STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2709–2722 (DOI: 10.1002/sim.738)

Exact inference for categorical data: recent advances and continuing controversies

Alan Agresti*,†

Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.

SUMMARY

Methods for exact small-sample analyses with categorical data have been increasingly well developed in recent years. A variety of exact methods exist, primarily using the approach that eliminates unknown parameters by conditioning on their sufficient statistics. In addition, a variety of algorithms now exist for implementing the methods. This paper briefly summarizes the exact approaches and describes recent developments. Controversy continues about the appropriateness of some exact methods, primarily relating to their conservative nature because of discreteness. This issue is examined for two simple problems in which discreteness can be severe – interval estimation of a proportion and the odds ratio. In general, adjusted exact methods based on the mid-*P*-value seem a reasonable way of reducing the severity of this problem. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

The recent development of various algorithms for implementing 'exact' small-sample analyses has been a major advance in categorical data analysis [1–5]. With these methods one can make probability calculations such as *P*-values using exactly specified distributions rather than with approximate large-sample ones. Exact methods guarantee that the size of a hypothesis test is no greater than the nominal level and that the coverage probability for a confidence interval is at least the nominal confidence coefficient. A variety of exact methods exist, using both conditional and unconditional approaches. This article surveys the exact approaches and summarizes some of the recent developments.

Many statisticians have been critical of some exact methods. Sometimes this is for philosophical reasons but more often it is because, although the probability calculations are exact, they lead to conservative inferences when used as the basis of hypothesis tests and confidence intervals. The conservatism, meaning that actual error probabilities are less than nominal

*Correspondence to: Alan Agresti, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.
† E-mail: aa@stat.ufl.edu

Contract/grant sponsor: NIH
Contract/grant sponsor: NSF

levels, reflects complications due to discreteness. We illustrate these issues for interval estimation of two basic parameters – the proportion and the odds ratio. For highly discrete data it seems sensible to use adjustments of exact methods based on smoothings of P-values, such as the mid-P-value.

2. THE EXACT CONDITIONAL APPROACH

This section reviews the conditional approach to exact inference for categorical data. This utilizes the distribution of the sufficient statistic for the parameter of interest, conditional on sufficient statistics for the other model parameters.

First, consider a two-way contingency table having r rows and c columns, cross-classifying a row explanatory variable X and column response variable Y. In biomedical applications, probably the most common sampling scheme is independent multinomial samples within the rows; if instead the sampling is full multinomial over the entire table, the independent multinomial scheme applies after conditioning on totals at the levels of X. Let $\pi_{j|i} = P(Y = j|X = i)$, $i = 1, \ldots, r$, $j = 1, \ldots, c$, and denote the cell counts by $\{n_{ij}\}$, with $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_j n_{ij}$ and $n = \sum_{i,j} n_{ij}$. Under independence of Y and X, for each j, $\pi_{j|1} = \cdots = \pi_{j|r}$. For 2×2 tables, this is equivalent to the odds ratio $\theta = 1.0$. One can test this hypothesis by comparing $\{n_{ij}\}$ to the maximum likelihood fit $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$ using chi-squared statistics such as the Pearson and likelihood-ratio statistics. These have asymptotic null chi-squared distributions, but the approximation may be poor if many $\{\hat{\mu}_{ij}\}$ are small (for example, less than about 5). Asymptotic behaviour may also be poor when the overall sample size is large but the contingency table has a large number of cells. Difficulties can occur when the table contains a mixture of relatively large counts and relatively small ones or when an asymptotic framework applies in which the number of cells (and hence the number of parameters) grows with the sample size [6].

The actual null distribution of chi-squared statistics depends on $\{\pi_{j|i}\}$. When one conditions on the sufficient statistics $\{n_{+j}\}$ for $\{\pi_{j|i}\}$ as well as the multinomial sample sizes $\{n_{i+}\}$, the resulting distribution no longer depends on those parameters. R. A. Fisher, who had introduced the concept of sufficient statistics, noted this for 2×2 tables, in which case the conditional distribution is the hypergeometric. In the non-null case, this distribution depends also on θ and is

$$P(n_{11} = k|\{n_{i+}\}, \{n_{+j}\}; \theta) = \frac{\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}\theta^k}{\sum_u \binom{n_{1+}}{u}\binom{n-n_{1+}}{n_{+1}-u}\theta^u}$$

The P-value for the one-sided test of H_0:$\theta = 1$ against H_1:$\theta > 1$, based on sample odds ratio $\hat{\theta}_{\text{obs}}$ and count $n_{11,\text{obs}}$, is

$$P\text{-value} = P_{H_0}[\hat{\theta} \geqslant \hat{\theta}_{\text{obs}}|\{n_{i+}\}, \{n_{+j}\}] = P(n_{11} \geqslant n_{11,\text{obs}}|\{n_{i+}\}, \{n_{+j}\}; \theta = 1)$$

This test is called *Fisher's exact test*, reflecting its basis on an exact probability calculation rather than a large-sample approximation. Non-null inference about θ, such as confidence intervals, uses this distribution for all θ. An exact confidence interval for θ results from

inverting two separate one-sided exact tests [7]; for example, the lower endpoint of a 95 per cent confidence interval is the θ_0 value for which P-value $= 0.025$ in testing $H_0{:}\theta = \theta_0$ against $H_1{:}\theta > \theta_0$.

