*Statistical Methods in Medical Research* 1994; **3**: 135–156

# The exact analysis of contingency tables in medical research

**Cyrus R Mehta** Harvard School of Public Health, and Cytel Software Corporation

A unified view of exact nonparametric inference, with special emphasis on data in the form of contingency tables, is presented. While the concept of exact tests has been in existence since the early work of RA Fisher, the computational complexity involved in actually executing such tests precluded their use until fairly recently. Modern algorithmic advances, combined with the easy availability of inexpensive computing power, has renewed interest in exact methods of inference, especially because they remain valid in the face of small, sparse, imbalanced, or heavily tied data. After defining exact $p$-values in terms of the permutation principle, we reference algorithms for computing them. Several data sets are then analysed by both exact and asymptotic methods. We end with a discussion of the available software.

## 1 Introduction

Modern statistical methods rely heavily on nonparametric techniques for comparing two or more populations. These techniques generate $p$-values without making any distributional assumptions about the populations being compared. However they rely on asymptotic theory that is valid only if the sample sizes are reasonably large and well balanced across the populations. For small, sparse, skewed, or heavily tied data, the asymptotic theory may not be valid. See Agresti and Yang[1] for some empirical results, and Read and Cressie[2] for a more theoretical discussion.

One way to make valid statistical inferences in the presence of small, sparse or imbalanced data is to compute exact $p$-values, based on the permutational distribution of the test statistic. This approach was first proposed by RA Fisher,[3] and has been used extensively for the single $2 \times 2$ contingency table. In the past, exact tests were rarely attempted for tables of higher dimension than $2 \times 2$, primarily because of the formidable computing problem involved in their execution. Two developments over the past ten years have removed this obstacle. First, the easy availability of immense quantities of computing power in homes and offices has revolutionized our thinking about what is computationally affordable. Second, many new, fast and efficient algorithms for exact permutational inference have recently been published. Thus problems that would previously have taken several hours or even days to solve now take only a few minutes. It only remained to incorporate these algorithms into friendly, well documented statistical packages. Now this step also has been accomplished. In the present paper we present a unified framework for exact nonparametric inference, anchored in the permutation principle. We demonstrate that exact statistical inference for a very broad class of nonparametric problems can be accomplished by permuting the entries in a contingency table subject to fixed margins. Exact and Monte Carlo algorithms for solving these permutation problems are referenced but not described. We then apply these algorithms to several data sets in the form of unordered, singly ordered and doubly ordered

Address for correspondence: Cyrus R Mehta, Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139, USA.

contingency tables. Both exact and asymptotic $p$-values are computed for these data so that one may assess the accuracy of the asymptotic methods. Finally we discuss the availability of software to implement the algorithms.

## 2 Nonparametrics and the permutation principle

For a broad class of statistical tests the data can be represented in the form of the $r \times c$ contingency table $\mathbf{x}$ displayed as Table 1:

**Table 1** Layout for a generic $r \times c$ contingency table

| Rows | Col 1 | Col 2 | ... | Col $c$ | Row total |
|---|---|---|---|---|---|
| Row 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1c}$ | $m_1$ |
| Row 2 | $x_{21}$ | $x_{22}$ | ... | $x_{2c}$ | $m_2$ |
| ⋮ | ⋮ | ⋮ | ... | ⋮ | ⋮ |
| Row $r$ | $x_{r1}$ | $x_{r2}$ | ... | $x_{rc}$ | $m_r$ |
| Column total | $n_1$ | $n_2$ | ... | $n_c$ | $N$ |

The entry in each cell of this $r \times c$ table is the number of subjects falling in the corresponding row and column classifications. The row and column classifications may be based on either *nominal* or *quantitative* variables. Nominal variables take on values which cannot be positioned in any natural order. An example of a nominal variable is profession—medicine, law, business. In some statistical packages, nominal variables are also referred to as *class* variables, or *unordered* variables. Quantitative variables take on values which can be ordered in a natural way. An example of a quantitative variable is drug dose—low, medium, high. Quantitative variables may of course assume numerical values as well (for example, the number of cigarettes smoked per day).

### 2.1 Unconditional sampling distributions

The exact probability distribution of $\mathbf{x}$ depends on the sampling scheme that was used to generate $\mathbf{x}$. When both the row and column classifications are categorical, Agresti[4] lists three sampling schemes that could give rise to $\mathbf{x}$; full multinomial sampling, product multinomial sampling, and Poisson sampling. Under all three schemes the probability distribution of $\mathbf{x}$ contains unknown parameters relating to the individual cells of the $r \times c$ table.

Under full multinomial sampling a total of $N$ items are sampled independently, and $x_{ij}$ of them are classified as belonging to row-category $i$ and column-category $j$, each with probability $\pi_{ij}$. Thus the probability of observing the table $\mathbf{x}$ is

$$P(\mathbf{x}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{N! \pi_{ij}^{x_{ij}}}{x_{ij}!}. \tag{2.1}$$

Full multinomial sampling might arise for example if one were to sample $N$ hospital patients and classify them according to their race (white, black, other) and their major medical insurance (Blue Cross, HMO, other). One would be interested in testing the null hypothesis that race and insurance plan were independent. Formally let $\pi_{i.}$ be the marginal probability of falling in row-category $i$, and $\pi_{.j}$ be the marginal probability of falling in column-category $j$. The null hypothesis assumes that $\pi_{ij} = \pi_{i.} \pi_{.j}$.

Under product multinomial sampling a predetermined number, $m_i$, of items are

sampled independently from population $i$, and $x_{ij}$ of them are classified as falling into category $j$. Let $\pi_{ij}$ be the conditional probability that an item will fall into category $j$ given that it was sampled from population $i$. Thus the probability of observing table $\mathbf{x}$ is

$$P(\mathbf{x}) = \prod_{i=1}^{r} \frac{m_i! \prod_{j=1}^{c} \pi_{ij}^{x_{ij}}}{\prod_{j=1}^{c} x_{ij}!}. \tag{2.2}$$

Product multinomial sampling might arise for example if $r$ drug therapies were being tested in a clinical trial, $m_i$ patients were treated with drug $i$, and each patient fell into one of $c$ possible categories of response. One would be interested in testing the null hypothesis that the probability of falling into response category $j$ was the same for all $i$, i.e., the drugs are all equivalent in terms of response. Formally let $\pi_{ij}$ be the probability that an individual treated with drug $i$ manifests the response $j$. The null hypothesis assumes that $\pi_{ij} = \pi_j$, for all $j = 1, 2, \ldots, c$, independent of $i$.

Under Poisson sampling cell $(i, j)$ of the contingency table accumulates events at a Poisson rate of $\pi_{ij}$ so that the probability of observing table $\mathbf{x}$ is

$$P(\mathbf{x}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{(N\pi_{ij})^{x_{ij}} e^{-N\pi_{ij}}}{x_{ij}!}. \tag{2.3}$$

Poisson sampling might arise for example if the entry in cell $(i, j)$ represented the number of induced abortions in district $i$ in year $j$. One would be interested in testing the null hypothesis that the abortion rate did not change from year to year within a district. Formally the null hypothesis would assume that the Poisson parameter $\pi_{ij} = \pi_{i.}\pi_{.j}$ where $\pi_{i.}$ is the marginal rate for district $i$ and $\pi_{.j}$ is the marginal rate for year $j$.

Notice that the above probability distributions for $\mathbf{x}$ depend on a total of $rc$ unknown parameters, $\pi_{ij}$, $(i = 1, 2, \ldots, r)$, $(j = 1, 2, \ldots, c)$. Since statistical inference is based on the distribution of $\mathbf{x}$ under the null hypothesis of independence of row and column classifications, the number of unknown parameters is reduced ($\pi_{ij}$ being replaced by $\pi_{i.}\pi_{.j}$ or $\pi_j$ depending on the sampling scheme) but not eliminated. Unknown nuisance parameters still remain in equations (2.1), (2.2) and (2.3) even after assuming that the null hypothesis is true. Asymptotic inference relies on estimating these unknown parameters by maximum likelihood and related methods. But in exact inference we eliminate nuisance parameters by conditioning on their sufficient statistics. This is discussed next.

