Exact Inference for Categorical Data

Cyrus R. Mehta and Nitin R. Patel

Harvard University and Cytel Software Corporation

January 1, 1997

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. They rely, however, 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 [5] for some empirical results, and Read and Cressie [31] for a more theoretical discussion.

One way to make valid statistical inferences in the presence of small, sparse or unbalanced data is to compute exact p-values and confidence intervals, based on the permutational distribution of the test statistic. This approach was first proposed by R. A. Fisher [11] and has been used extensively for the single 2×2 contingency table. Previously exact tests were rarely attempted for tables of higher dimension than 2×2, primarily because of the formidable computing problems involved in their execution. In recent years, however, the easy availability of immense quantities of computing power combined with many new, fast and efficient algorithms for exact permutational inference have revolutionized our thinking about what is computationally feasible. Problems that would previously have taken several hours or even days to solve now take only a few minutes. Exact inference is

now a practical proposition and has been incorporated into standard statistical software packages.

In the present paper we present a unified framework for exact inference, anchored in the permutation principle. We demonstrate that, for a very broad class of nonparametric problems, such inference 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. We then apply these algorithms to several data sets. 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 and cite an internet resource for performing exact permutational inference.

Exact Permutation Tests for $r \times c$ Contingency Tables

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

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 *ordered* 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. Ordered variables take on values which can be ordered in a natural way. An example of an ordered variable is Drug Dose—Low, Medium, High. Ordered variables may of course assume numerical values as well (for example, the number of cigarettes smoked per day).

The Exact Permutation Distribution of $\boldsymbol{x}$

The exact probability distribution of $\boldsymbol{x}$ depends on the sampling scheme that was used to generate $\boldsymbol{x}$. When both the row and column classifications are categorical, Agresti [1] lists three sampling schemes that could give rise to $\boldsymbol{x}$; full multinomial sampling, product multinomial sampling, and Poisson sampling. Under all three schemes the probability distribution of $\boldsymbol{x}$ contains unknown parameters, π_{ij}, relating to the individual cells of the $r \times c$ table (see LOGLINEAR MODELS). For instance, for full multinomial sampling, π_{ij} denotes the probability of classification in row i and column j, whereas for product multinomial sampling π_{ij} denotes the conditional probability of falling in column j given that the subject belongs to row i.

Consider the null hypothesis of no row by column interaction. Since statistical inference is based on the distribution of $\boldsymbol{x}$ under the null hypothesis of no row by column interaction, the number of unknown parameters is reduced (π_{ij} being replaced by $\pi_{i.}\pi_{.j}$ or $\pi_{.j}$ depending on the sampling scheme) but not eliminated. Asymptotic inference relies on estimating these unknown parameters by maximum likelihood and related methods. The key to exact permutational inference is getting rid of all nuisance parameters from the probability distribution of $\boldsymbol{x}$. 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 $\boldsymbol{x}$. Specifically, define the reference set

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

Then one can show that, under the null hypothesis of no row by column interaction, the probability of observing $\boldsymbol{x}$ conditional on $\boldsymbol{x} \in \Gamma$ is of the hypergeometric form

$$\Pr(\boldsymbol{x} \mid \boldsymbol{x} \in \Gamma) \equiv P(\boldsymbol{x}) = \prod_{i=1}^{r} \prod_{j=1}^{c} \frac{n_j! m_i!}{N! x_{ij}!} \ . \tag{2}$$

Equation (2), which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate $\boldsymbol{x}$ is full multinomial, product multinomial, or Poisson. (See, for example, Agresti [2].)

Since (2) contains no unknown parameters, exact inference is possible. The nuisance parameters were, however, eliminated by conditioning on the margins of the observed contingency table. Now some of 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. R.A. Fisher [11] first proposed this idea for exact inference on a single 2×2 contingency table. At various times since then prominent statisticians have commented on this approach. The principles most cited for conditioning are the *sufficiency principle*, the *ancillarity principle*, and the *randomization principle*. An informal intuitive explanation of these three principles is provided below.