The exact conditional approach is trustworthy compared to possibly poor large-sample approximations. It is also quite flexible, applying to any exponential family model with canonical link, which are the models for which nuisance parameters have reduced sufficient statistics. This includes all Poisson log-linear models and binomial logit models. For instance, independence in $r \times c$ tables can be formulated as a Poisson log-linear model with main effects. Conditioning on row and column totals to eliminate those parameters yields a multivariate hypergeometric distribution defined on the set of tables having the same row and column margins as the observed table. The P-value equals the total probability of those tables that have test statistic value at least as contradictory to the null as the observed value.

For another example, consider the logistic regression model. For subject i, denote the response by $y_i = 0$ or 1, and denote the explanatory variables by $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$. The model is

$$\log[P(Y_i = 1)/P(Y_i = 0)] = \sum_j \beta_j x_{ij}$$

with typically $x_{i0} = 1$ for an intercept term. Consider testing the effect of predictor j on the response and estimating the log-odds ratio β_j to which the effect pertains. The sufficient statistic for β_j is $T_j = \sum_i y_i x_{ij}$. Exact inference for β_j uses the distribution of T_j, conditional on $\{T_k,\ k \neq j\}$. To test $H_0{:}\beta_j = 0$ against $H_1{:}\beta_j > 0$

$$P\text{-value} = P(T_j \geqslant t_j | \{T_k = t_k,\ k \neq j\})$$

where $\{t_k\}$ denote observed values [1]. Exact inference after eliminating nuisance parameters with conditional logistic regression is especially useful in matched case-control studies [8, 9] and other applications in which the strata contain few observations and the number of strata (and hence the number of parameters) increases with the sample size. The method extends to certain multinomial logit models, such as ones using baseline-category logits.

For many years the use of exact methods was hindered by the lack of suitable software. The reference set of tables on which the conditional distribution is defined may be difficult to generate or it may be enormous. For instance, a 4×4 table with only 20 observations can have over 40000 tables with the same margins, and with 100 observations it can have about 7×10^9 such tables. Various computational solutions have been proposed. One of these computes the characteristic function of the statistic of interest using a recurrence relation and then inverts it using a Fourier transform to obtain the relevant distribution [10]. A series of papers in the past 15 years by Cyrus Mehta, Nitin Patel and various co-authors showed how to use an alternative approach, the network algorithm. They developed software that provides several exact methods (for example, *StatXact* and *LogXact*, distributed by Cytel Software in Cambridge, MA). Modules from StatXact are available in SAS (for example, PROC FREQ and PROC STATXACT) and in SPSS. A variety of problems can be handled, including tests of independence in two-way tables with ordered or unordered categories, tests of conditional independence and of homogeneous odds ratios in several 2×2 tables, and inferences for parameters in logistic regression. Exact inferences for multinomial logit models will apparently be available in an upcoming revision of LogXact.

Even with modern computing power, exact methods are sometimes computationally unfeasible. Algorithms that provide total enumeration of the conditional reference set of tables are time consuming and inadequate for large problems. However, Monte Carlo methods can estimate precisely the results of many exact inferences, by randomly sampling from the conditional distribution [11–14].

Although the conditional approach is versatile, it does have limitations. For binary data, for instance, it does not apply except for logit models. One cannot use it to conduct exact inference about the difference of proportions or other parameters that do not occur in logit modelling. Also, the relevant conditional distributions can be highly discrete, which leads to conservativeness in inference, as discussed in Section 5. Further details about this approach are available in recent survey articles [1, 2, 4, 5].

3. AN EXACT UNCONDITIONAL APPROACH

When a contingency table consists of independent samples within rows, only $\{n_{i+}\}$ are naturally fixed. Some find the conditional approach, which also fixes $\{n_{+j}\}$, artificial. An unconditional approach eliminates nuisance parameters using a 'worst-case' scenario. The P-value is a tail probability maximized over all possible values for the nuisance parameters. Such a test uses the original (unconditional) distributions.

We illustrate with the comparison of binomial parameters for two independent samples. The null hypothesis is $H_0{:}\pi_{1|1} = \pi_{1|2}$. Given the common value (under H_0) of $\rho = \pi_{1|1} = \pi_{1|2}$, let $P_\rho(T \geqslant t)$ denote the null probability that the test statistic T is at least as large as the observed value t. Since ρ is unknown, one can form a P-value using [15] $P\text{-value} = \sup_{0 \leqslant \rho \leqslant 1} P_\rho(T \geqslant t)$. For the usual decision framework, rejecting H_0 if $P\text{-value} \leqslant \alpha$, the test guarantees that the actual size is no greater than α. Because of this and since it does not require estimating unknown parameters, it can also be called an 'exact' method. It has been proposed in various forms (for example, references [16–18]).

In principle this approach is more general than the conditional one, since it does not require models with reduced sufficient statistics. However, it is a computational challenge to extend it to the wide variety of problems to which the conditional approach has been applied, particularly those with several nuisance parameters. Also, taking the supremum over unknown parameters itself leads to conservatism. This discussion illustrates that there is more than one possible 'exact' approach. Moreover, each method has quite different ways of performing it. For instance, one can define P-values in different ways, one might use a likelihood-ratio, Wald, or score statistic as the test statistic, and one can construct confidence intervals by inverting two separate one-tailed tests or a single two-tailed test.

Statisticians have been critical of both of these approaches to contingency table analysis. Criticisms of the conditional approach partly refer to using a sample space consisting only of tables having *exactly* the same response margins as the observed table. Proponents of it respond that it is unnatural to consider samples that are quite different from the observed one in terms of characteristics (such as marginal distributions) that provide little or no information about the association. In addition, in many biomedical applications one does not truly have multinomial or independent binomial samples. For the observed sample one can still then use Fisher's exact test through a permutational argument, considering all ways the subjects could have been assigned to the two samples, as a way of checking whether the observed results

are unusual. For instance, suppose a test refers to comparing drug with placebo and the null is true that the drug has no effect, providing the same success probability as placebo; then the same total number of successes would have occurred no matter how subjects were assigned to the two groups.