## 2.2 Exact conditional sampling distributions

The key to exact nonparametric inference is getting rid of all nuisance parameters from the probability distribution of $\mathbf{x}$. This is accomplished by restricting the sample space to the set of all $r \times c$ contingency tables that have the same marginal sums as the observed table $\mathbf{x}$. Specifically, define the references set

$$\Gamma = \left\{ \mathbf{y} : \mathbf{y} \text{ is } r \times c;\ \sum_{j=1}^{c} y_{ij} = m_i;\ \sum_{i=1}^{r} y_{ij} = n_j; \text{ for all } i, j \right\} \tag{2.4}$$

Then one can show that, under the null hypothesis of no row and column interaction, the probability of observing any $\mathbf{y} \in \Gamma$ is

$$P(\mathbf{y} \mid \mathbf{y} \in \Gamma) \equiv P(\mathbf{y}) = \frac{\prod_{j=1}^{c} n_j! \prod_{i=1}^{r} m_i!}{N! \prod_{j=1}^{c} \prod_{i=1}^{r} y_{ij}!}. \tag{2.5}$$

Equation (2.5), which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate $\mathbf{x}$ is full multinomial, product multinomial, or Poisson.[5]

Since (2.5) contains no unknown parameters, exact inference is possible. However the nuisance parameters were eliminated by conditioning on the margins of the observed contingency table. Now these margins were not fixed when the data were gathered. Thus it is reasonable to question the appropriateness of fixing them for purposes of inference. The justification for conditioning at inference time on margins that were not naturally fixed at data sampling time has a long history. RA Fisher[3] first proposed this idea for exact inference on a single $2 \times 2$ contingency table. At various times since then prominent statisticians have commented on this approach. The two reasons most cited for conditioning are *convenience* and *ancillarity*.

**Convenience** The margins of the contingency table do not contain any information about the hypothesis under test. Since they are the sufficient statistics for the nuisance parameters, conditioning affords a convenient way to eliminate nuisance parameters and thereby perform exact inference without loss of information.

**Ancillarity** The principle underlying hypothesis testing is to compare what was actually observed with what could have been observed in hypothetical repetitions of the original experiment, under the null hypothesis. In these hypothetical repetitions it is a good idea to keep all experimental conditions unchanged as far as possible. The margins of the contingency table are representative of the nuisance parameters. Fixing them in hypothetical repetitions is the nearest we can get to fixing the values of the nuisance parameters themselves in hypothetical repetitions, since the latter are unknown.

An excellent exposition of the conditional viewpoint is available in Yates.[6] For a theoretical justification refer to Cox and Hinkley.[7] Throughout the present paper we shall adopt the conditional approach. It provides us with a unified way to perform exact inference and thereby compute accurate $p$-values and confidence intervals even when the observed $r \times c$ contingency table has small cell counts.

### 2.3 Exact $p$-value computation

Having assigned an exact probability $P(\mathbf{y})$ to each $\mathbf{y} \in \Gamma$, the next step is to order each contingency table in $\Gamma$ by a test statistic or 'discrepancy measure' that quantifies the extent to which that table deviates from the null hypothesis of no row and column interaction. Let us denote the test statistic by a real valued function $D: \Gamma \rightarrow \mathcal{R}$ mapping $r \times c$ tables from $\Gamma$ onto the real line $\mathcal{R}$. The functional form of $D$ for some important nonparametric tests is specified in the next subsection.

The $p$-value is defined as the sum of null probabilities of all the tables in $\Gamma$ which are at least as extreme as the observed table, $\mathbf{x}$, with respect to $D$. In particular if $\mathbf{x}$ is the observed $r \times c$ table, the exact $p$-values are obtained by computing

$$p = \sum_{D(\mathbf{y}) \geq D(\mathbf{x})} P(\mathbf{y}) = p\{D(\mathbf{y}) \geq D(\mathbf{x})\}. \tag{2.6}$$

Classical nonparametric methods rely on the large-sample distribution of $D$ to estimate $p$. For $r \times c$ tables with large cell counts it is possible to show that $D$ converges to a chi-square distribution with appropriate degrees of freedom. Thus $p$ is usually estimated by $\hat{p}$, the chi-square tail area to the right of $D(\mathbf{x})$. Modern algorithmic

techniques have made it possible to compute $p$ directly instead of relying on $\hat{p}$, its asymptotic approximation. This is achieved by powerful recursive algorithms[8] that are capable of generating the actual permutation distribution of $D$ instead of relying on its asymptotic chi-square approximation. We shall see later that $p$ and $\hat{p}$ can differ considerably for contingency tables with small cell counts.

## 2.4 Choosing the test statistic

As stated previously, the reference set $\Gamma$ is ordered by the test statistic $D$. Here we define $D$ for three important classes of problems; general tests on $r \times c$ contingency tables, linear rank tests on $2 \times c$ contingency tables, and odds ratio tests on stratified $2 \times 2$ contingency tables.

### Tests on $r \times c$ contingency tables

Different test statistics are appropriate for different types of $r \times c$ contingency tables. When both the row and column classifications of the table are nominal the Fisher, Pearson and Likelihood ratio statistics are the most appropriate. Tests based on these three statistics are known as omnibus tests for they are powerful against any general alternative to the null hypothesis.

*Fisher* Fisher's exact test orders the tables in $\Gamma$ in proportion to their hypergeometric probabilities. Specifically, the test statistic for each $\mathbf{y} \in \Gamma$ is

$$D(\mathbf{y}) = -2 \log(\gamma P(\mathbf{y})) \tag{2.7}$$

where

$$\gamma = (2\pi)^{(r-1)(c-1)/2} (N)^{-(rc-1)} \prod_{i=1}^{r} (m_i)^{(c-1)/2} \prod_{j=1}^{c} (n_j)^{(r-1)/2}.$$

Fisher[3] originally proposed this test for the single $2 \times 2$ contingency table. The idea was extended to tables of higher dimension by Freeman and Halton.[9] Thus, this test is also referred to as the Freeman–Halton test. Asymptotically, under the null hypothesis of row and column independence, the Freeman–Halton statistic has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.[10]

*Pearson* The Pearson test orders the tables in $\Gamma$ according to their Pearson chi-squared statistics. Thus, for each $\mathbf{y} \in \Gamma$ the test statistic is

$$D(\mathbf{y}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - m_i n_j / N)^2}{m_i n_j / N}. \tag{2.8}$$

Asymptotically, under the null hypothesis of row and column independence, the Pearson statistic has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.[4]

*Likelihood ratio* The Likelihood ratio test[4] orders the tables in $\Gamma$ according to the likelihood ratio statistic. Specifically, for each $\mathbf{y} \in \Gamma$ the test statistic is

$$D(\mathbf{y}) = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} \log\left(\frac{y_{ij}}{m_i n_j / N}\right). \tag{2.9}$$

In many textbooks this statistic is denoted by $G^2$. Asymptotically, under the null hypothesis of row and column independence, $D(\mathbf{y})$ has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.[4]

When there is a natural ordering of the columns of the $r \times c$ table, but the row classifications are based on nominal categories, the appropriate test is the Kruskal–Wallis.[4] One can think of the Kruskal–Wallis test as the nonparametric version of one-way ANOVA. It is used to test the equality of $r$ populations with ordered outcomes. For example, suppose that the $r$ rows represent $r$ different drug therapies, and the $c$ columns represent $c$ distinct ordered responses (such as: no response, mild response, moderate response, severe response, etc). The Kruskal–Wallis statistic is more powerful than the Fisher, Pearson or Likelihood ratio statistics for detecting shifts in response among the $r$ populations. When there are only two rows in the contingency table, the Kruskal–Wallis test specializes to the Wilcoxon-rank-sum test.

*Kruskal–Wallis* The Kruskal–Wallis test orders the table in $\Gamma$ according to the Kruskal–Wallis statistic. Specifically, for each $\mathbf{y} \in \Gamma$ the test statistic is

$$D(\mathbf{y}) = \frac{12}{N(N+1)[1-(\lambda/(N^3-N))]} \sum_{i=1}^{r} [R_i(\mathbf{y}) - m_i(N+1)/2]^2/m_i, \qquad (2.10)$$

where $\lambda$ is the tie correction factor $\sum_{j=1}^{c}(n_j^3 - n_j)$, and

$$R_i(\mathbf{y}) = y_{i1}(n_1+1)/2 + y_{i2}[n_1 + (n_2+1)/2] + \ldots + y_{ic}\left[\sum_{j=1}^{c-1} n_j + (n_c+1)/2\right].$$

Asymptotically, under the null hypothesis that the $r$ populations are the same, $D(\mathbf{y})$ has a chi-squared distribution with $(r-1)$ degrees of freedom.

