Statistical Science
2005, Vol. 20, No. 4, 358–366
DOI 10.1214/088342305000000340
© Institute of Mathematical Statistics, 2005

Fuzzy and Randomized Confidence Intervals and P-Values

Charles J. Geyer and Glen D. Meeden

Abstract. The optimal hypothesis tests for the binomial distribution and some other discrete distributions are uniformly most powerful (UMP) one-tailed and UMP unbiased (UMPU) two-tailed randomized tests. Conventional confidence intervals are not dual to randomized tests and perform badly on discrete data at small and moderate sample sizes. We introduce a new confidence interval notion, called fuzzy confidence intervals, that is dual to and inherits the exactness and optimality of UMP and UMPU tests. We also introduce a new P-value notion, called fuzzy P-values or abstract randomized P-values, that also inherits the same exactness and optimality.

Key words and phrases: Confidence interval, P-value, hypothesis test, uniformly most powerful unbiased (UMP and UMPU), fuzzy set theory, randomized test.

1. INTRODUCTION

1.1 Bad Behavior of Conventional Confidence Intervals

It has long been recognized that conventional confidence intervals, which we also call crisp confidence intervals, using a term from fuzzy set theory, can perform poorly for discrete data. A recent article (Brown, Cai and DasGupta, 2001) reviewed *crisp* confidence intervals for binomial models. The authors and discussants of that paper do recommend some crisp confidence intervals (not all recommending the same intervals), and the crisp confidence intervals they recommend are indeed better than the intervals they dislike (for some definitions of "better"). However, even the best crisp confidence intervals behave very badly. The actual achieved confidence level oscillates wildly as a function of both the true unknown parameter value and the sample size. See our Figure 1, Figures 1–5, 10 and 11 in Brown, Cai and DasGupta (2001), Figures 4 and 5 in Agresti and Coull (1998) and Figure 1 in Casella (2001).

It is important to recognize that the behavior of all crisp intervals for discrete data must exhibit oscillatory behavior similar to that shown in Figure 1. The fundamental reason is discreteness. When the data x are discrete, so are the endpoints $l(x)$ and $u(x)$ of possible crisp confidence intervals. As the parameter θ passes from just inside to just outside a possible confidence interval $(l(x), u(x))$, the coverage probability jumps discontinuously by $\Pr_\theta(X = x)$. This flaw is unavoidable—an irreconcilable conflict between crisp confidence intervals and discrete data.

Standard asymptotic theory says that as the sample size goes to infinity, the oscillations get smaller in small neighborhoods of one fixed parameter value θ in the interior of the parameter space. For the binomial distribution, this means the oscillations get smaller for success probability θ not near 0 or 1, but the oscillations remain large for shockingly large sample sizes (Brown, Cai and DasGupta, 2001) and remain large for all sample sizes for θ sufficiently near 0 and 1. The inherent flaws of the crisp confidence interval idea suggest there should be a better approach to this problem.

1.2 Randomized Tests and Confidence Intervals

The testing problem for discrete models was solved long ago by the introduction of randomized tests. For the binomial distribution and many other discrete distributions there exist uniformly most powerful (UMP) one-tailed tests and UMP unbiased (UMPU) two-tailed tests (Lehmann, 1959, Chapters 3 and 4). These tests are optimal procedures.

Charles J. Geyer and Glen D. Meeden are Professors, School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, USA (e-mail: charlie@stat.umn.edu; glen@stat.umn.edu).

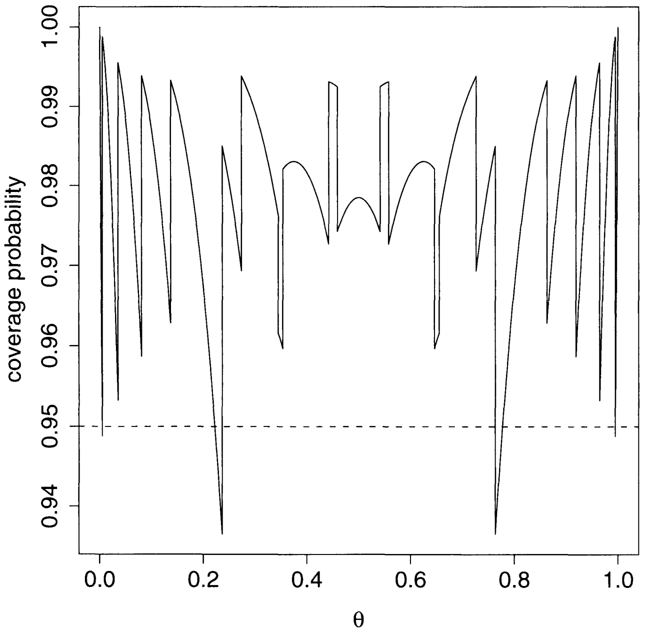


FIG. 1. *Coverage probability of the nominal* 95% *confidence interval for the binomial distribution with sample size* $n = 10$ *calculated by the function* `prop.test` *in the R statistical computing language* (*R Development Core Team*, 2004). *This is the Wilson* (*also called, score*) *interval with continuity correction and modifications when* $x = 0$ *or* $x = n$. *The dashed line is the nominal level. The solid line is the coverage probability of the interval as a function of the success probability* θ.

