LENGTH OF CONFIDENCE INTERVALS*

JOHN W. PRATT
Harvard University

The expected length of a confidence interval is shown to equal the integral over false values of the probability each false value is included. Thus two desiderata for choosing among confidence procedures lead to the same measure of desirability. Furthermore, by common definitions of "optimum," a procedure is optimum as regards including false values if and only if it is optimum as regards expected length. However, the procedure with minimum expected length ordinarily depends on the true value of the parameter. The possibility is explored of minimizing the average expected length, averaging according to some weighting on the possible parameter values. (This is not the same as assuming a prior distribution and using Bayes' Theorem.) The ideas are applied to the mean and variance of a normal distribution and the probability of success in binomial trials.

INTRODUCTION AND SUMMARY

To measure the desirability of a confidence interval procedure, one would like to measure in some way the extent to which the interval includes false values. Since at most one point of the interval is the true value, one natural measure is the length of the interval (if it is fixed) or the expected length (if it is random; assume that both end-points are finite). That is, the expected length of the interval is a measure of the "average extent" of the false values included.

There is another approach to the problem. A natural measure of the extent to which the confidence interval procedure includes a particular false value is the probability of including that particular value. To "average" this over all false values, one might simply integrate it over the false values. This gives an apparently different measure of the "average extent" of the false values included.

It turns out, however, that the two measures are equal. To be more specific, let $L \leq \tau \leq U$ be a confidence interval for a parameter τ. The expected length of the interval is $E\{U-L\}$. The probability that the interval includes the false value τ is $P\{L \leq \tau \leq U\}$. The fact stated above is that integrating this probability over τ (omitting the true value τ_0) gives exactly the expected length:

$$E\{U - L\} = \int_{\tau \neq \tau_0} P\{L \leq \tau \leq U\} d\tau.$$

In Section 1 we draw some conclusions from this fact, after stating it more carefully and proving it. The well-known relation of confidence intervals to tests permits another interpretation of expected length (Section 2) and shows how to minimize it (Section 3). However, the minimizing procedure generally depends on the true value of the parameter. In Section 4, accordingly, the ex-

* This research has been supported by the United States Navy through the Office of Naval Research, under contract Nonr-1866(37). Reproduction in whole or in part is permitted for any purpose of the United States Government.

pected length is integrated over the possible values of the parameter, and the resulting "average" expected length is minimized. Sections 5–9 deal with some examples. The mean of the normal distribution is discussed in Section 5. In binomial trials (Section 6) one is led to procedures previously suggested by Sterne [7] and Crow [4]. The corresponding randomized procedures may consist of two disjoint intervals. This pathology is discussed briefly in Section 7. Section 8 is concerned with the variance of the normal distribution, and illustrates the behavior of invariant procedures in the general context of this paper. Section 9 brings out the relation between the procedures of Section 8 and Bayes procedures. A generalization given in Section 10 covers one-sided problems, which are discussed in Section 11. The concluding remarks (Section 12) depend for intelligibility only slightly on Sections 1–11.

1. EXPECTED SIZE AND COVERING FALSE VALUES

Let X be the observed random variable, and suppose $R(x)$ is a confidence region for a "parameter" τ. (X need not be restricted in any way: it may be a single observation or a sample from a univariate or multivariate distribution or from a distribution on an arbitrary space. The problem need not be "parametric": for instance, the distribution sampled might be an arbitrary univariate distribution and τ its median.)

Suppose $\mu(R)$ is a measure of the size of the region R. The discussion below may be read at any of the following levels of generality.

A. τ is one-dimensional, $R(X)$ is an interval with end-points $L(X)$ and $U(X)$, and $\mu(R)=U-L$, the length of R. Then $d\mu(\tau)$ below may be interpreted as $d\tau$, so $\int_R d\mu(\tau)=\int_L^U d\tau=U-L=\mu(R)$.

B. τ is one-dimensional, R is a union of (one or more) disjoint intervals, $\mu(R)$ is the sum of their lengths, and $d\mu(\tau)=d\tau$.

C. $\tau=(\tau_1, \cdots, \tau_k)$ is k-dimensional, R is any region in k-space, $\mu(R)$ is the k-dimensional volume of R, and $d\mu(\tau)=d\tau_1 \cdots d\tau_k$.

D. τ is k-dimensional, R is any region in k-space, $\mu(R)=\int_R W(\tau_1, \cdots, \tau_k)d\tau_1 \cdots d\tau_k$, and $d\mu(\tau)=W(\tau_1, \cdots, \tau_k)d\tau_1 \cdots d\tau_k$, where W is a non-negative function.

E. τ is arbitrary, R is any region in τ-space, and μ is a σ-finite, non-atomic measure on τ-space. (Allowing atoms would change no essential conclusions of this paper, but would require the addition of another term on the right-hand side of (3) and similar equations below.)

E includes A–D. Discussion of measurability in C–E will be omitted.

In each case,

$$\mu(R) = \int_{\tau\in R} d\mu(\tau). \tag{1}$$

Since $R(X)$ is a random region, its size $\mu(R(X))$ is also random in general: hence, if size is the criterion, it is natural to consider the expected size. This turns out to be the probability that $R(X)$ includes τ integrated over the possible values of τ, that is

$$E\{\mu(R(X))\} = \int P\{R(X) \ni \tau\} d\mu(\tau). \tag{2}$$

The expectation on the left and the probability on the right must be computed under the same distribution of X.

To prove equation (2), simply substitute $R(x)$ in equation (1), integrate over the distribution P of X, and reverse the order of integration, thus:

$$E\{\mu(R(X))\} = \int\int_{\tau \in R(x)} d\mu(\tau) dP(x)$$
$$= \int\int_{R(x) \ni \tau} dP(x) d\mu(\tau) = \int P\{R(X) \ni \tau\} d\mu(\tau), \qquad \text{Q.E.D.}$$

Now the omission of a single value of τ from the range of integration on the right-hand side of equation (2) will not change the value of the integral. Therefore,

$$E\{\mu(R(X))\} = \int_{\tau \neq \tau_0} P\{R(X) \ni \tau\} d\mu(\tau). \tag{3}$$

If τ_0 is the true value of τ, this says the expected size is the probability that $R(X)$ includes τ integrated over the false values of τ.

