Statistical Science
2005, Vol. 20, No. 4, 375–379
DOI 10.1214/088342305000000395

Comment: Fuzzy and Randomized Confidence Intervals and P-Values

Lawrence D. Brown, T. Tony Cai and Anirban DasGupta

Professor Geyer and Professor Meeden have given us an intriguing article with much material for thought and exploration, and they deserve our congratulations. Although the idea of randomized procedures has long existed, this paper has revitalized the discussion on randomized confidence intervals and randomized P-values.

Interval estimation of a binomial proportion is a very basic but very important problem with an extensive literature. Brown, Cai and DasGupta (2001) revisited this problem and showed that the performance of the standard Wald interval, which is used extensively in textbooks and in practice, is far more erratic and inadequate than is appreciated. Several natural alternative confidence intervals for p were recommended in Brown, Cai and DasGupta (2001). See also Agresti and Coull (1998). These intervals are all what the authors call crisp intervals.

The coverage probability of these crisp confidence intervals contains significant oscillation, which is intrinsic in all crisp intervals due to the lattice structure of the binomial distributions. In the present paper, Geyer and Meeden introduce the notion of fuzzy confidence intervals with the goal to eliminate oscillation and to have the exact coverage probability. The confidence intervals are obtained by inverting families of randomized tests. In addition, the authors introduce the notion of fuzzy P-values. The introduction of the critical function ϕ as a function of three variables x, α and θ provides a unified description of fuzzy decision, fuzzy confidence interval and fuzzy P-values.

Our discussion here will focus on four issues: (1) What is new in this paper?; (2) exact versus approximate coverage; (3) expected length; (4) generalization of abstract randomized confidence intervals to simultaneous inference.

Lawrence D. Brown and T. Tony Cai are Professor and Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. Anirban DasGupta is Professor, Department of Statistics, Purdue University, West Lafayette, Indiana 47907, USA (e-mail: dasgupta@stat.purdue.edu).

1. WHAT IS NEW IN THIS PAPER?

As the authors observe, the notion of a randomized confidence interval has a long history. Such intervals are a natural consequence of the formulation of randomized tests in the Neyman–Pearson lemma, and appear in Lehmann (1959, page 81; 2nd ed., 1986, page 93), Blyth and Hutchinson (1960) and Pratt (1961). It is thus important to try to clarify which portions of the current paper are new, which represent a valuable new focus on a classical concept and which are an informative survey of key elements of that concept.

The earlier authors mentioned above, and others, realized that there are several ways to represent randomized confidence intervals. Most preferred versions in which the statistician produces a particular interval. In view of the discussion in the present Section 1.4, it appears this is what the authors would call a realized randomized interval, but some preferred what the present paper would refer to as an abstract randomized interval. For example, Lehmann (1959) created realized randomized intervals by introducing an auxiliary independent uniform random variable. Pratt (1961) created such intervals for the binomial problem by the equivalent device of constructing nonrandomized intervals on the basis of observation of $X + U$, where X is binomial and U is an independent uniform$(0, 1)$ random variable. However, the discussion in Cohen and Strawderman (1973), Brown and Cohen (1995) and Brown, Casella and Hwang (1995) is in terms of abstract randomized intervals.

From a formal mathematical perspective there seems to be nothing about the definition of abstract randomized intervals here that is different from the treatment in these earlier papers. Thus, while the descriptive language is different, the formal structure here for "fuzzy intervals" is the same as that for abstract randomized intervals. We find one feature in this new descriptive language to be very appealing: the pictorial representation in the figure near (1.2). This representation allows the user to think of abstract randomized intervals as a minor extension of ordinary ones, and helps the statistician in some circumstances to avoid the need for more precise but cumbersome statements like "the probability is 30% that I have 95% confidence in the

value $p = 0.75$."

We do not recall having seen before a definition of randomized P-values like that in the paper. The authors' clever definition flows in a nice way from their representation in (1.1) of ϕ as a function of three variables, whereas most authors in the earlier literature have fixed α and have then written ϕ as a function of only the two variables x and θ, depending implicitly on the fixed α. However, the notion of a P-value as a distribution or corresponding density, rather than as a number, is a conceptual complication that many users may find undesirable.