Tests and confidence intervals are dual notions. Hence randomized confidence intervals based on these tests can achieve their nominal coverage probability and inherit the optimality of these tests. For the binomial distribution Blyth and Hutchinson (1960) gave tables for constructing such randomized intervals (for sample sizes up to 50 and coverage probabilities 0.95 and 0.99). Due to the discreteness of the tables, the randomized intervals they produce are not close to exact, hence a computer should now be used instead of these tables (see Sections 2 and 3.3 below).

These randomized tests and intervals have been little used in practice, however, because users object to a procedure that can give different answers for the exact same data due to the randomization. It is annoying that two statisticians analyzing exactly the same data and using exactly the same procedure can nevertheless report different results. We can avoid the arbitrariness of randomization while keeping the beautiful theory of these procedures by a simple change of viewpoint to what we call *fuzzy* and *abstract randomized* concepts.

1.3 Fuzzy Set Theory

We actually use only some concepts and terminology of fuzzy set theory, which can be found in the most elementary of introductions to the subject (Klir, St. Clair and Yuan, 1997). We do not need the theory itself.

A *fuzzy set* A in a space S is characterized by its *membership function*, which is a map $I_A : S \to [0, 1]$. The value $I_A(x)$ is the "degree of membership" of the point x in the fuzzy set A or the "degree of compatibility ... with the concept represented by the fuzzy set" (Klir, St. Clair and Yuan, 1997, page 75). The idea is that we are uncertain about whether x is in or out of the set A. The value $I_A(x)$ represents how much we think x is in the fuzzy set A: The closer $I_A(x)$ is to 1.0, the more we think x is in A; the closer $I_A(x)$ is to 0.0, the less we think x is in A.

A fuzzy set whose membership function only takes on the values 0 or 1 is called *crisp*. For a crisp set, the membership function I_A is the same thing as the indicator function of an ordinary set A. Thus "crisp" is just the fuzzy set theory way of saying "ordinary," and "membership function" is the fuzzy set theory way of saying "indicator function." The *complement* of a fuzzy set A that has membership function I_A is the fuzzy set B that has membership function $I_B = 1 - I_A$ (Klir, St. Clair and Yuan, 1997, page 90).

If I_A is the membership function of a fuzzy set A, the γ*-cut* of A (Klir, St. Clair and Yuan, 1997, Section 5.1) is the crisp set

$$^{\gamma}I_A = \{x : I_A(x) \geq \gamma\}.$$

Clearly, knowing all the γ-cuts for $0 \leq \gamma \leq 1$ tells us everything there is to know about the fuzzy set A. The 1-cut is also called the *core* of A, denoted core(A) and the set

$$\text{supp}(A) = \bigcup_{\gamma > 0} {}^{\gamma}I_A = \{x : I_A(x) > 0\}$$

is called the *support* of A (Klir, St. Clair and Yuan, 1997, page 100). A fuzzy set is said to be *convex* if each γ-cut is convex (Klir, St. Clair and Yuan, 1997, pages 104–105).

1.4 Fuzzy and Abstract Randomized Procedures

Let ϕ be the critical function of a randomized test. This is a function from the sample space to the interval $[0, 1]$. Since it is a function on the sample space, it is usually written $\phi(x)$, but since the function also depends on the size of the test and the hypothesized value of the parameter, we prefer to write it $\phi(x, \alpha, \theta)$, where x is the data, α is the significance level (size) and θ is the hypothesized value of the parameter under the null hypothesis. A *randomized test* of size α

rejects $H_0 : \theta = \theta_0$ when data x are observed with probability $\phi(x, \alpha, \theta_0)$. If the test is one-tailed with compound null and alternative hypotheses, say $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, then it must be equivalent to a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ to fit into our scheme.

The only well-known examples come from UMP and UMPU theory, but the exact form of the critical function does not matter for the discussion in this section. Curious readers who have forgotten UMP and UMPU theory can look at (3.1) and (3.4) below.

Now we come to the two main ideas of this paper. The first is that the critical function ϕ can be viewed as three different functions:

$$x \mapsto \phi(x, \alpha, \theta_0), \tag{1.1a}$$

$$\theta \mapsto 1 - \phi(x, \alpha, \theta), \tag{1.1b}$$

$$\alpha \mapsto \phi(x, \alpha, \theta_0). \tag{1.1c}$$

- For fixed α and θ_0, the function (1.1a) is called the *fuzzy decision* or the *abstract randomized decision* for the size α test of $H_0 : \theta = \theta_0$.
- For fixed x and α, the function (1.1b) is called (the membership function of) the *fuzzy confidence interval* with coverage $1 - \alpha$.
- For fixed x and θ_0, the function (1.1c) is called (the membership function of) the *fuzzy P-value* or (the distribution function of) the *abstract randomized P-value* for the test of $H_0 : \theta = \theta_0$.

