct confidence spheres.

By plotting confidence circles for all treatments, all with the same α, we can group treatments. A confidence circle centred on (0, 0) can be used to test the hypothesis of no linear and no quadratic effects at the $100\alpha\%$ level.

Cigarette Example. Baba (1994) gave the data in Table A.1 for 20 consumers ranking eight cigarettes. The counts corresponding to these data are given in Table A.2.

Figure A.1: 95% Confidence Circle

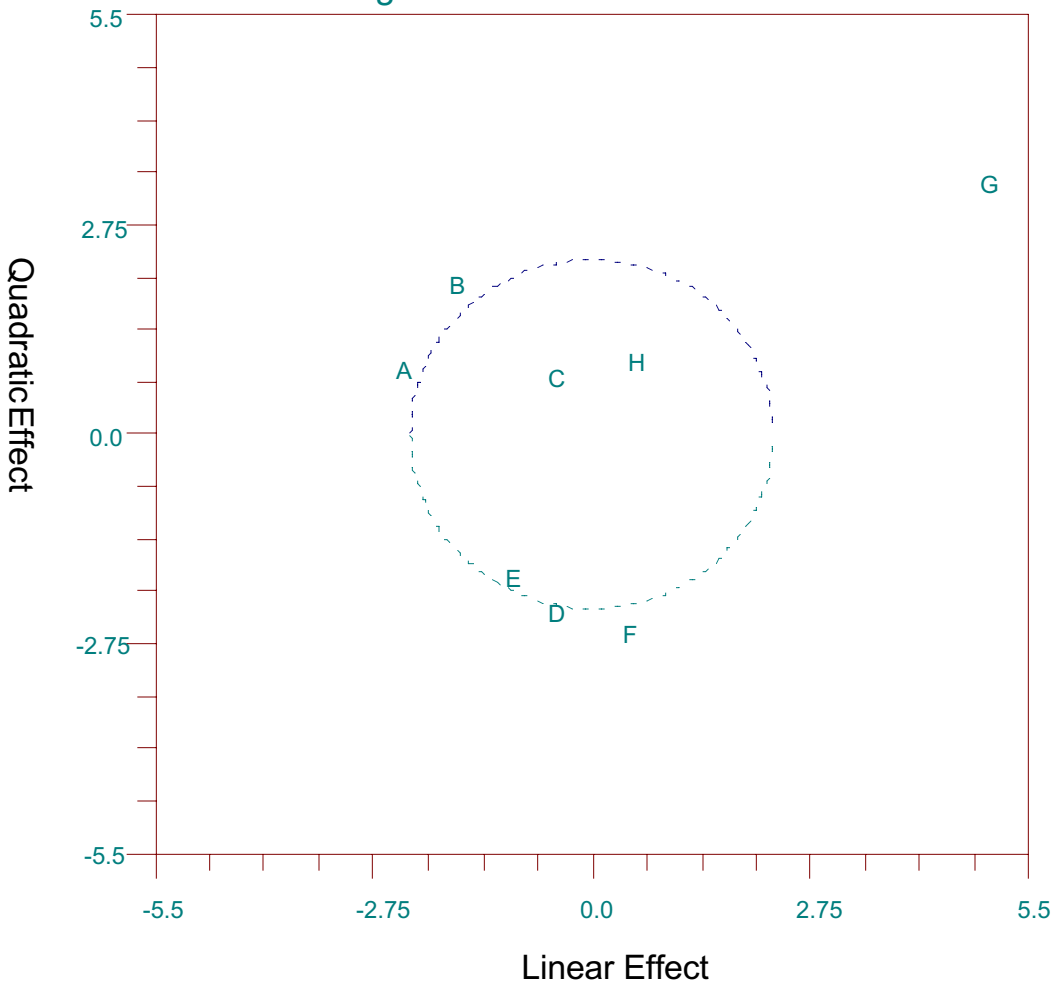


Table A.1 Rankings of cigarettes

Judge	Cigarette								Judge	Cigarette							
	A	B	C	D	E	F	G	H		A	B	C	D	E	F	G	H
1	2	5	8	6	3	4	7	1	11	6	7	1	5	4	2	8	3
2	1	5	8	4	6	3	7	2	12	3	1	2	4	6	5	8	7
3	1	5	7	3	6	4	8	2	13	8	1	7	6	5	3	4	2
4	4	2	5	6	1	7	8	3	14	3	1	6	2	4	5	8	7
5	1	4	2	6	3	7	8	5	15	3	1	6	4	2	5	8	7
6	1	7	5	4	2	6	8	3	16	3	1	6	2	5	4	8	7
7	4	1	2	3	6	5	7	8	17	8	5	4	2	1	3	6	7
8	4	1	2	3	5	7	6	8	18	1	7	2	8	4	5	6	3
9	3	2	1	4	6	5	8	7	19	3	8	2	5	4	7	6	1
10	4	1	7	3	2	5	8	6	20	1	6	2	5	4	3	8	7