2. EXACT VERSUS APPROXIMATE COVERAGE

It is traditionally required that confidence procedures should have coverage probability with a minimum at least the nominal level. We feel that this is an overly conservative requirement and take a different perspective. Most statistical models are only approximately correct and many inferential procedures are only asymptotically valid. So the coverage probability of confidence intervals can only be expected to be approximately the nominal value. Thus, when we claim a certain nominal level of coverage probability, we clearly intend to convey that the coverage of the confidence procedure is close to the nominal level, rather than to guarantee it is at least or exactly the nominal level. Once one accepts that all confidence level statements are only approximate, there seems to be no point in introducing the additional complications inherent in randomization.

As we noted in Brown, Cai and DasGupta (2001), it is true that the binomial model has a somewhat special feature relative to this general discussion. There are indeed practical situations where one is confident that the binomial model holds with very high precision, and asymptotics are not required to construct practical procedures or evaluate their properties, although asymptotic calculations can be useful in both regards. However, the lattice structure of the binomial distribution introduces a related barrier to the construction of exact confidence procedures.

In the Introduction the authors present the continuity-corrected Wilson interval as an example to show the "bad behavior" of conventional confidence intervals. However, this continuity-corrected interval is not a good choice of nonrandomized procedure. Figure 1 in the paper plots the coverage probability of the continuity-corrected Wilson interval for $n = 10$. It is clear that the nominal coverage probability of 95% is misleading as a statement about the average coverage probability and not very helpful as a statement about the minimum coverage. This example reiterates a point made in Brown, Cai and DasGupta (2001) and DasGupta and Zhang (2005) that these continuity-corrected Wilson intervals are not desirable from any perspective. They have extremely conservative coverage properties, though they are also not, in principle, guaranteed to be conservative everywhere. This lack of conservativity can be clearly seen from the two sharp downward spikes in Figure 1 of the paper. These confidence intervals are not precise in the sense that they are unnecessarily long in terms of the expected length. Even if one's goal is to produce conservative intervals, the continuity-corrected intervals will be very inefficient relative to Blyth–Still or even Clopper–Pearson.

Figure 1 herein plots the coverage probability for $n = 10$ of the ordinary 95% Wilson interval with modification at the boundaries as described in Brown, Cai and DasGupta (2001). The coverage probability of this interval is centered at around 95% and has an average value of 0.959, whereas the coverage probability of the

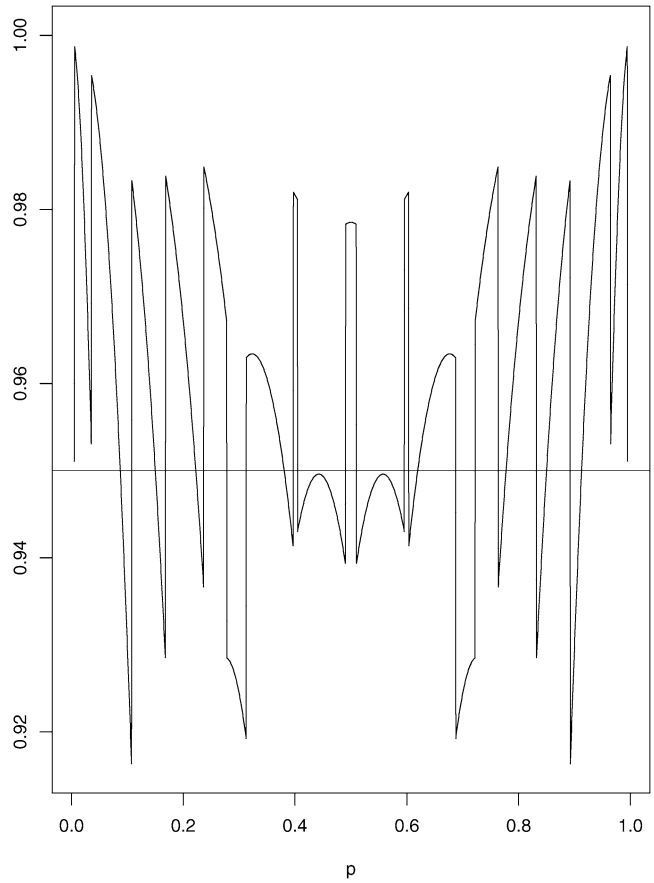

FIG. 1. *Coverage probability of the nominal* 95*% modified Wilson interval for* $n = 10$. *The boundary modifications are as described in Brown, Cai and DasGupta* (2001).

continuity-corrected version as given in Figure 1 of the paper centers around approximately 98% and has an average value of 0.982.

