Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2006/01

Reducing Conservatism of Exact Small-Sample Methods of Inference for Discrete Data

Alan Agresti, Anna Gottard

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Reducing Conservatism of Exact Small-Sample Methods of Inference for Discrete Data

Alan Agresti[1] and Anna Gottard[2]

[1] Department of Statistics, University of Florida, U.S.A. aa@stat.ufl.edu
[2] Department of Statistics, University of Florence, Italy. gottard@ds.unifi.it

1 Introduction

In recent years, considerable attention has been paid to ways of conducting exact small-sample inference for discrete data. Most of this has been in the context of the analysis of contingency tables. These methods use distributions determined exactly rather than as large-sample approximations. To achieve exactness, most common is a conditional inference approach whereby one focuses on the parameter of interest while eliminating nuisance parameters by conditioning on their sufficient statistics. For 2×2 tables, there is also some literature on an unconditional approach.

Software is now readily available for small-sample methods. Best known and most complete are StatXact for contingency table methods and LogXact for logistic regression, both marketed by Cytel Inc. (Cytel 2005). Although many statisticians are aware only of Fisher's exact conditional test for 2×2 tables, there is now a wide variety of methods available in such software. These include unconditional methods for comparing binomial proportions with tests and confidence intervals, inferences for $r \times c$ tables, inferences for stratified tables including tests of conditional independence and homogeneity of association, inferences for dependent samples and for clustered data, inferences about measures of association and measures of agreement, and inferences about parameters in logistic regression models and some of their multinomial extensions.

StatXact and LogXact utilize network algorithms. For any algorithm, computations become increasingly intensive as the sample size increases. The StatXact 7 manual (Cytel 2005, p. 13) notes that with current capabilities, almost all exact tests can be executed within a few seconds when the sample size does not exceed about 30. Even for a relatively small sample size, however, the number of contingency tables that contribute to an analysis can be huge when the number of categories is moderate. For example, the StatXact 7 manual (Cytel 2005, p. 12) notes that a 5×6 table with row margins (7, 7, 12, 4, 4) and column margins (4, 5, 6, 5, 7, 7) has a reference set of 1.6

billion contingency tables that have the same margins and contribute to exact conditional tests. For cases that are infeasible or that take a long time, fast and precise Monte Carlo approximations are available.

The terminology "exact" refers to the use of exactly determined, small-sample distributions, rather than normal or chi-squared approximations, to obtain P-values and confidence intervals. However, the inferences are *not* exact in the sense that error probabilities exactly equal the nominal values. Rather, the nominal values are upper bounds for the true error probabilities. This is well known for significance tests. For example, suppose a test of a simple hypothesis H_0 has nominal size 0.05, in the sense that H_0 is rejected when the P-value is no greater than 0.05. If the possible P-values for the exact discrete, small-sample distribution are 0.02, 0.06, 0.12, ..., then the actual size is 0.02.

The same phenomenon is true for confidence intervals. Consider intervals constructed by inverting a test (e.g., a 95% confidence interval consists of the set of parameter values not rejected at the 0.05 significance level in the family of tests). Inverting a test that has actual size no greater than 0.05 for each possible parameter value results in a confidence interval having coverage probability at least equal to 0.95. The actual coverage probability varies according to the parameter value, and so in practice it is unknown. Thus, conservatism of exact tests propagates to conservatism of exact confidence intervals. In fact, the situation is worse in the sense that one does not know the actual error probability, but merely its upper bound. See Agresti (2001) for a review and a discussion of issues that make exact inference awkward for discrete data.

Section 2 reviews small-sample inference for discrete exponential-family distributions and illustrates with the binomial. Section 3 surveys ways to reduce the conservatism. In theory, discreteness is not a problem if one uses supplementary randomization to achieve the desired error probability exactly. Section 3 also reviews this approach, which was fashionable for a time around 1950. Section 4 discusses a related approach for discrete data proposed by Geyer and Meeden (2005), fuzzy inference, which yields exactly the desired error rate. We then present a simpler way of conducting fuzzy inference for discrete exponential family distributions.

The randomized and fuzzy inference approaches have connections with inference based on the mid-P value. Section 5 reviews this approach and evaluates its performance for inference about a binomial parameter. We conclude that inference based on the mid-P value provides a sensible compromise that mitigates the effects of conservatism of exact methods yet is more useful in practice than randomized or fuzzy inference.

2 Small-Sample Inference for Discrete Distributions

Exact inference about a parameter θ requires the actual error probability to be no greater than the nominal level, which we denote by α. For a significance test of a hypothesis H_0, the actual size is no greater than α. That is, the P-value satisfies

$$P_\theta(\text{P-value} \leq \alpha | H_0) \leq \alpha$$

for all α and for all θ in H_0. For a confidence interval, the actual coverage probability must be at least $1 - \alpha$ for all possible values of θ.