If X_P^2 is the usual Pearson chi-squared statistic for testing homogeneity of the distributions of the ranks for the t products, then Schach (1979) showed that asymptotically $A = \{(t - 1)/t\} X_P^2$ has the $\chi^2_{(k - 1)(t - 1)}$ distribution, as is discussed in Chapters 6 and 7. In this case it is the components (see A.1.4) of A rather than of X_P^2 that are used to construct the confidence circle in Figure A.1.

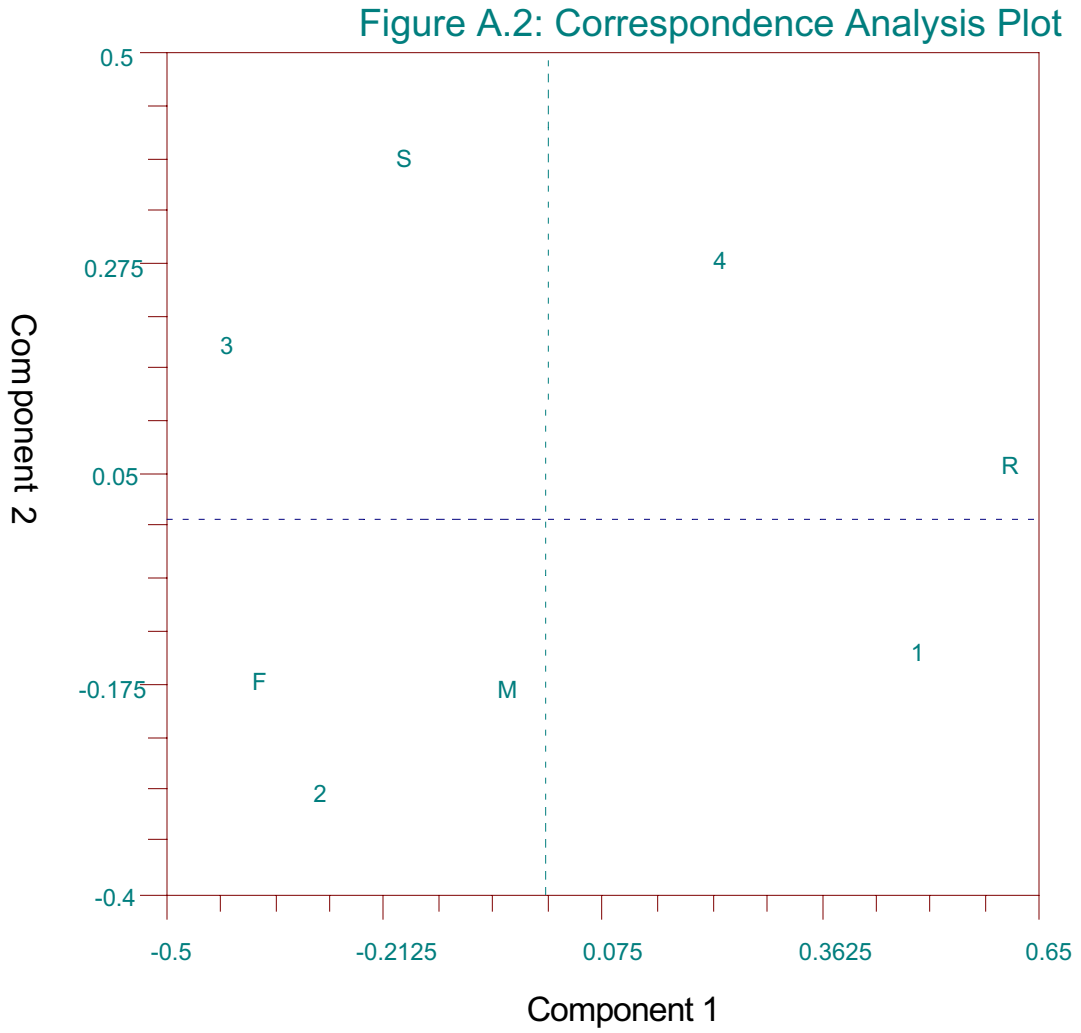
Figure A.1 shows a confidence circle around (0, 0) when $\alpha = 0.05$. This is essentially a two dimensional test of significance and in this case indicates product G is significantly different from what would be expected if there were no product differences. Products A, B, D and F are also significantly different from what would be expected if there were no differences, but only just.

