Statistical Science
2005, Vol. 20, No. 4, 372–374
DOI 10.1214/088342305000000377

Comment: Fuzzy and Randomized Confidence Intervals and P-Values

Roger L. Berger and George Casella

1. INTRODUCTION

We thank Professor Geyer and Professor Meeden for their thought-provoking article. We hope to also be thought-provoking in response, for we pretty much disagree with their position.

The fuzzy procedures proposed by the authors result from examining the test function, $\phi(x, \alpha, \theta)$ in three different ways, as a function of each of the three variables. This is an interesting exercise, which has not been done before in this way, and the authors are to be commended for their innovation. However, we think the resulting procedures will be of limited practical interest.

The authors start with the belief that discontinuous coverage probability functions are somehow inherently bad, saying that they "perform badly" and "behave very badly," and refer to their properties as "flaws." The new fuzzy procedures eliminate these flaws by having coverage probabilities that are exactly equal to $1 - \alpha$ and test sizes that are exactly equal to α. However, these flaws are merely the properties of discrete data, showing us the limit of the possible inference. To go beyond the inherent limitations of the data is to base inference on mathematical fictions. In particular, oscillations are just a feature of coverage probability with discrete data, and there is no principle that says coverage probability functions should be continuous. Although it is probably good if a coverage probability function stays close to $1-\alpha$, so the intervals of Blaker (2000) might be preferred to those of Clopper and Pearson, we do not see a need (or a way!) to eliminate discontinuities.

Procedures already exist that have coverage probabilities exactly equal to $1 - \alpha$ and sizes equal to α; they are classical randomized procedures. However, randomized procedures are unpalatable in actual data analysis, because, as the authors state, users object to a procedure that can give different answers for the exact same data. Unfortunately, the fuzzy procedures proposed by these authors are closely related to, in some cases almost indistinguishable from, randomized procedures. As such, we believe they will be equally unpalatable for practical inference. The fuzzy procedures do not give different answers for the same data; instead they give a single, different, harder to interpret answer for a given set of data. When the fuzzy procedures are used to produce confidence intervals and P-values in the usual sense, they simply result in classical randomized procedures.

2. WHAT IS FUZZY?

The description of the fuzzy confidence sets and Figure 2 are interesting. Instead of stating an interval of θ values as the inference, as classical nonrandomized and randomized confidence intervals do, the inference from a fuzzy confidence interval is a function, examples of which are shown in Figure 2. They are somewhat appealing, with their representation of the uncertainty of the inclusion probabilities of the endpoints, but will these functions be useful to or interpretable by researchers?

In classical confidence intervals there is one kind of uncertainty quantified by the confidence coefficient, $1 - \alpha$. This still is present in fuzzy intervals, but in fuzzy intervals there is a second uncertainty about the endpoints of the interval, represented by the ascending and descending portions of the functions in Figure 2. We all know the difficulty in teaching students the correct interpretation of the uncertainty quantified in $1 - \alpha$. How much more open to misinterpretation will be the uncertainty in the endpoints? The authors say that randomization is a "notoriously tricky concept," but are "partial coverage," "degree of membership" and "degree of compatibility" any less tricky?

To overcome this difficulty in interpretation, one can use the fuzzy interval to produce a realized randomized interval as described in Section 2.1. However, a realized randomized interval is just a classical randomized

Roger Berger is Professor and Chair, Department of Mathematical Sciences and Applied Computing, Arizona State University at the West campus, Phoenix, Arizona 85069, USA (e-mail: Roger.Berger@asu.edu). George Casella is Distinguished Professor and Chair, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA (e-mail: casella@stat.ufl.edu).

confidence interval. It has the fatal flaw: The same data can lead to different intervals, which, of course, violates every reasonable principle of inference.