We shall see that (except under pathological circumstances), for each distribution P of X, there is a confidence procedure which, for every τ, minimizes the probability under P of covering τ. By (2) or (3), such a procedure has minimum expected size under P for every measure μ of size. Conversely, for any particular measure μ of size, if R has minimum, and finite, expected size under P, then R uniformly minimizes the probability under P of covering false values (except possibly for a set of false values of size 0, that is, of μ-measure 0). The same statements hold if attention is restricted throughout to a suitable particular class of procedures (for instance, to "unbiased" procedures; see the next-to-last paragraph of Section 3). Thus, if there is an "optimum" procedure as regards including false values, it is also "optimum" as regards expected length, and vice versa.

2. EXPECTED SIZE AND TYPE II ERROR

Confidence regions and families of tests are associated by the relation given, for instance, by Lehmann [5, p. 174],

$$R(x) \ni \tau \quad \text{if and only if} \quad x \in A(\tau). \tag{4}$$

Here $A(\tau)$ is an acceptance region for the null hypothesis that τ is the true value. If a confidence procedure $R(x)$ is given, equation (4) defines an acceptance region $A(\tau)$ for each τ; conversely, if an acceptance region $A(\tau)$ is given for each τ, equation (4) defines a confidence procedure $R(x)$. For any τ and any distribution P of X, (4) implies

$$P\{R(X) \ni \tau\} = P\{X \in A(\tau)\}. \tag{5}$$

Thus, in particular, the probability under P that the confidence region $R(X)$ will include the false value τ equals the type II error probability at P of the associated test of the null hypothesis that τ is the true value. Substituting (5) in (2) and (3) gives another interpretation of the expected size:

$$
\begin{aligned}
E\{\mu(R(X))\} &= \int P\{X \in A(\tau)\} d\mu(\tau) \\
&= \int_{\tau \neq \tau_0} P\{X \in A(\tau)\} d\mu(\tau).
\end{aligned} \tag{6}
$$

3. MINIMIZING EXPECTED SIZE

It is now clear how to minimize the expected size under P for any particular distribution P of X. For each τ, choose a test of the null hypothesis that τ is the true value which is most powerful against the alternative P. The associated confidence region has minimum expected size under P, because it minimizes the integrand of equation (6), at each τ.

A most powerful test exists except in pathological circumstances [5, p. 81]. It can be found by the Neyman-Pearson Lemma [5, p. 65] if τ completely specifies the distribution of X. Specifically, if X has density or frequency function f under P and f_τ under τ, a most powerful test is

$$
\left.\begin{matrix} \text{reject } \tau \\ \text{accept } \tau \end{matrix}\right\} \text{ if } \frac{f(X)}{f_\tau(X)} \left\{\begin{matrix} > c(\tau), \\ < c(\tau), \end{matrix}\right. \tag{7}
$$

where $c(\tau)$ and the rejection rule at X's where $f(X)/f_\tau(X) = c(\tau)$ are chosen to give the test the required level at each τ. The confidence procedure with minimum expected size under f is that corresponding to (7), which includes or excludes τ according as (7) accepts or rejects τ.

If P is a uniform distribution, then $f(x)$ is constant and (7) becomes

$$
\begin{aligned}
&\text{reject } \tau \quad \text{if} \quad f_\tau(X) < c_1(\tau), \\
&\text{accept } \tau \quad \text{if} \quad f_\tau(X) > c_1(\tau),
\end{aligned} \tag{8}
$$

where $c_1(\tau)$ and the rule at equality are chosen to give the required level. This is an old (and perhaps old-fashioned) rule for choosing critical regions.

If P satisfies the null hypothesis that τ_0 (say) is the true value, the power against P and the size of a test of the null hypothesis that τ_0 is the true value are equal. One may select the test arbitrarily for this one value of τ, and indeed for any finite (or even denumerable) number of values of τ, since this will not affect (2) or (6).

In a particular problem, there is ordinarily a class of possible distributions of X. Nothing so far has required that P belong to that class. Even if this is required, however, under most assumptions about the class of possible distributions, the most powerful test against P and hence the confidence procedure with minimum expected size under P and minimum probability under P of including false values will not be the same for every possible distribution P.

One possibility is to take some kind of average over the possible distributions P of the expected size under P. This is discussed in the next section.

Another possibility is to restrict consideration to unbiased confidence regions, that is, regions for which the probability of including a false value is less than or equal to the probability of including the true value. By (5), a confidence region is unbiased if and only if the corresponding tests are all unbiased [5, p.

177]. Sometimes there is a uniformly most powerful unbiased test of the null hypothesis that τ is the true value for each τ. Then the associated confidence procedure will, among unbiased procedures, uniformly minimize the probability of including false values and the expected size of the confidence region for every possible distribution P and every measure μ of size, that is, uniformly in P and μ.

Another possibility in some problems is to restrict consideration to invariant confidence procedures [5, p. 243]. By (5), a confidence procedure is invariant if and only if the corresponding family of tests is an invariant family. Sometimes there is a uniformly most powerful invariant test of the null hypothesis that τ is the true value for each τ and these tests form an invariant family. Then the associated confidence procedure will, among invariant procedures, minimize the probability of including false values and the expected size of the confidence region for every possible distribution P and every measure μ of size. For a particular P, possible or not, except in pathological circumstances, there is for each τ a most powerful test against P among invariant tests of the null hypothesis that τ is the true value. If these tests form an invariant family, then the associated confidence procedure will, among invariant procedures, minimize the probability under P of including false values and the expected size under P of the confidence region for the particular P and every measure μ of size. If the most powerful invariant tests against P do not form an invariant family, then no invariant family minimizes the integrand of equation (6) at each τ, and a different method of minimizing is needed. The minimizing procedure may, in fact, depend on the measure of size used. This will be illustrated (Section 8), but not discussed in general because considerable machinery would be required.

