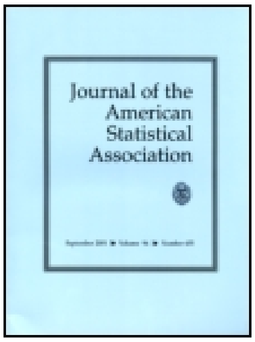

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Comment

Alan Agresti

To cite this article: Alan Agresti (1998) Comment, Journal of the American Statistical Association, 93:444, 1307-1310, DOI: 10.1080/01621459.1998.10473791

To link to this article: https://doi.org/10.1080/01621459.1998.10473791

Published online: 17 Feb 2012.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Mehta, C. R., and Patel, N. R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in $R \times C$ Contingency Tables," *Journal of the American Statistical Association*, 78, 427–434.

Mehta, C. R., Patel, N. R., and Gray, R. (1985), "Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2×2 Contingency Tables," *Journal of the American Statistical Association*, 80, 969–973.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (1988), "Importance Sampling for Estimating Exact Probabilities in Permutational Inference," *Journal of the American Statistical Association*, 83, 999–1005.

——— (1992), "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data," *Journal of Computational and Graphical Statistics*, 1, 21–40.

Mehta, C. R., and Walsh, S. J. (1992), "Comparison of Exact, Mid-p, and Mantel–Haenszel Confidence Intervals for the Common Odds Ratio Across Several 2×2 Tables," *The American Statistician*, 46, 146–151.

Munõz, A., and Rosner, B. (1984), "Power and Sample Size for a Collection of 2×2 Tables," *Biometrics*, 40, 995–1004.

Olver, F. W. J. (1974), *Asymptotics and Special Functions*, New York: Academic Press.

Pan, V. (1997), "Solving a Polynomial Equation: Some History and Recent Progress," *SIAM Review*, 39, 187–220.

Parlett, B. (1980), *The Symmetric Eigenvalue Problem*, Englewood Cliffs, NJ: Prentice Hall.

Pierce, D., and Peters, D. (1992), "Practical Higher Order Asymptotics for Multiparameter Exponential Families," *Journal of the Royal Statistical Society*, Ser. B, 54, 701–737.

Pitman, J. (1997), "Probabilistic Bounds on the Coefficients of Polynomials With Only Real Zeros," *Journal of Combinatorial Theory*, Ser. A, 77, 279–303.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes* (2nd ed.), Cambridge, U.K.: Cambridge University Press.

Robins, J., Breslow, N., and Greenland, S. (1986), "Estimators of the Mantel–Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models," *Biometrics*, 42, 311–325.

Schoenberg, I. J. (1955), "On The Zeros of the Generating Functions of Multiply Positive Sequences and Functions," *Annals of Mathematics*, 62, 447–471.

Strawderman, R. L., Casella, G., and Wells, M. T. (1996), "Practical Small-Sample Asymptotics for Regression Problems," *Journal of the American Statistical Association*, 91, 643–655.

Szego, G. (1975), *Orthogonal Polynomials*, Providence, RI: American Mathematical Society.

Temme, N. (1982), "The Uniform Asymptotic Expansion of a Class of Integrals Related to Cumulative Distribution Functions," *SIAM Journal of Mathematical Analysis*, 13, 239–253.

van Doorn, E. A. (1987), "Representations and Bounds for Zeros of Orthogonal Polynomials and Eigenvalues of Sign-Symmetric Tridiagonal Matrices," *Journal of Approximation Theory*, 51, 254–266.

Vollset, S. E., Hirji, K. F., and Elashoff, R. M. (1991), "Fast Computation of Exact Confidence Limits for the Common Odds Ratio in a Series of 2×2 Tables," *Journal of the American Statistical Association*, 86, 404–409.

Wang, S. (1993), "Saddlepoint Expansions in Finite Population Problems," *Biometrika*, 80, 583–590.