3. EXPECTED LENGTH

In this paper the authors focus on the coverage probability of the confidence intervals. In addition to the coverage probability, the expected length is also very important in evaluation of a confidence interval. For a crisp interval the length of the interval can be defined easily as the difference between the upper and lower limits. For an abstract randomized interval, as observed in Cohen and Strawderman (1973), a pseudo length can be defined similarly as

$$L_n(x,\alpha) = \int_0^1 \big(1 - \phi(x,\alpha,p)\big)\,dp$$

and the expected length is defined as

$$\begin{aligned}\text{expected length} &= E_{n,p} L_n(X,\alpha)\\ &= \sum_{x=0}^{n} L_n(x,\alpha)\binom{n}{x} p^x (1-p)^{n-x}.\end{aligned}$$

The average expected length is then just the integral $\int_0^1 E_{n,p} L_n(X,\alpha)\,dp$.

Even if one plans to use a nonrandomized interval, it is worth investigating the performance of randomized intervals to provide a benchmark for performance of potential nonrandomized procedures. Within the class of $(1-\alpha)$-level abstract randomized confidence intervals, it is natural to seek the one which minimizes the average expected length. The question is, "Which family of randomized tests should be used for the construction of abstract randomized confidence intervals?" Although the UMPU tests have the desirable unbiasedness property as hypothesis tests, their inversion does not yield shortest confidence intervals. A similar phenomenon occurs for point estimators. An optimal point estimator (e.g., the sample proportion $\hat{p}$) is not necessarily a good choice for the center of a confidence interval.

Pratt (1961) showed that the optimal confidence procedure which minimizes the average expected length is the inversion of the family of tests

$$H_0 : p = p_0 \quad \text{versus} \quad H_a : p \text{ is uniform}(0,1).$$

The abstract randomized interval given in the present paper is obtained by inversion of the two-sided UMPU test. We plot in Figure 2 herein for $n = 5$–20 the average expected lengths of three 95% confidence intervals: Pratt's interval, the modified Wilson interval, and the abstract randomized interval obtained from the inversion of the UMPU test.

The comparison is clear and consistent as n changes. For small n, the average expected length of the abstract randomized interval inverted from the UMPU test is noticeably larger than those of Pratt's interval and the modified Wilson interval. For example, for n up to 10, the average expected length of the interval inverted from the UMPU test is larger than that of Pratt's interval by 0.065 (at $n = 5$) to 0.025 (at $n = 10$), and this can be significant in practice. Thus if one insists on using a randomized confidence procedure so as to have precise coverage probability, Pratt's interval is much preferred to the inversion of the UMPU test. In addition, Figure 2 also shows that the modified Wilson interval dominates the UMPU randomized interval in terms of the expected length and is nearly competitive in this sense with Pratt's interval. We also note that its average coverage probability is 0.959, so that in this