Other statisticians have argued that the unconditional approach is artificial because it averages what happened in the observed sample with hypothetical response distributions, some of which are much different than observed. For instance, Fisher [19] argued that only the sampling distribution of samples of the same type can supply a rational test of significance. Opinions on this issue have been expressed in sufficient detail [16, 17, 20–28] that we will not further address the arguments here. In our opinion, much of the disagreement on this issue is inflamed by the quite different results the methods can provide when the relevant distribution is highly discrete. We return to this issue in Section 5.

4. RECENT RESEARCH ON EXACT INFERENTIAL METHODS

We now turn our attention to some recent developments. This section summarizes a few of the main areas that have seen significant progress in the past 5–10 years, primarily based on the exact conditional approach.

Three-way contingency tables are common in practice, for instance for cross-classifying a categorical response by some factor separately at levels of a possibly confounding factor or for the centres at which data are collected; for example, such a table might relate dosage of drug (placebo, low dose, high dose) and response (success, partial success, failure) for subjects stratified by clinic. Two hypotheses of interest are (i) conditional independence of treatment and response, and (ii) homogeneity of effects (for example, odds ratios) across levels of the stratifying factor. The StatXact software is limited to stratified 2×2 tables for the latter (Zelen's [29] test) and stratified $2 \times c$ tables with ordered columns [30, 31] for the former.

Many possible alternative hypotheses can apply for tests of conditional independence in stratified $r \times c$ tables. For instance, the test statistics for the large-sample chi-squared tests given by PROC FREQ in SAS can be derived as score statistics about conditional association parameters in log-linear models for ordinal or nominal variables or a mixture of the two [32]. Although software is not available for exact inference, it is straightforward to use Monte Carlo methods to approximate exact results for these alternatives by sampling from the product multivariate hypergeometric distribution that applies after conditioning on row and column totals in each stratum [32]. For ordinal variables a related approach uses test statistics obtained by expressing the alternative in terms of various types of monotone trends, such as uniformly non-negative values of ordinal odds ratios of various types [33]. When $c = 2$, this reduces to comparing r binomial parameters (in each stratum) against an alternative in which the parameters are monotone increasing [34].

For the null hypotheses of independence and conditional independence, the relevant conditional distribution is simple, which is why ordinary Monte Carlo applies so easily. This is not true of more complex hypotheses, such as quasi-independence, quasi-symmetry, and others for square and triangular contingency tables that were until recently given little attention in the exact literature. A group of statisticians at Southampton (U.K.) have developed alternative methods, such as Markov chain Monte Carlo (MCMC), to approximate precisely exact

P-values for such cases [35–40]. Their MCMC approach applies more generally to log-linear and logit models [41] including log-linear models for rates [42].

Booth and Butler [14] provided an alternative computational approach using a general simulation method for exact tests of goodness-of-fit for log-linear models. Their Monte Carlo approximation utilizes an importance sampling method based on a crude normal approximation to the Poisson. They illustrated their approach with a wide variety of log-linear models, including quasi-symmetry and related models for square tables, mutual independence in a three-way table and uniform association for ordinal tables. One can use their methodology, for instance, to generalize the Zelen [29] test of homogeneity of odds ratios for stratified 2×2 tables to stratified $r \times c$ tables, by treating the test as one of goodness-of-fit for the model of no three-factor interaction. Although their primary focus was testing goodness-of-fit, in which case large-sample chi-squared approximations are often poor for sparse data, they also showed how to simulate exact tests with unsaturated alternatives. In all their examples the test d.f. < 20, and they noted that the importance sampling method breaks down for large values of d.f.; in that case, MCMC methods [41] seem to be the method of choice.

Among the other areas in which the conditional approach has seen much recent attention are inference about an assumed common odds ratio in several 2×2 tables [43–48] including an improved confidence interval for a common odds ratio [49], inference for two-way tables with ordered categories [14, 50–53], log-linear models for multi-way tables [14, 54, 55], logistic regression [1, 41] and polytomous extensions of logistic regression [56], power and sample size calculations [45, 57–63], and alternative algorithms for implementing exact methods [3, 10, 54, 64–67]. The past decade has also seen much research on higher-order improvements of large-sample methods [68] to make results more closely agree with those using exact methods. This work includes a resurgence of interest in saddlepoint methods for approximating conditional distributions [69–74]. Another approximate method with substantial promise for improving on standard asymptotics is the iterated bootstrap [75]. When Monte Carlo is feasible, however, it has the advantage that estimated *P*-values necessarily converge to the actual exact ones as the number of simulations increases.

The unconditional approach has also seen new results. For testing equality of two binomial parameters, Berger and Boos [76] recently answered a primary criticism of it by restricting the supremum search over the unknown success probability ρ to those values compatible with the data. Letting C_γ denote a $100(1-\gamma)$ per cent confidence interval for ρ, where γ is very small (for example, 0.001) and arbitrary, they defined

$$P\text{-value} = \sup_{\rho \in C_\gamma} [P_\rho(T \geqslant t)] + \gamma$$

This approach also guarantees that the actual size is no greater than the nominal size. This adjusted unconditional approach has been generalized to an analogue of the Cochran–Mantel–Haenszel test of conditional independence for several 2×2 tables and the trend test for $2 \times c$ tables [77].