Let T be a discrete test statistic with probability mass function $f(t|\theta)$ and cumulative distribution function $F(t|\theta)$ indexed by the parameter θ. For each value θ_0 of θ let $A(\theta_0)$ denote the acceptance region for testing H_0: $\theta = \theta_0$. This is the set of values t of T for which the P-value exceeds α. Then, for each t, let $C(t) = \{\theta_0 : t \in A(\theta_0)\}$. The set of $\{C(t)\}$ for various t are the confidence regions with the desired property. In other words, having acceptance regions such that

$$P_{\theta_0}[T \in A(\theta_0)] \geq 1 - \alpha$$

for all θ_0 guarantees that the confidence level for $\{C(t)\}$ is at least $1 - \alpha$. For a typical θ_0, one cannot form $A(\theta_0)$ to achieve probability of Type I error exactly equal to α, because of discreteness. Hence, such significance tests and confidence intervals are conservative. The actual coverage probability of $C(T)$ varies for different values of θ but is bounded below by $1 - \alpha$ (Neyman 1935). In technical terms, the bound results from the distribution of $F(T|\theta)$ being stochastically larger than uniform when T is discrete (Casella and Berger 2001, pp. 77, 434).

2.1 One-parameter exponential families

In this article we will assume that the observations $x_1, x_2, ..., x_n$ are independent from a single-parameter exponential family distribution with probability mass function,

$$f(x|\theta) = h(x)c(\theta)\exp[w(\theta)t(x)].$$

The minimal sufficient (and complete) statistic is $T = \sum_i t(x_i)$. Let $F_T(t|\theta) = P(T \leq t|\theta)$. Below for specificity we discuss one-sided inference in terms of a significance test and two-sided inference in terms of confidence intervals.

Standard results found in statistical theory texts such as Casella and Berger (2001) include the following: If $w(\theta)$ is nondecreasing, the family of distributions has monotone likelihood ratio. This is true in the standard cases, and we'll assume it below. Then, for testing H_0: $\theta \leq \theta_0$ against H_a: $\theta > \theta_0$, for any t, the test that rejects H_0 if and only if $T \geq t$ is a uniformly most powerful (UMP) test of size $\alpha = P_{\theta_0}(T \geq t)$. With observed test statistic value t_{obs}, the P-value for the test is $P_{\theta_0}(T \geq t_{obs})$. If $F_T(t|\theta)$ is a decreasing function of θ for each t (which is true when there is monotone likelihood ratio), and if

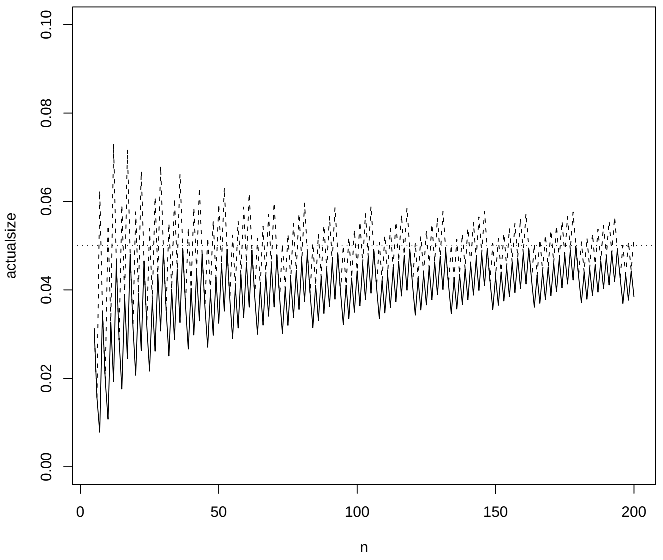
$$P(T \leq t|\theta_U(t)) = \alpha/2, \ P(T \geq t|\theta_L(t)) = \alpha/2, \tag{1}$$

then $[\theta_L(T), \theta_U(T)]$ is a $100(1 - \alpha)\%$ confidence interval for θ. That is, it has probability at least $1 - \alpha$ of containing θ. This method of forming a confidence interval is often called the *tail method.*

2.2 Illustration for the binomial distribution

For n independent, identically distributed Bernoulli observations with parameter θ, T is the "number of successes" and has binomial distribution with index n and parameter θ. To test $H_0 : \theta \leq \theta_0$ against $H_a : \theta > \theta_0$, the UMP test rejects for sufficiently large values of T.

Fig. 1. Actual sizes of exact (—) and mid-P (- - -) binomial tests of $H_0 : \theta \leq 0.50$ against $H_a : \theta > 0.50$, plotted as a function of n between 5 and 200



For the case $\theta_0 = 0.50$, which is most common in practice, Figure 1 shows the actual size of a nominal size $\alpha = 0.05$ test, plotted as a function of n for n between 5 and 200. The conservatism is quite marked for small n, which is precisely when one would not want to rely on large-sample asymptotics, but it persists even for moderately large n.

In a standard application of the above confidence interval theory, Clopper and Pearson (1934) proposed the following 100(1 - α)% confidence interval for the binomial parameter: The endpoints (θ_L, θ_U) satisfy

$$\sum_{k=t_{obs}}^{n} \binom{n}{k} \theta_L^k (1-\theta_L)^{n-k} = \alpha/2 \text{ and } \sum_{k=0}^{t_{obs}} \binom{n}{k} \theta_U^k (1-\theta_U)^{n-k} = \alpha/2,$$

except that $\theta_L = 0$ when $t_{obs} = 0$ and $\theta_U = 1$ when $t_{obs} = n$. This confidence interval is based on inverting two one-sided UMP binomial tests.