Comment

Alan AGRESTI

1. INTRODUCTION

The development of computational algorithms for exact inferential analyses has been a major advance of the past decade in contingency table analysis. The release of StatXact and the inclusion of many of its routines in SAS and SPSS now makes exact methods easy to use. Although certain exact analyses are still computationally infeasible, methods exist for their accurate approximation, such as Monte Carlo (e.g., Agresti, Wackerly, and Boyett 1979; Booth and Butler 1998; Mehta, Patel, and Senchaudhuri 1988), Markov chain Monte Carlo (e.g., Forster, McDonald, and Smith 1996), and the iterated bootstrap (Presnell 1996). The Strawderman and Wells article is a very helpful contribution in showing how to use saddlepoint methods to approximate exact inference for several 2×2 tables. As in many applications, such as ones for contingency tables discussed by Davison (1988), Pierce and Peters (1992) and Agresti, Lang, and Mehta (1993), saddlepoint approximations tend to work amazingly well.

Exact algorithms are usually much slower for interval estimation than for testing a null parameter value, so saddlepoint approximations for exact confidence intervals are especially useful. I also appreciated the emphasis in the Strawderman and Wells article on studying power. Although power approximations are also relatively simple to compute when using Monte Carlo generation of null and non-null distributions, this topic has received surprisingly little attention in the literature.

Strawderman and Wells focus primarily on inference for a common odds ratio. As discussed in the following sections, two-sided large-sample inference for the odds ratio often performs reasonably well for small samples. For several degree of freedom problems, on the other hand, such as testing fit of models using chi-squared statistics with sparse data, large-sample methods can break down severely. For many such problems, Monte Carlo methods are relatively simple, and I wonder how much potential exists for extending the Strawderman and Wells work to more complex problems of this type? Of course, meaningful power questions become more difficult to pose when the number of parameters is large.

2. WHICH EXACT ANALYSIS TO APPROXIMATE?

With exact methods, one can guarantee that the size of a test is no greater than some prespecified level and that

Alan Agresti is Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (E-mail: aa@stat.ufl.edu). The author thanks James Booth for discussions about this topic.

© 1998 American Statistical Association
Journal of the American Statistical Association
December 1998, Vol. 93, No. 444, Theory and Methods

the coverage probability for a confidence interval is at least the nominal level. Nonetheless, in using exact methods or in approximating them with discrete data, one must think carefully about which exact method should be the "gold standard." A variety of options exist. These include (1) exact conditional tests and confidence intervals based on inverting such tests for non-null parameter values, such as discussed by Strawderman and Wells, and (2) exact unconditional inferences that condition only on margins of the contingency table naturally fixed by the sampling design (e.g., Berger and Boos 1994). The latter approach has been studied mainly for 2×2 tables, but Freidlin and Gastwirth (1999) propose unconditional versions of the Mantel–Haenszel test for several 2×2 tables. The exact conditional approach eliminates the nuisance parameters by conditioning on their sufficient statistics. The exact unconditional approach eliminates nuisance parameters using a "worst-case" scenario. For instance, the p value is a tail probability maximized over all possible values for the nuisance parameters. Exact conditional methods have the advantage of versatility, applying to a wide variety of exponential family problems under several sampling situations, including randomization-based analyses when the samples are in no sense the result of binomial or multinomial sampling. On the other hand, when the sampling scheme is multinomial or independent binomial rather than hypergeometric, the restriction of the sample space to samples having exactly the same response margins as the one observed may seem artificial.

It is not my intention here to discuss the issues in the conditioning controversy, but I mention this to emphasize that in approximating exact inference, one must first decide which exact inference to approximate. The unconditional approach is much more computationally intensive than the conditional approach, and the development of approximate exact inference of this form is an interesting challenge for future work. Moreover, each type of exact inference has different ways of performing it. For instance, to define p values in different ways, one might use a likelihood-ratio, Wald, or score-test statistic, and confidence intervals can be constructed by inverting two separate one-tailed tests or a single two-tailed test. Even within a certain type of exact inference, results can vary considerably according to the choices of p value and test. For instance, Strawderman and Wells report an exact confidence interval of (1.08, 531.5) for Example 1 in Table 3; an alternative exact interval based on the same test statistic but inverting a single two-sided test (Kim and Agresti 1995) yields the interval (1.29, 261.5).