Sufficiency Principle The margins of the contingency table are sufficient statistics for unknown nuisance parameters. Thus conditioning on them affords a convenient way to eliminate nuisance parameters from the likelihood function. For example, if the data are generated by product multinomial sampling the row margins, m_i, would ordinarily be considered fixed but the column margins, n_j would be considered random variables. The null hypothesis of interest states that $\pi_{ij} = \pi_{\cdot j}$ for all i. Thus the probability of $\boldsymbol{x}$ depends on c unknown nuisance parameters, $(\pi_{\cdot 1}, \pi_{\cdot 2}, \ldots \pi_{\cdot c})$ even under the null hypothesis. By the sufficiency principle, these nuisance parameters are eliminated if we condition on $(n_1, n_2, \ldots n_c)$, their sufficient statistics. It follows that by restricting our attention to $r \times c$ tables in Γ we are implicitly conditioning on $(n_1, n_2, \ldots n_c)$, since the other set of margins, $(m_1, m_2, \ldots m_r)$, are

fixed naturally by the sampling scheme. Similar sufficiency arguments can be made for full multinomial and Poisson sampling.

Ancillarity Principle 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 unrelated to the null hypothesis unchanged as far as possible. The margins of the contingency table are representative of nuisance parameters whose values do not provide any information about the null hypothesis of interest. In this sense they are ancillary statistics. 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.

Randomization Principle The case for conditioning is especially persuasive if the r rows of the contingency tables represent r different treatments, with m_i subjects being assigned to treatment i by a randomization mechanism. Each subject provides a multinomial response that falls into one of the c columns. Thus n_j represents the total number of responses of the jth type. Now, under the null hypothesis, the r treatments are equally effective. Therefore the response that a patient provides is the same, regardless of the treatment to which that patient is randomized. Thus the value of n_j is predetermined and may be regarded as fixed. The statistical significance of the observed outcome is judged relative to its permutational distribution in hypothetical repetitions of the randomization rule for assigning patients to treatments.

An excellent exposition of the conditional viewpoint is available in Yates [40]. For a theoretical justification of the sufficiency and ancillarity principles refer to Cox and

Hinkeley [9], and Reid [32]. For a detailed exposition of the randomization principle, highlighting its applicability to a broad range of problems, refer to Edgington [10]. 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 for $r \times c$ contingency tables, stratified 2×2 contingency tables, stratified $2 \times c$ contingency tables, and logistic regression.

Exact P-Values

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

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

$$p = \sum_{D(\boldsymbol{y}) \geq D(\boldsymbol{x})} P(\boldsymbol{y}) = \Pr\{D(\boldsymbol{y}) \geq D(\boldsymbol{x})\} \ . \tag{3}$$

Classical nonparametric methods rely on the large-sample distribution of D to estimate p. For $r \times c$ tables with large cell counts and the usual forms for the function D it is possible to show that D converges in distribution to a chi-square with appropriate degrees of freedom. Thus p is usually estimated by $\tilde{p}$, the chi-square tail area to the right of $D(\boldsymbol{x})$. Modern algorithmic techniques have made it possible to compute p directly instead of relying on $\tilde{p}$, its asymptotic approximation. This is achieved by powerful recursive

algorithms 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 $\tilde{p}$ can differ considerably for contingency tables with small cell counts.

The main advantage of using p rather than $\tilde{p}$ is that it is guaranteed to bound the type-1 error rate of the hypothesis testing procedure to any desired level. Moreover, this guarantee is provided unconditionally even though each p-value, p, is calculated conditionally by restricting attention to a specific reference set Γ. To see this, let

$$\mathcal{S}(\Gamma) = \Pr(p \leq \alpha | \Gamma) \ . \tag{4}$$

That is, $\mathcal{S}(\Gamma)$ is the conditional type-1 error rate of a level-α hypothesis testing procedure in which you repeatedly generate $r \times c$ tables from the same reference set, Γ, under the null hypothesis, and reject whenever $p \leq \alpha$. Under the null hypothesis $\mathcal{S}(\Gamma) \leq \alpha$. Now the unconditional type-1 error rate, where Γ may be different each time you execute the test, is

$$\mathcal{S} = \sum \mathcal{S}(\Gamma) \Pr(\Gamma) \ , \tag{5}$$

the sum being taken over all possible reference sets, Γ. Notice that (5) is a weighted sum of terms of the form $\mathcal{S}(\Gamma)$, where each such term is less than or equal to α, the weights, $\Pr(\Gamma)$, are positive, and they sum to 1. Thus

$$\mathcal{S} \leq \alpha \ .$$

That is, the guaranteed protection against the type-1 error of an exact conditional hypothesis test also applies unconditionally. Note, however, that this guarantee does not hold if you use $\tilde{p}$ rather than p in the decision to reject the null hypothesis, since $\Pr(\tilde{p} \leq 0.05 | \Gamma) \leq \alpha$ holds only asymptotically.

Choosing the Test Statistic

As stated previously, the reference set Γ is ordered by the test statistic D. Here we define D for three important classes of problems; tests on unordered $r \times c$ contingency tables, tests on singly ordered $r \times c$ contingency tables and tests on doubly ordered $r \times c$ contingency tables.