For instance, the 95% confidence interval when $x = 5$ in $n = 5$ trials is (0.478, 1.000). This means that θ_0 must be below 0.478 in order for the binomial right-tail probability in testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$

to fall below 0.025. In fact, when $n = 5$ this exact 95% confidence interval contains 0.50 for *every* value of x. Thus, the actual coverage probability of this exact interval when $\theta = 0.50$ is 1.0, not 0.95.

Various evaluations have shown that the Clopper–Pearson confidence interval tends to be extremely conservative for small to moderate n. See, for instance, Newcombe (1998), Agresti and Coull (1998), and Brown et al. (2001). When $t_{obs} = 0$, it equals $[0, 1 - (\alpha/2)^{1/n}]$. The actual coverage probability necessarily exceeds $1 - \alpha/2$ for θ below $1 - (\alpha/2)^{1/n}$ and above $(\alpha/2)^{1/n}$. This is the entire parameter space when $n \le \log(\alpha/2)/\log(.5)$, for instance $n \le 5$ for $\alpha = 0.05$.

Fig. 2. Actual coverage probabilities of Clopper–Pearson (—) and mid-P (- - -) confidence intervals for binomial parameter θ, plotted for n between 5 and 200 when $\theta = 0.50$

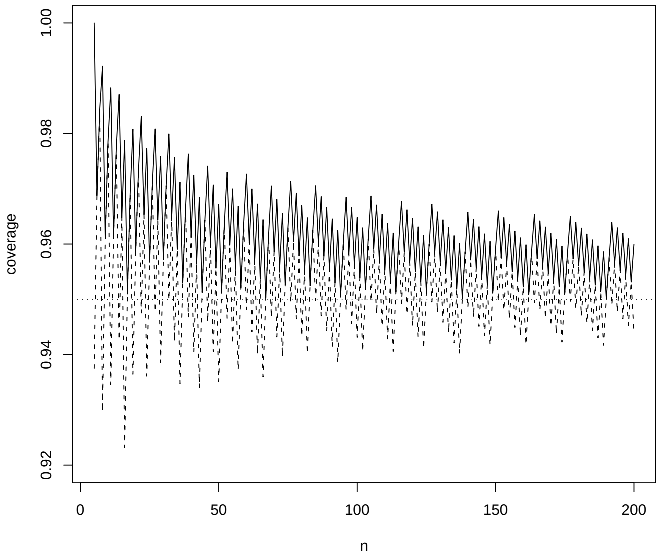


Figure 2 plots the actual probability of coverage of the 95% Clopper–Pearson confidence interval, as a function of n, when the actual parameter value is 0.50. Again, the degree of conservatism is quite severe, even when n is moderately large.

3 Ways of Reducing Conservatism

This section mentions some ways that have been proposed of reducing the degree of conservatism of exact, small-sample inference. We'll illustrate these for the case of two-sided interval estimation of the binomial parameter.

3.1 Confidence intervals not based on the tail method

Inverting a family of tests corresponds to forming the confidence region from the set of θ_0 for which the test's P-value exceeds α. The tail method (1) requires the stronger condition that the probability be no greater than $\alpha/2$ that T falls below $A(\theta_0)$ and no greater than $\alpha/2$ that T falls above $A(\theta_0)$. The interval for this method is the set of θ_0 for which each one-sided P-value exceeds $\alpha/2$. One disadvantage of the tail method is that for sufficiently small and sufficiently large θ, the lower bound on the coverage probability is actually $1-\alpha/2$ rather than $1-\alpha$. For sufficiently small θ, for instance, the interval can never exclude θ by falling below it.

Alternatives to the tail method exist for which intervals tend to be shorter and coverage probabilities tend to be closer to the nominal level. One approach inverts a single two-sided test instead of two equal-tail one-sided tests. For instance, a possible two-sided P-value is $\min[P_{\theta_0}(T \geq t_{obs}), P_{\theta_0}(T \leq t_{obs})]$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. The confidence intervals based on inverting such a test necessarily are contained in confidence intervals obtained with the tail method. Blaker (2000) used this approach for the binomial parameter and gave S-plus functions for implementing it. See Agresti (2003) for an example of the improvement this provides over the Clopper–Pearson method.

Another two-sided approach forms the acceptance region $A(\theta_0)$ by entering the test statistic values t in $A(\theta_0)$ in order of their null probabilities, starting with the highest, stopping when the total probability is at least $1-\alpha$; that is, $A(\theta_0)$ contains the smallest possible number of most likely outcomes (under $\theta=\theta_0$). In its crudest partitioning of the sample space, the corresponding P-value is the sum of null probabilities that are no greater than the probability of the observed result. When inverted to form confidence intervals, this approach satisfies the optimality criterion of minimizing total length. Sterne (1954) proposed this approach for interval estimation of a binomial proportion.

Yet another way to invert a two-sided test orders points for the acceptance region and forms P-values according to a statistic that describes the distance of the observed data from H_0. One could use a statistic T based on a standard large-sample criterion, such as the likelihood-ratio statistic, the Wald statistic, or the score statistic.

These various two-sided approaches do not have the tail method disadvantage of a lower bound of $1-\alpha/2$ for the coverage probability over part of the parameter space. However, some methodologists find discomforting the lack of information about how each tail contributes to the analysis.