5. COMPLICATIONS FROM DISCRETENESS

Although some statisticians have philosophical objections to the conditional approach, the root of most objections is their conservatism, because of discreteness. In terms of practical

performance the degree of discreteness is the determinant more so than whether one uses conditioning [78].

To illustrate the basic problem, consider Fisher's exact test with a significance level of $\alpha = 0.05$. For the data

3	1
1	3

the P-value for a one-sided test of $H_0 : \theta = 1$ against $H_1 : \theta > 1$ is 0.243. With these margins, the only possible P-values are 0.014, 0.243, 0.757, 0.986, 1.00, for the observed counts $n_{11} = 4, 3, 2, 1, 0$, respectively. Using $\alpha = 0.05$ as a nominal size in a test, one rejects H_0 only when $n_{11} = 4$, but under the null, the probability this happens is 0.014, so the actual P (type I error) $= 0.014$. In theory, this is not a problem. One can use supplementary randomization to achieve a desired size. If one rejects H_0 when $n_{11} = 4$ and with probability 0.16 when $n_{11} = 3$, then the actual size achieves 0.05. With such randomization, conditional tests in exponential family models have optimality properties; for instance, the one-sided exact conditional test for an odds ratio in a 2×2 table is uniformly most powerful out of the unbiased tests.

In the real world, data-unrelated randomization is unacceptable, and it is rarely possible to achieve an arbitrary size such as 0.05. Some argue that fixing an unachievable α-level is artificial and that one should either merely report the P-value, or if a decision must be made, choose α to be a possible P-value. The discreteness has more disturbing implications, however, for unconditional power calculations and for confidence intervals. The usual exact confidence interval consists of parameter values not rejected in corresponding exact tests. Since the family of tests all have actual size $\leqslant 0.05$ with a high percentage having actual size much less than 0.05, the true confidence coefficient at any particular parameter value is bounded below by 0.95 and may be substantially larger [79]. In constructing a nominal 95 per cent confidence interval, we know only that the actual confidence level is at least that high; we do not know the level, since we do not know the true parameter value.

For 2×2 contingency tables the degree of conservativeness is usually more severe for conditional than for unconditional inference, because the extra conditioning increases the severity of discreteness [17, 25]. Generally, it may be non-negligible unless n is large or the table has many cells with the data spread thinly. In some cases, the conditional distribution is even degenerate, with mass concentrated at one point, so that necessarily the exact P-value $= 1$ and this approach is uninformative. This is common in logistic regression with one or more continuous predictors. Then the observed sequence of 0 and 1 responses may be the only one that can have the observed values of the sufficient statistics that are fixed for the conditional inference.

Large-sample methods do not have the guarantee of bounds on error probabilities. They can be conservative or liberal, and thus their results can appear quite different than exact methods. For example, for the 2×2 table discussed above, the P-value for the Pearson chi-squared test equals 0.157, compared to 0.486 for the two-sided exact test. A 95 per cent large-sample confidence interval for the odds ratio, based on the formula given below, is $(0.4, 221)$, compared to Cornfield's exact interval of $(0.2, 626)$. Although normally one would prefer an exact method over an approximate one, when the conditional distribution is highly discrete the choice is not so obvious. The next section illustrates this for inference about odds ratios and proportions.

Table I. Behaviour of 95 per cent confidence intervals for the odds ratio, with 10000 randomly generated pairs of binomial parameters, when $n_{1+} = n_{2+} = 10$.

Criterion	Large-sample	Exact	Mid-P
Mean coverage probability	0.977	0.990	0.970
Minimum coverage probability	0.941	0.970	0.938
Minimum coverage probability, $\theta < 20$	0.949	0.970	0.938
Median expected length	61	228	118

6. EXAMPLES OF CONSERVATISM OF EXACT METHODS

For 2×2 tables, a large-sample confidence interval for the log-odds ratio is

$$\log(n_{11}n_{22}/n_{12}n_{21}) \pm z(n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1})^{1/2}$$

where z is the appropriate standard normal percentile (for example, 1.96 for 95 per cent confidence). Table I illustrates the behaviour for small samples of this interval (transformed to the odds ratio scale) and Cornfield's exact conditional interval. For 10000 pairs of parameter values randomly selected over the triangular region $0 \leqslant \pi_{1|2} \leqslant \pi_{1|1} \leqslant 1$ in which $\theta \geqslant 1.0$ and for binomial sampling with $\{n_{i+} = 10\}$, for both methods we calculated the true coverage probability and the expected length of the nominal 95 per cent confidence interval. Table I reports the minimum and the mean coverage probability and the median of the expected lengths. To illustrate the behaviour for the portion of the parameter space usually dealt with in practice, Table I also reports the minimum coverage probability for the probability pairs for which $\theta < 20$.

Table I shows that the exact method can be quite conservative for small n. Here, the large-sample interval performs surprisingly well. Because of tail behaviour, differences in coverage probability can translate to large differences in expected interval length. (This is especially so on the odds ratio scale; by contrast, the median expected *log* lengths were 4.5, 5.9 and 5.1.) Based on similar evaluations [80] for various sample sizes, it seems that if one can tolerate the minimum coverage probability dipping slightly below the nominal confidence level, then the large-sample interval is adequate. This interval has defects and can be improved. For instance, if one $n_{ij} = 0$, it covers the entire parameter space but it is more appropriate then to provide a lower or an upper bound. Such evaluations show, however, that in highly discrete problems the choice between exact and approximate methods is not necessarily obvious.