When both the row and column classifications of the table are nominal the table is said to be unordered and 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 of no row by column interaction (see CHI-SQUARED TESTS).

Fisher's exact test orders each table, $\boldsymbol{y} \in \Gamma$, in proportion to its hypergeometric probability, $P(\boldsymbol{y})$, given by equation (2). Fisher [11] originally proposed this test for the single 2×2 contingency table. The idea was extended to tables of higher dimension by Freeman and Halton [12]. Thus, this test is also referred to as the Freeman-Halton test. Asymptotically, under the null hypothesis of no row by column interaction, $-2 \log \gamma P(\boldsymbol{y})$ has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom, where γ is a normalizing constant (Mehta and Patel [22]).

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

$$D(\boldsymbol{y}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - m_i n_j / N)^2}{m_i n_j / N} \ . \tag{6}$$

Asymptotically, under the null hypothesis of no row by column interaction, the Pearson statistic has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

The Likelihood Ratio test orders the tables in Γ according to the likelihood ratio statistic.

Specifically, for each $\boldsymbol{y} \in \Gamma$ the test statistic is

$$D(\boldsymbol{y}) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij} \log(\frac{y_{ij}}{m_i n_j / N}) \ . \tag{7}$$

In many textbooks this statistic is denoted by G^2. Asymptotically, under the null hypothesis of no row by column interaction, $D(\boldsymbol{y})$ has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

When there is a natural ordering of the columns of the $r \times c$ table, but the row classifications are based on nominal categories, appropriate tests are the Kruskal-Wallis test (Landis, Heyman and Koch [17]), and its generalization, the one-way ANOVA test (Miller [27]). 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.). One is interested in testing the null hypothesis that the r drugs have the same multinomial response rates. The Kruskal-Wallis and generalized one-way ANOVA tests are more powerful than the Fisher, Pearson or Likelihood Ratio tests for testing this null hypothesis against ordered alternatives which imply that some of these r drugs are more responsive than others. These tests take advantage of the natural ordering of the columns by assigning a rank or column score to all the observations in a column. The test statistic is obtained as a quadratic function of an r-dimensional vector whose components are formed by summing the column scores of the observations in each of the r rows and standardizing each sum. For the Kruskal-Wallis test the observations in a column are assigned their mid-rank and the special case, $r = 2$, yields the Wilcoxon rank-sum test. For the generalized one-way ANOVA test any monotone scores may be assigned. By suitable choice of these scores, one can construct a large number of tests, including the normal scores, exponential scores and logrank tests as special cases. The test statistics for all these tests are given in Chapter 18 of the StatXact-3 User Manual [37]. Asymptotically they are all distributed as chi-square with

$(r-1)$ degrees of freedom under the null hypothesis of no row by column interaction.

When the $r \times c$ contingency table has a natural ordering along both its rows and its columns, the Jonckheere-Terpstra test (Hollander and Wolfe [16]) and the Linear-by-Linear association test (Agresti, Mehta and Patel [3]) have more power than the Kruskal-Wallis test or the various $(r-1)$ degree of freedom generalized ANOVA tests. For example suppose the r rows represent r distinct drug therapies at progressively increasing 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. The Jonckheere-Terpstra statistic, is the normalized sum of $r(r-1)/2$ Wilcoxon rank-sum statistics formed by taking all possible pairs of rows from the r rows of the observed $r \times c$ contingency table, and computing a Wilcoxon rank-sum statistic for each resulting $2 \times c$ contingency table. The Linear-by-linear association statistic is obtained by standardizing $\sum_{i,j} u_i v_j y_{ij}$, where the u_i's are arbitrary row scores and the v_j's are arbitrary column scores. The row scores often represent progressively increasing doses of a treatment while the column scores often represent progressively increasing levels of response to treatment. If the u_i's and v_j's 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 Wilcoxon mid-rank scores, we have a test of Spearman's correlation coefficient. Refer to Chapter 19 of the StatXact-3 User Manual [37] for the precise functional forms of the Jonckheere-Terpstra and the Linear-by-linear test statistics. Under the null hypothesis of no row by column interaction these test

statistics are normally distributed. The special case, $r = 2$, yields the family of two-sample linear rank tests. For these tests, row scores are irrelevant but a large number of different column scores, covering most of the important nonparametric tests, are listed in Chapters 9 and 15 the StatXact-3 User Manual [37].

Extension to Continuous Data

The methods described above 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 displayed in Table 2. The two groups are 'males' and 'females'. The continuous variable being compared in the two groups is 'monthly income'.