3.2 Confidence intervals based on less discrete statistics or P-values

In constructing a test or a confidence interval based on a test, the test statistic should not be any more discrete than necessary. For instance, a sample proportion of $\hat{\theta} = 0.40$ gives less evidence in testing $H_0 : \theta = 0.50$ than in testing $H_0 : \theta = 0.30$, because the null standard error is smaller in the second case. It is better to base tests and subsequent confidence intervals on a standardization, such as by dividing the difference between the sample proportion and its null value by the null standard error, or the relative likelihood values.

Likewise, it is sometimes possible to reduce conservativeness by using a less discrete form of P-value. For instance, instead of including the probabilities of all relevant samples having $T = t_{obs}$ in the P-value, Kim and Agresti (1995) included only probabilities of those samples that are no more likely to occur than the observed one. For an example of estimating a common odds ratio in 18 2×2 tables for which the tail method gave a 95% confidence interval of (0.05, 1.16), the interval based on this less discrete P-value was (0.09, 0.99).

3.3 Confidence intervals based on an unconditional approach with nuisance parameters

For comparing parameters from two discrete distributions, the conditional approach eliminates nuisance parameters by conditioning on their sufficient statistics. This approach, however, increases the degree of discreteness. Moreover, it is limited to the natural parameter for exponential family distributions.

An alternative approach to eliminating the nuisance parameter is unconditional. For a nuisance parameter ψ, let $p(\theta_0; \psi)$ denote the P-value for testing $H_0 : \theta = \theta_0$ for a given value of ψ. The unconditional approach takes P-value $= \sup_{\psi} p(\theta_0; \psi)$. This is a legitimate P-value (Casella and Berger 2001, p. 397). If $p(\theta_0; \psi)$ is relatively stable in ψ, this method has the potential to improve on conditional methods. See, for instance, Suissa and Shuster (1985), who showed improvement in power over Fisher's exact test for testing equality of two independent binomials. Agresti and Min (2001) used the unconditional approach to form a confidence interval for the difference of proportions, based on inverting the score test. Agresti and Min (2002) used the unconditional approach for interval estimation of the odds ratio.

3.4 Randomized tests and confidence intervals

In the statistical theory of hypothesis testing, for discrete problems one can achieve the exact size by randomizing appropriately on the boundary of the critical region (e.g., Lehmann 1986, p. 71-76). One uses a critical function $\phi(t)$ for the probability of rejecting the null hypothesis. It equals 1.0 for t in the interior of the rejection region, 0.0 outside that region, and a value between 0 and 1 on the boundary of the rejection region determined so that the size equals the desired value. For testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$ for an exponential family with test statistic T and observed value t_{obs}, this corresponds to using P-value

$$P_{\theta_0}(T > t_{obs}) + \mathcal{U} \times P_{\theta_0}(T = t_{obs}) \tag{2}$$

where U is a uniform(0,1) random variable (Cox and Hinkley 1974, p. 101).

To construct a confidence interval that achieves exactly (a priori) probability $(1-\alpha)$ of covering the unknown parameter value, one can invert two such randomized tests. The upper and lower endpoints of the confidence interval are the solutions to the equations

$$P_{\theta_U}(T < t_{obs}) + \mathcal{U} \times P_{\theta_U}(T = t_{obs}) = \alpha/2 \tag{3}$$

and

$$P_{\theta_L}(T > t_{obs}) + (1 - \mathcal{U}) \times P_{\theta_L}(T = t_{obs}) = \alpha/2. \tag{4}$$

This was suggested by Stevens (1950) for the binomial parameter, but the same argument works for other exponential family distributions. This confidence interval inverts tests for which (as in the case of continuous random variables) the one-sided P-values sum to 1 and each have a uniform null distribution, unlike the ordinary one-sided P-values used in the tail-method confidence interval.

In order to achieve the nominal size exactly, a randomized confidence interval must have some counterintuitive behavior at the boundary T values. When T takes its minimum possible value, the lower bound exceeds the smallest parameter value when $\mathcal{U} > 1 - \alpha/2$; when T takes its maximum possible value, the upper bound is less than the largest parameter value when $\mathcal{U} < \alpha/2$.

These days statisticians regard randomized inference as a tool for the mathematical convenience of achieving exactly the desired size or confidence level with discrete data, but in practice no one seriously considers using it. However, this method was originally thought to have considerable promise. For example, Pearson (1950) suggested that statisticians may come to accept randomization after performing an experiment just as they had gradually come to accept randomization for the experiment itself. Stevens (1950) stated "We suppose that most people will find repugnant the idea of adding yet another random element to a result which is already subject to the errors of random sampling. But what one is really doing is to eliminate one uncertainty by introducing a new one. The uncertainty which is eliminated is that of the true probability that the parameter lies within the calculated interval. It is because this uncertainty is eliminated that we no longer have to keep 'on the safe side', and can therefore reduce the width of the interval."