When the $r \times c$ contingency table has a natural ordering along both its rows and its columns, the Jonckheere–Terpstra test[11] and the Linear-by-Linear association test[4] have more power than the Kruskal–Wallis test. For example suppose the $r$ rows represent $r$ distinct drug therapies at progressively increasing drug doses and the $c$ columns represent $c$ ordered responses. Now one would be interested in detecting alternatives to the null hypothesis in which drugs administered at larger doses are more responsive than drugs administered at smaller doses. The Jonckheere–Terpstra and Linear-by-linear association test statistics cater explicitly to such alternatives for they are better able to pick up departures from the null hypothesis in which the response distribution shifts progressively towards the right as we move down the rows of the contingency table.

*Jonckheere–Terpstra* The tables in $\Gamma$ are ordered to the Jonckheere–Terpstra statistic, which is really just a sum of $r(r-1)/2$ Wilcoxon–Mann–Whitney statistics. Specifically, for each $\mathbf{y} \in \Gamma$ the test statistic is

$$D(\mathbf{y}) = \sum_{i=2}^{r} \sum_{j=1}^{i-1} \sum_{k=1}^{c} [w_{ijk} y_{ik} - m_i(m_i+1)/2], \qquad (2.11)$$

where the $w_{ijk}$ values are the Wilcoxon scores corresponding to a $2 \times c$ table formed from rows $i$ and $j$ of the full $r \times c$ table. Thus, for $k = 1, \ldots, c$,

$$w_{ijk} = [(y_{i1} + y_{j1}) + \ldots + (y_{i,k-1} + y_{j,k-1}) + (y_{i,k} + y_{j,k} + 1)/2].$$

Under the null hypothesis that the $r$ populations are the same, the Jonckheere–Terpstra statistic has a mean

$$E(D(\mathbf{y})) = \left(N^2 - \sum_{i=1}^{r} m_i^2\right)/4 \qquad (2.12)$$

and a variance

$$\begin{aligned}\operatorname{var}(D(\mathbf{y})) = \frac{1}{72}\Bigg[N(N-1)(2N+5) - \sum_{i=1}^{r} m_i(m_i-1)(2m_i+5) \\ - \sum_{j=1}^{c} n_j(n_j-1)(2n_j+5)\Bigg] \\ + \frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^{r} m_i(m_i-1)(m_i-2)\right] \\ \times\left[\sum_{j=1}^{c} n_j(n_j-1)(n_j-2)\right] \\ + \frac{1}{8N(N-1)}\left[\sum_{i=1}^{r} m_i(m_i-1)\right]\left[\sum_{j=1}^{c} n_j(n_j-1)\right].\end{aligned}$$

The asymptotic distribution of

$$Z = \frac{D(\mathbf{y}) - E(D(\mathbf{y}))}{\sqrt{\operatorname{var}(D(\mathbf{y}))}} \tag{2.13}$$

is normal with mean 0 and variance 1.

*Linear-by-linear association* The tables in $\Gamma$ are ordered according to the linear rank statistic

$$D(\mathbf{y}) = \sum_{i=1}^{r}\sum_{j=1}^{c} u_i v_j y_{ij}, \tag{2.14}$$

where $u_i$, $i = 1, 2, \ldots, r$, are arbitrary row scores, and $v_j$, $j = 1, 2, \ldots, c$, are arbitrary column scores. Under the null hypothesis of no row by column interaction the test statistic has a mean

$$E(D(\mathbf{y})) = N^{-1}\left(\sum_{i=1}^{r} u_i m_i\right)\left(\sum_{j=1}^{c} v_j n_j\right), \tag{2.15}$$

and a variance

$$\operatorname{var}(D(\mathbf{y})) = (N-1)^{-1}\left[\sum_i u_i^2 m_i - \frac{(\sum_i u_i m_i)^2}{N}\right]\left[\sum_j v_j^2 n_j - \frac{(\sum_j v_j n_j)^2}{N}\right]. \tag{2.16}$$

See page 284 and page 303 (problem 8.29) of Agresti[4] for more information. The asymptotic distribution of

$$Z = \frac{D(\mathbf{y}) - E(D(\mathbf{y}))}{\sqrt{\operatorname{var}(D(\mathbf{y}))}} \tag{2.17}$$

is normal with mean 0 and variance 1.

The freedom to select the $u_i$ and $v_j$ scores arbitrarily is a powerful feature of the Linear-by-linear test.[12] If $u$ and $v$ represent the original raw data, the Linear-by-linear test is a test of significance for Pearson's correlation coefficient. On the other hand if the raw data are replaced by ridit or mid-rank scores, we have a test of Spearman's correlation coefficient. For the special case of the $2 \times c$ contingency table the

Linear-by-linear test statistic yields a rich class of linear rank tests. These are defined next.

**Linear rank tests**

For the special case of the $2 \times c$ contingency table, the Linear-by-linear association test reduces to the family of linear rank tests

$$D(\mathbf{y}) = \sum_{j=1}^{c} v_j y_{1j}. \tag{2.18}$$

Since we are conditioning on the column sums, it is not necessary to sum over the second row. The score $u_i$ has therefore been dropped from the expression for $D$ without any loss of generality.

The mean and variance of $D$, under the null hypothesis of no row and column interaction, and conditional on $\mathbf{y} \in \Gamma$, can be derived from equations (2.5) and (2.18). The mean is

$$E(D) = \left(\frac{m_1}{N}\right) \sum_{j=1}^{c} v_j n_j. \tag{2.19}$$

The variance is

$$\sigma^2 = \left[\frac{m_1 m_2}{N(N-1)}\right] \sum_{j=1}^{c} \left[v_j - \frac{E(D)}{m_1}\right]^2 n_j. \tag{2.20}$$

By the Chernoff–Savage theorem,[13] the standardized test statistic

$$Z = \frac{D - E(D)}{\sigma} \tag{2.21}$$

converges in distribution to the standard normal distribution with a mean of 0 and unit variance, under suitable regularity conditions on the scores.

Different choices of scores $v_j$ yield different linear rank tests. These scores and the conditions under which to use each test are specified below:

*Wilcoxon scores* The Wilcoxon scores are the ranks (mid-ranks in the case of tied observations) of the underlying responses.

$$v_j = n_1 + \cdots n_{j-1} + (n_j + 1)/2. \tag{2.22}$$

The Wilcoxon rank-sum test[14] is one of the most popular nonparametric tests for detecting a shift in location between two populations. It has an asymptotic relative efficiency of 95.5%, relative to the $t$ test when the underlying distributions are normal. If there is censoring in the data, the scores defined by equation (2.22) are replaced by the generalized Wilcoxon–Gehan scores, as discussed in Kalbfleisch and Prentice.[15] In particular, let $a_1, a_2, \ldots, a_g$, be the $g$ distinct death times. Let $n_1, n_2, \ldots, n_g$, be the number of deaths and $r_1, r_2, \ldots, r_g$, be the number at risk at these death times. The score assigned to all $n_j$ subjects who die at time $a_j$ is

$$v_{aj} = 1 - \frac{2}{n_j}\left[\sum_{i=c_{j-1}+1}^{c_j} \prod_{l=1}^{i} \left(\frac{N-l+1}{N-l+2}\right)\right], \tag{2.23}$$

where $c_j = n_1 + n_2 + \cdots + n_j$. For all subjects who are censored between the two death times $a_j$ and $a_{j+1}$, the corresponding scores are

$$v_{a_{j+}} = 1 - \prod_{l=1}^{c_j} \left( \frac{N-l+1}{N-l+2} \right) \Bigg\}. \tag{2.24}$$

Scores for all subjects censored prior to the first failure time are zero. Scores for all subjects censored past the last failure time are computed by (2.23). This convention ensures that the sum of Wilcoxon–Gehan scores over all subjects, and hence the expected value of the Wilcoxon–Gehan statistic, is always zero.