These data can be represented by the 2×8 contingency table, displayed as Table 3, which may then be permuted in the usual way for exact inference.

The same idea extends to continuous K-sample data with or without stratification, and with or without censoring.

Stratified 2×2 Contingency Tables

A very important class of exact nonparametric tests and confidence intervals is defined on data in the form of s 2×2 contingency tables. The ith such table is displayed in Table 4 below.

We may regard the two columns of each table as arising from two independent binomial distributions. Specifically, let (x_{i1}, x_{i2}) represent the number of successes in (n_{i1}, n_{i2})

Bernoulli trials, with respective success probabilities (π_{i1}, π_{i2}). The odds ratio for the ith table is defined as

$$\Psi_i = (\frac{\pi_{i2}}{1-\pi_{i2}})/(\frac{\pi_{i1}}{1-\pi_{i1}}) \ . \tag{8}$$

Stratified 2×2 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 columns of the 2×2 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 (present or absent) and the two columns represent the exposure status (not-exposed or exposed) in the ith of s matched sets.

We shall be interested in deriving an exact test for the null hypothesis that

$$\Psi_1 = \Psi_2 = \cdots = \Psi_s = \Psi \ . \tag{9}$$

This is known as the homogeneity test. Next, under the assumption of homogeneity, we shall be interested in computing an exact confidence interval for the common odds ratio, Ψ, and in testing that it equals 1.

Homogeneity Test

Let $\boldsymbol{x}$ denote the observed collection of s 2×2 contingency tables, where the ith table in this collection is displayed in Table 4, and define

$$t = x_{11} + x_{21} + \cdots + x_{s1} \ . \tag{10}$$

Let Ω denote a reference set of collections of s 2×2 contingency tables whose margins are fixed at the values that were actually observed:

$$\Omega = \left\{ \begin{array}{ll} \boldsymbol{y}\text{:} & y_{i1} + y_{i2} = m_{i1};\ y'_{i1} + y'_{i2} = m_{i2} \\ & y_{i1} + y'_{i1} = n_{i1};\ y_{i2} + y'_{i2} = n_{i2} \ . \end{array} \right\} \tag{11}$$

Define the more restricted reference set

$$\Omega_t = \{\boldsymbol{y} \in \Omega\colon\ y_{11} + y_{21} + \ldots y_{s1} = t\}\ . \tag{12}$$

Zelen [41] has shown that under the null hypothesis of homogeneity (9)

$$\Pr(\boldsymbol{x}|\boldsymbol{x} \in \Omega_t) = \frac{\prod_{i=1}^{s}\prod_{j=1}^{2}\binom{n_{ij}}{x_{ij}}}{\sum_{\boldsymbol{y}\in\Omega_t}\prod_{i=1}^{s}\prod_{j=1}^{2}\binom{n_{ij}}{y_{ij}}}\ . \tag{13}$$

An exact test for the homogeneity of odds ratios can thus be constructed by ordering all elements $\boldsymbol{y} \in \Omega_t$ according to the test statistic

$$D(\boldsymbol{y}) = -\log \Pr(\boldsymbol{y}|\boldsymbol{y} \in \Omega_t)$$

and computing the exact p-value

$$p = \sum_{D(\boldsymbol{y})\geq D(\boldsymbol{x})} \Pr(\boldsymbol{y}|\boldsymbol{y} \in \Omega_t)\ .$$

This test is known as Zelen's exact test. A statistic proposed by Breslow and Day [7] is approximately distributed as chi-square with $(s-1)$ degrees of freedom under the null hypothesis (see BRESLOW-DAY TEST).

Common Odds Ratio Inference

Exact inference about Ψ, the common odds ratio, is based on the conditional distribution of

$$T = y_{11} + y_{21} + \cdots + y_{s1} \tag{14}$$

given $\boldsymbol{y} \in \Omega$. It is shown in Mehta, Patel and Gray [23] that

$$\Pr(T = t|\boldsymbol{y} \in \Omega) = \frac{C_t\Psi^t}{\sum_u C_u\Psi^u}\ , \tag{15}$$

where

$$C_t = \sum_{\boldsymbol{y}\in\Omega_t}\prod_{i=1}^{s}\prod_{j=1}^{2}\binom{n_{ij}}{y_{ij}}\ , \tag{16}$$

and the denominator of equation (15) is simply the normalizing constant obtained by summing over all possible values of u in the range $t_{min} \leq u \leq t_{max}$.