4 Fuzzy Inference using Discrete Data

To address the conservativism issue with randomized procedures but without the arbitrariness of actually picking a uniform random variable, Geyer and Meeden (2005) suggested using fuzzy inference. For testing $H_0 : \theta = \theta_0$ with a desired size α, they defined a fuzzy decision to be a critical function $\phi(t, \alpha, \theta_0)$ having that size, viewed as a function of the value t of the test statistic T. For given t, they regarded ϕ as a function of α and called it a *fuzzy P-value*. For fixed t and α, the function $[1 - \phi(t, \alpha, \theta)]$ is the *fuzzy confidence interval.* With T treated as a random variable (for given θ), it has unconditional coverage probability $(1 - \alpha)$. We focus on the fuzzy confidence interval here.

Geyer and Meeden defined the *core* of the fuzzy confidence interval to be the set of θ for which $[1 - \phi(t, \alpha, \theta) = 1]$. They defined the *support* to be the set of θ for which $[1 - \phi(t, \alpha, \theta) > 0]$. Given t, rather than performing the randomization, they recommended merely plotting the fuzzy confidence interval. This is a way of portraying the inference about where θ falls while guaranteeing achieving exactly the appropriate coverage probability (unconditionally).

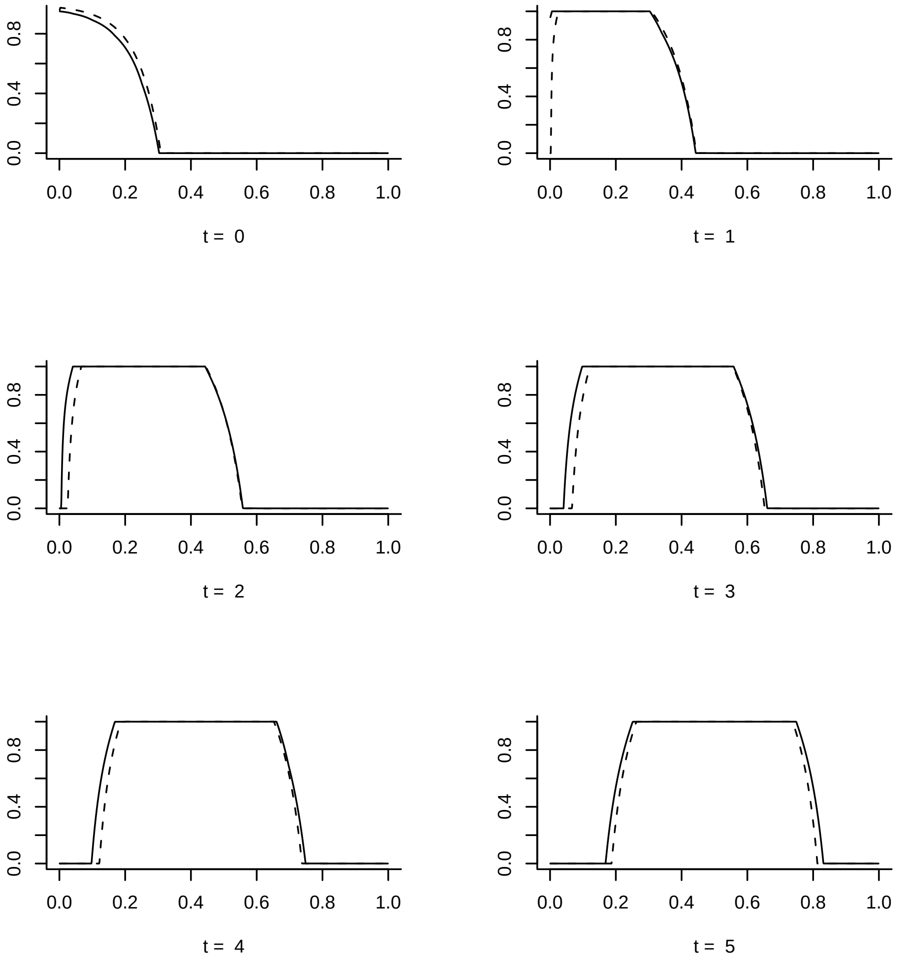
Geyer and Meeden proposed fuzzy inferences that are UMP in the one-sided case and UMPU in the two-sided case, based on standard exponential family theory. Their two-sided inference is complex to conduct. Details were not given in their article, but a companion website (http://www.stat.umn.edu/geyer/fuzz/) shows that computations are complex even for simple cases such as a single binomial parameter.

In the discussion of Geyer and Meeden (2005), Agresti and Gottard suggested a simpler way to construct two-sided fuzzy inferences directly uses the randomized tests and randomized confidence interval described in Section 3.4. We illustrate here for a fuzzy confidence interval. Consider the set of possible randomized intervals with endpoints determined by (3) and (4). As U increases from 0 to 1, the lower and upper endpoints are monotonically increasing. Substituting $\mathcal{U} = 0$ in equations (3) and (4) gives the bounds for a randomized interval having as lower bound the lower bound from the conservative confidence interval (1). Substituting $\mathcal{U} = 1$ gives the bounds for a randomized interval having as upper bound the upper bound from the conservative confidence interval (1). Thus, the support of the fuzzy confidence interval is the ordinary conservative confidence interval (e.g., the Clopper–Pearson interval for the binomial parameter). The core of the fuzzy confidence interval is the set of θ values that fall in every one of the possible randomized confidence intervals. This core goes from the lower bound of the randomized confidence interval with $\mathcal{U} = 1$ to the upper bound of the randomized confidence interval with $\mathcal{U} = 0$.