Table A.2 Response frequencies of rankings of cigarettes

Cigarette	Rank 1	2	3	4	5	6	7	8
A	6	1	6	4	0	1	0	2
B	8	2	0	1	4	1	3	1
C	2	7	0	1	2	3	3	2
D	0	3	4	5	3	4	0	1
E	2	3	2	5	3	5	0	0
F	0	1	4	3	7	1	4	0
G	0	0	0	1	0	4	3	12
H	2	3	4	0	1	1	7	2

A correspondence analysis plot is a frequently used alternative approach to graphically presenting data of this kind. We will not develop the details of this approach. Instead we give an example so that a correspondence analysis plot may be compared with the given confidence circle.

Consider the 4 by 4 data from Best (1993) which is presented in Table A.3. These data were collected as part of a large CSIRO Food Research project designed to improve the flavour of Australian grown tomatoes. A total of 24 taste testers were asked to rank, in order of preference, four varieties of tomato.



A two dimensional correspondence analysis plot for these data is given in Figure A.2. This shows that all varieties of tomato are different in terms of their taste, and that the rankings assigned are quite different. Rutgers is the highest ranking variety according to taste, while the remaining order of preference is Momotaro, Floradade and the lowest ranked is Summit. This outcome is also apparent as the data in Table A.3 are scrutinised. However note that while Rutgers has most first preferences, it has eight last preferences. In the correspondence analysis plot it is closest to rank 1, but also quite close to rank 4.

A third type of plot using *uncertainty circles* is given in Best, Rayner and O'Sullivan (2000). This paper also gives more details of the two plots

we have just illustrated.

Table A.3 Rankings of four varieties of tomato according to taste

Variety	Rank 1	2	3	4	Total
Floradade	4	9	8	3	24
Momotaro	6	8	5	5	24
Summit	3	4	9	8	24
Rutgers	11	3	2	8	24

A.9 Permutation and Bootstrap Methods

A.9.1 Permutation p-values

Exact p-values for most of the common nonparametric tests can be calculated using permutation theory. Conover (1980, 1998) gave many examples of such calculations. An advantage of such an approach to p-value calculation is that the nonparametric nature of the test is assured and approximate asymptotic p-values need not be used.

Permutation tests were introduced by Fisher and Pitman in the 1930s to demonstrate that the usual t-test for independent samples was a sensible procedure. See, for example, Pitman (1937). Suppose that two samples have sizes m and n and that m + n = N. The null hypothesis is that the samples come from the same population. It follows that under the null hypothesis all rearrangements or permutations of the data into samples of m and n are equally likely. There are ${}^{N}C_{n}$ such permutations. We calculate some relevant statistic, such as $t^* = \bar{x} - \bar{y}$, where $\bar{x}$ is the mean of the original sample of size m and $\bar{y}$ is the mean of the original sample of size n. Choose $B \leq {}^{N}C_{n}$ random permutations and recalculate the difference of the means, t, for all of them. Calculate the proportion of these recalculated t values greater than t*. This is the permutation or

Monte Carlo estimate of the exact p-value. Its routine use has only become possible with increased computer power.

In the Introduction we noted that data for many common nonparametric tests could be exhibited in contingency table form. It turns out that permutation p-values can be calculated by generating random contingency tables with fixed margins. The algorithm of Patefield (1981) is particularly useful here and is used by the nonparametric statistical inference software package *StatXact* (1995). This package gives permutation p-values for most of the common tests.

The recent book by Good (1994) gave more discussion of permutation tests.

Although this book is concerned with nonparametric methods, we do introduce appropriate models to guide or justify our choice of nonparametric tests.

In this text we usually use the permutation approach, but we can also calculate p-values using bootstrap methods, which we now briefly introduce. The p-values calculated by bootstrap methods may differ from the permutation test p-values, particularly in small samples.