*Normal scores* The scores for the Normal scores (or Van der Waerden) test are the percentiles of the standard normal distribution:

$$v_j = \frac{1}{n_j} \left[ \sum_{i=c_{j-1}+1}^{c_j} \Phi^{-1} \left( \frac{i}{N+1} \right) \right] \tag{2.25}$$

where $c_j = n_1 + n_2 + \cdots + n_j$ and $\Phi^{-1}(\alpha)$is the $100\alpha$th percentile of the standard normal distribution. The Normal scores test[14] is an alternative to the Wilcoxon rank sum test for comparing two populations. It is a nonparametric test with 100% asymptotic relative efficiency relative to the $t$ test when the underlying distributions are normal with shifted means. If the tails of the distributions are diffuse, this test is less powerful than the Wilcoxon.

*Savage scores* The scores for the Savage test, also known as the exponential scores test, are defined by:

$$v_j = \frac{1}{n_j} \left[ \sum_{i=c_{j-1}+1}^{c_j} \sum_{l=1}^{i} \left( \frac{1}{N-l+1} \right) \right] - 1 \tag{2.26}$$

where $c_j = n_1 + n_2 + \cdots + n_j$. The Savage test is a locally most powerful test.[16]

*Logrank scores* Logrank scores are used for censored survival data.[15] They are defined as follows. Let $a_1, a_2, \ldots, a_g$, be the $g$ distinct death times. Let $n_1, n_2, \ldots, n_g$, be the number of deaths and $r_1, r_2, \ldots, r_g$, be the number at risk at these death times. The score assigned to all $n_j$ subjects who die at time $a_j$ is

$$v_{a_j} = \frac{1}{n_j} \left[ \sum_{i=c_{j-1}+1}^{c_j} \sum_{l=1}^{i} \frac{1}{N-l+1} \right] - 1, \tag{2.27}$$

where $c_j = n_1 + \cdots + n_j$. For all subjects who are censored between the two death times $a_j$ and $a_{j+1}$, the corresponding scores are

$$v_{a_{j+}} = \sum_{l=1}^{c_j} \frac{1}{N-l+1}. \tag{2.28}$$

Scores for all subjects censored prior to the first failure time are zero. Scores for all subjects censored past the last failure time are computed by (2.27). This convention ensures that the sum of Logrank scores over all subjects, and hence the expected value of the Logrank statistic, is always zero. It can easily be seen that for uncensored data the Logrank scores specialize to the Savage scores defined previously. The Logrank test is a competitor to the Wilcoxon–Gehan test for censored data. It is the optimal test against proportional hazard alternatives. However for non-proportional hazards, with

early differences in the hazard rates, or crossing hazard functions, the Wilcoxon–Gehan test is more powerful.

*Trend* The Trend test[17] uses the equally spaced scores

$$v_j = j. \tag{2.29}$$

It is also known as the Cochran–Armitage trend test and is a very popular test of a dose-response relationship among $c$ binomial populations, where the $j$the population is sampled $n_j$ times and each member of the sample is exposed to dose $w_j$. The probability of a response for each sample is $\pi_j$. The null hypothesis is that

$$\pi_1 = \pi_2 = \cdots = \pi_c.$$

The alternative hypothesis is that there is a trend whereby the binomial probabilities, $\pi_j$, increase with increasing dose $w_j$. A variant of the Cochran–Armitage trend test uses the actual doses, $w_j$, or their logarithms, as the scores instead of replacing them by the equally spaced scores.

**Tests on stratified $2 \times 2$ contingency tables**

A very important class of exact nonparametric tests and confidence intervals is defined on data in the form of several $2 \times 2$ contingency tables. The $i$th table is of the form, shown in Table 2,

**Table 2** Layout for the $i$th of $s$ $2 \times 2$ contingency tables

| Rows | Col 1 | Col 2 | Row total |
|---|---|---|---|
| Row 1 | $y_i$ | $x_i$ | $m_i$ |
| Row 2 | $y_i'$ | $x_i'$ | $m_i'$ |
| Column total | $N_i - n_i$ | $n_i$ | $N_i$ |

for $i = 1, 2, \ldots, s$. We may regard the two rows of each table as arising from two independent binomial distributions. Specifically, let $(x_i, x_i')$represent the number of successes in $(m_i, m_i')$ Bernoulli trials, with respective success probabilities $(\pi_i, \pi_i')$. The odds ratio for the $i$th table is defined as

$$\Psi_i = \left(\frac{\pi_i}{1-\pi_i}\right) \Big/ \left(\frac{\pi_i'}{1-\pi_i'}\right). \tag{2.30}$$

Stratified $2 \times 2$ contingency tables arise commonly in prospective studies with binary end points as well as in retrospective case-control studies. Thus although we have specified that the two rows of the $2 \times 2$ table represent two independent binomial distributions, this is just a matter of notational convenience. We could equivalently assume that the two rows represent the disease status and the two columns represent the exposure status in a case-control setting.

We shall be interested in testing the null hypothesis that

$$\Psi_i = \Psi \text{ for } i = 1, 2, \ldots, s.$$

This is known as the homogeneity test. Next, under the assumptions of homogeneity, we shall be interested in estimating the common odds ratio, $\Psi$. In order to formulate these two problems we need to extend the notation developed previously for the reference set $\Gamma$ of $r \times c$ contingency tables with fixed margins. Accordingly let $\tau$ denote a

generic set of $s$ $2 \times 2$ tables. Let $\tau_0$ denote a specific realization of $\tau$. Exact inference, both for testing that the odds ratio across $s$ $2 \times 2$ tables is constant as well as for estimating the common odds ratio, is based on determining how extreme the observed $\tau_0$ is relative to other $\tau$'s that could have been observed in some reference set. Different reference sets are used for testing the homogeneity of odds ratios and for estimating the common odds ratio. Define the reference set

$$\Omega = \left\{ \tau: \begin{array}{l} x_i + y_i = m_i;\ x_i' + y_i' = m_i';\\ x_i + x_i' = n_i;\ y_i + y_i' = N_i - n_i \end{array} \right\}. \tag{2.31}$$

Also define the more restricted reference set

$$\Omega_t = \{\tau \in \Omega: x_1 + x_2 + \cdots + x_s = t\}. \tag{2.32}$$

An exact test for homogeneity of the odds ratios is based on ordering the $\tau$'s in $\Omega_t$, while exact inference about the common odds ratio is based on ordering the $\tau$'s in $\Omega$. These two exact procedures are discussed next. For completeness, a corresponding asymptotic procedure is also provided next to each exact procedure.

*Homogeneity test* Zelen[18] developed an exact test for the null hypothesis

$$H_0: \Psi_i = \Psi,\ i = 1, 2, \ldots, s.$$

Zelen's test is based on the fact that under $H_0$ the probability of observing any $\tau$ from the conditional reference set $\Omega_t$ is a product of hypergeometric probabilities which does not depend on the nuisance parameter $\Psi$. Specifically, the conditional probability of obtaining any $\tau \in \Omega_t$ is

$$P(\tau \mid t) = \frac{\prod_{i=1}^{s} \binom{m_i}{x_i}\binom{m_i'}{x_i'} \Big/ \binom{N_i}{n_i}}{\sum_{\tau \in \Omega_t} \prod_{i=1}^{s} \binom{m_i}{x_i}\binom{m_i'}{x_i'} \Big/ \binom{N_i}{n_i}}. \tag{2.33}$$

In addition to its probability interpretation, equation (2.33) may be used to order each $\tau \in \Omega_t$ so as to determine how extreme or discrepant the observed $\tau_0$ is under $H_0$. Thus, $P(\tau \mid t)$ is also the test statistic for the homogeneity test. Its observed value, $P(\tau_0 \mid t)$, defines the critical region of the exact two-sided $p$-value. Let

$$\Omega_t^* = \{\tau \in \Omega_t: P(\tau \mid t) \leq P(\tau_0 \mid t)\}. \tag{2.34}$$

The $p$-value for Zelen's test of homogeneity is

$$p = \sum_{\tau \in \Omega_t^*} P(\tau \mid t). \tag{2.35}$$

There is no well accepted large-sample theory for this problem. Breslow and Day[17] propose the statistic

$$\chi^2_{BD} = \sum_{i=1}^{s} \frac{[X_i - A_i(\hat{\Psi})]^2}{\operatorname{var}(X_i \mid \hat{\Psi})}, \tag{2.36}$$

where $A_i(\hat{\Psi})$ is the positive root of the quadratic equation

$$\frac{A_i(N_i - m_i - n_i + A_i)}{(m_i - A_i)(n_i - A_i)} = \hat{\Psi}, \tag{2.37}$$

formed by expressing the $i$th table as

$$\begin{matrix} m_i - A_i & A_i \\ N_i - m_i - n_i + A_i & n_i - A_i \end{matrix},$$

and equating its empirical odds ratio to the Mantel–Haenszel common odds ratio

$$\hat{\Psi} = \frac{\sum_{i=1}^{s} x_i(N_i - m_i - n_i + x_i)/N_i}{\sum_{i=1}^{s}(n_i - x_i)(m_i - x_i)/N_i}. \tag{2.38}$$