4. AVERAGE EXPECTED SIZE

One might, instead of invoking unbiasedness or invariance, average the expected size under the different possible distributions P according to some system of weights. Specifically, let θ index the possible distributions p_θ of X, and consider the "average" expected size $\int E_\theta\{\mu(R(X))\}d\nu(\theta)$, where $d\nu(\theta)$ may be interpreted for one-dimensional θ as $d\theta$ (compare A and B, p. 550), for m-dimensional θ as $d\theta_1 \cdots d\theta_m$ (compare C) or $W(\theta_1, \cdots, \theta_m)d\theta_1 \cdots d\theta_m$ with $W \geq 0$ (compare D), or ν may be a general measure, even one with atoms (compare E). Substituting P_θ for P and $\tau(\theta)$ for τ_0 in equation (3) and integrating gives

$$\int E_\theta\{\mu(R(X))\}d\nu(\theta) = \int\int_{\tau \neq \tau(\theta)} P_\theta\{R(X) \ni \tau\}d\mu(\tau)d\nu(\theta). \tag{9}$$

This says the average expected size is the probability of including a false value integrated over false values and over the possible values of θ.

Integrating (2) and substituting (5) gives

$$\begin{aligned} \int E_\theta\{\mu(R(x))\}d\nu(\theta) &= \int\int P_\theta\{R(X) \ni \tau\}d\mu(\tau)d\nu(\theta) \\ &= \int\int P_\theta\{X \in A(\tau)\}d\nu(\theta)d\mu(\tau). \end{aligned} \tag{10}$$

Let P be the "distribution" obtained by "mixing" the distributions P_θ according to the weighting ν. In other words, let P be the marginal "distribution" of X when θ has marginal "distribution" ν and X has conditional distribution P_θ given θ. (P has total weight 1 and hence is a true probability distribution only if the same holds for ν.) Let E be the "expectation" corresponding to P. Then (2) is exactly the first equality of (10) and the first equality of (6) is exactly the second equality of (10).

The average expected size is minimized by the confidence procedure associated with a family of tests $A(\tau)$ such that, for each τ, among tests of the null hypothesis that τ is the true value, $A(\tau)$ is a most powerful test against the alternative "distribution" P just described; that is, the acceptance region $A(\tau)$ minimizes

$$P\{X \in A(\tau)\} = \int P_\theta\{X \in A(\tau)\} d\nu(\theta). \tag{11}$$

This is just like the situation of Section 3 except that P may not be a true probability distribution. A minimizing $A(\tau)$ exists except in pathological circumstances. It can be found by the Neyman-Pearson Lemma when the null hypothesis that τ is the true value is simple. Specifically, if $\tau(\theta)=\theta$ and X has a density or frequency function $f_\theta(x)$, a minimizing procedure is:

$$\left.\begin{matrix}\text{reject } \tau \\ \text{accept } \tau\end{matrix}\right\} \text{ if } \frac{\int f_\theta(X) d\nu(\theta)}{f_\tau(X)} \left\{\begin{matrix} > c(\tau), \\ < c(\tau), \end{matrix}\right. \tag{12}$$

where $c(\tau)$ and the rejection rule when equality holds are chosen to give the procedure the required level at each τ.

Even in an apparently reasonable problem, it can turn out that the minimizing region is not always an interval (Section 6).

The minimizing procedure for a particular weighting ν may have infinite average expected size for some measure μ of size. Then the average expected size of every procedure is infinite for this μ. However, the minimizing procedure for the weighting ν minimizes the average expected size for all μ, including trivially those μ for which the minimum is infinite. It will also nearly minimize the average expected size for a weighting ν' near ν, and the minimum for a particular μ may be finite under ν' though infinite under ν.

Infinite minima arise most naturally when the total weight of ν is infinite, but the preceding paragraph applies even if the total weight is finite, and even (with trivial alterations of terminology) to Section 3.

5. EXAMPLE. MEAN OF A NORMAL DISTRIBUTION

If X has a normal distribution with a known variance of 1, the usual level $1-\alpha$ confidence interval for the mean τ is

$$X - \xi_{\alpha/2} \leq \tau \leq X + \xi_{\alpha/2}, \tag{13}$$

where $\xi_{\alpha/2}$ is the upper $\alpha/2$-point of a normal distribution with mean 0 and variance 1. The length of this interval is constant at $2\xi_{\alpha/2}$.

We will now find the procedure with minimum expected length when the mean is 0. For $\tau \neq 0$, a most powerful test of the null hypothesis τ against the alternative 0 accepts τ for

$$\begin{aligned} X &\leq \tau + \xi_\alpha \quad \text{if} \quad \tau < 0; \\ X &\geq \tau - \xi_\alpha \quad \text{if} \quad \tau > 0. \end{aligned} \tag{14}$$

At $\tau = 0$, any test may be used; for convenience, let all x be in the acceptance region for the null hypothesis $\tau = 0$. The interval of all τ which would be accepted by the above tests minimizes the expected length when the mean is 0. This interval is (see Figure 1)

$$\begin{aligned} X - \xi_\alpha \leq \tau \leq 0 \qquad & \text{if} \quad X \leq -\xi_\alpha; \\ X - \xi_\alpha \leq \tau \leq X + \xi_\alpha \qquad & \text{if} \ -\xi_\alpha < X < \xi_\alpha; \\ 0 \leq \tau \leq X + \xi_\alpha \qquad & \text{if} \quad \xi_\alpha \leq X. \end{aligned} \tag{15}$$

This can be written $\min\{0, X - \xi_\alpha\} \leq \tau \leq \max\{0, X + \xi_\alpha\}$.

Although this interval may be arbitrarily long, when the mean is 0 it will have length $2\xi_\alpha$ with probability $1-2\alpha$ and expected length

$$2[(1-\alpha)\xi_\alpha + (2\pi)^{-1/2}e^{-\xi_\alpha^2/2}].$$

Table 1 shows how much is gained in terms of expected length, when the mean is 0, by using (15) instead of (13). The last line gives the efficiency of (13) with respect to (15) in the sense of the ratio of the sample sizes needed to achieve the same expected length when the mean is 0.