A.9.2 Bootstrap p-values

Suppose we are interested in a statistic T, such as the sample mean or variance, calculated from a given sample of n observations. A *bootstrap value* of the statistic T is obtained by sampling with replacement to obtain a sample of size n, and calculating from that sample a value T* of the statistic. If this is done B times, a *bootstrap sample of size B,* or *bootstrap estimate of the null distribution of T*, $T_1^*, \ldots, T_B^*$, is obtained. A *bootstrap p-value*, $\widehat{P_B}$, is the proportion of the T_i^* that is at least as large as T_0, the value of the statistic T based on the original sample. There are $M = n^n$ equally likely bootstrap values $T_1^*, \ldots, T_M^*$ and so $P_M = \{\text{number of } T_1^*, \ldots, T_M^* \geq T_0\}/M$ can be computed. In practice $B << M$ sets of random samples are drawn, and $\widehat{P_B}$ is used as an estimate of P_M. Values of B of the order of 200 work well in practice.

In the *parametric bootstrap* the statistic of interest, T, depends on

some parameter θ, so $T = T(\theta)$. An estimate $\widehat{\theta}$ of θ is calculated from the original sample and a bootstrap sample $T_1^*(\widehat{\theta}), \ldots, T_B^*(\widehat{\theta})$ found and used as before. See Efron and Tibshirani (1993).

References

Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions*. New York: Dover.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Anderson, R.L. (1959). Use of contingency tables in the analysis of consumer preference studies. *Biometrics*, 15, 582-590.

Anderson, T.W. (1958). *An Introduction to Multivariate Analysis*. New York: Wiley.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386.

AS2542.2.1 (1982). *Sensory Analysis of Foods - Specific Methods - Paired Comparison Test*. North Sydney: Standards Australia.

Baba, Y. (1994). *New approaches based on ranking in sensory evaluation.* In *New Approaches in Classification and Data Analysis.* Editors Diday, E., Lechevallier, Y., Schader, M., Bertrand, P. and Burtschy, B. New York: Springer-Verlag.

Beh, Eric J. and Davy, Pamela J. (1998). Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *Austral. N.Z. J. Statist.*, 40(4), 465-477.

Beh, Eric J. and Davy, Pamela J. (1999). Partitioning Pearson's chi-squared statistic for a partially ordered three-way contingency table. *Austral. N.Z. J. Statist.*, 41(2), 233-246.

Beh, Eric J. and Davy, Pamela J. (2000). A non-iterative alternative to ordinal log-linear models. *University of Wollongong, School of Mathematics and Applied Statistics,* Preprint 8/2000.

Bera, A.K. and McKenzie, C.R. (1986). Tests for normality with stable alternatives. *J. Statist. Comp. Simul.*, 25, 37-52.

Best, D.J. (1990). Multiple comparisons for ranked data. *J. Food Sci.*, 55, 1168-1169.

Best, D.J. (1993). Extended analysis for ranked data. *Austral. J Statist.*, 35, 257-262.

Best, D.J. (1994). Nonparametric comparison of two histograms. *Biometrics*, 50, 538-541.

Best, D.J. (1995). Consumer data - statistical tests for differences in dispersion. *Food Quality and Preference*, 6, 271-280.

Best, D.J. and Rayner, J.C.W. (1985). Lancaster's test of normality. *J. Statist. Planning Infer.*, 12, 395-400.

Best, D.J. and Rayner, J.C.W. (1987). Goodness-of-fit for grouped data using components of Pearson's X^2. *Computational Statistics and Data Analysis*, 5, 53-57.

Best, D.J. and Rayner, J.C.W. (1988). Testing for bivariate normality. *Statistics and Probability Letters*. 6, 407-412.

Best, D.J. and Rayner, J.C.W. (1996). Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics*, 52, 1153-1156.

Best, D.J. and Rayner, J.C.W. (1997a). Goodness of fit for the ordered categories discrete uniform distribution. *Commun. Statist.-Theor. Meth.*, 26(4), 899-909.

Best, D.J. and Rayner, J.C.W. (1997b). Product maps for ranked preference data. *J.R.S.S.*, Series D (*The Statistician*), 46(3), 347-354.

Best, D.J. and Rayner, J.C.W. (1997c). Goodness of fit for the binomial distribution. *Aust. J. Statist.*, 39(3), 355-364.

Best, D.J. and Rayner, J.C.W. (1997d). Crockett's Test of fit for the bivariate Poisson. *Biom. J.*, 39(4), 423-430.

Best, D.J. and Rayner, J.C.W. (1998). Nonparametric analysis of ordinal categorical response data with factorial structure. *Appl. Statist.*, 46, 439-446.

Best, D.J. and Rayner, J.C.W. (1999a). Goodness of fit for the Poisson distribution. *Statistics and Probability Letters,* 44, 259-265.