The variance of $X_i$ is estimated by

$$\text{var}(X_i|\hat{\Psi}) = \left[\frac{1}{A_i(\hat{\Psi})} + \frac{1}{m_i - A_i(\hat{\Psi})} + \frac{1}{n_i - A_i(\hat{\Psi})} + \frac{1}{N_i - m_i - n_i + A_i(\hat{\Psi})}\right]^{-1}. \tag{2.39}$$

In large samples, $\chi^2_{BD}$ is chi-squared distributed with $s-1$ degrees of freedom, and the $p$-value for testing $H_0$ is

$$p_{BD} = P(\chi^2_{BD} \geq \chi^2_0), \tag{2.40}$$

where $\chi^2_0$ is the observed value of $\chi^2_{BD}$. The chi-squared approximation to the $\chi^2_{BD}$ statistic is rather poor for skewed or sparse contingency tables.

*Common odds ratio estimation* Exact inference about the common odds ratio, $\Psi$, is based on the fact that the probability of any $\tau \in \Omega$ is a product of noncentral hypergeometric probabilities in which $\Psi$ is the only unknown parameter. As shown in Gart[19] this probability is

$$P(\tau) = \frac{\prod_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{n_i - x_i}\Psi^{x_i}}{\sum_{\tau\in\Omega}\prod_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{n_i - x_i}\Psi^{x_i}}. \tag{2.41}$$

To make inferences about $\Psi$, we require the distribution of its sufficient statistic

$$T = X_1 + X_2 + \cdots + X_s. \tag{2.42}$$

This distribution can be derived from (2.41) as

$$P(T = t|\Psi) = \frac{C_t\Psi^t}{\sum_{u=t_{\min}}^{t_{\max}} C_u\Psi^u}, \tag{2.43}$$

where

$$C_t = \sum_{\tau\in\Omega_t}\prod_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{n_i - x_i}, \tag{2.44}$$

$$t_{\min} = \sum_{i=1}^{s}\max(0, n_i - m_i), \tag{2.45}$$

$$t_{\max} = \sum_{i=1}^{s}\min(m_i', n_i). \tag{2.46}$$

It is straightforward to test the hypothesis $\Psi = \Psi_0$ based on the conditional

distribution (2.43). The test has critical regions of the form $T \geq t (T \leq t)$ for alternatives of the form $\Psi > \Psi_0 (\Psi < \Psi_0)$. An exact confidence interval for $\Psi$ may be constructed by inverting this test, as discussed in Cox and Snell.[20] An efficient numerical algorithm for generating the distribution (2.43) is given in Mehta *et al.*[21]

An asymptotic confidence interval for $\Psi$ is usually computed by the Mantel–Haenszel[22] method. The Mantel–Haenszel point estimate, $\hat{\Psi}$, is computed by equation (2.38). The inference is based on the large-sample approximation to the distribution of $\log \hat{\Psi}$. This distribution is normal, with mean $\log \Psi$. There has been a great deal of research on the appropriate variance for $\log \hat{\Psi}$. The most satisfactory candidate is the Robins, Breslow and Greenland (RBG) variance.[23] This variance estimator is known to perform well both when $s$ is small but $(m_i, n_i)$ are large, and when $s$ is large but $(m_i, n_i)$ are small. The RBG variance is

$$\text{var}(\log \hat{\Psi}) = \sum_{i=1}^{s} \left( \frac{a_i c_i}{2c_+^2} + \frac{a_i d_i + b_i c_i}{2c_+ d_+} + \frac{b_i d_i}{2d_+^2} \right) \tag{2.47}$$

where $a_i = (x_i + y_i')/N_i$, $b_i = (x_i' + y_i)/N_i$, $c_i = (x_i y_i')/N_i$, $d_i = (x_i' y_i)/N_i$, $c_+ = \sum_{i=1}^{k} c_i$, and $d_+ = \sum_{i=1}^{k} d_i$. A $100(1-\alpha)\%$ confidence interval for $\log \Psi$ is then

$$CI_{RBG} = \log \hat{\Psi} \pm z_{\alpha/2} [\text{var}(\log \hat{\Psi})]^{1/2}. \tag{2.48}$$

### 2.5 Computational issues

Computing equation (2.6) is a nontrivial task. The size of the reference set grows exponentially so that explicit enumeration of all the tables in $\Gamma$ soon becomes computationally infeasible. For example, the reference set of all $5 \times 6$ tables with row sums of (7, 7, 12, 4, 4) and column sums of (4, 5, 6, 5, 7, 7) contains 1.6 billion tables. Yet, the tables in this reference set are all rather sparse and unlikely to yield accurate $p$-values based on large sample theory. Network algorithms have been developed by Mehta and Patel[8,10,21,24,25] to enumerate the tables in $\Gamma$ implicitly. This makes it feasible to compute exact $p$-values for tables with the above margins. A different approach to implicit enumeration is provided by Pagano and Halvorsen,[26] Pagano and Trichler,[27] Baglivo *et al.*[28] and Streitberg and Rohmel.[29] Sometimes a data set is too large even for implicit enumeration, yet it is sufficiently sparse that the asymptotic results are suspect. For such situations a Monte Carlo estimate and associated 99% confidence interval for the exact $p$-value may be obtained. In the Monte Carlo method, tables are sampled from $\Gamma$ in proportion to their hypergeometric probabilities (2.5), and a count is kept of all the sampled tables that are more extreme than the observed table. For details, refer to Agresti and Wackerly,[30] Patefield[31] and Mehta *et al.*[32]

## 3 Analysis of data sets

In this section we will illustrate the techniques developed in the previous section with some data analysis. Each example will highlight the different conclusions one might draw if an asymptotic analysis were performed instead of an exact analysis.

### 3.1 Unordered contingency tables

Data were obtained on the location of oral lesions, in house-to-house surveys in three geographic regions of rural India, by Gupta *et al.*[33] Consider a hypothetical subset of

these data in the form of a $9 \times 3$ contingency table in which the counts are the number of patients with oral lesions per site and geographic region, Table 3.

**Table 3** Oral lesions data

| Site of lesion | Kerala | Gujarat | Andhra |
|---|---|---|---|
| Labial mucosa | 0 | 1 | 0 |
| Buccal mucosa | 8 | 1 | 8 |
| Commissure | 0 | 1 | 0 |
| Gingiva | 0 | 1 | 0 |
| Hard palate | 0 | 1 | 0 |
| Soft palate | 0 | 1 | 0 |
| Tongue | 0 | 1 | 0 |
| Floor of mouth | 1 | 0 | 1 |
| Alveolar ridge | 1 | 0 | 1 |

The question of interest is whether the distribution of the site of the oral lesion is significantly different in the three geographic regions. The row and column classifications for this $9 \times 3$ table are clearly unordered, making it an appropriate data set for either the Fisher, Pearson or Likelihood ratio tests. The contingency table is so sparse that the usual chi-squared asymptotic distribution with 16 degrees of freedom is not likely to yield accurate $p$-values. The exact and asymptotic $p$-values are displayed in Table 4.

**Table 4** Exact and asymptotic $p$-values for oral lesions data

| | Three tests of independence | | |
|---|---|---|---|
| Type of inference | Pearson | Fisher | Likelihood ratio |
| Value of $D(\mathbf{x})$ | 22.1 | 19.72 | 23.3 |
| Asymptotic $p$-value | 0.1400 | 0.2331 | 0.1060 |
| Exact $p$-value | 0.0269 | 0.0101 | 0.0356 |

For each test the asymptotic $p$-value was obtained by looking up the tail area to the right of $D(\mathbf{x})$ (displayed on the first line of the table) from a chi-square distribution with 16 degrees of freedom. The exact $p$-value was obtained by actually permuting the observed $9 \times 3$ table in all possible ways, subject to fixed margins, and summing the probabilities of permutations $\mathbf{y}$ for which $D(\mathbf{y}) \geq D(\mathbf{x})$. There are striking differences between the exact and asymptotic $p$-values. The exact analysis suggests that the row and column classifications are highly dependent, but the asymptotic analysis fails to show this.