Another example where small n has a severe impact of discreteness is interval estimation for a binomial parameter, π. The most commonly cited exact method, the Clopper–Pearson interval [81], is based on inverting two single-tailed binomial tests. For nominal 95 per cent confidence based on binomial outcome x with index n and sample proportion $\hat{\pi} = x/n$, the endpoints are the values of π that satisfy

$$\sum_{k=x}^{n} \binom{n}{k} \pi^k (1-\pi)^{n-k} = 0.025 \qquad \text{and} \qquad \sum_{k=0}^{x} \binom{n}{k} \pi^k (1-\pi)^{n-k} = 0.025$$

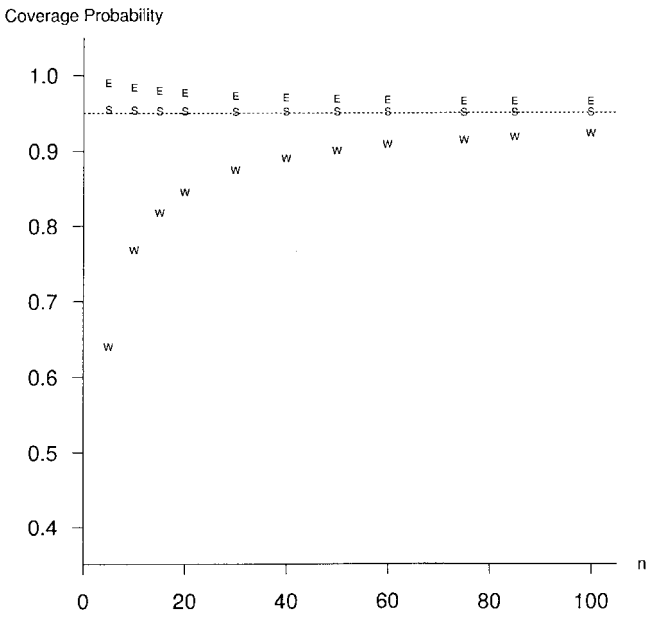

Figure 1. Mean coverage probability as a function of sample size for the nominal 95 per cent exact (E), score (S) and Wald (W) confidence intervals.

except that the lower bound is 0 when $x=0$ and the upper bound is 1 when $x=n$. Large-sample intervals utilize the approximate normality of $\hat{\pi}$. A good 95 per cent interval has endpoints that are the values of π that satisfy $1.96=(\hat{\pi}-\pi)/\sqrt{\{\pi(1-\pi)/n\}}$. The test inverted for this interval is the *score test*, which uses the null standard error. A much poorer interval results from inverting the Wald test (W) with estimated standard error (that is, giving $\hat{\pi}\pm 1.96\sqrt{\{\hat{\pi}(1-\hat{\pi})/n\}}$, which is the method commonly taught in elementary statistics textbooks).

Let $C_n(\pi)$ be the actual coverage probability, for a binomial with n trials and parameter π. The exact method guarantees $\inf_\pi C_n(\pi)\geqslant 0.95$, but $C_n(\pi)$ may seriously exceed 0.95 for most π. For large-sample methods, $\inf_\pi C_n(\pi)\geqslant 0.95$ is not true even as $n\to\infty$. Here, we compare methods by plotting in Figure 1, as a function of n, the mean $\int_0^1 C_n(\pi)\,\mathrm{d}\pi$ of the coverage probabilities for the Clopper–Pearson exact interval (E), the interval based on the score test (S), and the interval based on the Wald test (W). This figure shows the poor behaviour of the large-sample Wald interval but the slow disappearance of conservatism for the exact one as n increases. According to this criterion, the large-sample score interval performs well even for small n.

To be fair, reporting only the mean coverage does not reveal how poorly large-sample intervals can behave. For instance, the score interval has narrow regions for π near 0 and near 1 where $C_n(\pi)$ is poor (about 0.84) even for large n, but, the coverage probability for that method is closer to 0.95 than for the exact method over more than 90 per cent of the parameter space, for a variety of sample sizes [82]. Better methods exist of forming exact

intervals for a proportion than the Clopper–Pearson method [83], but the conservativeness issue persists.

Most analyses with categorical data are more complex than these single-parameter problems just discussed, typically referring to a multi-way table with many nuisance parameters. In such cases, the degree of discreteness is often negligible and the adequacy of large-sample methods is highly questionnable, so an exact method is preferable. An example is testing fit of several degree-of-freedom log-linear or logit models using chi-squared statistics with sparse data. Exact methods are also preferred when it is necessary to ensure the nominal size as a lower bound, as it might be in a drug approval process or a lawsuit. Thus, regardless of the occasional problems caused by discreteness, there is even then an important niche for exact methods.

7. A COMPROMISE – METHODS USING THE MID-*P*-VALUE

For highly discrete data when large-sample methods are questionable but exact methods may be overly conservative, one could alternatively use adjustments of exact methods based on the mid-*P*-value [84]. In a test, the mid-*P*-value adds *half* (rather than *all*) of the observed probability to the more extreme probabilities. Similarly, in constructing confidence intervals, one inverts the test using the mid-*P*-value [85]. Although no longer guaranteed to have error probabilities not exceeding the nominal level, this method usually comes closer than the exact method to the desired level. Table I shows its performance as the basis of confidence intervals for the odds ratio. Numerical evaluations [8, 9, 70, 85–91] for a variety of cases show that it usually has coverage probability slightly exceeding the nominal value, but it tends to be less conservative than ordinary exact methods. It has the advantage, compared to large-sample methods, that it is guaranteed to work well (being 'nearly exact') as the degree of discreteness diminishes.