To test the null hypothesis that $\Psi = 1$ and to compute an exact confidence interval for this common odds ratio we need the coefficients C_t for all possible values of t. Network algorithms for this and related computations are described in Mehta, Patel and Gray [23]. Once these coefficients have been obtained, the conditional distribution of t for any value of Ψ can be generated by equation (15) and hypothesis tests and confidence intervals may thereby be obtained as shown in the above references.

Asymptotic inference for Ψ is usually based on the popular Mantel-Haenszel [20] method (see MANTEL-HAENSZEL TEST).

Extension to Stratified $2 \times c$ Contingency Tables

In this section we discuss inference on stratified $2 \times c$ tables where the ith of s such tables is displayed as Table 5.

This collection of s $2 \times c$ tables, denoted by $\boldsymbol{x}$, can accomodate two situations; two multinomial populations, and c binomial populations. For both cases we assume that data are stratified into s independent strata. Inference is conditional on ordering all three-way collections of s $2 \times c$ tables in the conditional reference set

$$\Lambda = \{\mathbf{y}\colon\ y_{ij} + y'_{ij} = n_{ij}, \forall ij;\ \sum_{j=1}^{c} y_{ij} = m_{i1},\ \sum_{j=1}^{c} y'_{ij} = m_{i2}, \forall i\} \tag{17}$$

according to some discrepancy measure $D(\boldsymbol{y})$. We shall be concerned in this section with the special case where the c columns of each $2 \times c$ contingency table have a natural ordering. In this case an appropriate (unstandardized) discrepancy measure is the linear rank test statistic

$$t(\boldsymbol{y}) = \sum_{i=1}^{s} \sum_{j=1}^{c} v_{ij} y_{ij} \tag{18}$$

where the v_{ij}'s are arbitrary column scores.

Two Multinomial Populations The two rows of stratum i represent two independent multinomial populations. Each observation falls into one of c ordinal response categories. Thus x_{ij} is the number of stratum-i observations, out of a total of m_{i1}, falling into ordered category j for population 1, and x'_{ij} is the number of stratum-i observations out of a total of m_{i2} falling into ordered category j for population 2. The Wilcoxon rank-sum test, the Normal scores test, the Savage test, and the logrank test are examples of tests that are applicable to such data. The v_{ij} scores for these tests are defined in Chapter 15 of StatXact-3 [37].

Several Binomial Populations The c columns of stratum i represent c independent binomial populations with row 1 representing successes and row 2 representing failures. For population j in stratum i there are x_{ij} successes and x'_{ij} failures in n_{ij} independent Bernoulli trials. The Cochran-Armitage Trend test and the Permutation test with arbitrary scores are applicable to such data, and determine whether the success rates of the c populations are the same, as against the alternative that they follow an increasing or decreasing trend. The scores, $v_{i1}, v_{i2}, \ldots v_{ic}$ typically represent doses, or levels of exposure, affecting the success rates of the c binomial populations. Often one uses the equally spaced scores $v_{ij} = j$ for all i.

We shall assume throughout that there exists no three-factor interaction between rows, columns and strata. An exact test of this hypothesis requires us to include interaction terms in a general logistic regression framework and test that they equal zero. This is an extension of the exact test of homogeneity for s 2×2 contingency tables and will not be discussed here. Given that there is no three-factor interaction, however, we are interested in testing the null hypothesis that the row and column classifications in each stratum are

independent. This is known as the hypothesis of conditional independence. One can show that, for both the two multinomial and the c binomial settings under the null hypothesis of conditional independence, the probability of observing $\boldsymbol{y}$ given $\boldsymbol{y} \in \Lambda$ is

$$\Pr(\boldsymbol{y}|\boldsymbol{y} \in \Lambda) = \frac{\prod_{i=1}^{s} \prod_{j=1}^{c} \binom{n_{ij}}{y_{ij}}}{\prod_{i=1}^{s} \binom{N_i}{m_{i1}}} \ . \tag{19}$$

The exact one-sided p-value for testing the null hypothesis of conditional independence is therefore

$$p_1 = \sum_{t(\boldsymbol{y}) \geq t(\boldsymbol{x})} \Pr(\boldsymbol{y}|\boldsymbol{y} \in \Lambda) \ . \tag{20}$$

The exact two-sided p-value is defined by reflecting the observed value of the test statistic an equal distant away from its mean in the opposite tail. See StatXact-3 [37] for details.