### 3.2 Singly ordered contingency tables

The tumour regression rates of five chemotherapy regimens, Cytoxan (CTX) alone, Cyclohexyl-chloroethyl nitrosourea (CCNU) alone, Methotrexate (MTX) alone, CTX + CCNU and CTX + CCNU + MTX were compared in a small clinical trial. Tumour regression was measured on a three-point scale: no response, partial response, or complete response. The results are shown in Table 5.

**Table 5** Chemotherapy pilot study data

| Chemotherapy | No response | Partial response | Complete response |
|---|---|---|---|
| CTX | 2 | 0 | 0 |
| CCNU | 1 | 1 | 0 |
| MTX | 3 | 0 | 0 |
| CTX + CCNU | 2 | 2 | 0 |
| CTX + CCNU + MTX | 1 | 1 | 4 |

Small pilot studies like this one are frequently conducted as a preliminary to planning a large-scale randomized clinical trial. The columns of the observed $5 \times 3$ contingency table are ordered by the magnitude of the response. However the rows of the table do not have any natural ordering, but simply represent five different treatments. For such data the Kruskal–Wallis test may be used to determine whether or not the five drug regimens are significantly different with respect to their tumour regression rates. The observed value of the Kruskal–Wallis statistic for this table is 8.682. Referring this value to a chi-square distribution with 4 degrees of freedom yields an asymptotic $p$-value of 0.0695 which is not significant at the 0.05 level. However, the exact test based on the permutation distribution of equation (2.10) reveals that the exact $p$-value is 0.039, which is statistically significant. The small sample size and the presence of ties caused the asymptotic approximation to be nearly twice as large as the exact $p$-value.

### 3.3 Doubly ordered contingency tables

#### Dose response example

Patients were treated with a drug at four dose levels (100mg, 200mg, 300mg, 400mg) and then monitored for toxicity. The data are given in Table 6.

**Table 6** Dose-response drug toxicity data

| | | | Drug toxicity | | |
|---|---|---|---|---|---|
| Drug dose | Mild | Moderate | Severe | Drug death | Row score |
| 100mg | 100 | 1 | 0 | 0 | $u_1$ |
| 200mg | 18 | 1 | 1 | 0 | $u_2$ |
| 300mg | 50 | 1 | 1 | 0 | $u_3$ |
| 400mg | 50 | 1 | 1 | 1 | $u_4$ |
| Column score | $v_1$ | $v_2$ | $v_3$ | $v_4$ | |

Notice that there is a natural ordering along the rows as well as the columns of this $4 \times 4$ contingency table. Thus the Jonckheere–Terpstra test and the Linear-by-linear association test are appropriate for determining if the increase in drug dose leads to greater toxicity.

We first perform the Jonckheere–Terpstra test. The exact two-sided $p$-value of 0.1134 closely matches the corresponding asymptotic two-sided $p$-value of 0.1210, indicating that the dose-response relationship between drug dose and toxicity is not statistically significant.

Next we perform the Linear-by-linear association test, using the equally spaced scores, $u = i$, $v_j = j$, for $i, j = 1, 2, \ldots, 4$. Now the exact two-sided $p$-value is 0.0866 and the corresponding asymptotic two-sided $p$-value is 0.0812, confirming that the dose-response relationship is at best marginally statistically significant. The Linear-by-linear association test does give us some added flexibility over the Jonckheere–Terpstra test however. We are free to choose the row and column scores arbitrarily. Suppose for instance that the toxic event 'Drug death' was deemed to be catastrophic, and orders of magnitude more serious than a 'Severe toxicity'. In that case it might be reasonable to maintain the equally spaced row scores, $u_i = i$, $i = 1, 2, \ldots, 4$, but assign unequally spaced column scores $v_1 = 1$ for 'Mild toxicity', $v_2 = 2$ for 'Moderate toxicity', $v_3 = 3$ for 'Severe toxicity', and $v_4 = 10000$ for 'Drug death'. Because of this severe discontinuity in the column scores, the asymptotic theory breaks down. Now the two-sided asymptotic $p$-value is 0.1604, implying that there is no association between drug-dose

and toxicity, while the two-sided exact $p$-value is 0.0372, implying that the dose-response relationship is indeed statistically significant.

### Space shuttle Challenger example

Professor Richard Feynman in his delightful book *What do you care what other people think*[34] recounted at great length his experiences as a member of the Presidential Commission, formed to determine the cause of the explosion of the space shuttle Challenger, in 1986. He suspected that the low temperature at take-of caused the O-rings to fail. On page 137 of his book, he has published the data on temperature versus the number of O-ring incidents, on 24 previous space shuttle flights. These data are shown in Table 7.

**Table 7** Space-shuttle O-ring data

| O-ring incidents | Temperature (Fahrenheit) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| None | 66 | 67 | 67 | 67 | 68 | 68 | 70 | 70 | 72 |
| | 73 | 75 | 76 | 76 | 78 | 79 | 80 | 81 | |
| One | 57 | 58 | 63 | 70 | 70 | | | | |
| Two | **75** | | | | | | | | |
| Three | **53** | | | | | | | | |

These data may be represented as a contingency table whose rows are the numbers of O-ring incidents, and whose columns are the temperatures at take-off. Thus both the rows and columns are ordered, and the Jonckheere–Terpstra test is an appropriate one for determining if take off temperature is correlated with O-ring failures. The exact $p$-value is 0.0241, while the asymptotic $p$-value is 0.0262. Both are indicative of a significant association between take off temperature and O-ring incidents.

The linear-by-linear association test may also be used to test the association between temperature and O-ring incidents. Using the number of O-ring incidents as the row scores and the take-off temperature as the column scores, the exact $p$-value is 0.0272. The corresponding asymptotic $p$-value is 0.0175. These results confirm the conclusions of the Jonckheere–Terpstra test.

## 3.4 Linear rank tests

A cohort of Hiroshima atomic bomb survivors was followed to determine the relationship between deaths from leukemia during 1950–1970 and estimated radiation dosage from the bombing. Subjects were stratified according to their age at the time of the bombing. In Table 8 we tabulate a subset of the data; children in the 0–9 age group exposed to radiation doses ranging from 0 to 99 rads. Cases are subjects who died from leukemia during the follow-up. Controls are subjects who did not die from leukemia during the follow-up.

**Table 8** Leukemia deaths among Hiroshima/Nagasaki atomic bomb survivors

| | Radiation dose (rads) | | | |
|---|---|---|---|---|
| Survival status | 0 | 1–9 | 10–49 | 50–99 |
| Case | 0 (0%) | 7 (0.07%) | 3 (0.1%) | 1 (0.14%) |
| Control | 5015 | 10752 | 2989 | 694 |
| Total | 5015 | 10759 | 2992 | 695 |

Two additional dose groups, 100–199 rads and 200+ rads, are excluded from the present analysis. Their inclusion increased the standardized value of the test statistic

from 3 to 16, strongly suggesting that their effect on the risk of leukemia is nonlinear and should be considered in a more general model. The full data set is on page 285 of Agresti.[4]

In absolute terms, the leukemia death rates are rather low. Only 11 deaths were observed in a cohort of size 19461, amounting to a death rate of 0.06%. However the rates increase from 0% in the lowest dose group to 0.14% in the highest. It is therefore interesting to ask whether this increasing trend is real, or merely due to chance fluctuations in the data. Our intuition cannot help much with these extremely low death rates, and we must resort to a formal statistical test of significance.