Fuzzy decision and abstract randomized decision are different interpretations of the same mathematical object (Section 1.4.1), and similarly for fuzzy P-value and abstract randomized P-value (Section 1.4.3). There does not seem to be a unique abstract randomized confidence interval (Section 2.1).

The second main idea is that statistical analysis should stop with these functions. They are what a statistician or scientist should report. No additional and arbitrary randomization should be done.

1.4.1 *Fuzzy decisions.* In a situation where a classical randomized test would be done, where α and θ_0 are fixed, we think a statistician using a randomized test should just report the value $\phi(x, \alpha, \theta_0)$. We call this a *fuzzy test* and the reported value a *fuzzy decision.* (When one does not want to test at fixed α, use the fuzzy P-value described in Section 1.4.3 below.)

A statistician preferring a classical randomized test can always generate his or her own Uniform(0, 1) random variate U, and reject H_0 if $U < \phi(x, \alpha, \theta_0)$ and accept H_0 otherwise.

Of course, if an actual immediate decision is required, then the randomized test must be used. However, in scientific inference the decision is often merely metaphorical, a way to discuss results that have no effect other than whatever impression they make on readers of a paper. Such metaphorical decisions more accurately describe the data when they are left fuzzy.

If one prefers, one can also call a fuzzy decision an *abstract randomized decision*, emphasizing the distinction between an *abstract* random variable (a mathematical object that has a probability distribution) and a *realization* of the random variable (data assumed to be generated according to that probability distribution). The random variable D that takes the value reject H_0 with probability $\phi(x, \alpha, \theta)$ and takes the value accept H_0 with probability $1 - \phi(x, \alpha, \theta)$ is an abstract randomized decision. Generating a realization of D and carrying out the indicated decision is what is usually called a randomized test, but we recommend stopping with the description of D, leaving it to readers to generate a realization if they find it helpful.

1.4.2 *Fuzzy confidence intervals.* The fuzzy confidence interval (1.1b) is a function taking values between 0 and 1, and is to be interpreted as (the membership function of) a fuzzy set, the fuzzy complement of $\theta \mapsto \phi(x, \alpha, \theta)$.

As with any mathematical function, a fuzzy confidence interval is best visualized by plotting its graph. Figure 2 shows three different fuzzy confidence intervals.

The dashed curve in Figure 2 is the (graph of the membership function of) the fuzzy confidence interval for $n = 10$ and $x = 4$. It is not very different from an indicator function: the rise and fall at the edges are fairly steep; its core is the interval $0.169 \leq \theta \leq 0.660$ and its support is the interval $0.098 < \theta < 0.749$. So this fuzzy interval is not so different from (the indicator function of) a crisp interval. The amount of fuzziness is fairly small and gets smaller still for larger sample sizes.

The solid and dotted curves in the figure do not look much like (the indicator functions of) conventional confidence intervals. In particular, the core of the interval represented by the solid curve is empty. The cases $x = 0$, 1, $n - 1$ and n for the binomial are unusual (more on this in Section 3.2 below), but conventional procedures also treat these data values as special cases.

Many scientists like "error bars," which are essentially confidence intervals indicated by bars on a plot. Error bars can be made fuzzy by giving them the shape of the fuzzy confidence interval as shown below for the

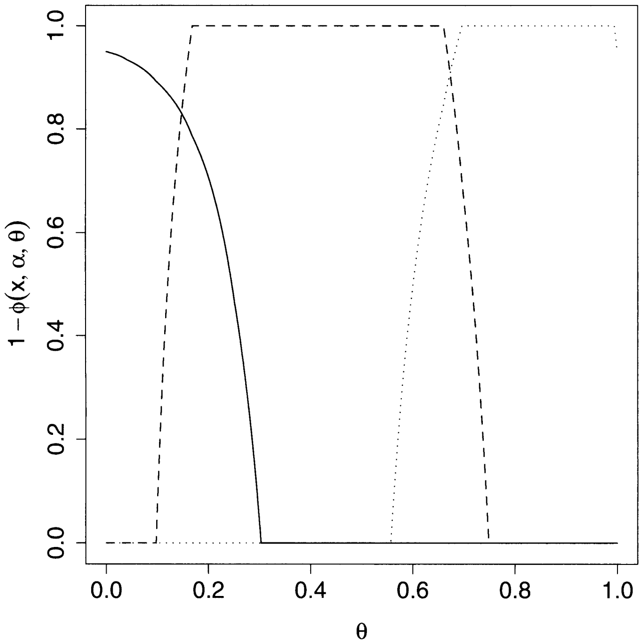
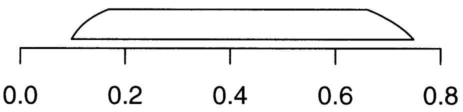


FIG. 2. *Fuzzy confidence intervals for binomial data with sample size* $n = 10$, *confidence level* $1-\alpha = 0.95$, *and observed data* $x = 0$ (*solid curve*), $x = 4$ (*dashed curve*) *and* $x = 9$ (*dotted curve*). *Note that the* $x = 0$ *curve starts at* $1-\alpha$ *at* $\theta = 0$ *and the* $x = 9$ *curve ends at* $1-\alpha$ *at* $\theta = 1$ (*Section* 3.2 *explains this behavior*). *The parameter* θ *is the probability of success.*

$x = 4$, $n = 10$ confidence interval shown as the dashed curve in Figure 2.



