Statistical Science
2005, Vol. 20, No. 4, 367–371
DOI 10.1214/088342305000000403

Comment: Randomized Confidence Intervals and the Mid-P Approach

Alan Agresti and Anna Gottard

We enjoyed reading the interesting, thought-provoking article by Geyer and Meeden. In our comments we will try to place their work in perspective relative to the original proposals for exact and randomized confidence intervals for the binomial parameter. We propose a fuzzy version of the original binomial randomized confidence interval, due to Stevens (1950). Our approach motivates an existing nonrandomized confidence interval based on inverting a test using the mid-P value. The mid-P confidence interval provides a sensible compromise that mitigates the effects of conservatism of exact methods, yet provides results that are more easily understandable to the scientist.

1. HISTORICAL PERSPECTIVE

Clopper and Pearson (1934) proposed the following $100(1-\alpha)\%$ confidence interval for a binomial parameter θ: The bounds $[\theta_L, \theta_u]$ are the solutions to the equations

$$\sum_{k=0}^{x} \binom{n}{k} \theta_U^k (1-\theta_U)^{n-k} = \alpha/2$$

and

$$\sum_{k=x}^{n} \binom{n}{k} \theta_L^k (1-\theta_L)^{n-k} = \alpha/2.$$

(One takes $\theta_L = 0$ when $x = 0$ and $\theta_U = 1$ when $x = n$.) This confidence interval is based on inverting two one-sided binomial tests. Because of discreteness, the method is conservative; the actual confidence level is bounded below by $1-\alpha$ (Neyman, 1935).

To eliminate the conservativeness, Stevens (1950) suggested instead solving the binomial-probability equations

$$\Pr_{\theta_U}(X < x) + U \times \Pr_{\theta_U}(X = x) = \alpha/2$$

and

$$\Pr_{\theta_L}(X > x) + (1-U) \times \Pr_{\theta_L}(X = x) = \alpha/2,$$

where U is a Uniform$(0, 1)$ random variable. This confidence interval is based on inverting tests for which (as in the case of continuous random variables) the one-sided P-values have a uniform null distribution and sum to 1, unlike the ordinary one-sided P-values used in the Clopper–Pearson confidence interval. We will refer to this as the *Stevens randomized confidence interval*. Anscombe (1948) made the analogous one-sided proposal of inverting a randomized one-sided binomial test so as to obtain an upper or lower randomized confidence bound. Blyth and Hutchinson (1960) provided tables for implementing a slightly different randomized confidence interval (proposed by M. W. Eudey in a 1949 technical report at the University of California, Berkeley) that has the property of being Neyman shortest unbiased.

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 they do not consider actually implementing it in practice. 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. He predicted that randomized confidence intervals "will not meet with strong objection." 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." He argued that "it is the statistician's duty to be wrong the stated proportion of times, and failure to reach this proportion is equivalent to using an inefficient in place of an efficient method of estimation." He noted, though, the apparent paradox

Alan Agresti is Distinguished Professor Emeritus, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, USA (e-mail: aa@stat.ufl.edu). Anna Gottard is Assistant Professor, Department of Statistics, University of Florence, Florence, Italy 50134 (e-mail: gottard@ds.unifi.it).

that two people could analyze the same data, yet a 90% confidence interval for one person could be contained in an 89% confidence interval for the other person.

2. FUZZY CONFIDENCE INTERVALS BASED ON THE ANSCOMBE AND STEVENS RANDOMIZED CONFIDENCE INTERVALS

Let us consider first the one-sided lower confidence bound for Anscombe's (1948) method. A $100(1-\alpha)\%$ randomized lower confidence bound is obtained by solving the equation

$$\Pr_{\theta_L}(X > x) + U \times \Pr_{\theta_L}(X = x) = \alpha/2.$$

Without actually implementing the randomization, this would correspond to a UMP testing procedure and to the Geyer–Meeden fuzzy confidence interval (in the form of a bound). Solving this equation with $U = 1$ yields an interval $(\theta_L, 1)$ that is the support of the Geyer–Meeden fuzzy confidence interval. Solved with $U = 0$, this equation yields an interval $(\theta_L, 1)$ that is the core of the Geyer–Meeden fuzzy confidence interval. Solving with $U = u$ provides the u-cut of the fuzzy interval recommended by Geyer and Meeden in Section 2.1 for those who prefer a realized randomized confidence interval.