It is also not clear to us why we should not interpret the fuzzy confidence interval as the conditional probability of including θ given that $X = x$ is observed. The authors state "fuzzy set theorists have taken great pains to distinguish their subject from probability theory," but then look at (1.3) or the second display in Section 2.1. These say to us that $1 - \phi(x, \alpha, \theta)$ is the conditional probability of including θ in the interval given $X = x$ is observed. A few lines below (1.3) the authors say any interpretation that reflects (1.3) is correct. Why not the conditional probability interpretation?

Thus, it is not clear to us (notice that we did not say, "It is fuzzy to us...") if fuzzy = probability here. Although the fuzzy set theorists claim to not do probability, Geyer and Meeden admit they are not using the full blown fuzzy set theory, only some definitions. A principle aim of fuzzy confidence sets is to achieve a coverage probability of exactly $1-\alpha$. To do this, the fuzzy confidence set is interpreted as giving conditional probabilities. With the few definitions the authors have used, they have not distinguished themselves from conditional probability and randomized procedures very much.

3. A P-VALUE BY ANY OTHER NAME...

Fuzzy P-values are the companion to fuzzy inclusion probabilities. The function (1.1c) is given two interpretations by the authors. It can be interpreted as the membership function of a fuzzy set, the set of α values for which H_0 can be rejected. Although this interpretation seems more in line with the topic of this paper, the authors do not discuss this interpretation much in Section 1.4.3. They spend most of the discussion on the second interpretation, and so will we.

The second interpretation of (1.1c) is that it is the distribution function of a random variable, P, called the abstract randomized P-value. Again, we find the interpretation of this distribution, and its distinction from classical randomized P-values, as difficult as the interpretation of fuzzy confidence intervals.

The authors' interpretation of abstract randomized P-values and how to use them is given in the last two paragraphs of Section 1.4.3. They say, "The null hypothesis is to be rejected for all $\alpha \geq P$." Then in Section 2.2 the authors describe the use of the realized randomized P-value as "the test that rejects H_0 when $P \leq \alpha$ is the traditional randomized test." We see virtually no distinction between these two descriptions. Apparently, using their preferred abstract randomized P-value, the authors want the researcher to define the probability distribution of P, but not actually sample from it to determine if α is greater than or equal to the observed value. Somehow, one is to "interpret" this probability distribution to provide some kind of degree of evidence against H_0. We can not imagine the common researcher doing anything but one of two things:

1. Reject only if α is greater than the complete support of P. This is the classical nonrandomized P-value.
2. Generate a realized value $P = p$ from the distribution and reject only if $\alpha \geq p$. This is the classical randomized P-value.

Unfortunately, fuzzy P-values or abstract randomized P-values do not add new insight in these cases.

The notion of interpreting the abstract randomized P-value as degree of evidence against H_0, without making an accept H_0 or reject H_0 decision, seems quite Fisherian. As Christensen (2005) interpreted Fisher, "an α level should never be chosen, a scientist should simply evaluate the evidence embodied in the p value." However, this seems to ignore the large number of practical situations in which accept or reject decisions need to be made. We are thinking of quality control decisions or drug approval decisions by the Food and Drug Administration (FDA), for example. Any attempt to objectively operationalize such decisions using abstract randomized P-values will result in a classical randomized or nonrandomized test, as we described in the previous paragraph. However, a randomized test will not be used in any practical situation. (We hope the FDA is not flipping coins!) So in the end, some nonrandomized decision rule will be defined. Thus, it seems that the effort of defining an abstract randomized P-value will have little impact.

4. IN THE END

Although constructs such as randomization and fuzziness may provide more eye-pleasing curves of coverage probabilities and P-values, in the end they do not provide the experimenter with an inference that is an improvement over procedures which treat the data as discrete. Is not improving the inference what it is all about?

ACKNOWLEDGMENT

This research was supported by National Science Foundation Grant DMS-04-05543.

REFERENCES

BLAKER, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canad. J. Statist.* **28** 783–798. [Correction **29** (2001) 681.]

CHRISTENSEN, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *Amer. Statist.* **59** 121–126.