To interpret fuzzy confidence intervals we need a little theory. The test that has critical function ϕ is *exact* if

$$E_\theta\{\phi(X, \alpha, \theta)\} = \alpha \quad \text{for all } \alpha \text{ and } \theta. \tag{1.2}$$

Note that this trivially implies

$$E_\theta\{1-\phi(X, \alpha, \theta)\} = 1-\alpha \quad \text{for all } \alpha \text{ and } \theta. \tag{1.3}$$

The left-hand side of (1.3) is called the *coverage probability* of the fuzzy confidence interval. This makes the fuzzy confidence interval inherit the exactness of the corresponding test. Since UMP and UMPU tests are exact, so are the corresponding fuzzy confidence intervals, such as those in Figure 2.

Any interpretation of fuzzy confidence intervals that accurately reflects the mathematics embodied in (1.3) is correct. As with conventional confidence intervals, the hardest thing for naive users to absorb is that only the lucky intervals cover the unknown true parameter value: It is called a 95% confidence interval because it misses 5% of the time. Whether any particular interval covers or misses can never be known. We claim the fuzziness at the edges of a fuzzy interval is a minor part of the interpretative problem, but some readers will nevertheless want a precise interpretation. We say that when the true parameter value θ happens to be in the fuzzy edge of an interval, this only counts as partial coverage and the degree to which it counts is the degree to which θ is considered to be in the fuzzy interval, which is $1-\phi(X, \alpha, \theta)$, and this is reflected in the stated confidence level (1.3).

Our definition makes conventional confidence intervals a special case of fuzzy confidence intervals (the fuzzy intervals that just happen to be crisp). Thus our fuzzy theory is a generalization of current theory. It includes all current results. In particular, fuzzy confidence intervals based on UMP and UMPU tests for *continuous* data are automatically crisp because those UMP and UMPU tests are not randomized. So our theory only says new things about discrete data. For continuous data, it is the same old story.

There does not seem to be any simple way to treat the function (1.1b) as an abstract random variable (see also Section 2.1 below).

1.4.3 *Fuzzy and abstract randomized P-values.* For conventional P-values to even be definable, a test must have nested critical regions. For fuzzy and randomized P-values to even be definable, a test must have the fuzzy analog of nested critical regions, which is

$$\alpha_1 \le \alpha_2 \quad \text{implies} \quad \phi(x, \alpha_1, \theta) \le \phi(x, \alpha_2, \theta) \quad \text{for all } x \text{ and } \theta. \tag{1.4}$$

When the data are discrete and (1.4) holds, it can easily be shown that the fuzzy P-value (1.1c) is for each x and θ a continuous nondecreasing function that goes from 0 to 1 as α goes from 0 to 1. Thus (1.1c) has two possible interpretations: the membership function of a fuzzy set called the fuzzy P-value or the distribution function of a random variable called the abstract randomized P-value (for brevity just randomized P-value for the rest of this section). For example, consider the UMPU (two-tailed) test with binomial data $x = 10$ and $n = 10$, and null hypothesis $\theta = 0.7$. Then this function (which can be considered either the membership function of the fuzzy P-value or the distribution function of the randomized P-value) is

$$F(\alpha) = \begin{cases} 0, & \alpha \le 0, \\ 24.8 \cdot \alpha, & 0 \le \alpha \le 0.00002, \\ 0.00002 + 23.6 \cdot \alpha, & 0.00002 \le \alpha \le 0.00043, \\ 0.00066 + 22.1 \cdot \alpha, & 0.00043 \le \alpha \le 0.00429, \\ 0.0088 + 20.2 \cdot \alpha, & 0.00429 \le \alpha \le 0.0253, \\ 0.0073 + 17.7 \cdot \alpha, & 0.0253 \le \alpha \le 0.0524, \\ 1, & 0.0524 \le \alpha. \end{cases} \tag{1.5}$$

When using the alternative interpretation of (1.5) as the distribution function of a randomized P-value, perhaps neither (1.5) nor a graph of the function it defines is the best way to display the function. As with any continuous random variable, a randomized P-value is best visualized by plotting its probability density function

$$\alpha \mapsto \frac{\partial}{\partial\alpha}\phi(x, \alpha, \theta_0),$$

which is shown in Figure 3. Since (1.5) is piecewise linear, its derivative is a step function. As we show below (Section 3.2), every probability density function of a randomized P-value that corresponds to a UMP or UMPU test is a step function.