TABLE 1. COMPARISON OF TWO CONFIDENCE INTERVALS FOR A NORMAL MEAN

Confidence level, $1-\alpha$	.80	.90	.95	.98	.99	.999
Expected length of interval (15) when the mean is 0	1.907	2.658	3.332	4.122	4.659	6.181
Length of the usual interval	2.563	3.290	3.920	4.653	5.152	6.581
Ratio squared	.55	.65	.72	.79	.82	.88

Of course, if the mean is not 0, interval (15) will not have minimum expected length, and if the mean is far from 0, (15) will have greater expected length than the usual procedure (13). I am not advocating the use of interval (15). My real objection to (15) cannot be stated in the present framework, but is indicated in Section 12.

The usual procedure (13) is of the form (8) and hence minimizes the expected length when X is uniform on $(-\infty, \infty)$, that is, it minimizes the ordinary (unweighted) integral over x of the length as a function of x.

The usual procedure also minimizes the ordinary integral over τ of the expected length when the mean is τ. This follows immediately from (12) and amounts to the statement of the previous paragraph, since X is uniform if τ is uniform. It also follows from the fact that the usual two-sided equal-tailed

Normal distribution, mean τ, variance 1.
Horizontal section of shaded region is acceptance region of most powerful test against alternative $\tau = 0$.
Vertical section is confidence interval with minimum expected length when $\tau = 0$.

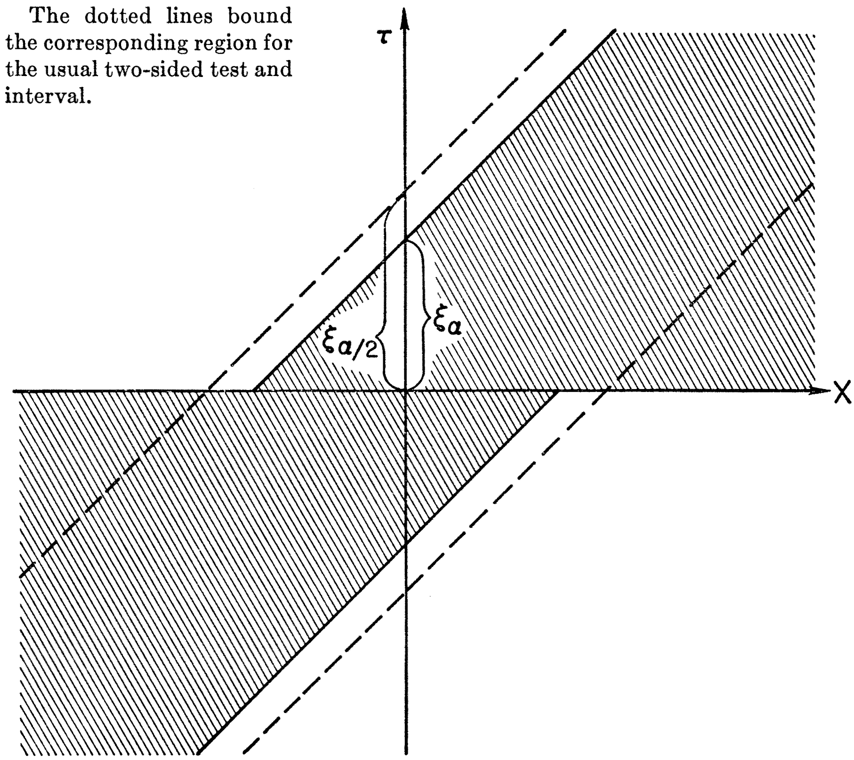

FIG. 1. A family of tests and the corresponding confidence procedure for the mean of a normal distribution.

test is uniformly most powerful invariant (under the transformation of x carrying $x-\tau$ into $\tau-x$) and hence minimizes (10) for $d\nu(\theta) = d\theta$.

The previous two paragraphs apply, of course, to any other measure of size of the sort considered in this paper.

6. EXAMPLE. PROBABILITY OF SUCCESS IN BINOMIAL TRIALS

Let S be binomially distributed with parameters n, p. Four confidence procedures for p have been tabled. Clopper and Pearson's graphs [3] and most tables are based on tests for each possible value of p having probability at most $\alpha/2$ in each tail under the null hypothesis. Tables of the unbiased confidence intervals, which are randomized (see below) have recently been prepared by Blyth and Hutchinson [1].

Sterne's rule [7] is to take the smallest interval containing the region corresponding to tests of the form (8). If this region is an interval for each possible value of S, it follows Sterne's procedure minimizes the expected length of the confidence interval when all values of S are equally likely. This expected length is just the arithmetic average of the lengths of the intervals (one for each possible value of S), and will be called the "average length" below.

Crow [4] notes that, for some values of n, α, and S, the regions given here by (8) are not intervals, and he gives a geometric argument amounting to the relevant case of the first equality of (6) to show that the average length will be minimized by any set of confidence intervals corresponding to tests which minimize, for each p, the number of values of S in the acceptance region $A(p)$. In the cases he tabled, Crow was able to choose acceptance regions containing a minimum number of values of S so that the corresponding confidence regions were always intervals.

Thus Crow's tables, and probably Sterne's also, minimize the average length of the interval. Crow also points out that this amounts to minimizing the expected length when p has a uniform "prior" distribution, since then all values of S are equally likely.

A test or confidence procedure could depend, at least in part, on a random variable drawn independently of the observations from a known distribution; if it does not, it is "non-randomized." Crow's tables, and probably Sterne's also, minimize the average length among non-randomized procedures. If one minimizes the average size among all procedures, there are only a finite number of values of p at which there is a choice of acceptance region, and the minimizing confidence regions are not always intervals, as we shall see.

An easy way to incorporate the possibility of randomization into the discussion is to include with the original observations one observation on a random variable U uniformly distributed between 0 and 1. In any problem, randomized procedures based on the original observations Y are equivalent to non-randomized procedures based on $X=(Y, U)$.

In the binomial situation at hand, we can even let $X=S+U$; then S (the binomial variable) and U are the integral and fractional parts of X, respectively.

If S is uniform on $0, 1, \cdots, n$, where n is the number of binomial trials, then $X=S+U$ is uniform on $(0, n+1)$, and the first equality of (6) becomes

$$E\{\mu(R(X))\} = \int\int_{x\in A(p)} dx d\mu(p). \tag{16}$$

This is minimized if and only if $A(p)$ is as short as possible for every p (except possibly for a set of p's of μ-measure 0). Here it happens the minimizing $A(p)$ can always be taken to be an interval.