The figure for this fuzzy confidence interval is easily constructed, especially when t is not at its minimum or maximum value. Consider an arbitrary value $\mathcal{U} = u$ for the uniform random variable. The value that is the lower bound of the randomized confidence interval with $\mathcal{U} = u$ is contained only in all the randomized confidence intervals with $\mathcal{U}$ less than or equal to u. So, for the given t, the probability $1 - \phi(t, \alpha, \theta)$ of containing that value is u. So, at the value θ that is the lower bound of the randomized confidence interval with $\mathcal{U} = u$, the height of the curve to display the fuzzy confidence interval is u. Likewise, the value that is the upper bound of the randomized confidence interval with $\mathcal{U} = u$ is contained only in all the randomized confidence intervals with U greater than or equal to u. So, for the given t, the probability $1 - \phi(t, \alpha, \theta)$ of containing that value is $1 - u$. So, at the value θ that is the upper bound of the randomized confidence interval with $\mathcal{U} = u$, the height of the curve to display the fuzzy confidence interval is $1 - u$.

Figure 3 illustrates both fuzzy 95% confidence intervals for the binomial parameter θ when $n = 10$. For $t = 0, 1, \ldots, 5$, this plots $1 - \phi(t, 0.05, \theta)$ as a function of θ; by symmetry, analogous plots apply for $t = 6, \ldots, 10$. Averaged over t for a given θ, the fuzzy confidence interval has coverage probability 0.95. Our experience shows that the fuzzy confidence interval we presented above typically has better performance than the Geyer and Meeden UMPU fuzzy interval, in terms of a more restricted core and support, except when t is at or very near the boundary.

Fig. 3. Fuzzy confidence intervals (Geyer and Meeden (—) , Agresti and Gottard (- - -) for binomial data with sample size $n = 10$, confidence level $1 - \alpha = 0.95$, and observed test statistic $t = 0, 1, 2, 3, 4, 5$.



5 The Mid-P Quasi-Exact Approach

Our focus in this article has been on exact methods for which the nominal error probability α is an upper bound for the actual value. In practice, it is often reasonable to relax this requirement slightly. Conservativeness can be reduced if the error probability is allowed to go slightly above α for some θ values.

5.1 The mid-P-value for significance tests

One way to reduce conservatism while continuing to use the exact probabilities from the small-sample distribution uses the *mid-P-value* (Lancaster 1949, 1961). This replaces $P_{\theta_0}(T = t_{obs})$ in the P-value by $(1/2)P_{\theta_0}(T = t_{obs})$. For instance, a one-sided right-tail P-value has form

$$P_{\theta_0}(T > t_{obs}) + (1/2)P_{\theta_0}(T = t_{obs}).$$

This type of P-value results from forming the usual type of P-value but with Parzen's (1997) *mid-distribution function*, which is $F_{mid}(t) = P(T \leq t) - 0.5P(T = t)$. The mid-P-value $= 1 - F_{mid}(t_{obs})$.

The mid-P-value depends only on the data, unlike the randomized P-value (2). The randomized P-value corresponds to a test that achieves the nominal size, and the mid-P-value replaces $\mathcal{U}$ in it by its expected value. Under the null hypothesis, with discrete distributions the ordinary P-value is stochastically larger than a uniform random variable. By contrast, the mid-P-value has null expected value equal to 1/2 (see, e.g., Berry and Armitage 1995). Also, for the ordinary P-value the sum of the right-tail and left-tail P-values is $1 + P_{\theta_0}(T = t_{obs})$; for the mid-P-value, this sum is 1. Lancaster's (1949) original motivation for proposing the mid-P-value was to create a statistic that, like the uniform P-value for a continuous random variable, could easily be combined for several independent samples.

Unlike the P-values discussed previously in this article, the mid-P-value does not necessarily satisfy $P_{\theta_0}(\mathrm{P-value} \leq \alpha) \leq \alpha$. With it, it is possible to exceed the nominal size. However, evaluations of the mid-P-value in a significance testing format have been encouraging, as summarized next:

Haber (1986) showed that a modification of Fisher's exact test using the mid-P-value has actual size near the nominal size, and the power of the modified test is usually close to that of the randomized UMPU exact test. Hirji, Tan, and Elashoff (1991) and Seneta and Phipps (2001) had similar size results for this case in comparisons with various classical tests. Hirji (1991) showed that the mid-P test worked well for conditional logistic regression (which can be highly discrete). Hwang and Yang (2001) presented an optimality theory for mid-P-values in 2×2 contingency tables, showing how this P-value is the expected value of an optimal P-value resulting from a decision-theoretic approach. Strawderman and Wells (1998) showed that ordinary P-values obtained with higher-order asymptotic methods without continuity corrections for discreteness yield performance similar to that of the mid-P-value.

An awkward aspect of exact conditional inference in logistic regression is that the relevant conditional distribution can be highly discrete. It can even be degenerate when an explanatory variable is continuous. Potter (2005) proposed a permutation test that is also a small-sample method but does not have this disadvantage. The predictor of interest is replaced by residuals from a linear regression of it on the other explanatory variables. Logistic regressions are done for permutations of these residuals, and a P-value is computed by comparing the resulting likelihood-ratio statistics to the original observed value. Potter noted that in small data sets, this permutation P-value is usually similar to the mid-P-value for the exact conditional approach.