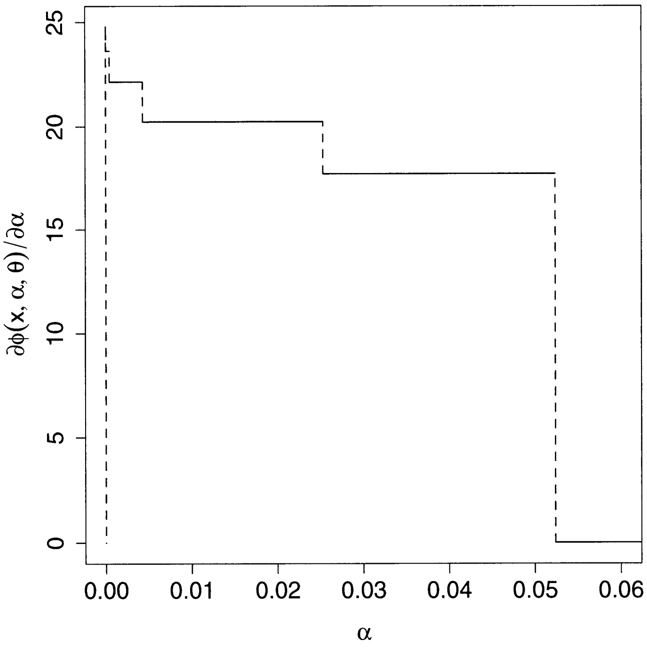


FIG. 3. *The density of the randomized P-value for the UMPU (two-tailed) test with binomial data* $x = 10$ *and* $n = 10$ *and null hypothesis* $\theta = 0.7$. *The plotted step function has five steps. The intervals of constancy are the intervals in* (1.5). *The heights are the coefficients of* α *in* (1.5).

Sometimes the step function has just one step, that is, the randomized P-value is uniformly distributed on an interval. This always happens when the test is UMP (one-tailed). For example, the upper-tailed UMP test for the same null hypothesis ($\theta_0 = 0.7$) and the same data ($x = 10$ and $n = 10$) used for Figure 3 has its randomized P-value uniformly distributed on the interval $(0, 0.028)$. This can also happen in a UMPU (two-tailed) test when the observed data x are in the long thin tail of its null distribution. For example, for the UMPU test for the same null hypothesis ($\theta_0 = 0.7$) and the same sample size ($n = 10$) used for Figure 3, the randomized P-value is uniformly distributed when $x = 0$, 1, 2, or 3, and these are the data values in the lower tail that the UMPU test always rejects at the 0.05 level. The complicated behavior seen in Figure 3 arises because the data $x = 10$ are in the short fat tail of the null distribution.

As with conventional P-values, the hardest thing for naive users to absorb is that a P-value is not a probability. Only in the special case of a point null hypothesis can a conventional P-value be interpreted as a probability. A better interpretation of a conventional P-value, at least better for our purposes here, is the least α at which the null hypothesis can be rejected.

When the P-value becomes fuzzy or randomized, there is no longer a sharp cutoff between acceptance and rejection. The fuzzy P-value gives the range of α for which the null hypothesis can be rejected as a fuzzy set. The "fuzzy edge" of this range, where the membership function is strictly between 0 and 1, allows both acceptance and rejection to varying degrees. The degree to which α is considered to be in the fuzzy P-value is the probability $\phi(x, \alpha, \theta_0)$ that the classical randomized test rejects (so the connection between fuzziness and probability is very close here).

The theory of abstract randomized P-values is also simple. By the definition of distribution function, if P is a randomized P-value, then

$$\Pr_\theta\{P \le \alpha | X\} = \phi(X, \alpha, \theta) \tag{1.6}$$

for all α and θ. Hence by iterated conditional expectation,

$$\begin{aligned} \Pr_\theta\{P \le \alpha\} &= E_\theta\{\Pr_\theta\{P \le \alpha | X\}\} \\ &= E_\theta\{\phi(X, \alpha, \theta)\} \qquad (1.7) \\ &= \alpha. \end{aligned}$$

Thus P is Uniform(0, 1) distributed marginally (not conditioning on X), and this is the sense in which a randomized P-value inherits the exactness of the corresponding randomized decision (the probability that the test rejects at level α is exactly α).

Any interpretation of abstract randomized P-values that accurately reflects the mathematics embodied in (1.7) is correct. The null hypothesis is to be rejected for all $\alpha \geq P$, but P is now random rather than deterministic. So this means rejected for all $\alpha \geq P(\omega)$, where ω ranges over some underlying probability space. The α for which we reject are random (depend on ω). Property (1.7) assures that this test is exact when the randomness in both X and P is accounted for.