Computational Issues

Computing equation (3) is a nontrivial task. The size of the reference set grows exponentially so that explicit enumeration of all the tables in Γ soon becomes computationally infeasible. For example, the reference set of all 5×6 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, Patel, and co-workers, [22], [23], [21], [26], [25], to enumerate the tables in Γ implicitly. In these algorithms the reference set is represented by a network of nodes and arcs. A sequence of connected arcs from the starting to the terminal node constitutes a path through the network. Each such path represents one and only one table in Γ. The length of a path equals the value of the test statistic for the table to which that path corresponds. The probability of the path equals the probability of the corresponding table. Thus the problem of computing an exact p-value is equivalent to the problem of

identifying paths whose lengths equal or exceed a specified value, and summing the probabilities of all these paths. This can be accomplished by well-known operations research techniques such as backward induction and forward probing through the network. These methods are very efficient and make it feasible to compute exact p-values.

Alternate approaches are provided by Pagano and Halvorsen [28], Pagano and Tritchler [29], Streitberg and Rohmel [38], Baglivo, Olivier and Pagano [6], Vollset, Hirji and Elashoff [39], and Cheung and Klotz [8]. 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 Γ in proportion to their hypergeometric probabilities (2), and a count is kept of all the sampled tables that are more extreme than the observed table. For details, refer to Agresti, Wackerly and Boyett [4], Patefield [30], Mehta, Patel and Senchaudhuri [24], and Senchaudhuri, Mehta and Patel [34].

Analysis of Data Sets

In this section we will illustrate the techniques developed in the previous sections with some data analyses. Each example will highlight the different conclusions one might draw if an asymptotic analyis were performed instead of an exact analysis. A large number of additional examples are available at the Cytel web site `http://www.cytel.com`. All results were obtained by the StatXact-3 software package [37].

An Unordered Contingency Table

Data were obtained on the location of oral lesions, in house to house surveys in three geographic regions of rural India, by Gupta, Mehta and Pindborg [15]. Consider a hypothetical subset of these data displayed by Table 6 as a 9×3 contingency table in which the counts are the number of patients with oral lesions per site and geographic region.

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 9×3 table are clearly unordered, making it an appropriate data set for either the Fisher, Pearson or Likelihood Ratio tests. The exact and asymptotic p-values are displayed in Table 7. There are striking differences between the two methods. The exact analysis suggests that the row and column classifications are dependent, but the asymptotic analysis fails to show this.

A Singly Ordered Contingency Table

The tumor regression rates of five chemotherapy regimens, Cytoxan (CTX) alone, Cyclohexyl-chloroethyl nitrosurea (CCNU) alone, Methotrexate (MTX) alone, CTX + MTX, and CTX + CCNU + MTX were compared in a small clinical trial. Tumor regression was measured on a three-point scale: no response, partial response, or complete response. The results are tabulated in Table 8

Small pilot studies like this one are frequently conducted as a preliminary to planning a large-scale randomized clinical trial. 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 tumor 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, based on the permutation distribution of the Kruskal-Wallis statistic, the exact p-value is 0.039, which is statistically significant.

Analysis of Stratified 2×2 Contingency Tables

Testing the Homogeneity of Odds Ratios

The binary response data tabulated in Table 9 compare a new drug with a control drug at 22 hospital sites.

The data can be thought of as twenty-two 2×2 contingency tables, one for each site. If you examine the 2×2 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 (Gastwirth [13]) involved the minority hiring data displayed in Table 10.

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 [20] 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.

Test of Trend in Stratified $2 \times c$ Contingency Tables

The data for this example were provided by the U.S. Food and Drug Administration (FDA). Animals were treated with four dose levels of a carcinogen and then observed (at necropsy) for the presence or absence of a tumor type. The data were stratified by survival time (in weeks) into the four time-intervals 0–50, 51–80, 81–104, and terminal sacrifice. Since there were no tumors found in the first time-interval, this stratum may be excluded from data entry. The data for the remaining three strata are displayed in Table 11.

We use the **stratified** Cochran-Armitage trend test (Breslow and Day [7], page 148) to determine if there is a dose-response relationship between the level of carcinogen and the presence of tumors. The test statistic is defined by equation (18), where v_{ij} is the dose-level of carcinogen and y_{ij} is the number of animals with tumors, at the jth dose level in the ith stratum. The results are tabulated in Table 12.

There are large differences between the exact and asymptotic one-sided p-values, and they lead to different conclusions about the significance of the dose-response relationship. They also show that the usual practice of doubling the one-sided p-value is unnecessarily conservative with asymmetric distributions. But the most interesting finding of all is that the distribution of the linear rank statistic (18) has multiple towers. A normal approximation would be seriously misleading. This is shown in Figure 1.