Inference based on the mid-*P*-value seems to be a sensible compromise between the conservativeness of exact methods and the uncertain adequacy of large-sample methods. It has some appealing properties. Its null expected value is 0.5, as is true for the *P*-value when test statistics have a continuous distribution. It takes the *P*-value for a test with supplementary randomization, which is the probability of a test statistic more extreme than observed plus a uniform (0,1) random variable multiplied by the probability of the observed value of the statistic, and replaces the uniform multiple by its expected value. Recent research showed that it is an optimal *P*-value in terms of estimating a truth indicator of the null hypothesis [92].

Similar benefits can accrue from alternative proposed *P*-values, but these do not have the simplicity of the mid-*P*-value. One approach, useful when several tables in the conditional sample space have a particular value for a test statistic, uses a secondary, finer partitioning of that space; for tables having the observed value of the test statistic, only those contribute to the *P*-value that are at least as contradictory to the null in terms of the secondary statistic [49, 51]. Methods using ordinary *P*-values obtained with 'approximate conditioning' techniques may yield similar performance [93].

Some undoubtedly would point out that another solution is to adapt a Bayesian perspective. For Bayesians the discreteness issue is not a problem; the posterior distribution for the parameter is continuous when the prior distribution is.

8. CONCLUSIONS

The effect of discreteness on exact methods raises a dilemma for the practising statistician. In interval estimation, for instance, is it better to use an approach that guarantees that the coverage probabilities are *at least* 0.95 yet may have actual coverage probabilities of about 0.97 or 0.98 (such as an exact interval when substantial discreteness occurs), or an approach giving narrower intervals for which the actual coverage probability could be less than 0.95 but is usually *close* to 0.95? Traditionally statisticians evaluate interval estimators by the infimum of their coverage probabilities. Rather than insisting that the coverage probability be *at least* 0.95 over the entire parameter space, perhaps it is sufficient for the coverage probability to *average* at least 0.95 so long as it rarely falls more than some substantive amount (say 0.02) below that level. Acceptance of this weaker criterion suggests adaptations of exact methods using the mid-*P*-value, and in some cases (for example, inference for proportions and odds ratios) to use of certain large-sample methods.

As pointed out above, most criticisms of exact methods disappear as discreteness does, and the advances of recent years are welcome. The complications due to discreteness point out, though, that the choice of a statistical method is not always obvious. However, if statisticians cannot agree how to analyse even a 2×2 table, with no one approach being obviously 'best', what hope is there for a consensus on more complex analyses? The probability is probably decreasing to 0 with time, but we should keep in mind that in most cases different methods provide the same substantive conclusions. Distinctions such as (exact conditional/exact unconditional) and (Bayesian/frequentist) that may seem fundamentally important to us do not seem all that striking to an outsider.

Finally, since submission of this article, the author has become aware of other articles of interest for exact inference. These include work on a new and relatively simple approach for interval estimation of a binomial proportion [94], improved confidence intervals for the difference between two proportions [95,96], a review of methods for comparing proportions [97], and a comparison of ways of forming confidence intervals in discrete data problems [98].

ACKNOWLEDGEMENTS

This work was partially supported by grants from NIH and NSF.

REFERENCES

1. Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Statistics in Medicine* 1995; **14**: 2143–2160.
2. Agresti A. A survey of exact inference for contingency tables (disc: P153–177). *Statistical Science* 1992; **7**:131–153.
3. Cytel Software. *StatXact. A Statistical Package for Exact Nonparametric Inference* (*version* 4.0). Cytel Software: Cambridge, MA, 1998.
4. Mehta CR. The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research* 1994; **3**:135–156.
5. Mehta CR. Exact inference for categorical data. *Encyclopedia of Biostatistics* 1998; **2**:1411–1422.
6. Haberman SJ. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association* 1988; **83**:555–560.
7. Cornfield J. A statistical problem arising from retrospective studies. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1956; **4**:135–148.
8. Vollset SE, Hirji KF, Afifi AA. Evaluation of exact and asymptotic interval estimators in logistic analysis of matched case-control studies. *Biometrics* 1991; **47**:1311–1325.