Figure 2 gives these shortest acceptance regions for $n=5$, $\alpha=.10$. The corresponding confidence regions are not always intervals, though they have minimum average size (for any measure of size). To illustrate the calculation, for $.317<p<.401$, 1 and 2 are both more likely than 3, 3 is more likely than 0, 0 than 4, and 4 than 5. Hence 1, 2, and 3 would go into the acceptance region first. We have the following probabilities:

p	.32	.34	.36	.38	.40
$P_p\{S=1, 2, \text{ or } 3\}$	.81560	.82613	.83284	.83583	.83520
$P_p\{S=0\}$	.14539	.12523	.10737	.09161	.07776

p = binomial parameter
$S = [X]$ is binomial ($n = 5$)
$U = X - [X]$ is independent of X and uniform on [0, 1]

$\alpha = .10$
Horizontal section is shortest acceptance region.
Vertical section is corresponding confidence region.

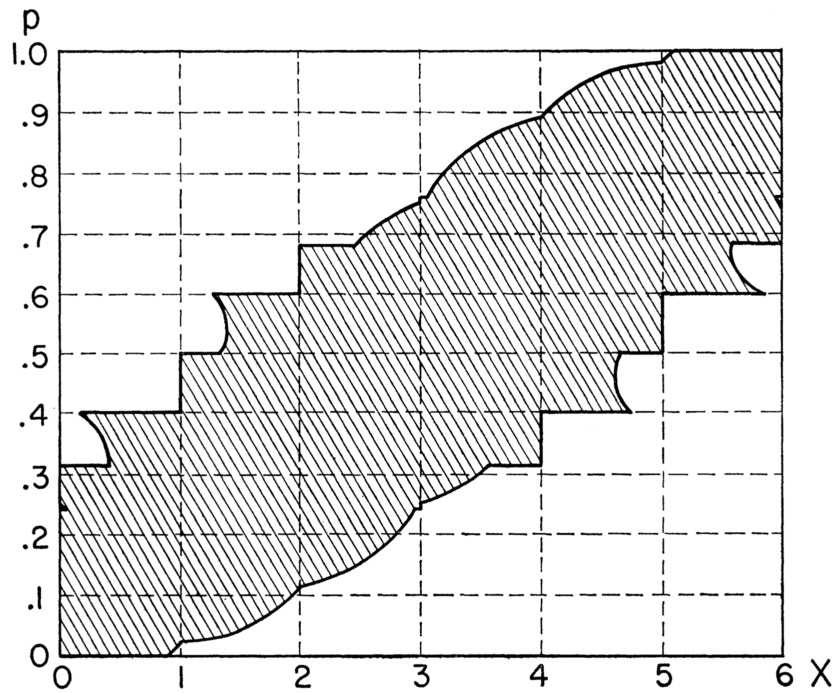

FIG. 2. A family of tests and the corresponding confidence procedure for the binomial parameter.

Thus at $p = .32$, for instance, $S = 1$, 2, and 3 are in the acceptance region; further probability of $.9 - .81560 = .08440$ is needed. This is obtained by accepting at $S = 0$ with probability $.08440/.14539 = .581$, say by accepting when $U > .419$. Similarly at $p = .40$, further probability of .06480 must be obtained by accepting at $S = 0$ with probability $.06480/.07776 = .833$; thus it is impossible that the corresponding confidence region at $S = 0$ should always be an interval—there must be probability at least $.833 - .581 = .252$ that .40 is in the region but .32 is not.

7. DISCONNECTED REGIONS

If the failure of the foregoing region to be an interval is regarded as a drawback, it must be held equally against minimizing expected length and against minimizing the probability of covering false values since these principles amount to the same thing, as it has been the main purpose of this paper to show. If one judges principles not by prior reasonableness but by how good the resulting procedures are (How does one judge this "goodness" without *some* prior values? One must avoid the fallacy of over-naive pragmatism.), one may feel this failure is a strike against integrating (2) and (3) (pp. 550–1) to get (9) and (10), and that it would be preferable to invoke some other principle. Unbiasedness, for instance, leads to randomized confidence regions for the binomial parameter which are intervals [5, p. 179], but different from even the randomized form of the intervals of [3], [4], or [7]. (Unbiasedness here, as so often, picks out a single procedure from a class of "good" procedures, though few find it compelling in itself.)

Insomuch as integrating (2) and (3) to get (9) and (10) seems Bayesian, this may seem a strike against Bayesianism. It is not. A Bayesian would be interested in a posterior probability region, not a confidence region. If the posterior distribution were not unimodal, the smallest region of posterior probability .90 might not be an interval, but the Bayesian would not find this objectionable and would feel that it was just an attribute of the particular problem. In the binomial problem, taking p a priori uniformly distributed between 0 and 1 (corresponding to the weighting used in Section 6), the posterior distribution is unimodal anyway, having the beta density $p^S(1-p)^{n-S}/B(S+1,\ n-S+1)$. Thus the region minimizing (10) may fail to be an interval even when the posterior distribution (with the same ν as prior) is unimodal.

8. EXAMPLE. VARIANCE OF A NORMAL DISTRIBUTION

The problem of estimating a variance τ under normal models often reduces to that of estimating τ on the basis of a statistic X such that X/τ has a chi-square distribution with (say) n degrees of freedom. The chi-square density is

$$k_n(y) = y^{\frac{1}{2}n-1}e^{-\frac{1}{2}y}/2^{\frac{1}{2}n}\Gamma(\tfrac{1}{2}n), \qquad y > 0.$$

We will be especially concerned with confidence intervals

$$\frac{X}{b} < \tau < \frac{X}{a}, \tag{17}$$

where a and b are determined by the conditions

$$\int_a^b k_n(x)dx = 1 - \alpha, \tag{18}$$

$$k_{n+r}(a) = k_{n+r}(b). \tag{19}$$

For $r>2-n$, it follows from the properties of k_{n+r} (it is continuous, strictly unimodal, and vanishes at 0 and ∞) that (18) and (19) determine a and b uniquely. Let I_r be the confidence interval determined by (17)–(19), $r>2-n$.