Suppose we replace U in the Anscombe randomized confidence bound by its expected value, $1/2$. This corresponds to inverting the result of the test using the *mid-P value*, which is the null probability of more extreme results plus half the probability of the observed result (Lancaster, 1961). In the one-sided case, it provides a confidence bound that is a compromise between the support and the core of the fuzzy confidence bound. It is the u-cut with $u = 1/2$.

In the two-sided case (i.e., confidence intervals rather than confidence bounds), the Stevens (1950) randomized confidence interval without actually performing the randomization is the basis of a fuzzy confidence interval that we propose as an alternative to the Geyer–Meeden fuzzy confidence interval. As U increases from 0 to 1, the lower and upper endpoints of the Stevens randomized confidence interval are monotonically increasing. Substituting $U = 0$ in Stevens equations gives the bounds for a realized Stevens interval, having as lower bound the Clopper–Pearson lower bound. Substituting $U = 1$ gives the bounds for a realized Stevens interval having as upper bound the Clopper–Pearson upper bound. Thus, the support of the proposed fuzzy confidence interval is the Clopper–Pearson interval. The core of the fuzzy confidence interval is the set of θ values that fall in every one of the possible realized Stevens confidence intervals. This core goes from the lower bound of the realized Stevens confidence interval with $U = 1$ to the upper bound of the realized Stevens confidence interval with $U = 0$.

The figure for this proposed fuzzy confidence interval is easily constructed. This is simplest to describe when $1 \le x \le n - 1$, because the endpoints are then strictly monotone in U. Consider an arbitrary value $U = u$ for the uniform random variable. The value that is the lower bound of the randomized confidence interval with $U = u$ is contained only in all the randomized confidence intervals with U less than or equal to u. So, for the given x, the probability $1 - \phi(x, \alpha, \theta)$ of containing that value is u. Thus, at the value θ that is the lower bound of the randomized confidence interval with $U = u$, 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 $U = u$ is contained only in all the randomized confidence intervals with U greater than or equal to u. So, for the given x, the probability $1 - \phi(x, \alpha, \theta)$ of containing that value is $1 - u$. Thus, at the value θ that is the upper bound of the randomized confidence interval with $U = u$, the height of the curve to display the fuzzy confidence interval is $1 - u$.

This method of forming a fuzzy confidence interval by inverting two single-tailed randomized tests applies to interval estimation with any of the popular discrete one-parameter exponential family distributions. Unlike the Geyer–Meeden fuzzy confidence interval, this type of fuzzy confidence interval is not UMPU. There is no reason, however, that a statistical procedure needs to be unbiased to have good practical performance. Using the Geyer–Meeden approach to specify the realized confidence interval by the u-cut and thus putting the parameter values in the confidence interval that are "least contradictory" seems analogous to the Sterne (1954) approach for forming nonrandomized intervals by inverting tests in which acceptance regions are formed using the most likely outcomes. By contrast, the approach we have presented seems analogous to the tail method of forming a confidence interval, inverting two one-sided tests with equal tail probabilities.

Let us consider the example from Geyer and Meeden of a binomial distribution with $n = 10$. Table 1 shows the support and core of the 95% fuzzy confidence intervals for the possible x outcomes (analogous results follow for x between 6 and 10, by symmetry). We were surprised that the support of the Geyer–Meeden fuzzy

TABLE 1
Core and support of Geyer–Meeden and proposed fuzzy confidence intervals for binomial distribution with $n = 10$

	Observed outcome x					
Fuzzy conf. int.	**0**	**1**	**2**	**3**	**4**	**5**
Core						
Geyer–Meeden	(0.000, 0.000)	(0.006, 0.303)	(0.041, 0.443)	(0.098, 0.558)	(0.169, 0.660)	(0.251, 0.749)
Proposed	(0.000, 0.000)	(0.025, 0.308)	(0.067, 0.445)	(0.122, 0.556)	(0.187, 0.652)	(0.262, 0.738)
Support						
Geyer–Meeden	(0.000, 0.303)	(0.000, 0.443)	(0.006, 0.558)	(0.041, 0.660)	(0.098, 0.749)	(0.169, 0.831)
Proposed	(0.000, 0.308)	(0.003, 0.445)	(0.025, 0.556)	(0.067, 0.652)	(0.122, 0.738)	(0.187, 0.813)

confidence interval often contains the Clopper–Pearson confidence interval, which is the support of our proposed fuzzy confidence interval. In our experience this is typical when x is near the middle of the range. The Clopper–Pearson confidence interval is known to be notoriously conservative. We do not find it desirable to use a fuzzy confidence interval that has support even wider than the Clopper–Pearson interval to satisfy the goal of achieving the nominal coverage probability exactly.