5.2 Mid-P confidence intervals

One can form confidence intervals that are less conservative than the traditional discrete one (1) by inverting tests using the mid-P-value. For example, the upper endpoint of the 95% mid-P confidence interval is the solution to

$$P_{\theta_U}(T < t_{obs}) + 0.5 \times P_{\theta_U}(T = t_{obs}) = 0.025.$$

Berry and Armitage (1995) reviewed this approach. Unlike a randomized confidence interval, the mid-P confidence interval necessarily has lower endpoint equal to the smallest value in the parameter space when T takes its minimum value and upper endpoint equal to the largest value in the parameter space when T takes its maximum value. Mid-P-based inference has the advantage over other approximate methods, such as large-sample methods, that it uses the exact distribution.

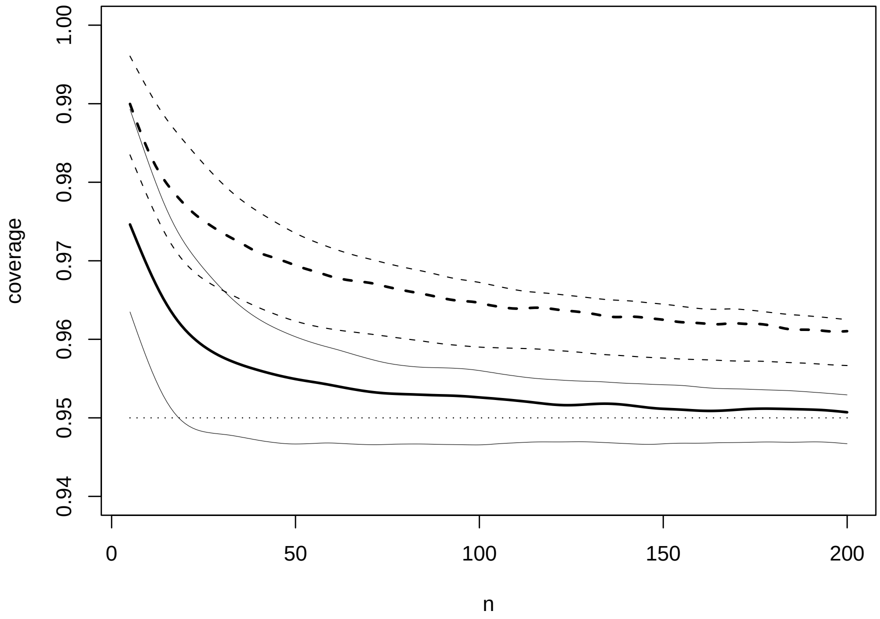
Confidence intervals based on inverting tests using the mid-P-value cannot guarantee that coverage probabilities have at least the nominal level. However, evaluations for a variety of cases have shown that this method still tends to be somewhat conservative, although necessarily less so than using the ordinary P-value. For details, see Vollset (1993), Agresti and Coull (1998), and Newcombe (1998) for the binomial parameter, Agresti (1999) for the odds ratio, Mehta and Walsh (1992) for a common odds ratio in several 2×2 tables, Vollset, Hirji and Afifi (1991) for parameters in conditional logistic regression, and Cohen and Yang (1994) for the Poisson parameter.

Brown, Cai and DasGupta (2001) stated that the mid-P interval for the binomial parameter approximates closely the most popular interval for the Bayesian approach, which uses the Jeffreys prior distribution (beta with parameters 0.5 and 0.5). This relates to work of Routledge (1994), who showed that for a test of H_0: $\theta \geq 0.5$ against H_a: $\theta < 0.5$, the Bayesian P-value given by the posterior probability $P(\theta \geq 0.5|y)$ approximately equals the one-sided mid-P-value for the frequentist binomial test when one uses the Jeffreys' prior.

5.3 Performance of mid-P methods for binomial parameter

We illustrate the behavior of mid-P inference for the binomial case. For testing $\theta = 0.50$ against $\theta > 0.50$, Figure 1 plots the actual size of a nominal size $\alpha = 0.05$ test as a function of n for the ordinary exact binomial test and for the adaptation using the mid-P-value. For $\theta = 0.50$, Figure 2 plots the actual coverage probability of nominal 95% confidence intervals as a function of n, for the Clopper–Pearson exact approach and for the mid-P adaptation. In either case, the actual error probability for the mid-P-based inference tends to fluctuate around the nominal value.

Fig. 4. Average and first and third quartile coverage probabilities (using a uniform distribution for θ) of Clopper–Pearson (- - -) and mid-P (—-) confidence intervals for binomial parameter θ, plotted for n between 5 and 200.



Likewise, for fixed n and varying θ, the actual error probabilities for mid-P-based inferences tend to fluctuate around the nominal value, with the variability of the fluctuations diminishing as n increases. As a consequence, if we average error probabilities uniformly across the parameter space, the average tends to be quite close to the nominal level. Figure 4 shows such an average coverage probability as a function of n, for the ordinary and the mid-P-based confidence intervals. In this average sense, the ordinary exact interval is very conservative (even for moderately large n) while the mid-P-based interval is slightly conservative. This suggests the mid-P approach is an excellent one to adopt if one hopes to achieve close to the nominal level in using a method repeatedly for various studies in which θ itself varies. For this, one must tolerate the actual coverage probability being, for some θ, slightly below the nominal level.