Best, D.J. and Rayner, J.C.W. (1999b). Nonparametric tests for randomised blocks with ordered alternatives. *Journal of Applied Mathematics and Decision Sciences*, Volume 3(2), 143-153.

Best, D.J. and Rayner, J.C.W. (2000). Tests of Fit for the Geometric Distribution. In preparation.

Best, D.J., Rayner, J.C.W. and O'Sullivan, M.G. (2000). Product maps for consumer categorical data. *Food Quality and Preference*, 11, 91-97.

Best, D.J., Rayner, J.C.W. and Stephens, L.G. (1998). Small sample comparison of McCullagh and Nair analyses for nominal-ordinal categorical data. *Computational Statistics and Data Analysis*, 28, 217-223.

Bhapkar, V.P. (1970). On Cochran's Q test and its modifications. In *Random Counts in Scientific Work*, Volume 2, Editor G.P. Patil, University Park: Pennsylvania State University Press.

Bliss, C.I. (1967). *Statistics in Biology*. Vol. 1, New York: McGraw-Hill.

Boos, D. (1986). Comparing k populations with linear rank statistics. *Journal of the American Statistical Association*, 81, 1018-1025.

Boulerice, B. and Ducharme, G. R. (1995). A note on smooth tests of goodness of fit for location-scale families. *Biometrika*, 82, 437-438.

Box, G. and Jones, S. (1986). Discussion of Nair, V., Testing in industrial experiments with ordered categorical data, *Technometrics*, 28, 283-294; *Technometrics,* 28, 295-301.

Bradley, R.A. (1984). Paired comparisons: some basic procedures and examples. Chapter 14 in *Handbook of Statistics*, Volume 4, Editors P.R. Krishnaiah and P.K. Sen, Amsterdam: North-Holland.

Bradley, R.A., Katti, S.K. and Coons, I.J. (1962). Optimal scaling for ordered categories. *Psychometrika*, 27, 355-374.

Bross, J. (1958). How to use ridit analysis. *Biometrics*, 14, 18-38.

Brown, G.H. (1988). The statistical comparison of reproduction rates for groups of sheep. *Aust. J. Agric. Res*., 39, 899-905.

Buse, A. (1982). The likelihood-ratio, Wald, and Lagrange multiplier tests: an expository note. *Amer. Statistician*, 36, 153-157.

Carolan, A.M. (2000). *Partially Parametric Testing*. Unpublished PhD thesis. University of Wollongong.

Carolan, A.M. and Rayner, J.C.W. (2000a). One sample score tests for modes of nonnormal data. To appear in the *Journal of Applied Mathematics and Decision Sciences*.

Carolan, A.M. and Rayner, J.C.W. (2000b). One sample tests of location for nonnormal symmetric data. *Commun. Statist.-Theor. Meth.*, 29(7), 1569-1581.

Carolan, A.M. and Rayner, J.C.W. (2000c). Wald tests of location for symmetric nonnormal data. *Biometrical Journal*, 42(5), 77-92.

Carolan, A.M. and Rayner, J.C.W. (2000d). A Note on the Asymptotic Behaviour of Smooth Tests of Goodness of Fit. Submitted.

Carolan, A.M. and Rayner, J.C.W. (2000e). Interpreting the Components of a Smooth Goodness of Fit Test for Normality. Submitted.

Coakley, C.W. and Heise, M.A. (1996). Versions of the sign test in the presence of ties. *Biometrics*, 52, 1242-1251.

Cochran, W. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-266.

Conover, W.J. (1980). *Practical Nonparametric Statistics* (2nd ed.). New York: Wiley.

Conover, W.J. (1998). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cressie, N. (1978). Power results for tests on high order gaps. *Biometrika*, 65, 214-218.

D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-fit techniques*. New York: Marcel Dekker.

Daniel, W. (1990). *Applied Nonparametric Statistics* (2nd ed.). Boston: PWS Kent.

David, H.A. (1988). *The Method of Paired Comparisons*. London: Griffith.

Davy, Pamela J., Rayner, J.C.W. and Beh, Eric J. (2000). Generalised Correlations. Submitted.

Devroye, L. (1986). *Non-uniform Random Variate Generation*. New York: Springer-Verlag.

Durbin, J. (1951). Incomplete blocks in ranking experiments. *British Journal of Psychology (Statistical Section)*, 4, 85-90.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Emerson, P.L. (1968). Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24, 695-701.