Figure 1 shows our proposed fuzzy confidence interval for the case $x = 4$, comparing it with the Geyer–Meeden fuzzy confidence interval plotted in Figure 2 of their article. In various examples we considered, the proposed fuzzy confidence interval gave more precise results when x is near the middle of the range, and the Geyer–Meeden fuzzy confidence interval did better for observations at the boundary values. This is similar to what Blyth and Hutchinson (1960) reported in comparing the Stevens randomized confidence interval to the one they tabulated having the Neyman shortest unbiased property.

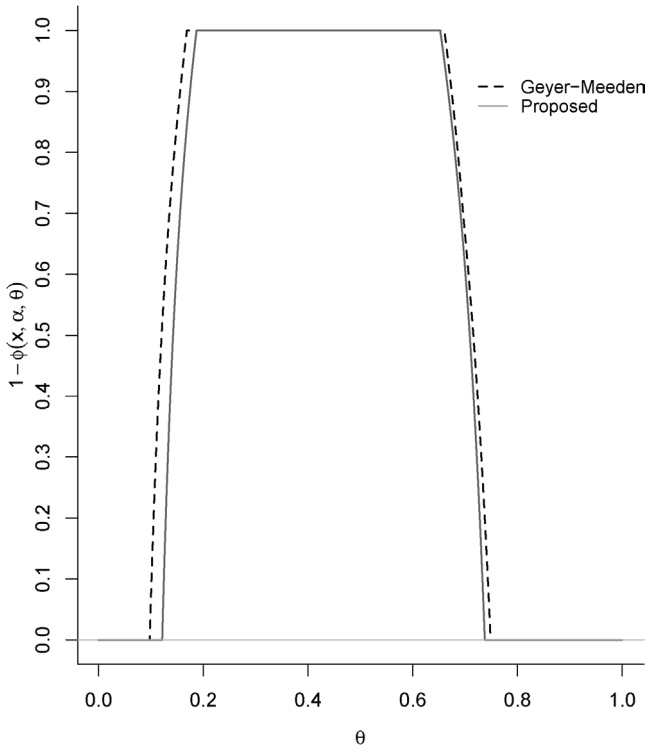

FIG. 1. *Fuzzy confidence intervals for binomial distribution with sample size $n = 10$, confidence level $1 - \alpha = 0.95$ and observed $x = 4$.*

3. THE MID-P CONFIDENCE INTERVAL

With $U = 1/2$, the Stevens randomized confidence interval reduces to the mid-P confidence interval. For $n = 10$ with $x = 4$, this interval is (0.142, 0.709). By comparison, the 0.5-cut for the Geyer–Meeden fuzzy confidence interval is (0.120, 0.716).

The mid-P confidence interval has lower endpoint of 0 when $x = 0$ and upper endpoint of 1 when $x = n$. This is not necessarily the case for randomized confidence intervals. For example, for the Stevens method, when $x = 0$ the lower bound exceeds 0 when $U > 1 - \alpha/2$ and when $x = n$ the upper bound is less than 1 when $U < \alpha/2$. This apparently relates to the remark in Section 3.2 of Geyer and Meeden about $(1 - \phi(x, \alpha, \theta))$ converging to $1 - \alpha$ at one of the boundaries of θ values for extreme values of x. To us, this is an unappealing property of randomized confidence intervals and fuzzy confidence intervals.

Inference based on the mid-P value is a useful, general-purpose method for discrete data. It applies directly to interval estimation with any of the common discrete one-parameter distributions. The mid-P confidence interval does sacrifice the guarantee of having coverage probability equal to exactly $1 - \alpha$ or at least $1 - \alpha$. However, results in Vollset (1993), Agresti and Coull (1998), Newcombe (1998) and Brown, Cai

and DasGupta (2001) suggest that for practical purposes it performs quite well for interval estimation of a binomial proportion. It is less conservative than the Clopper–Pearson interval, but it still usually has actual coverage probability greater than $1-\alpha$; when it is less, it rarely seems to be enough less to be of practical importance. Brown, Cai and DasGupta (2001) showed that it approximates closely an interval obtained with a Bayesian approach using the Jeffreys prior distribution [beta with parameters 0.5 and 0.5, but their formula (17) for the mid-P confidence interval in terms of beta quantiles is incorrect].