As with conventional crisp P-values, there is no difficulty interpreting the extreme cases. If a randomized P-value is concentrated below 0.01, then this is strong evidence against H_0. If a randomized P-value is concentrated above 0.2, then this is evidence against H_0 so weak as to be inconsequential. If a randomized P-value is uniformly distributed on $(0.04, 0.06)$, then one has an interpretative problem, but really no more severe a problem than for crisp P-values like $P = 0.045$ or $P = 0.055$. The equivocalness of P-values of moderate size is not made worse by making them fuzzy.

2. REALIZED RANDOMIZED PROCEDURES

To each fuzzy or abstract randomized concept in the trio of decisions, confidence intervals and P-values, there is an analogous realized randomized concept. The first two of these concepts have existing literature cited in the Introduction. We do not actually recommend any of these realized randomized procedures, preferring their fuzzy or abstract randomized analogs, but we need to be clear about the relationship between fuzzy, abstract randomized and realized randomized procedures, if for no other reason than to avoid confusion.

By a *realized randomized decision* we mean the decision of a conventional randomized test. Since this is well known, we need say no more about it.

2.1 Randomized Confidence Intervals

Let

$$I_x(\theta) = 1 - \phi(x, \alpha, \theta)$$

be the fuzzy confidence interval with coverage $1 - \alpha$ for observed data x. Then the *realized randomized confidence interval* we suggest is the U-cut ${}^U I_X$ of the fuzzy interval, where U is a Uniform(0, 1) random variate.

By construction

$$\Pr_\theta\{\theta \in {}^U I_X | X\} = E_\theta\{I_X(\theta) | X\} = 1 - \phi(X, \alpha, \theta),$$

so

$$\Pr_\theta\{\theta \in {}^U I_X\} = E_\theta\{I_X(\theta)\} = 1 - \alpha$$

and the randomized confidence interval inherits exactness from the fuzzy confidence interval.

Interestingly, this is not the randomized confidence interval suggested by Blyth and Hutchinson (1960). Their intervals can be related to fuzzy confidence intervals as follows. Generate two randomized confidence intervals, which in our notation are ${}^U I_X$ and ${}^{1-U} I_X$. Then construct a new interval by taking the left endpoint from one of these and the right endpoint from the other. This only works when the fuzzy confidence interval is convex, but that is the usual case.

Yet a third recipe for generating randomized confidence intervals that also only works when the fuzzy interval is convex also generates two randomized confidence intervals, which in our notation are ${}^U I_X$ and ${}^V I_X$, where U and V are independent Uniform(0, 1) random variates. Then construct a new interval by taking the left endpoint from one of these and the right endpoint from the other.

There is, of course, no difference in performance between these three recipes, and many other recipes with identical performance are possible. Since we do not expect that randomized procedures will find much use, there is little point in trying to justify any particular recipe, but the first does have a simpler relationship to fuzzy intervals.

2.2 Randomized P-Values

A *realized randomized P-value* is a number P generated by a mechanism that gives it the probability distribution of the abstract randomized P-value, that is, the distribution with distribution function (1.1c). Property (1.7) assures us that the test that rejects H_0 when $P \leq \alpha$ is the traditional randomized test.

We show in Section 3.2 below that, for UMP and UMPU tests, the fuzzy P-value is a continuous random variable that has piecewise constant density, hence a mixture of uniforms and trivial to simulate given a uniform random number generator.

3. UMP AND UMPU FUZZY PROCEDURES

The most important case of our theory is UMP and UMPU procedures. To understand some of their behavior we need to look more deeply at that theory.

3.1 UMP

Lehmann (1959, pages 68–69) said for a one-parameter model with likelihood ratio monotone in the statistic $T(X)$ there exists a UMP test that has null hypothesis $H_0 = \{\vartheta : \vartheta \le \theta\}$, alternative hypothesis $H_1 = \{\vartheta : \vartheta > \theta\}$, significance level α and critical function ϕ defined by

$$\phi(x, \alpha, \theta) = \begin{cases} 1, & T(x) > C, \\ \gamma, & T(x) = C, \\ 0, & T(x) < C, \end{cases} \tag{3.1}$$

where the constants γ and C are determined by

$$E_\theta\{\phi(X, \alpha, \theta)\} = \alpha.$$

The description of the analogous lower-tailed test is the same except that all inequalities are reversed.

The constant C is clearly any $(1 - \alpha)$th quantile of the distribution of $T(X)$ for the parameter value θ. If C is not an atom of this distribution, then the test is effectively not randomized and the value of γ is irrelevant. Otherwise

$$\gamma = \frac{\alpha - \Pr_\theta\{T(X) > C\}}{\Pr_\theta\{T(X) = C\}}. \tag{3.2}$$

In considering the distribution function of the randomized P-value, we look at $\phi(x, \alpha, \theta)$ as a function of α for fixed x and θ; hence we look at (3.2) in the same way. Now $T(x)$ will be a $(1 - \alpha)$th quantile if

$$\Pr_\theta\{T(X) > T(x)\} \le \alpha \le \Pr_\theta\{T(X) \ge T(x)\}. \tag{3.3}$$

Since (3.2) is linear in α, so is the distribution function of the randomized P-value (where it is not 0 or 1). Hence the randomized P-value is uniformly distributed on the interval (3.3).