9. Hirji KF. A comparison of exact, mid-*p*, and score tests for matched case-control studies. *Biometrics* 1991; **47**:487–496.
10. Hirji K. A review and a synthesis of the fast Fourier transform algorithms for exact analysis of discrete data. *Computational Statistics and Data Analysis* 1997; **25**:321–336.
11. Agresti A, Wackerly D, Boyett JM. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* 1979; **44**:75–84.
12. Mehta CR, Patel NR, Senchaudhuri P. Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association* 1988; **83**:999–1005.
13. Senchaudhuri P, Mehta CR, Patel NR. Estimating exact *p*-values by the method of control variates or Monte Carlo rescue. *Journal of the American Statistical Association* 1995; **90**:640–648.
14. Booth J, Butler R. Monte Carlo approximation of exact conditional tests for log-linear models. *Biometrika* 1999; **86**:321–332.
15. Barnard GA. A new test for 2×2 tables. *Nature* 1945; **156**:177.
16. Suissa S, Shuster JJ. Are uniformly most powerful unbiased tests really best? *American Statistician* 1984; **38**:204–206.
17. Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society*, *Series A*, *General* 1985; **148**:317–327.
18. Suissa S, Shuster JJ. The 2×2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* 1991; **47**:361–372.
19. Fisher RA. A new test for 2×2 tables (letter to the editor). *Nature* 1945; **156**:388.
20. Little RJA. Testing the equality of two independent binomial proportions. *American Statistician* 1989; **43**: 283–288.
21. Routledge RD. Resolving the conflict over Fisher's exact test. *Canadian Journal of Statistics* 1992; **20**:201–209.
22. Greenland S. On the logical justification of conditional tests for two-by-two contingency tables. *American Statistician* 1991; **45**:248–251.
23. Upton GJG. Fisher's exact test. *Journal of the Royal Statistical Society*, *Series A*, *General* 1992; **155**: 395–402.
24. Martín Andrés A, Tejedor IH. Is Fisher's exact test very conservative? *Computational Statistics and Data Analysis* 1995; **19**:579–591.
25. Reid N. The roles of conditioning in inference (disc: P173–199). *Statistical Science* 1995; **10**:138–157.
26. Howard JV. The 2×2 table: a discussion from a Bayesian viewpoint. *Statistical Science* 1998; **13**:351–367.
27. Cormack RS, Mantel N. Fisher's exact test: the marginal totals as seen from two different angles. *Statistician* 1991; **40**:27–34.
28. Yates F. Tests of significance for 2×2 contingency tables (with discussion). *Journal of the Royal Statistical Society*, *Series A*, *General* 1984; **147**:426–463.
29. Zelen M. The analysis of several 2×2 contingency tables. *Biometrika* 1971; **58**:129–137.
30. Hirji KF, Vollset SE. [algorithm AS 293] Computing exact distributions for several ordered $2 \times K$ tables. *Applied Statistics* 1994; **43**:541–548.
31. Mehta CR, Patel N, Senchaudhuri P. Exact stratified linear rank tests for ordered categorical and binary data. *Journal of Computational and Graphical Statistics* 1992; **1**:21–40.
32. Kim D, Agresti A. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Computational Statistics and Data Analysis* 1997; **24**:89–104.
33. Agresti A, Coull B. Order-restricted inference for monotone trend alternatives in contingency tables. *Computational Statistics and Data Analysis* 1998; **28**:139–155.
34. Agresti A, Coull BA. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* 1996; **52**:1103–1111.
35. Smith PWF, Forster JJ, McDonald JW. Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society*, *Series A*, *General* 1996; **159**:309–321.
36. McDonald JW, Smith PWF. Exact conditional tests of quasi-independence for triangular contingency tables: estimating attained significance levels. *Applied Statistics* 1995; **44**:143–151.
37. Smith PWF, McDonald JW. Exact conditional tests for incomplete contingency tables: estimating attained significance levels. *Statistics and Computing* 1995; **5**:253–256.
38. Smith PWF, McDonald JW. Simulate and reject Monte Carlo exact conditional tests for quasi-independence. *COMPSTAT. Proceedings in Computational Statistics*, *11th Symposium* 1994; 509–514.
39. Smith PWF, McDonald JW, Forster JJ, Berrington AM. Monte Carlo exact methods used for analysing interethnic unions in Great Britain. *Applied Statistics* 1996; **45**:191–202.
40. McDonald JW, De Roure DC, Michaelides DT. Exact tests for two-way symmetric contingency tables. *Statistics and Computing* 1998; **8**:391–399.
41. Forster JJ, McDonald JW, Smith PWF. Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1996; **58**:445–453.
42. McDonald JW, Smith PWF, Forster JJ. Exact tests of goodness of fit of log-linear models for rates. *Biometrics* 1999; **55**:620–624.