Eplett, W.J.R. (1982). The distributions of Smirnov type two-sample rank tests for discontinuous distribution functions. *J. Roy. Statist. Soc. B*, 44, 361-369.

Eubank, R. L. and LaRiccia, V.N. (1990). Components of Pearson's Phi-squared distance measure for the k-sample problem. *J. Amer. Stat. Assoc.*, 85, 441-445.

Eubank, R.L., La Riccia, V.N. and Rosenstein, R.B. (1987). Test statistics derived as components of Pearson's phi-squared distance measure. *Journal of the American Statistical Association*, 82, 816-825.

Everitt, B.S. (1992). *The Analysis of Contingency Tables* (2nd ed.). London: Chapman and Hall.

Fienberg, S.E. (1982). Contingency tables. In *Encyclopedia of Statistical Sciences*, 2, Editors S. Kotz and N.L. Johnson, pp.161-171, New York: Wiley.

Fisher, R.A. and Yates, F. (1970). *Statistical Tables for Biological Agricultural and Medical Research*. Edinburgh: Oliver and Boyd.

Gacula, M.C. (1993). *Design and Analysis of Sensory Optimization*. Trumbull: Food and Nutrition Press.

Gacula, M.C. and Singh, J. (1984). *Statistical Methods in Food and Consumer Research*. Orlando, FL: Academic Press.

Gart, J.J. (1969). An exact test for comparing matched proportions in crossover designs. *Biometrika*, 56, 75-80.

Gibbons, J.D. and Chakraborti, S. (1992). *Nonparametric Statistical Inference* (2nd ed.). New York: Marcel Dekker.

Good, P. (1994). *Permutation Tests, A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag: New York.

Graubard, B.J. and Korn, E.L. (1987). Choice of column scores for testing independence in ordered 2XK contingency tables. *Biometrics*, 43, 471-476.

Haberman, S.J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics* 30, 589-600.

Hall, P. (1985). Tailor-made tests of goodness of fit. *J. R. Statist. Soc.*, B, 47, 125-131.

Hamada, M. and Wu, C.F.J. (1990). A critical look at accumulation analysis and related methods. *Technometrics*, 32, 119-130.

Hamdan, M.A. (1974). The use of orthogonal polynomials in the calculation of the noncentrality parameter of chi-squared. *Comm. Statist.*, 3, 157-166.

Hanson, H., Kline, L. and Lineweaver, H. (1951). Applications of balanced incomplete block design to scoring of ten dried egg samples. *Food Technology*, 5: 9-13.

Henze, N. and Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Austral. J. Statist.*, 38(1), 61-74.

Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods* (2nd ed.). New York: Wiley.

Horswell, R. L. and Looney, S.W. (1993). Diagnostic limitations of skewness coefficients in assessing departures from univariate and multivariate normality. *Commun. Statist. B - Simul. Comp.*, 22, 437-459.

IMSL Library (1989). *Users Manual Version 1.1 for PCs*. Vol 1. Houston: IMSL.

ISO5495.2 (1979). *Sensory Analysis - Methodology - Paired Comparison.* Paris: International Organization for Standardization.

Jonckheree, G.R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41, 133-145.

Kallenberg, W.C.M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, 92, 1094-1104.

Kallenberg, W. C. M. and Ledwina, T. (1999). Data-driven rank tests for independence. *Journal of the American Statistical Association*, 94, 285-301.

Kallenberg, W. C. M., Ledwina, T. and Rafajlowicz, E. (1997). Testing bivariate independence and normality. *Sankhya*, A, 59, 42-59.

Kepner, J.L. and Robinson, D.H. (1984). A distribution free rank test for ordered alternatives in randomised complete block designs. *Journal of the American Statistical Association*, 79, 212-217.

Kiefer, J. (1958). K-Sample analogues of the Kolmogorov-Smirnov and Cramer-v. Mises tests. *Ann. Math. Statist.*, 29, 420-447.

Lancaster, H.O. (1953). A reconciliation of χ^2 from metrical and enumerative aspects. *Sankhya*, 13, 1-10.

Lancaster, H.O. (1965). The Helmert matrices. *American Mathematical Monthly*, 72, 4-12.

Lancaster, H.O. (1969). *The Chi-squared Distribution*. New York: Wiley.

Landis, J.R., Cooper, M.M., Kennedy, T. and Koch, G.G. (1979). A computer program for testing average partial association in three-way contingency tables (PARCAT). *Computer Programs in Biomedicine*, 9, 223-246.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.