Condition (18) makes the confidence level exactly $1-\alpha$ for all τ. In the following discussion, "an interval of the form (17)" means one satisfying (18) as well.

Tate and Klett [8] show that (i) I_4 is the interval of the form (17) with minimum length. They also mention that (ii) for samples from a normal distribution with unknown mean and variance, the likelihood-ratio criterion leads to I_3, (iii) I_2 is the only unbiased procedure of the form (17), (iv) I_2 is the interval of the form (17) with minimum b/a, and (v) among all unbiased regions I_2 minimizes the probability of covering false values uniformly for all false values and all true values.

Statements (i) and (iv) may be generalized as follows. Let us call a confidence procedure for τ invariant if it satisfies

τx is in the confidence region when $X=x$ if and only if τ is in the confidence region when $X=1$. (20)

A confidence procedure $R(X)$ for τ is invariant if and only if, for some R' not depending on X,

$$R(X) = \{\tau;\, \tau/X \in R'\}. \tag{21}$$

When R' is the interval $(1/b,\, 1/a)$, (21) reduces to (17).

Corresponding to the confidence region $R(X)$ for τ is a confidence region for $\lambda=\tau^u$, $u\neq 0$, namely $R_u(X)=\{\lambda:\tau\in R(X)\}$. $R_u(X)$ is invariant for λ (in the natural sense) if and only if $R(X)$ is invariant for τ; then $R_u(X)$ has size

$$\int_{\lambda\in R_u(X)} d\lambda = u\int_{\tau\in R(X)} \tau^{u-1}d\tau = uX^u\int_{z\in R'} z^{u-1}dz. \tag{22a}$$

For $u=0$, let $\lambda=\log\tau$. The size of the region $R_0(X)$ in λ-space corresponding to $R(X)$ is

$$\int_{\lambda\in R_0(X)} d\lambda = \int_{\tau\in R(X)} \tau^{-1}d\tau = \int_{z\in R'} z^{-1}dz. \tag{22b}$$

It is proved below (Section 8.2) that among invariant procedures with confidence level $1-\alpha$, I_{2u+2} minimizes (22a or b) for every X. Thus the corresponding confidence region for $\lambda=\tau^u$ if $u\neq 0$ and $\lambda=\log\tau$ if $u=0$, has minimum size for every X (in the ordinary sense of size), among invariant procedures.

In particular I_{2u+2} minimizes (22a or b) among intervals of the form (17), which is statement (i) above when $u=1$ and statement (iv) when $u=0$. For $u=\frac{1}{2}$, we find the likelihood ratio procedure I_3 minimizes the length (or size, in the ordinary sense) of the confidence region for σ among invariant procedures with level $1-\alpha$.

Note that, among invariant regions for τ, I_{2u+2} has minimum size for every X and the specific measure of size $\int_{R(X)}\tau^{u-1}d\tau$. (Compare the middle term of (22a or b).) This illustrates the possibility, mentioned at the end of Section 3, that among invariant procedures, the minimizing procedure may depend on the measure of size used.

The following discussion answers to a large extent some questions raised by Tate and Klett [8] about the length of confidence intervals.

A computation like that leading to (15) shows that, for any measure of size, the level α region having minimum expected size when in fact $\tau=\tau_1$ is the interval

$$\min\{\tau_1,\, X/k''\} < \tau < \max\{\tau_1,\, X/k'\}, \tag{23}$$

where k' and k'' are the lower and upper α-points of k_n. This interval is not of the form (17) and depends on τ_1.

Since (23) depends on τ_1, let us consider the average expected size

$$\int E_\tau\{\mu(R(X))\}W(\tau)d\tau.$$

The density of X under τ is $(1/\tau)k_n(x/\tau)$, so by (12), a procedure minimizing the average expected size for every measure μ of size and the particular weight function $W(\tau)$ is

$$\left.\begin{matrix}\text{reject } \tau\\ \text{accept } \tau\end{matrix}\right\} \text{ if } \frac{\int \frac{1}{\theta} k_n\left(\frac{X}{\theta}\right) W(\theta)d\theta}{\frac{1}{\tau} k_n\left(\frac{X}{\tau}\right)} \quad \begin{cases} > c(\tau), \\ < c(\tau). \end{cases} \tag{24}$$

It is proved below (Section 8.2) that, in particular, I_{2-2u} is a minimizing procedure for $W(\tau)=\tau^{u-1}$.

If one has a sample from a normal distribution with mean ξ and variance τ, both unknown, then all procedures may, without loss of generality, be based on the sufficient statistic $(\overline{X}, X)$, where $\overline{X}$ is the sample mean and X is a multiple of the sample variance such that X/τ has a chi-square distribution. If one also asks that a confidence procedure for τ remain unchanged upon the addition of a constant to all observations, one is led to procedures depending only on X, and the foregoing discussion applies. Nevertheless, the interval of shortest length based on $(\overline{X}, X)$ might not depend only on X, and Tate and Klett [8] ask whether in fact it does. The region with minimum expected size under the particular values ξ_1, τ_1 does not depend only on X, since the most powerful test of the null hypothesis that τ is the true value against the alternative ξ_1, τ_1 does not depend only on X for $\tau > \tau_1$ [5, p. 95]. The same kind of argument shows that if ϕ is any normal density, the region with minimum

$$\int E_{\xi,\tau_1}\{\mu(R(\overline{X}, X))\}\phi(\xi)d\xi$$

does not depend only on X. However, in what might be considered the limit as the variance of ϕ approaches ∞, the region with minimum $\int E_{\xi,\tau_1}\{\mu(R(\overline{X},X))\}d\xi$ is (23), which does depend only on X. This follows from the fact that the usual upper-tailed (if $\tau_1 > \tau$) or lower-tailed (if $\tau_1 < \tau$) test based on X is the most powerful test of the null hypothesis that τ is the true value against the alternative that ξ has a uniform "distribution" on $(-\infty, \infty)$ and τ_1 is the true value of τ. Similarly, (24) is the region with minimum $\int E_{\xi,\tau}\{\mu(R(\overline{X},X))\}W(\tau)d\xi d\tau$, and in particular I_{2-2u} minimizes $\int E_{\xi,\tau}\{\mu(R(\overline{X},X))\}\tau^{u-1}d\xi d\tau$.