One way to determine if there is a statistically significant association between leukemia deaths and radiation exposure is to perform the Cochran–Armitage trend test.[14] The test statistic is given by equation (2.18) with $v_j$ being the mid-range of the $j$th radiation dose. For these data $v_1 = 0$ rads, $v_2 = 4.5$ rads, $v_3 = 30$ rads and $v_4 = 75$ rads. Previously the only way to perform this trend test was to assume that the linear rank statistic, $D$, is normally distributed. Figure 1 displays the true distribution of $D$. It is not even close to

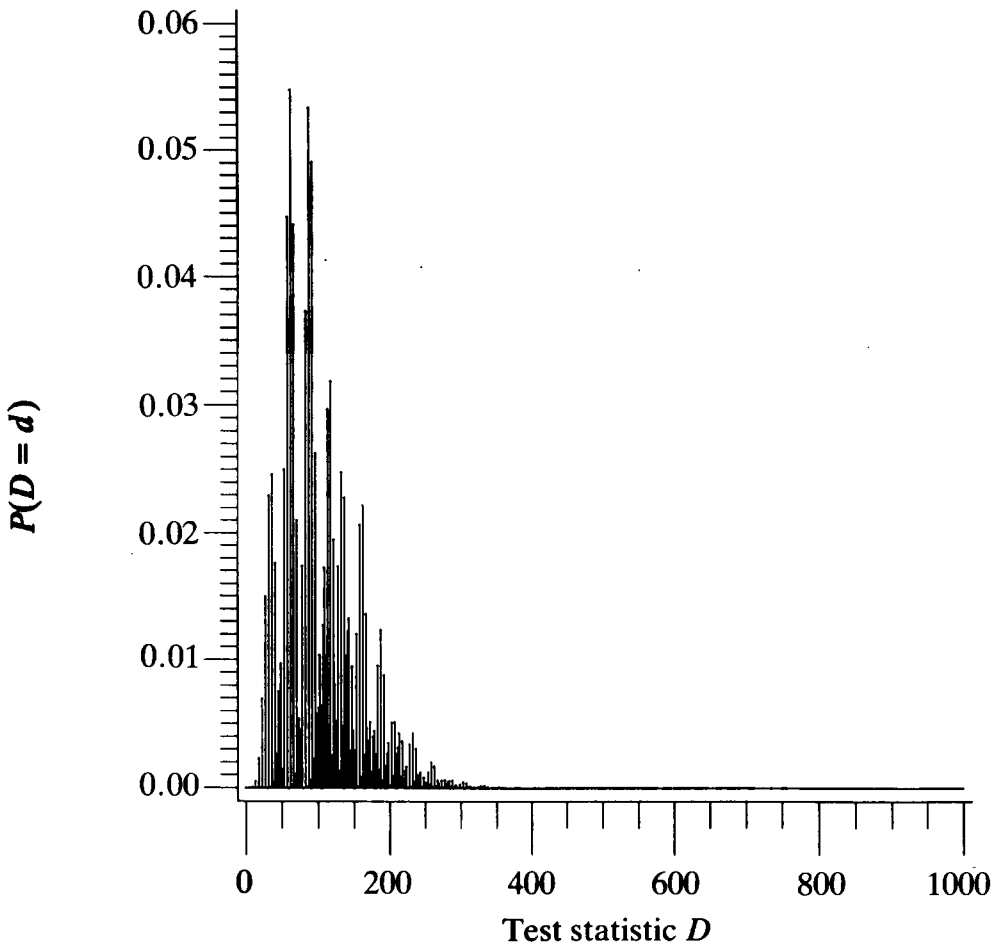


**Figure 1** Exact probability density for Hiroshima data

normal. Its distinct values are unequally spaced, the distribution has an unusually long right tail, extending all the way out to $D = 825$ even though $E(D) = 107.6$. In addition the distribution is multimodal. Not surprisingly the exact and asymptotic $p$-values for the Cochran–Armitage trend test differ. The results are given in Table 9.

**Table 9** Exact and asymptotic $p$-values for Hiroshima data

| $p$-values | One-sided | Two-sided |
|---|---|---|
| Exact | 0.0653 | 0.0682 |
| Asymptotic | 0.0465 | 0.0929 |

## 3.5 Stratified $2 \times 2$ tables

We present two examples in this section, one for a test of homogeneity of odds ratios and one for estimating the common odds ratio.

### Homogeneity of odds ratios

The binary response data in Table 10 compare a new drug at 22 hospital sites. (At the request of the drug company conducting the study, the names of the two agents are not reported here.)

**Table 10** Site by treatment interaction

| | New drug | | Control drug | |
|---|---|---|---|---|
| Test site | Response | No | Response | No |
| 1 | 0 | 15 | 0 | 15 |
| 2 | 0 | 39 | 6 | 32 |
| 3 | 1 | 20 | 3 | 18 |
| 4 | 1 | 14 | 2 | 15 |
| 5 | 1 | 20 | 2 | 19 |
| 6 | 0 | 12 | 2 | 10 |
| 7 | 3 | 49 | 10 | 42 |
| 8 | 0 | 19 | 2 | 17 |
| 9 | 1 | 14 | 0 | 15 |
| 10 | 2 | 26 | 2 | 27 |
| 11 | 0 | 19 | 2 | 18 |
| 12 | 0 | 12 | 1 | 11 |
| 13 | 0 | 24 | 5 | 19 |
| 14 | 2 | 10 | 2 | 11 |
| 15 | 0 | 14 | 11 | 3 |
| 16 | 0 | 53 | 4 | 48 |
| 17 | 0 | 28 | 0 | 20 |
| 18 | 0 | 21 | 0 | 21 |
| 19 | 1 | 50 | 1 | 48 |
| 20 | 0 | 13 | 1 | 13 |
| 21 | 0 | 13 | 1 | 13 |
| 22 | 0 | 21 | 0 | 21 |

The data can be thought of as twenty-two $2 \times 2$ contingency tables, one for each site. If you examine the $2 \times 2$ tables carefully, you notice that site 15 appears to be different from the others. Whereas all the other sites have a low response rate for both the new drug and the control drug, the response rate of the control drug is 79% at site 15. The Homogeneity test can tell you whether the observed difference at site 15 is a real difference or whether it is just a chance fluctuation due to a small sample. Because of the sparseness in the data, the asymptotic (Breslow–Day) statistic might not yield an

accurate $p$-value. The exact (Zelen) test is preferred. The exact $p$-value is 0.0135. Thus we reject the null hypothesis that there is a common odds ratio across the 22 sites. The data strongly suggest that the odds ratio at site 15 is different from the other odds ratios. The asymptotic (Breslow–Day) $p$-value is much larger (0.0785) and is only marginally significant.

### Estimating the common odds ratio

The court case of Hogan v. Pierce[35] involved the hiring data presented in Table 11, by race.

**Table 11** Minority hiring data

| | Whites | | Blacks | |
|---|---|---|---|---|
| Date of hire | Hired | Not | Hired | Not |
| 7/74 | 4 | 16 | 0 | 7 |
| 8/74 | 4 | 13 | 0 | 7 |
| 9/74 | 2 | 13 | 0 | 8 |
| 4/75 | 1 | 17 | 0 | 8 |
| 5/75 | 1 | 17 | 0 | 8 |
| 10/75 | 1 | 29 | 0 | 10 |
| 11/75 | 2 | 29 | 0 | 10 |
| 2/76 | 1 | 30 | 0 | 10 |
| 3/76 | 1 | 30 | 0 | 10 |
| 11/77 | 1 | 33 | 0 | 13 |

The most notable feature of these data is that at each hiring opportunity not a single black was hired, whereas small numbers of whites were hired. This makes it impossible to use the usual large-sample maximum likelihood or Mantel–Haenszel[22] methods for estimating the odds of being hired for whites relative to blacks. These methods simply fail to converge. Only the exact method provides a valid answer and it shows that the odds of being hired for a white relative to a black are no lower than 2.3 to 1, with 95% confidence.

## 4 Concluding remarks

We have presented the essential idea behind exact nonparametric inference, referenced numerical algorithms and software for its implementation, and shown through several examples that exact inference is a valuable supplement to corresponding asymptotic methods.

The methods described here extend naturally to continuous data. In principle, such data can also be represented as contingency tables but the columns of these tables will sum to 1. Thus these methods provide a unified approach to handling nonparametric data both for the categorical case and the more traditional continuous case. For example consider the two-sample problem involving continuous data shown in Table 12. The two groups are 'males' and 'females'. The continuous variable being compared in the two groups is 'monthly income'.

**Table 12** Two-sample continuous data represented in the traditional way

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M | M | M | M | F | F | F | F |
| 2010 | 3100 | 2555 | 2095 | 1990 | 2122 | 1875 | 2550 |

These data can be represented by the $2 \times 8$ contingency table, Table 13, which may then be permuted in the usual way for exact inference.