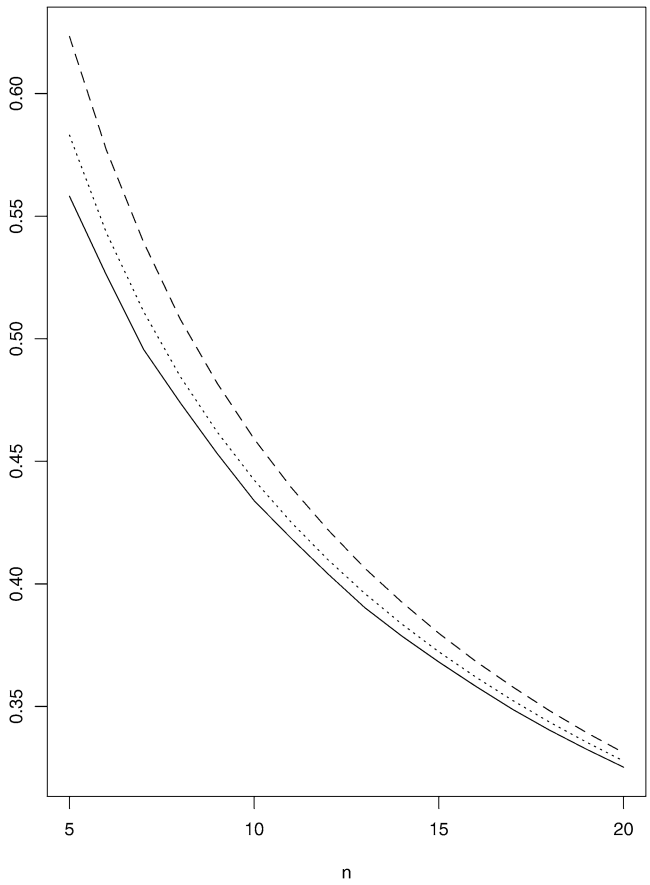

FIG. 2. *The average expected lengths of Pratt's interval* (*bottom solid line*), *the modified Wilson interval* (*middle dotted line*) *and the abstract randomized interval inverted from the UMPU test* (*top dashed line*) *for* $n = 5$–20 *and* $\alpha = 0.05$.

sense it is more conservative than either the Pratt or UMPU intervals.

[Pratt noted that his intervals can fail to have the intuitively desirable property that $1-\phi(x,\alpha,\theta)$ is a unimodal function in θ for every fixed x and α. However, it is possible to make small modifications to his intervals so that they have this property, and are only very slightly conservative in terms of coverage and average length.]

This example demonstrates that nonrandomized intervals can be very competitive in performance with the best choice of randomized intervals as long as the statistician adopts the objective that confidence levels are only approximate rather than exact. Given this fact, we see no point in using randomized procedures either in their original form or in the fuzzy interpretation introduced in the present paper.

4. GENERALIZATION TO SIMULTANEOUS INFERENCE

The idea of abstract randomized confidence intervals can be extended to that of a simultaneous confidence region for multiple binomial proportions. Suppose that $X_i \sim \text{Binom}(n_i, p_i)$ for $i = 1, 2$, and X_1 and X_2 are not necessarily independent. We wish to construct a $(1-\alpha)$-level confidence region for the proportions (p_1, p_2). The idea of abstract randomized confidence intervals can be used via the randomization version of Bonferroni's inequality to construct such an abstract randomized confidence region for (p_1, p_2). Indeed, the randomized confidence region can be simply set via

$$\left(1-\phi_{n_1}\left(X_1, \frac{\alpha}{2}, p_1\right)\right)\left(1-\phi_{n_2}\left(X_2, \frac{\alpha}{2}, p_2\right)\right),$$

where $\phi_{n_i}(X_i, \frac{\alpha}{2}, p_i)$, $i = 1, 2$, are the critical functions for the corresponding one-parameter problem. Here we use the subscript n_i to indicate the dependence of the critical function on the parameters n_i. Note that individually $\phi_{n_i}(X_i, \frac{\alpha}{2}, p_i)$ are $\frac{\alpha}{2}$-level critical functions for $i = 1, 2$. Since $E_{p_i}\phi_{n_i}(X_i, \frac{\alpha}{2}, p_i) = \frac{\alpha}{2}$ for $i = 1, 2$, the coverage probability of the abstract randomized confidence region satisfies

$$\begin{aligned} &E\left\{\left(1-\phi_{n_1}\left(X_1, \frac{\alpha}{2}, p_1\right)\right)\left(1-\phi_{n_2}\left(X_2, \frac{\alpha}{2}, p_2\right)\right)\right\} \\ &\quad = 1-\alpha+E\phi_{n_1}\left(X_1, \frac{\alpha}{2}, p_1\right)\phi_{n_2}\left(X_2, \frac{\alpha}{2}, p_2\right) \\ &\quad \geq 1-\alpha \end{aligned}$$

and thus the confidence region has coverage probability of at least $1-\alpha$. Let us look at one example.