For significance testing in discrete problems, many statisticians prefer the mid-P value to the ordinary P-value. Under the null hypothesis, the ordinary P-value is stochastically larger than a uniform random variable. By contrast, the mid-P value has null expected value equal to $1/2$. For the ordinary P-value, the sum of the right-tail and left-tail P-values is $1+\Pr_{\theta_0}(x)$; for the mid-P value, this sum is 1. Ordinary P-values obtained with higher order asymptotic methods without continuity corrections for discreteness yield performance similar to that of the mid-P value (Pierce and Peters, 1992; Strawderman and Wells, 1998). 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 a P-value resulting from a decision-theoretic approach. See also Routledge (1994) for views that promote use of the mid-P value.

4. WHY MID-P AND NOT FUZZY CONFIDENCE INTERVALS?

Geyer and Meeden tell us that statisticians and scientists should stop after finding the fuzzy confidence interval and report it. We are skeptical about being able to convince scientists to adopt fuzzy confidence intervals. However, it does seem to us that scientists could be convinced of the superiority of a confidence interval based on the mid-P value compared to the ordinary conservative confidence interval. Why? Because of properties such as a shorter interval and correspondence with a test having null expected P-value equal to $1/2$ and one-tailed P-values that sum to 1. Also, compared to randomized and fuzzy confidence intervals, the mid-P interval does not have the unappealing behavior mentioned in the previous section of not containing the smallest (largest) θ values when x takes its smallest (largest) value.

With many scientists, of course, it can be difficult to wean them away from merely reporting a P-value, and to report instead estimates of parameters and confidence intervals. We do not see much hope of getting them to present anything more complex than an ordinary confidence interval. If we want to progress beyond that, perhaps we as statisticians should be spending more time educating scientists about the likelihood function and supplying software with which they can look at the entire likelihood function (or profile likelihood) to learn more about plausible values of relevant parameters.

Geyer and Meeden also state that the fuzzy confidence interval could be taught in elementary statistics courses. Much as we enjoyed reading this paper, we cannot imagine teaching this concept in Statistics 101 at the typical American or Italian university. For that matter, we would not even try to teach the mid-P confidence interval in such a course. The student has enough difficulty understanding the P-value concept, and it is simpler to avoid the issue of discreteness in such a course by concentrating on large-sample approximations. One reason Agresti and Coull (1998) suggested their adjusted Wald confidence interval based on adding two outcomes of each type was because this *was* something simple enough to handle in an elementary course. It attacked a discrete problem by adapting a continuous large-sample solution, the ordinary Wald interval taught in such courses, in such a way that it provided much better performance.

We do not intend these comments to detract from a very interesting and provocative paper. The fuzzy confidence interval is an intriguing way to summarize results, and this paper makes an important contribution toward helping statisticians understand the difficulties in having methods for discrete variables match the same characteristics as corresponding methods for continuous variables. However, just as Stevens and Pearson were overly optimistic in believing there was a good future for the randomized test and confidence interval, we believe that Geyer and Meeden may be overly optimistic in their hopes for the adoption of fuzzy inferences in basic teaching and application of statistics.

ACKNOWLEDGMENTS

The authors thank Glen Meeden and Charlie Geyer for personal communications about their research. A. Agresti would like to thank Professor Matilde Bini for arranging his visiting appointment at the University of Firenze during which these comments were prepared.

REFERENCES

AGRESTI, A. and COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52** 119–126.

ANSCOMBE, F. (1948). The validity of comparative experiments (with discussion). *J. Roy. Statist. Soc. Ser. A* **111** 181–211.

BLYTH, C. R. and HUTCHINSON, D. W. (1960). Table of Neyman—shortest unbiased confidence intervals for the binomial parameter. *Biometrika* **47** 381–391.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16** 101–133.

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.

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

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

NEWCOMBE, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* **17** 857–872.

NEYMAN, J. (1935). On the problem of confidence limits. *Ann. Math. Statist.* **6** 111–116.

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

PIERCE, D. A. and PETERS, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 701–737.

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

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 (with discussion). *J. Amer. Statist. Assoc.* **93** 1294–1320.

VOLLSET, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine* **12** 809–824.