**Table 13** Two-sample continuous data represented as a $2 \times 8$ contingency table

| Rows | Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 | Col 7 | Col 8 | Row total |
|---|---|---|---|---|---|---|---|---|---|
| Male | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 4 |
| Female | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 4 |
| Column total | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 |
| Column score | 1875 | 1990 | 2010 | 2095 | 2122 | 2550 | 2555 | 3100 | |

In conclusion, exact methods are now an integral part of nonparametric inference. Software support for these methods is available in many standard packages including SAS. Some of the newer textbooks on nonparametric methods, for example Sprent[36] (1993), devote considerable space to exact methods. Thus one expects that exact methods will replace corresponding asymptotic ones as the standard approach for small, sparse or unbalanced data sets.

## Acknowledgement

This research was supported in part by grant CA61050 from the National Cancer Institute.

## References

1 Agresti A, Yang M. An empirical investigation of some effects of sparseness in contingency tables. *Communications in Statistics* 1987; **5**: 9–21.
2 Read RC, Cressie NA. *Goodness of fit statistics for discrete multivariate data*. New York: Springer-Verlag, 1988.
3 Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1925.
4 Agresti A. *Categorical data analysis*. New York: John Wiley & Sons, 1990.
5 Agresti A. *Analysis of ordinal categorical data*. New York: John Wiley & Sons, 1984.
6 Yates F. Test of significance for $2 \times 2$ contingency tables. *Journal of the Royal Statistical Society Series A* 1984; **147**: 426–63.
7 Cox DR, Hinkley DV. *Theoretical Statistics*. London: Chapman and Hall, 1974.
8 Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* 1983; **78**(382): 427–34.
9 Freeman GH, Halton JH. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 1951; **38**: 141–49.
10 Mehta CR, Patel NR. A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables. *Communications in Statistics* 1986; **15**(2): 387–403.
11 Hollander M, Wolfe DA. *Nonparametric statistical methods*. New York: John Wiley, 1973.
12 Agresti A, Mehta CR, Patel NR. Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association* 1990; **85**: 410, 453–58.
13 Chernoff H, Savage IR. Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematics* 1958; **29**: 972–94.
14 Gibbons JD. *Nonparametric statistical inference*, 2nd edition. New York: Marcel Dekker, 1985.
15 Kalbfleish JD, Prentice RL. *The statistical analysis of failure time data*. New York: John Wiley & Sons, 1980.
16 Hettmansperger TP. *Statistical inference based on ranks*. New York: John Wiley & Sons, 1984.
17 Breslow NE, Day NE. The analysis of case-control studies. *IARC Scientific Publications No. 32*. France: Lyon, 1980.
18 Zelen M. The analysis of several $2 \times 2$ contingency tables. *Biometrika* 1971; **58**(1): 129–37.

19 Gart J. Point and interval estimation of the common odds ratio in the combination of $2 \times 2$ tables with fixed marginals. *Biometrika* 1970; **57**: 471–75.
20 Cox DR, Snell EJ. *The analysis of binary data*, 2nd edition. New York: Chapman and Hall, 1989.
21 Mehta CR, Patel NR, Gray R. On computing an exact confidence interval for the common odds ratio in several $2 \times 2$ contingency tables. *Journal of the American Statistical Association* 1985; **80**(392): 969–73.
22 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**: 719–48.
23 Robins J, Breslow N, Greenland S. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**: 311–23.
24 Mehta CR, Patel NR, Tsiatis AA. Exact significance testing to establish treatment equivalence for ordered categorical data. *Biometrics* 1984; **40**: 819–25.
25 Mehta CR, Patel NR. FEXACT: A Fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* 1986; **12**(2): 154–61.
26 Pagano M, Halvorsen K. An algorithm for finding exact significance levels of $r \times c$ contingency tables. *Journal of the American Statistical Association* 1981; **76**: 931–34.
27 Pagano M, Tritchler D. On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association* 1983; **78**: 435–41.
28 Baglivo J, Olivier D, Pagano M. Methods for the analysis of contingency tables with large and small cell counts. *Journal of the American Statistical Association* 1988; **83**: 1006–13.
29 Streitberg B, Rohmel R. Exact distributions for permutation and rank tests. *Statistical Software Newsletter* 1986; **12**: 10–17.
30 Agresti A, Wackerly D. Some exact conditional tests of independence for $r \times c$ cross-classification tables. *Psychometrika* 1977; **42**: 111–25.
31 Patefield WM. An efficient method of generating $r \times c$ tables with given row and column totals. (Algorithm AS 159). *Applied Statistics* 1981; **30**: 91–97.
32 Mehta CR, Patel NR, Senchaudhuri P. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association* 1988; **83**(404): 999–1005.
33 Gupta PC, Mehta FR, Pindborg J. *Community Dentistry and Oral Epidemiology*, 1980; **8**: 287–333.
34 Feynman RP. *What do you care what other people think?* New York: WW Norton, 1988.
35 Gastworth JL. Combined tests of significance in EEO cases. *Industrial and Labor Relations Review*, 1984; **38**(1).
36 Sprent P. *Applied nonparametric statistical methods*. Second edition. London: Chapman and Hall, 1993.
37 EGRET user manual. Statistics and Epidemiology Research Corporation, Seattle, WA, USA, 1989.
38 *Epi Info manual*. Centers for Disease Control, Atlanta, GA, USA, 1989.
39 *SAS/Stat guide for personal computers*. Version 6 edition. The SAS Institute, Cary, NC, USA, 1987.
40 *StatXact Version 3. Software for exact nonparametric inference*. Cytel Software Corporation, Cambridge, MA 02139, USA, 1993.
41 Lehmann EL. *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden–Day, 1975.
42 Seigel S, Castellan NJ. *Nonparametric statistics for the behavioral sciences*. Second edition. New York: McGraw-Hill, 1988.
43 *LogXact. Software for exact logistic regression*. Cytel Software Corporation, Cambridge, MA, USA, 1993.
44 *Testimate Version 5.1*. IDV, Datenanalyse und Versuchsplanung, Munich, Germany, 1993.

## Appendix Software for exact inference

So far as we are aware there are only five statistical packages meeting commercial standards of reliability and documentation that offer exact inference capabilities beyond the single $2 \times 2$ contingency table.

**EGRET (1989)** The EGRET[37] package is available from Statistical and Epidemiology Research Corporation, 1107 NE 45, Suite 520, Seattle, WA 98105, USA. It offers exact inference for stratified $2 \times 2$ contingency tables and for the Pearson test for a $2 \times c$ contingency table. Exact inference for the general $r \times c$ problem is not provided.

**Epi Info (1989)** Epi Info[38] is a series of programs used to create and analyse questionnaires and perform other common epidemiological tasks. One of the statistical capabilities provided by Epi Info is exact inference for the common odds ratio in stratified $2 \times 2$ contingency tables. It is available from the Division of Surveillance and Epidemiologic Studies, Epidemiology Program Office, Centers for Disease Control, Atlanta, GA 30333, USA.

**SAS (1987)** SAS[39] is available from the SAS Institute, 100 SAS Campus Drive, Cary, North Carolina 27513, USA. Versions 6 and up offer the exact $p$-value capability for Fisher's exact test for $r \times c$ tables, but not for any of the other tests described here. A special module, StatXact for SAS (1993), developed by Cytel Software Corporation, Cambridge, MA, extends the exact capabilities of SAS by making it possible to call the StaXact package (described below) from within SAS, read in SAS data sets, and take the results back into SAS so as to avail oneself of SAS's powerful graphics and report generation capabilities.

**StatXact (1993)** The StatXact[40] package is available from Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139, USA. Version 2 was released in 1991. Version 3 is currently in beta test. It is a complete nonparametrics package with exact tests for one-sample, 2-sample, and $k$-sample problems, measures of association, $r \times c$ contingency tables, stratified $2 \times 2$ and $2 \times c$ contingency tables, multiple comparisons, exact one and two-sample Hodges–Lehmann confidence intervals, and exact confidence intervals for odds ratios, risk ratios and differences in two binomial parameters. It provides software support for standard textbooks on nonparametric statistics like Lehmann,[41] Hollander and Wolfe,[11] Gibbons,[14] Seigel and Castellan[42] and Sprent.[36] A companion package, LogXact,[43] provides exact inference capabilities for logistic regression.

**Testimate (1992)** The Testimate[44] package is available from IDV, Datenanalyse und Versuchsplanung, Wessobrunner Strasse 6, D-8035 Gauting, Munich, Germany. It offers exact one and two-sample tests, and Hodges–Lehmann confidence intervals. Fisher's exact test is provided for the $2 \times c$ contingency table. Only asymptotic tests are available for $r \times c$ contingency tables where $r > 2$.