The variability in results that can occur with different methods mainly reflects the complications that result in exact inference due to discreteness. With the use of supplementary randomization to achieve a desired size, one-sided exact conditional tests for an odds ratio in a 2×2 table or a common odds ratio in several such tables are uniformly most powerful unbiased. In practice, data-unrelated randomization is unacceptable, and it is rarely possible to achieve an arbitrary size such as .05. This is not problematic if one does not treat .05 as sacred and merely uses a p value to summarize the evidence against the null hypothesis. However, this discreteness has more disturbing implications for unconditional power calculations and for confidence intervals. With exact methods for interval estimation, the actual coverage probability can be much larger than the nominal confidence level and is unknown (Neyman 1935). The implication is conservativeness; in other words, as the relevant distribution becomes more highly discrete, exact tests lose power and exact confidence intervals tend to be overly wide. The degree of conservativeness is usually more severe for conditional than for unconditional inference, because the extra conditioning increases the severity of discreteness (Suissa and Shuster 1985).

3. APPROXIMATE ALTERNATIVES TO EXACT ANALYSES

If one is willing to use a method having actual size close to the nominal level but not necessarily bounded by it, then one can consider approximate as well as "exact" methods. Two-sided inference for single parameters based on large-sample normal approximations often performs reasonably well in this regard. I illustrate with confidence intervals for the odds ratio in a 2×2 table, based on independent binomial samples. For cell counts $\{a, b, c, d\}$, a large-sample 95% confidence interval based on the delta method is $\exp[\log(ad/bc) \pm 1.96(a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2}]$. I randomly sampled 10,000 pairs of binomial parameters (p_1, p_2) from the uniform distribution over the unit square and evaluated coverage probabilities of nominal 95% intervals, for binomial samples of size 10 each for each combination of parameters. The mean coverage probability was .977 for the large-sample method and .986 for the exact; the minimum coverage probability was .941 for the large-sample method and .970 for the exact. For each pair of parameters, I computed expected lengths conditional on the event that all four counts are positive, so that both intervals have finite length. Because of tail behavior, such differences in coverage probability can translate to large differences in expected interval length. With each of the 10,000 probability pairs oriented so that $p_1 \geq p_2$, the median of the expected lengths was 59 for the large-sample method and 228 for the exact.

In similar evaluations with many other small sample sizes, the large-sample confidence interval for the odds ratio rarely has actual coverage probability much below the nominal level, whereas the exact interval usually has coverage probability considerably above the nominal level and has length considerably longer than the large-sample interval. An approximate method may be preferred to an exact method if its actual coverage probability is never much less than the nominal confidence level. Situations exist, however, in which it is imperative to guarantee a bound on the actual coverage probability and/or in which large-sample methods are highly questionable (such as Example 2 of Strawderman and Wells, in which each of 18 partial tables has only one observation in one of the rows). In addition, the conservativeness problem disappears as the sample size and table size increase. Thus there will always be an important niche for exact methods. However, the complications due to discreteness suggest that statisticians should perhaps reconsider how to evaluate statistical procedures. For instance, the confidence coefficient is traditionally defined to be the infimum of the coverage probabilities over the parameter space. Is it better to use an approach that guarantees that the coverage probabilities are $\geq .95$, yet may have actual coverage probabilities of $\sim .97$ or .98 (such as the usual exact interval), or an approach giving narrower intervals for which the actual coverage probability could be $<.95$ but is usually quite close to .95?

If statisticians are willing to relax requirements about the bounding of coverage probabilities, then besides large-sample methods the possible methods include adaptations of exact methods based on the mid-p value. This seems to be a reasonable compromise for discrete data between the conservativeness of exact methods and the uncertain adequacy of a large-sample method. It is already available in StatXact, and it has some appealing properties. For instance, its null expected value is 0.5, as is true for the p value for test statistics having 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. A recent unpublished work by Hwang and Yang (1998) shows that it is an optimal p value in terms of estimating a truth indicator of the null hypothesis. Numerical evaluations (Mehta and Walsh 1992; Vollset 1993) for interval estimation of proportions and odds ratios based on inverting tests using the mid-p value show that it tends to have coverage probabilities slightly exceeding the nominal value, but it tends to be less conservative than exact methods using the ordinary p value. 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.

For cases in which exact mid-p value calculation is infeasible, the saddlepoint approximation to the mid-p value is likely to perform very well (e.g., Agresti et al. 1993). The Strawderman and Wells alternative p value of $P_{\theta_0}\{W(\theta) \geq w_{\text{obs}} + \frac{1}{2}\}$ is intriguing, since for most distributions it would be less conservative than the mid-p value. Can the authors provide further justification for it? Methods using ordinary p values obtained with "approximate conditioning" techniques may yield similar performance.