Software and Related Resources for Exact Inference

We have presented the essential idea behind exact permutational inference, described one numerical algorithm, referenced others, and shown through several examples that exact inference is a valuable supplement to corresponding asymptotic methods.

Software support for these methods is available in many standard packages including StatXact-3 [37], LogXact-2 [18], SPSS Exact Tests [36] and SAS Version 6.11 [33]. A brief description of the StatXact-3 and LogXact-2 software packages is given elsewhere (see STATXACT).

Some of the newer textbooks on nonparametric methods, for example, Manly [19], Sprent [35], Good [14], Edgington [10], and Agresti [1] devote considerable space to exact and Monte Carlo methods of inference for categorical data. A useful survey paper, in which a unified treatment of exact inference for categorical data is presented through the log linear model, was recently published by Agresti [2]. A complete collection of references to statistical methodology, numerical algorithms, commercial software, shareware, and textbooks on exact permutational inference can be obtained by visiting the Exact-Stats world-wide web site on the Internet. The address is `http://www.mailbase.ac.uk/lists-a-e/exact-stats`.

References

1 Agresti A (1990). *Categorical Data Analysis.* John Wiley & Sons, New York.

2 Agresti A (1992). A survey of exact inference for contingency tables (with discussion). *Statistical Science* 7(1):131-177.

3 Agresti A, Mehta CR, Patel NR (1990). Exact inference for contingency tables with ordered categories. *J. Amer. Stat. Assoc.*, 85, 410, 453-458.

4 Agresti A, Wackerly D and Boyett (1979). Exact conditional tests for cross-classifications. *Psychometrika* 44:75-83.

5 Agresti A, Yang M (1987). An empirical investigation of some effects of sparseness in contingency tables. *Comm. Stat.*, 5:9-21.

6 Baglivo J, Olivier D, Pagano M (1988). Methods for the analysis of contingency tables with large and small cell counts. *J. Amer. Stat. Assoc.*, 83, 1006-1013.

7 Breslow NE, Day NE (1980). The analysis of case-control studies. *IARC Scientific Publications No. 32* Lyon, France.

8 Cheung YK, Klotz JH (1996). The Mann Whitney Wilcoxon distribution using linked lists. *Statistica Sinica* (in press).

9 Cox DR and Hinkley DV (1974). *Theoretical Statistics.* Chapman and Hall, London.

10 Edgington ES (1995). *Randomization tests*, 3rd edition. Marcel Dekker, New York.

11 Fisher RA (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

12 Freeman GH, Halton JH (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38:141-149.

13 Gastwirth JL (1984). Combined tests of Significance in EEO cases. *Industrial and Labor Relations Review* 38(1).

14 Good P (1993). *Permutation Tests*. Springer Verlag, New York.

15 Gupta PC, Mehta FR, Pindborg J (1980). *Comm. Dent. and Oral Epid.*, 8:287- 333.

16 Hollander M, Wolfe DA (1973). *Nonparametric Statistical Methods*. John Wiley, New York.

17 Landis R, Heyman ER, Koch GG (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Int. Stat. Rev.* 46:237-254.

18 LogXact-2 for Windows (1996). *Software for Exact Logistic Regression*. Cytel Software Corporation, Cambridge, MA.

19 Manly BFJ (1991). *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.

20 Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22:719-748.

21 Mehta CR, Patel NR (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Stat. Assoc.* 78(382):427-434.

22 Mehta CR, Patel NR (1986). A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables. *Comm. Stat.* 15(2):387-403.

23 Mehta CR, Patel NR, Gray R (1985). On computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *J. Amer. Stat. Assoc.* 80(392):969-973.

24 Mehta CR, Patel NR, Senchaudhuri P (1988). Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Stat. Assoc.* 83(404):999-1005.

25 Mehta CR, Patel NR, Senchaudhuri P (1992). Exact stratified linear rank tests for ordered categorical and binary data. *J. Computational and Graphical Statistics* 1:21-40.

26 Mehta CR, Patel NR, Tsiatis AA (1984). Exact significance testing to establish treatment equivalence for ordered categorical data. *Biometrics* 40:819-825.

27 Miller RG (1981). *Simultaneous Statistical Inference*. Springer-Verlag, New York.

28 Pagano M, Halvorsen K (1981). An algorithm for finding exact significance levels of $r \times c$ contingency tables. *J. Amer. Stat. Assoc.*, 76:931-934.

29 Pagano M, Tritchler D (1983). On obtaining permutation distributions in polynomial time. *J. Amer. Stat. Assoc.* 78:435-441.

30 Patefield WM (1981). An efficient method of generating $r \times c$ tables with given row and column totals. (Algorithm AS 159). *Applied Statistics* 30:91-97.