3.2 UMPU

Lehmann (1959, pages 126–127) said for a one-parameter exponential family model with canonical statistic $T(X)$ and canonical parameter θ there exists a UMPU test that has null hypothesis $H_0 = \{\vartheta : \vartheta = \theta\}$, alternative hypothesis $H_1 = \{\vartheta : \vartheta \ne \theta\}$, significance level α and critical function ϕ defined by

$$\phi(x, \alpha, \theta) = \begin{cases} 1, & T(x) < C_1, \\ \gamma_1, & T(x) = C_1, \\ 0, & C_1 < T(x) < C_2, \\ \gamma_2, & T(x) = C_2, \\ 1, & C_2 < T(x), \end{cases} \tag{3.4}$$

where $C_1 \le C_2$ and the constants γ_1, γ_2, C_1 and C_2 are determined by

$$E_\theta\{\phi(X, \alpha, \theta)\} = \alpha, \tag{3.5a}$$

$$E_\theta\{T(X)\phi(X, \alpha, \theta)\} = \alpha E_\theta\{T(X)\}. \tag{3.5b}$$

If $C_1 = C_2 = C$ in (3.4), then the test only depends on $\gamma_1 + \gamma_2 = \gamma$. This occurs only in a very special case. Define

$$p = \Pr_\theta\{T(X) = C\}, \tag{3.6a}$$

$$\mu = E_\theta\{T(X)\}. \tag{3.6b}$$

Then, to satisfy (3.5a) and (3.5b), we must have

$$1 - (1 - \gamma)p = \alpha,$$
$$\mu - C(1 - \gamma)p = \alpha\mu,$$

which solved for γ and C gives

$$\gamma = 1 - \frac{1 - \alpha}{p}, \tag{3.7a}$$

$$C = \mu. \tag{3.7b}$$

Thus this special case occurs only when μ is an atom of the distribution of $T(X)$ for the parameter value θ, and then only for very large significance levels: $\alpha > 1 - p$. Hence this special case is of no practical importance, although it is of some computational importance to get every case right.

Returning to the general case, assume for a second that we have particular C_1 and C_2 that work for some α and θ. With μ still defined by (3.6b) and with the definitions

$$p_i = \Pr_\theta\{T(X) = C_i\}, \quad i = 1, 2, \tag{3.8a}$$

$$p_{12} = \Pr_\theta\{C_1 < T(X) < C_2\}, \tag{3.8b}$$

$$m_{12} = E_\theta\{T(X) I_{(C_1, C_2)}[T(X)]\}, \tag{3.8c}$$

(3.5a) and (3.5b) solved for γ_1 and γ_2 become

$$\gamma_1 = 1 - \frac{(1 - \alpha)(C_2 - \mu) + m_{12} - C_2 p_{12}}{p_1(C_2 - C_1)}, \tag{3.9a}$$

$$\gamma_2 = 1 - \frac{(1 - \alpha)(\mu - C_1) - m_{12} + C_1 p_{12}}{p_2(C_2 - C_1)}. \tag{3.9b}$$

These equations are valid over the range of α (if any) such that both equations give γ_1 and γ_2 values between 0 and 1.

Since (3.9a) and (3.9b) are linear in α (when they are valid), the distribution function of the randomized P-value is piecewise linear and the density is a step function. Further analysis of this phenomenon given in the documentation for the R implementation (Geyer and Meeden, 2004) shows that the support of the randomized P-value is connected.

The UMPU test is not well defined when the null hypothesis is on the boundary of the parameter space, but (3.4), (3.5a) and (3.5b) still make sense and define a test. Since the probability and the expectation in those equations are continuous in θ, this also characterizes the behavior as θ converges to a boundary point (which we need to know to calculate fuzzy confidence intervals, which involve all θ in the parameter space).

Suppose that, in addition to the setup for UMPU tests described at the beginning of this section, the canonical statistic $T(X)$ of the exponential family has a topologically discrete distribution (concentrated on a countable set of atoms that are topologically isolated, integer valued, e.g.). Suppose the range of $T(X)$ is bounded below (as with the binomial or Poisson distribution). As the canonical parameter goes to $-\infty$, the critical function $\phi(x, \alpha, \theta)$ converges to α for x such that $T(x)$ is equal to either of its *two* smallest values and converges to 1 for all other x. A proof is given in the documentation for the R implementation (Geyer and Meeden, 2004). By symmetry, the analogous thing happens for the *two* largest values when there is an upper bound. This endpoint behavior is clearly shown for the binomial distribution in Figure 2.