Although much discussion has occurred about the appropriateness of conditioning in analyzing contingency tables, in terms of practical performance of methods, the degree of discreteness is the determinant more than whether one uses conditioning (Mehta and Hilton 1993). Problems due to discreteness can arise even when no conditioning is involved, such as in interval estimation for a single binomial parameter (Agresti and Coull 1998). So that I do not seem too alarmist regarding ordinary exact conditional methods, I would like to temper my remarks and point out that the approximations of Strawderman and Wells are of greatest use for cases in which exact methods are themselves computationally infeasible. Such infeasible cases have a huge number of tables in the reference set and usually display little discreteness. Cautions in using exact conditional methods then lose their relevance, and mid-p approaches are essentially identical to the ordinary exact approaches.

4. MORE FLEXIBLE MODELS FOR SEVERAL 2 × 2 TABLES

The common odds ratio model is a natural first step for modeling several 2 × 2 tables. More complex models that permit heterogeneity among the odds ratios are now also being used. For instance, for a multi-center study for comparing two treatments on a binary response, the simple logit model having treatment and center main effects assumes a common odds ratio between treatment and response for each center. An extension treats center effects as random and adds random effects for center-by-treatment interaction, resulting in conditional log odds ratios between treatment and response that form an independent sample from a normal distribution. Fitting the model provides an estimated average log odds ratio and an estimate of variability about the average. For related analyses, see Givens, Smith, and Tweedie (1997), Liu and Pierce (1993), and Skene and Wakefield (1990). For random effects models it is difficult to construct even good approximate analyses (Breslow and Clayton 1993), and solutions that are exact in any sense provide a strong challenge for the future.

ADDITIONAL REFERENCES

Agresti, A., and Coull, B. A. (1998), "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Parameters," *The American Statistician*, 52, 119–126.

Agresti, A., Lang, J. B., and Mehta, C. (1993), "Some Empirical Comparisons of Exact, Modified Exact, and Higher-Order Asymptotic Tests of Independence for Ordered Categorical Variables," *Communications in Statistics: Simulation and Computation*, 22, 1–18.

Agresti, A., Wackerly, D., and Boyett, J. (1979), "Exact Conditional Tests for Cross-Classifications: Approximation of Attained Significance Level," *Psychometrika*, 44, 75–83.

Berger, R. L., and Boos, D. D. (1994), "*P* Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Booth, J., and Butler, R. (in press), "Monte Carlo Approximation of Exact Conditional Tests for Log-Linear Models," *Biometrika*, 93.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Forster, J. J., McDonald, J. W., and Smith, P. W. F. (1996), "Monte Carlo Exact Conditional Tests for Log-Linear and Logistic Models," *Journal of the Royal Statistical Society*, Ser. B, 58, 445–453.

Freidlin, B., and Gastwirth, J. L. (in press), "Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables," *Biometrics*, 55.

Givens, G. H., Smith, D. D., and Tweedie, R. L. (1997), "Publication Bias in Meta-Analysis: A Bayesian Data-Augmentation Approach To Account for Issues Exemplified in the Passive Smoking Debate," *Statistical Science*, 12, 221–250.

Liu, Q., and Pierce, D. A. (1993), "Heterogeneity in Mantel–Haenszel-type Models," *Biometrika*, 80, 543–556.

Mehta, C. R., and Hilton, J. F. (1993), "Exact Power of Conditional and Unconditional Tests: Going Beyond the 2 × 2 Contingency Table," *The American Statistician*, 47, 91–98.

Mehta, C. R., Patel, N. R., and Senchaudhuri, P. (1988), "Importance Sampling for Estimating Exact Probabilities in Permutational Inference," *Journal of the American Statistical Association*, 83, 999–1005.

Neyman, J. (1935), "On the Problem of Confidence Limits," *Annals of Mathematical Statistics*, 6, 111–116.

Presnell, B. (1996), "Bootstrap Unconditional *P*-Values for the Sign Test With Ties and the 2 × 2 Matched-Pairs Trial," *Journal of Nonparametric Statistics*, 7, 47–55.

Skene, A. M., and Wakefield, J. C. (1990), "Hierarchical Models for Multicentre Binary Response Studies," *Statistics in Medicine*, 9, 919–929.