31 Read RC, Cressie NA (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

32 Reid N (1995). The roles of conditioning in inference (with discussion). *Statistical Science* 10(2):138-157.

33 SAS Institute Inc. (1995). SAS/Stat User's Guide, Version 6.11. Cary, NC. The SAS Institute, Cary, NC.

34 Senchaudhuri P, Mehta CR, Patel NR (1995). Estimating exact p-values by the method of control variates, or Monte Carlo rescue. *J. Amer. Stat. Assoc.* 90(430):640- 648.

35 Sprent P (1993). *Applied Nonparametric Statistical Methods*. Second edition. Chapman and Hall, London.

36 SPSS Exact Tests for Windows (1995). SPSS Inc., Chicago, IL.

37 StatXact-3 for Windows (1995). *Software for Exact Nonparametric Inference*. Cytel Software Corporation, Cambridge, MA.

38 Streitberg B, Rohmel R (1986). Exact distributions for permutation and rank tests. *Statistical Software Newsletter* 12:10-17.

39 Vollset SE, Hirji KF, Elashoff RM (1991). Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *J. Amer. Stat. Assoc.* 86: 404-409.

40 Yates F (1984). Test of significance for 2×2 contingency tables. *J. Roy. Stat. Soc. Series A* 147:426-463.

41 Zelen M (1971). The analysis of several 2×2 contingency tables. *Biometrika* 58(1):129-137.

Tables and Figures

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
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
Row_r	x_{r1}	x_{r2}	...	x_{rc}	m_r
Col_Tot	n_1	n_2	...	n_c	N

Table 2: Two-sample Continuous Data Represented the Traditional Way

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

Table 3: Two-sample Continuous Data Represented as a 2×8 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
Col_Tot	1	1	1	1	1	1	1	1	8
Col_Score	1875	1990	2010	2095	2122	2550	2555	3100	

Table 4: Layout for the ith of s 2×2 Contingency Tables

Rows	Col_1	Col_2	Row_Total
Row_1	x_{i1}	x_{i2}	m_{i1}
Row_2	x'_{i1}	x'_{i2}	m_{i2}
Col_Tot	n_{i1}	n_{i2}	N_i

Table 5: Layout for the ith of s $2 \times c$ Contingency Tables

Rows	Col_1	Col_2	...	Col_c	Row_Total
Row_1	x_{i1}	x_{i2}	...	x_{ic}	m_{i1}
Row_2	x'_{i1}	x'_{i2}	...	x'_{ic}	m_{i2}
Col_Tot	n_{i1}	n_{i2}	...	n_{ic}	N_i
Col_Score	v_{i1}	v_{i2}	...	v_{ic}	

Table 6: 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

Table 7: Exact and Asymptotic P-values for Oral Lesions Data

TYPE OF INFERENCE	THREE TESTS OF INDEPENDENCE		
	PEARSON	FISHER	LIKELIHOOD RATIO
Value of $D(\boldsymbol{x})$	22.1	19.72	23.3
Asymptotic p-value	.1400	.2331	.1060
Exact p-value	.0269	.0101	.0356

Table 8: Chemotherapy Pilot Study Data

CHEMO	No Resp.	Partial Resp.	Complete Resp.
CTX	2	0	0
CCNU	1	1	0
MTX	3	0	0
CTX+CCNU	2	2	0
CTX+CCNU+MTX	1	1	4

Table 9: Site by Treatment Interaction Data

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

Table 10: Minority Hiring Data

Date of Hire	Whites		Blacks	
	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

Table 11: FDA Animal Toxicology Data

Stratum 1: 51–80 weeks of survival					
	Dose of Carcinogen				
Disease Status	None	1 unit	5 units	50 units	Total
Tumor Present	0	0	0	1	1
Tumor Absent	7	10	6	8	31
Stratum 2: 81–104 weeks of survival					
	Dose of Carcinogen				
Disease Status	None	1 unit	5 units	50 units	Total
Tumor Present	0	1	0	1	2
Tumor Absent	11	9	13	14	47
Stratum 3: Sacrificed at end of 104 weeks					
	Dose of Carcinogen				
Disease Status	None	1 unit	5 units	50 units	Total
Tumor Present	1	1	1	2	5
Tumor Absent	29	26	28	20	103

Table 12: One and Two-sided P-values for FDA Data

P-VALUES	One-Sided	Two-Sided	Double One-Sided
Exact	.0651	.0769	.1302
Asymptotic	.0410	.0820	.0820

Figure 1: Distribution of Trend Test Statistic for FDA Data