3.3 Computation

To present a fuzzy or abstract randomized decision, confidence interval or P-value, one needs to be able to compute $\phi(x, \alpha, \theta)$ for every set of possible values. We have written an R package `ump` that does this for the binomial model (Geyer and Meeden, 2004). This package is available from the authors' web site `www.stat.umn.edu/geyer/fuzz/` or from the Comprehensive R Archive Network `cran.r-project.org/`.

4. DISCUSSION

We claim that fuzzy set theory combined with UMPU testing theory gives an elegant and simple solution to a well recognized problem with conventional confidence intervals for discrete data (Brown, Cai and DasGupta, 2001, and discussion). Admittedly, our solution requires a picture like our Figure 2, but conventional confidence intervals also need a picture like our Figure 1 to accurately describe their performance. Those who object to statistics that requires graphics could at least report the core and support of the fuzzy interval (see the example in Section 1.4.2). This is a crude approximation to the fuzzy interval, but is still more informative than any crisp interval.

Although randomized confidence intervals may be more familiar to statisticians than fuzzy intervals, there are two reasons why fuzzy intervals are preferable. First, nonstatistician users may find them more understandable, randomization being a notoriously tricky concept. Second, randomized intervals are not unique, as we explained in Section 2.1, whereas the fuzzy interval (1.1b) is unique.

We also claim that fuzzy and abstract randomized P-values are a solution to a problem that, although not yet widely recognized, is just as important. We have no preference between the two (one of us prefers fuzzy, the other prefers abstract randomized). Abstract randomized P-values do have the nice property that they are uniform on an interval and hence can be described by two numbers all the time for UMP one-tailed tests and some of the time for UMPU two-tailed tests.

The picture for a fuzzy confidence interval or fuzzy or abstract randomized P-value is no more complicated than a histogram and just as easy to produce using a computer. They could be taught in elementary courses. In our experience most students have a very hard time understanding conventional confidence intervals and P-values. It is not obvious that fuzzy intervals and P-values are harder to understand. The fuzzy edges of a fuzzy confidence interval may help the student understand that the confidence interval does not capture the unknown θ in an all-or-nothing way. The fuzzy edge of a fuzzy P-value may help the student understand that the number 0.05 has no magical properties.

Statisticians, especially subjective Bayesians, naturally assume that fuzzy set theory can be replaced by probability theory. However, from the origins of the subject more than 30 years ago, fuzzy set theorists have taken great pains to distinguish their subject from probability theory, and the least acquaintance with the manipulations done in fuzzy set theory reveals no resemblance at all to those of probability theory. We stress this point because it is so natural for statisticians (certain statisticians anyway) to try to find priors or posteriors somewhere in our discussion. Let us assure all readers that this paper is entirely non-Bayesian and that whatever fuzzy confidence intervals and P-values may be, they are not Bayesian. We do not say this because we are anti-Bayes. We have looked for the Bayes angle and satisfied ourselves that it just is not there. This should not be surprising. There are few less Bayesian areas of statistics than confidence intervals and P-values. Making them fuzzy does not make them Bayesian.

It is important to emphasize that the fuzzy or abstract randomized approach is not restricted to the binomial case. There is a UMP or UMPU test for any one-parameter exponential family, for example, for Poisson and negative binomial data. In multiparameter exponential families, in which the parameter of interest is a canonical parameter, one gets a UMP or UMPU conditional test based on the one-parameter exponential family obtained by conditioning on the canonical statistics for the nuisance parameters. Thus there are UMP and UMPU tests and the analogous fuzzy and abstract randomized procedures for comparison of two independent binomials or two independent Poissons or two independent negative binomials. In large contingency tables, there is not usually a single parameter of interest, but in two-by-two tables, there are the UMP and UMPU competitors of Fisher's exact test and McNemar's test.

There is nothing that says you can not use fuzzy confidence intervals and P-values whenever you have discrete data. We do not know how to extend the UMPU construction outside of exponential families, but the idea of randomized tests and their associated fuzzy tests, confidence intervals and P-values is perfectly general. In principle, they can be applied to any discrete data.

Finally, let us say that, although in this paper we have been very pedantic about distinguishing *fuzzy* from *abstract randomized* (and our first draft was a mess because it failed to be so pedantic), we do not expect anyone else to be so pedantic and are not ourselves in less formal situations. We usually do not distinguish, calling both *fuzzy*.

ACKNOWLEDGMENT

The research of G. D. M. was partially supported by National Science Foundation Grant DMS-99-71331.

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.

CASELLA, G. (2001). Comment on "Interval estimation for a binomial proportion," by Brown, Cai and DasGupta (2001). *Statist. Sci.* **16** 120–122.

GEYER, C. J. and MEEDEN, G. D. (2004). ump: An R package for UMP and UMPU tests. Available at www.stat.umn.edu/geyer/fuzz/.

KLIR, G. J., ST. CLAIR, U. H. and YUAN, B. (1997). *Fuzzy Set Theory: Foundations and Applications*. Prentice Hall, Upper Saddle River, NJ.

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

R DEVELOPMENT CORE TEAM (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at www.R-project.org.