Clearly the situation is similar in many more complicated normal models.

8.1. To prove that I_{2u+2} minimizes (22a or b) among invariant confidence procedures with confidence level $1-\alpha$, note that when τ is the true value, X/τ has density $k_n(y)$ and τ/X therefore has density $z^{-2}k_n(1/z)$. Accordingly, when $R(X)$ is given by (21),

$$\begin{aligned} P_\tau\{\tau \in R(X)\} &= P_\tau\{\tau/X \in R'\} \\ &= \int_{R'} z^{-2}k_n(1/z)dz. \end{aligned} \tag{25}$$

We want to minimize (22a or b) subject to the condition that the value of (25) be $1-\alpha$. According to the Neyman-Pearson Lemma, we may choose R' to be the region where

$$z^{-2}k_n(1/z)/z^{u-1} < c. \tag{26}$$

or, since $y^{u+1}k_n(y)$ is a multiple of $k_{n+2u+2}(y)$,

$$k_{n+2u+2}(1/z) < c', \tag{27}$$

where the constants c and c' are chosen to make (25) $1-\alpha$. By the properties of k_{n+2u+2}, R' must be an interval, say $(1/b,\ 1/a)$. At the end-points equality must hold in (27), giving (19). Therefore I_{2u+2} is the minimizing procedure.

8.2. To show that (24) reduces to I_{2-2u} if $W(\tau)=\tau^{u-1}$, note that for this $W(\tau)$ the numerator in (24) is a multiple of X^{u-1}. Hence (24) becomes

$$\left.\begin{array}{l}\text{reject } \tau\\ \text{accept } \tau\end{array}\right\} \text{ if } \left(\frac{X}{\tau}\right)^{1-u} k_n\left(\frac{X}{\tau}\right) \quad \left\{\begin{array}{l}<c_1(\tau),\\ >c_1(\tau).\end{array}\right. \tag{28}$$

Now $c_1(\tau)$ is to be chosen so that the probability under τ of rejecting τ is just α. Since the distribution under τ of X/τ is the same for every τ, the required $c_1(\tau)$ is the same for every τ. Since also $y^{1-u}k_n(y)$ is a constant multiple of $k_{n+2-2u}(y)$, (28) is of the form

$$\left.\begin{array}{l}\text{reject } \tau\\ \text{accept } \tau\end{array}\right\} \text{ if } k_{n+2-2u}\left(\frac{X}{\tau}\right) \quad \left\{\begin{array}{l}<c,\\ >c.\end{array}\right. \tag{29}$$

By the properties of k_{n+2-2u}, (29) is of the form

$$\text{accept } \tau \quad \text{if} \quad a < \frac{X}{\tau} < b, \qquad \text{reject otherwise.} \tag{30}$$

(What we do at the end-points does not affect the probability.) Equation (30) is equivalent to (17). Equation (19) follows from the fact that equality must hold in (29) at the end-points of (30), and (18) is just the condition determining c in (29). Thus I_{2-2u} is the region (24) when $W(\tau)=\tau^{u-1}$.

The result of 8.1 follows from the one just proved, since, for any region of the form (21),

$$\begin{aligned}\int E_\theta\{\mu(R(X))\}\theta^{u-1}d\theta &= \int\int\int_{\tau\in R(x)} d\mu(\tau)k_n\left(\frac{X}{\theta}\right)\frac{dx}{\theta}\,\theta^{u-1}d\theta \\ &= \int_{z\in R'} z^{-u-1}dz\int y^{-u}k_n(y)dy\int \tau^u d\mu(\tau),\end{aligned} \tag{31}$$

where $z=\tau/x$ and $y=x/\theta$.

9. RELATION TO BAYES PROCEDURES

The confidence intervals of the previous section can also be interpreted as posterior probability intervals for appropriate prior distributions of τ. Specifically, suppose τ is distributed a priori with density proportional to τ^{v-1}, that is, τ^v (or $\log\tau$, if $v=0$) is uniformly distributed a priori. Then the posterior density of τ^u given X is (by Bayes' Theorem and straightforward calculation) proportional to $k_{n+2u-2v+2}(X/\tau)$. In particular, taking $u=-1$, it follows that the posterior distribution of X/τ given X is chi-square with $n-2v$ degrees of freedom. This means that a confidence region at level $1-\alpha$ computed as though there were $n-2v$ degrees of freedom is a region of posterior probability $1-\alpha$. If $v=0$, that is, if $\log \tau$ is uniformly distributed a priori, ordinary confidence regions at level $1-\alpha$ have posterior probability $1-\alpha$. These facts are not new, and are included only to set the stage for statement of the relation between the procedures of Section 8 and Bayes procedures.

Since the posterior density of τ^u given X is proportional to $k_{n+2u-2v+2}(X/\tau)$, the smallest region in τ^u-space of posterior probability $1-\alpha$ will be an interval with end-points corresponding to values of X/τ at which $k_{n+2u-2v+2}(X/\tau)$ has the same value.

Thus we find that, if τ^v (or $\log \tau$, if $v=0$) is uniformly distributed a priori, then the smallest region of posterior probability $1-\alpha$ for τ^u (or $\log \tau$ if $u=0$) is I_{2u+2} computed as though there were $n-2v$ degrees of freedom. Thus the effect of these prior distributions is to change the degrees of freedom, while the effect of looking for small regions in τ^u-space is the same for posterior probability regions as for invariant confidence regions. When average expected size is minimized, as at (24), the effect of the weight function $W(\tau)$ on the confidence region is different from the effect on the posterior probability region of the same weight function used either as prior distribution or as measure of size.

Lindley has pointed out [6, Introduction] that the unbiased procedure I_2 gives the shortest region of posterior probability $1-\alpha$ for $\log \tau$ if $\log \tau$ is uniformly distributed a priori. (If instead τ^v is uniform a priori, the degrees of freedom need only be reduced by $2v$.) Lindley also points out that integrating by parts shows I_0 is the same as I_2 on two fewer degrees of freedom, so that tables of I_2 can be used for I_0, which has minimum size for σ^{-2} (not σ^2, as stated [6, Section 1, last paragraph]).