5.4 Software and mid-P inference

For some basic inferences for discrete data, such as tests for a binomial parameter and Fisher's exact test for 2×2 tables, StatXact (Cytel 2005) reports the probability of the observed result as well as the exact P-value. Thus, it is possible to use its output to obtain the mid-P-value for tests. For inference about a parameter of a logistic regression model, LogXact can determine the mid-P-value using a score test or likelihood-ratio test with the exact conditional distribution. However, currently neither software supplies confidence intervals based on the mid-P-value.

We have prepared an R function for finding the mid-P confidence interval for a binomial parameter. It is available at www.stat.ufl.edu/∼aa/cda/software.html.

Acknowledgement

A. Agresti would like to thank Prof. Matilde Bini for arranging a visit to the Univ. of Firenze in Italy during which this work was initiated.

References

Agresti, A. 1999. On logit confidence intervals for the odds ratio with small samples, *Biometrics*, (1999), 55, 597-602.

Agresti A. Exact inference for categorical data: Recent advances and continuing controversies. *Statistics in Medicine* 2001; **20**: 2709-2722.

Agresti, A. 2003. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact, *Statistical Methods in Medical Research*, 12, 3-21.

Agresti, A., and B. A. Coull. 1998. Approximate is better than exact for interval estimation of binomial parameters. *Amer. Statist.* **52**: 119–126.

Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**: 963-971.

Agresti A, Min Y. Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics*, 2002,3, 379-386.

Berry, G. and Armitage, P. (1995) Mid-*P* confidence intervals: A brief review

The Statistician, 44, 417-423

Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* 2000; **28**: 783-98.

Brown, L. D., T. T. Cai, and A. Das Gupta. 2001. Interval estimation for a binomial proportion. *Statist. Sci.* **16**: 101–133.

Casella G, and Berger R.L. *Statistical Inference, 2nd ed.* Pacific Grove, CA: Wadsworth, 2001.

Clopper, C. J., and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26** 404–413.

Cohen, Geoffrey R. and Yang, Shu-Ying (1994) Mid-p confidence intervals for the Poisson expectation *Statistics in Medicine*, 13, 2189-2203

Cox, D. R., and D. V. Hinkley Cytel (2005). *StatXact 7 User Manual*, volumes 1 and 2, and *LogXact 7 User Manual.* Cambridge, Massachusetts: Cytel Inc.

Geyer, C. J. and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and P-values. *Statistical Science*

Haber, M. (1986) A modified exact test for $2x2$ contingency tables Biometrical Journal, 28, 455-463

Hirji, K. F. 1991. A comparison of exact, mid-P, and score tests for matched case-control studies. *Biometrics* **47**: 487–496.

Hirji, Karim F., Tan, Shu-Jane and Elashoff, Robert M. (1991) A quasi-exact test for comparing two binomial proportions, Statistics in Medicine, 10, 1137-1153

Hirji, Karim F., Tang, Man-Lai, Vollset, Stein E. and Elashoff, Robert M. (1994) Efficient power computation for exact and mid-p tests for the common odds ratio in several $2x2$ tables, Statistics in Medicine, 13, 1539-1549

Hwang, J. T., and Yang, M. C. (2001). An optimality theory for mid p-values in 2×2 contingency tables. *Statist. Sinica* **11** 807–826.

Kim D, Agresti A. Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association* 1995; **90**: 632-9.

Lancaster, H. O. (1949) The combination of probabilities arising from data in discrete distributions. *Biometrika* 36, 370-382.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56** 223–234.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses.* New York: Wiley.

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. *The American Statistician* 1992; **46**: 146-50.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statist. Medic.* **17** 857–872.

Neyman J. On the problem of confidence limits. *Annals of Mathematical Statistics* 1935; **6**: 111-6.

Parzen, E. 1997. Concrete statistics. Pp. 309-332 in *Statistics in Quality.* New York: Marcel Dekker.

Pearson, E. S. (1950). On questions raised by the combination of tests based on discontinuous distributions. *Biometrika* **37** 383–398.

Potter, D. M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Statistics in Medicine* 24: 693-708.

Routledge, R. D. (1994). Practicing safe statistics with the mid-p^*. *Canad. J. Statist.* **22** 103–110.

Seneta, Eugene and Phipps, Mary C. (2001) On the comparison of two observed frequencies Biometrical Journal, 43, 23-43

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41** 275–278.

Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37** 117–129.

Strawderman, R. L., and Wells, M. T. (1998). Approximately exact inference for the common odds ratio in several 2×2 tables. *J. Amer. Statist. Assoc.* **93** 1294–1307.

Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society* 1985; **A 148**: 317-27.

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statist. Medic.* **12** 809–824.

Vollset, Stein E., Hirji, Karim F. and Afifi, Abdelmonem A. (1991) Evaluation of exact and asymptotic interval estimators in logistic analysis of matched case-control studies Biometrics, 47, 1311-1325