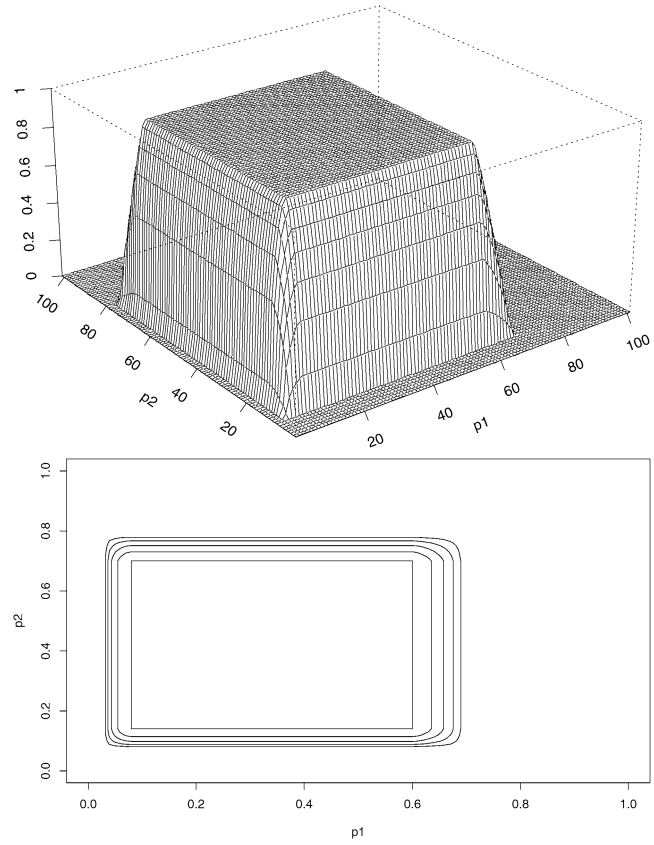

FIG. 3. *Perspective plot* (*top*) *and contour plot* (*bottom*) *of the* 95% *abstract randomized confidence region for* (p_1, p_2) *with* $n_1 = n_2 = 10$, $X_1 = 3$ *and* $X_2 = 4$.

Suppose $n_1 = n_2 = 10$, $X_1 = 3$ and $X_2 = 4$. Figure 3 presents a perspective plot of the 95% confidence region for (p_1, p_2) as constructed above as well as a corresponding contour plot of the fuzzy set.

Note that if the univariate functions $p_i \to 1-\phi_{n_i}(X_i, \frac{\alpha}{2}, p_i)$, $i = 1, 2$, are both concave in p_i over their support, then it is easy to show that the level sets of the abstract randomized confidence region $(1-\phi_{n_1}(X_1, \frac{\alpha}{2}, p_1))(1-\phi_{n_2}(X_2, \frac{\alpha}{2}, p_2))$ are convex. Consequently, all realized randomized confidence regions are convex. See, for example, the contour plot in Figure 3. This is an appealing property in applications.

ACKNOWLEDGMENTS

The research of Lawrence Brown and Tony Cai was supported in part by National Science Foundation Grants DMS-04-05716 and DMS-03-06576, respectively.

REFERENCES

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

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.

BROWN, L. D., CASELLA, G. and HWANG, J. T. G. (1995). Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *J. Amer. Statist. Assoc.* **90** 880–889.

BROWN, L. D. and COHEN, A. (1995). Complete classes for confidence set estimation. *Statist. Sinica* **5** 291–301.

COHEN, A. and STRAWDERMAN, W. E. (1973). Admissibility implications for different criteria in confidence estimation. *Ann. Statist.* **1** 363–366.

DASGUPTA, A. and ZHANG, T. (2005). Inference for binomial and multinomial parameters. In *Encyclopedia of Statistical Sciences*, 2nd ed. (N. Balakrishnan, C. Read and B. Vidakovic, eds.). Wiley, New York. To appear.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York. (2nd ed., Wiley, 1986; Springer, 1997.)

PRATT, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 549–567.