Suissa, S., and Shuster, J. J. (1985), "Exact Unconditional Sample Sizes for the 2 by 2 Binomial Trial," *Journal of the Royal Statistical Society*, Ser. A, 148, 317–327.

Vollset, S. E. (1993), "Confidence Intervals for a Binomial Proportion," *Statistics in Medicine*, 12, 809–824.

Comment

James G. Booth and Ronald W. Butler

1. INTRODUCTION

Exact conditional inference for contingency tables is an important statistical issue with the added benefit of being somewhat controversial. We congratulate Strawderman and Wells for a very interesting and timely article on exact conditional inference concerning a common odds ratio. We agree with the authors that saddlepoint approximations are often so accurate and computationally efficient that they essentially eliminate the need for exact computation, even if such computations are possible.

We have four points that we would like to discuss. The first is a relatively minor one concerning the authors' use of "sequential" rather than "single" saddlepoint approximation. We argue that the latter terminology more correctly describes their approximations. Our second point concerns the view expressed by the authors, and many others in the literature, that distributions must be sums of independent random variables to justify the use of saddlepoint methods. Numerical accuracy in a wide range of examples suggests that the sum need involve only a single random variable. Third, we compare the authors' approximation with Skovgaard's (1987) double-saddlepoint approximation and show that the latter is considerably simpler and virtually as accurate. We make some further comments on the more general comparison of single- versus double-saddlepoint approximations. Our last comment addresses our concern that the results of this article are limited to a rather narrow class of problems. In contrast, Monte Carlo methods are now available that can be applied much more generally and with increasing speed. We describe how one method based on importance sampling, which we have successfully applied in a wide variety of log-linear models to obtain exact conditional p values, can be easily extended to conditional power calculations.

2. SINGLE VERSUS SEQUENTIAL

For purposes of discussion, it is convenient to simplify matters to a single table of data. The noncentral hypergeometric distribution from this table has moment-generating function (MGF) $\psi(\theta e^s)/\psi(\theta)$, where θ is the odds ratio and ψ is a special case of the ${}_2F_1$ hypergeometric function given in (1). The main contribution of the article is in recognizing that the polynomial function ψ is characterized by its negative roots, allowing simple repeated computations of hypergeometric probabilities using a single-saddlepoint approximations based on cumulant-generating function (CGF) $\ln\psi(\theta e^s) - \ln\psi(\theta)$. The sequential-saddlepoint method was popularized by Fraser, Reid, and Wong (1991) and successfully used by Butler et al. (1992), but was originated by Bartlett (1938). The essential idea is to use one single-saddlepoint approximation to approximate the conditional MGF from its associated joint MGF; this approximate conditional MGF is subsequently used as a surrogate for the true MGF in a second single-saddlepoint approximation for conditional probability computation.

3. $n = 1$ ASYMPTOTICS

Saddlepoint methods are often referred to as "small" sample size asymptotics, but we point out that $n = 1$ generally suffices, as originally pointed out by Skovgaard (1987). The authors seem to use this as a reason for not addressing the applicability of double-saddlepoint procedures applied by Davison (1988) and Pierce and Peters (1992) for inference about a common odds ratio.

To illustrate this point, we computed both single- and double-saddlepoint approximations for central hypergeometric probabilities ($\theta = 1$) in Table 1, taking the first example from Skovgaard. This is admittedly not entirely a sample size 1 context, as the distribution theory follows from four independent Poisson variables that are infinitely divisible; however, it suffices for making our point.

All three approximations are extremely accurate and sufficiently so for most applications. The last single-saddlepoint approximation, which is given in (10), is slightly better. Its computation for the last example, however, as suggested by the authors, requires finding the 20 roots that characterize ψ. In contrast, the double-saddlepoint approximations given by Skovgaard are simple explicit computations requiring no root finding.

An enormous range of examples reflects the same accuracy seen in this example, with accuracy defying what might be explained through the asymptotics. The relevant question seems to be whether such approximations should be trusted in situations where exact computation cannot verify their accuracy. This is where several different approximations be-

James G. Booth is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611 (E-mail: jbooth@stat.ufl.edu). Ronald W. Butler is Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80525.

Journal of the American Statistical Association
December 1998, Vol. 93, No. 444, Theory and Methods