Lehmann, E.L. (1975). *Nonparametrics*. San Francisco: Holden-Day.

LogXact (1996). LogXact for Windows - Users Manual. Cambridge, MA: CYTEL Software Corporation.

Mack, G.A. and Wolfe, D.A. (1981). K-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association*, 76, 175-181.

McBride, R.L. (1986). Hedonic rating of food: single or side-by-side sample presentation. *J. Food Technology*, 21, 355-363.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, Series B, 42, 109-142.

McCullagh, P. and Nelder, J. (1989). *Generalised Linear Models*. London: Chapman and Hall.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157.

Mehta, C.R. and Patel, N.R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, 14, 2143-2160.

Meilgaard, M., Civille, G.V. and Carr, T.B. (1999). *Sensory Evaluation Techniques* (3rd ed.). Boca Raton: CRC Press.

Miettinen, O.S. (1969). Individual matching with multiple controls in the case of all-or-none response. *Biometrics*, 25, 339-355.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics*. Tokyo: McGraw-Hill Kogakusha.

Nair, V.N. (1986). Testing in industrial experiments with ordered categorical data. *Technometrics*, 28, 283-311.

Nair, V. N. (1987). Chi-squared-type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association*, 82, 283-291.

Nair, V. (1990). Discussion of Hamada, M. and Wu, C.F.J. A critical look at accumulation analysis and related methods. *Technometrics*, 32, 119-130; *Technometrics*, 32, 151-152.

Nam, J. (1971). On two tests for comparing matched proportions. *Biometrics*, 27, 945-959.

Nation, J. R., Bourgeois, A. E., Clark, D. E., Baker, D. M. and Hare, M. F. (1984). The effects of oral cadmium exposure on passive avoidance performance in the adult rat. *Toxicology Letters*, 20, 41-47.

Newell, G.J. (1986). Are rating scales linear? *Australian Marketing Researcher*, 10, 53-63.

Neyman, J. (1937). 'Smooth' test for goodness of fit. *Skand. Aktuarietidskr.*, 20, 150-199.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20, 175-240 and 263-294.

O'Mahony, M. (1986). *Sensory Evaluation of Food: Statistical Methods and Procedures*. New York: Marcel Dekker.

Page, E.B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216-230.

Patefield, W.M. (1981). AS159. An efficient method of generating random rXc tables with given row and column totals. *Appl. Statist.*, 30, 91-97.

Pearson, E.S. and Hartley, H.O. (1970). *Biometrika Tables for Statisticians*. Vol. 1 (3rd ed.). New York: Cambridge University Press.

Pirie, W.R. (1985). Page test for ordered alternatives. In *Encyclopedia of Statistical Sciences*. Vol 6 (editors S. Kotz and N.L. Johnson), pp.553-555. New York: Wiley.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population. Supplement to the *Journal of the Royal Statistical Society*, 4, 119-130 (225).

Putter, J. (1955). The treatment of ties in some nonparametric tests. *Annals of Mathematical Statistics*, 26, 368-386.

Rao, C.R. (1948). Tests of significance in multivariate analysis. *Biometrika*, 35, 58-79.

Rao, C.R. (1965). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Rayner, J.C.W. (1986). Constructing a usable overlapping cells X^2 goodness of fit test. *Statist. Prob. Letters*, 6, 257-261.

Rayner, J.C.W. (1997). The Asymptotically Optimal Tests. *J.R.S.S.*, Series D (*The Statistician*), 46(3), 337-346.

Rayner, J.C.W. and Best, D.J. (1986). Neyman-type smooth tests for location-scale families. *Biometrika*, 73, 437-446.

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

Rayner, J.C.W. and Best, D.J. (1989b). Components and power comparisons of some rank tests used in taste-testing. *University of Otago Department of Mathematics and Statistics Report Series*, Research Paper No. 20, October 1989.

Rayner, J.C.W. and Best, D.J. (1990a). Smooth tests of goodness of fit: an overview. *International Statistical Review*, 58, 1, 9-17.

Rayner, J.C.W. and Best, D.J. (1990b). A comparison of some rank tests used in taste-testing. *J. Royal Society of N.Z.*, 30, 2, 269-272.

Rayner, J.C.W. and Best, D.J. (1995a). Smooth tests for the bivariate Poisson. *Aust. J. Statist.*, 37(2), 233-245.

Rayner, J.C.W. and Best, D.J. (1995b). Extensions to the Friedman and Durbin Rank tests. *IAPP Biometrics Unit Report Series*, 95/1.