43. Yao Q, Tritchler D. An exact analysis of conditional independence in several 2×2 contingency tables. *Biometrics* 1993; **49**:233–236.
44. Gastwirth JL, Mehta CR. The usefulness of exact statistical methods in equal employment litigation. *Computational Statistics. Proceedings of 10th Symposium on Computational Statistics* 1992; **2**:91–95.
45. Hirji KF, Tang M-L, Vollset SE, Elashoff RM. Efficient power computation for exact and mid-p tests for the common odds ratio in several 2×2 tables. *Statistics in Medicine* 1994; **13**:1539–1549.
46. Hirji KF, Vollset SE, Reis IM, Afifi AA. Exact tests for interaction in several 2×2 tables. *Journal of Computational and Graphical Statistics* 1996; **5**:209–224.
47. Vollset SE, Hirji KF, Elashoff RM. Fast computation of exact confidence limits for the common odds ratio in a series of 2×2 tables. *Journal of the American Statistical Association* 1991; **86**:404–409.
48. Reis I, Hirji K, Afifi A. Exact and asymptotic tests for homogeneity in several 2×2 tables. *Statistics in Medicine* 1999; **18**:893–906.
49. Kim D, Agresti A. Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* 1995; **90**:632–639.
50. Agresti A, Mehta CR, Patel NR. Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association* 1990; **85**:453–458.
51. Cohen A, Sackrowitz HB. An evaluation of some tests of trend in contingency tables. *Journal of the American Statistical Association* 1992; **87**:470–475.
52. Berger V. Admissibility of exact conditional tests of stochastic order. *Journal of Statistical Planning and Inference* 1998; **66**:39–50.
53. Berger VW, Permutt T, Ivanova A. Convex hull test for ordered categorical data. *Biometrics* 1998; **54**: 1541–1550.
54. Morgan WM, Blumenstein BA. Exact conditional tests for hierarchical models in multidimensional contingency tables. *Applied Statistics* 1991; **40**:435–442.
55. Zelterman D, Chan IS-F, Mielke PWJ. Exact tests of significance in higher dimensional tables. *American Statistician* 1995; **49**:357–361.
56. Hirji KF. Computing exact distributions for polytomous response data. *Journal of the American Statistical Association* 1992; **87**:487–492.
57. Fu YX, Arnold J. A table of exact sample sizes for use with Fisher's exact test for 2×2 tables. *Biometrics* 1992; **48**:1103–1112.
58. Hilton JF, Mehta CR. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* 1993; **49**:609–616.
59. Thomas RG, Conlon M. Sample size determination based on Fisher's exact test for use in 2×2 comparative trials with low event rates. *Controlled Clinical Trials* 1992; **13**:134–147.
60. Royston P. Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide. *Statistics in Medicine* 1993; **12**:699–712.
61. Lui K-J. Sample size for the exact conditional test under inverse sampling. *Statistics in Medicine* 1996; **15**: 671–678.
62. Mehta CR, Patel NR, Senchaudhuri P. Exact power and sample-size calculations for the Cochran-Armitage trend test. *Biometrics* 1998; **54**:1615–1621.
63. Gordon I. Sample size for two independent proportions: A review: *Australian Journal of Statistics* 1994; **36**:199–209.
64. Hirji KF, Johnson TD. A comparison of algorithms for exact analysis of unordered $2 \times K$ contingency tables. *Computational Statistics and Data Analysis* 1996; **21**:419–429.
65. Baglivo J, Olivier D, Pagano M. Methods for exact goodness-of-fit tests. *Journal of the American Statistical Association* 1992; **87**:464–469.
66. Baglivo J, Pagano M, Spino C. Permutation distributions via generating functions with applications to sensitivity analysis of discrete data. *Journal of the American Statistical Association* 1996; **91**:1037–1046.
67. Baglivo J, Olivier D, Pagano M. Analysis of discrete data: rerandomization methods and complexity. *Computational Statistics and Data Analysis* 1993; **16**:175–184.
68. Mudholkar G, Hutson A. Continuity corrected approximations for and 'exact' inference with Pearson's X^2. *Journal of Statistical Planning and Inference* 1997; **59**:61–78.
69. Pierce DA, Peters D. Practical use of higher order asymptotics for multiparameter exponential families (disc: P725–737). *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1992; **54**:701–725.
70. Agresti A, Lang JB, Mehta C. Some empirical comparisons of exact, modified exact, and higher-order asymptotic tests of independence for ordered categorical variables. *Communications in Statistics*, *Part B – Simulation and Computation* 1993; **22**:1–18.
71. Kolassa JE, Tanner MA. Approximate conditional inference in exponential families via the Gibbs sampler. *Journal of the American Statistical Association* 1994; **89**:697–702.
72. Strawderman RL, Wells MT. Approximately exact inference for the common odds ratio in several 2×2 tables (with discussion). *Journal of the American Statistical Association* 1998; **93**:1294–1320.

73. Bedrick EJ, Hill JR. An empirical assessment of saddlepoint approximations for testing a logistic regression parameter. *Biometrics* 1992; **48**:529–544.
74. Kolassa J, Tanner M. Approximate Monte Carlo conditional inference in exponential families. *Biometrics* 1999; **55**:246–251.
75. Presnell B. Bootstrap unconditional p-values for the sign test with ties and the 2×2. *Journal of Nonparametric Statistics* 1996; **7**:47–55.
76. Berger RL, Boos DD. p-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**:1012–1016.
77. Freidlin B, Gastwirth JL. Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics* 1999; **55**:264–267.
78. Mehta CR, Hilton JF. Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. *American Statistician* 1993; **47**:91–98.
79. Neyman J. On the problem of confidence limits. *Annals of Mathematical Statistics* 1935; **6**:111–116.
80. Agresti A. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 1999; **55**:597–602.
81. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404–413.
82. Agresti A, Coull BC. Approximate is better than 'exact' for interval estimation of binomial parameters. *American Statistician* 1998; **52**:119–126.
83. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**:108–116.
84. Lancaster HO. Significance test in discrete distributions (corrections **57**:919). *Journal of the American Statistical Association* 1961; **56**:223–234.
85. Berry G, Armitage P. Mid-p confidence intervals: A brief review. *Statistician* 1995; **44**:417–423.
86. Vollset SE. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993; **12**:809–824.
87. Mehta CR, Walsh SJ. Comparison of exact, mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *American Statistician* 1992; **46**:146–150.
88. Hirji KF, Tan S-J, Elashoff RM. A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine* 1991; **10**:1137–1153.
89. Cohen GR, Yang S-Y. Mid-p confidence intervals for the Poisson expectation. *Statistics in Medicine* 1994; **13**:2189–2203.
90. Routledge RD. Practicing safe statistics with the mid-p^*. *Canadian Journal of Statistics* 1994; **22**:103–110.
91. Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
92. Hwang JTG, Yang M-C. Evaluate the p-values for testing the independence in 2×2 contingency tables using the estimated truth approach – one way to resolve the controversy relating to fisher's exact test: *Statistica Sinica* 2001; to appear.
93. Pierce DA, Peters D. Improving on exact tests by approximate conditioning. *Biometrika* 1999; **86**:265–277.
94. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **86**:783–798.
95. Coe PR, Tamhane AC. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics, Part B – Simulation and Computation* 1993; **22**:925–938.
96. Santer TJ, Yamagami S. Invariant small sample cofidence-intervals for difference of 2 success probabilities. *Communications in Statistics, Part B – Simulation and Computation* 1993; **22**:33–59.
97. Martin Andrés A. A review of classic non-asymptotic methods for comparing two proportions by means of independent samples. *Communications in Statistics, Part B – Simulation and Computation* 1991; **20**:551–583.
98. Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001;**57**: to appear.