Rayner, J.C.W. and Best, D.J. (1996a). Smooth extensions of Pearson's product moment correlation and Spearman's rho. *Statistics and Probability Letters*, 30(2), 171-177.

Rayner, J.C.W. and Best, D.J. (1996b). Extensions to some important nonparametric tests. In *Proceedings of the A.C. Aitken Centenary Conference, Dunedin 1995*, Editors L. Kavalieris, F.C. Lam, L.A. Roberts and J.A. Shanks, pp.257-266. Dunedin: University of Otago Press.

Rayner, J.C.W. and Best, D.J. (1997a). How order affects the sign test. *Biometrics*, 53(4), 1416-1421.

Rayner, J.C.W. and Best, D.J. (1997b). Extensions to the Kruskal-Wallis test and a generalised median test with extensions. *Journal of Applied Mathematics and Decision Sciences*, 1(1), 13-25.

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. (2000). Analysis of singly ordered two-way contingency tables. *Journal of Applied Mathematics and Decision Sciences*, 4(1), 83-98.

Rayner, J.C.W., Best, D.J. and Dodds, K.G. (1985). The construction of the simple X^2 and Neyman smooth goodness of fit tests. *Statist. Neerl.*, 39, 35-50.

Rayner, J.C.W., Best, D.J. and Mathews, K.L. (1995). Interpreting the skewness coefficient. *Commun. Statist. A - Theor. Meth.*, 24, 593-600.

Rayner, J.C.W. and McAlevey, L.G. (1990). Smooth goodness of fit tests for categorised composite null hypotheses. *Statistics and Probability Letters*, 9, 423-429.

Rayner, J.C.W. and Rayner, G.D. (1997). S-sample smooth goodness of fit tests. *Mathematical Scientist*, 22(2), 106-116.

Rayner, J.C.W. and Rayner, G.D. (1998). S-sample smooth goodness of fit tests: Rederivation and Monte Carlo Assessment. *Biom. J.*, 40, 651-663.

Risebrough, R.W. (1972). Effects of environmental pollutants upon animals other than man. *Proceedings of the 6th Berkeley Symposium on Mathematics and Statistics VI*, Berkeley: Univ. of Calif. Press, pp.443-463.

Roy, S.N. and Mitra, S.K. (1956). An introduction to some non-parametric generalisations of analysis of variance and multivariate analysis. *Biometrika*, 43, 361.

Schach, S. (1979). An alternative to the Friedman test with certain optimality properties. *Annals of Statistics*, 7, 537-550.

Scholz, F.W. and Stephens, M.A. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82, 918-924.

Seber, G.A.F. (1984). *Multivariate Observations*. New York: Wiley.

Shah, A.K. and Claypool, P.L. (1985). Analysis of binary data in randomized complete block designs. *Commun. Statist. - Theor. Meth.*, 14, 1175-1179.

Siegel, S. and Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioural Sciences* (2nd ed.). Singapore: McGraw-Hill.

Skillings, J.H. and Mack, G.A. (1981). On the use of a Friedman type statistic in balanced and unbalanced block designs. *Technometrics*, 23, 171-177.

Sprent, P. (1993). *Applied Nonparametric Statistical Methods* (2nd ed.). London: Chapman and Hall.

Sprent, P. (1998). *Data Driven Statistical Methods*. London: Chapman and Hall.

StatXact (1995). *Statistical Software for Exact Nonparametric Inference*. Cambridge, MA: CYTEL Software Corporation.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416.

Stuart, A. and Ord, J.K. (1991). *Kendall's Advanced Theory of Statistics*. Vol. 2 (5th ed.), London: Edward Arnold.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*. Vol. 1 (6th ed.), London: Edward Arnold.

Taguchi, G. (1966). *Statistical Analysis* (in Japanese). Tokyo: Maruzen.

Testimate (1994). *Test and Estimation*. Gauting/Munich. IDV-Datenanalyse und Versuchsplanung.

Van der Laan, P. (1988). The use of Durbin's rank test. *The American Statistician*, 42, 165-166.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, 54, 426-482.

Wittkowski, K.M. (1989). An asymptotic UMP sign test for discretized data. *The Statistician*, 38, 93-96.

Wittkowski, K.M. (1998). Versions of the sign test in the presence of ties. *Biometrics*, 54, 89-91.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, 35, 176-181.

Zar, J.H. (1984). *Biostatistical Analysis* (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall Inc.