10. MEASURE OF SIZE DEPENDING ON THE TRUE VALUE

In some confidence problems, in particular, one-sided ones, the undesirability of covering a false value depends on the true value. Sections 1–4 are easily modified to allow this.

Specifically, suppose θ indexes the possible distributions and we allow μ to depend on θ. This excludes A–C, p. 550, but under D (E is no more interesting), we would have, instead of (1),

$$\mu_\theta(R) = \int_{\tau\in R} W(\tau, \theta) d\tau, \tag{32}$$

where $d\tau = d\tau_1 \cdots d\tau_k$ and W is non-negative. Similarly, (2) and (3) become

$$\begin{aligned} E\{\mu_\theta(R(X))\} &= \int P\{R(X) \ni \tau\} W(\tau, \theta) d\tau \\ &= \int_{\tau\neq\tau(\theta)} P\{R(X) \ni \tau\} W(\tau, \theta) d\tau. \end{aligned} \tag{33}$$

(10) becomes

$$\begin{aligned}\int E_\theta\{\mu_\theta(R(X))\}d\nu(\theta) &= \int\int P_\theta\{R(X) \ni \tau\}W(\tau, \theta)d\tau d\nu(\theta) \\ &= \int\int P_\theta\{X \in A(\tau)\}W(\tau, \theta)d\nu(\theta)d\tau,\end{aligned} \tag{34}$$

and, as in (9), the integral is unchanged if one integrates only over false values, that is, over the region $\tau \neq \tau(\theta)$.

(34) is minimized by choosing, for each τ, the most powerful acceptance region $A(\tau)$ against the alternative $P^{(\tau)}$ given by

$$P^{(\tau)}\{X \in A\} = \int P_\theta\{X \in A\}W(\tau, \theta)d\nu(\theta). \tag{35}$$

(Compare (11).) If $\tau(\theta) = \theta$ and X has density or frequency function $f_\theta(x)$, a minimizing procedure is:

$$\left.\begin{matrix}\text{reject } \tau \\ \text{accept } \tau\end{matrix}\right\} \text{ if } \frac{\int f_\theta(x)W(\tau, \theta)d\nu(\theta)}{f_\tau(x)} \quad \left\{\begin{matrix} > c(\tau), \\ < c(\tau),\end{matrix}\right. \tag{36}$$

as at (12).

This is just like the situation of Section 4, except that P may depend on τ and on W. In contrast to the earlier situation, the average expected size (34) for a particular ν is not minimized simultaneously for all measures of size (32), but only for all $W(\tau, \theta) = W_1(\tau)W_2(\tau, \theta)$ with a particular W_2 (and arbitrary W_1).

11. ONE-SIDED PROBLEMS

Suppose τ is one-dimensional and the problem is one-sided. Suppose, for definiteness, it does not matter if values of τ below the true value are included in the confidence region. This suggests a measure of size of the form

$$\mu_\theta(R) = \int_{\tau > \tau(\theta), \tau \in R} d\mu(\tau). \tag{37}$$

If $d\mu(\tau) = d\tau$, and R is the region of all τ below the value U, say, then (37) becomes, for $U > \tau(\theta)$, $\mu_\theta(R) = U - \tau(\theta)$, the amount by which U overestimates the true value, and for $U \leq \tau(\theta)$, $\mu_\theta(R) = 0$. If $d\mu(\tau) = W_1(\tau)d\tau$, (37) is equivalent to (32) with $W(\tau, \theta) = W_1(\tau)W_2(\tau, \theta)$ where $W_2(\tau, \theta) = 1$ if $\tau > \tau(\theta)$ and 0 otherwise.

Under (37), (2) and (3) become

$$E\{\mu_\theta(R(X))\} = \int_{\tau > \tau(\theta)} P\{R(X) \ni \tau\}d\mu(\tau). \tag{38}$$

(10) becomes

$$\begin{aligned}\int E_\theta\{\mu_\theta(R(X))\}d\nu(\theta) &= \int\int_{\tau > \tau(\theta)} P_\theta\{R(X) \ni \tau\}d\mu(\tau)d\nu(\theta) \\ &= \int\int_{\tau > \tau(\theta)} P_\theta\{X \in A(\tau)\}d\nu(\theta)d\mu(\tau).\end{aligned} \tag{39}$$

If, for each τ, there is a uniformly most powerful test of the null hypothesis that τ is the true value against alternatives p_θ with $\tau(\theta) < \tau$, then the corresponding confidence procedure uniformly minimizes the probability of covering false values larger than the true value and uniformly minimizes $E_\theta\{\mu_\theta(R(X))\}$ for all θ and all measures (37) of the amount of overestimation. Failing this, unbiasedness or invariance may be invoked. Alternatively one may minimize (39) simultaneously for all measures (37) by including τ in the confidence region if and only if it would be accepted by the most powerful test of the null hypothesis that τ is the true value against the alternative $P^{(\tau)}$ given by

$$P^{(\tau)}\{X \in A\} = \int_{\tau(\theta)>\tau} P_\theta\{X \in A\} d\nu(\theta). \tag{40}$$

In particular, if $\tau(\theta) = \theta$ and X has density or frequency function $f_\theta(x)$, such a procedure is:

$$\left.\begin{array}{l}\text{reject } \tau\\ \text{accept } \tau\end{array}\right\} \text{ if } \frac{\int_{\theta>\tau} f_\theta(x) d\nu(\theta)}{f_\tau(x)} \left\{\begin{array}{l} > c(\tau), \\ < c(\tau). \end{array}\right. \tag{41}$$

12. REMARKS

An argument sometimes made against expected length as a measure of the desirability of a confidence interval procedure is that, as Lehmann says [5, p. 182], "Short intervals are desirable when they cover the true parameter value but not necessarily otherwise." If a short confidence interval is taken to indicate accurate information about the parameter, then it may be preferable that the interval be long when it is far from the true parameter value. Considering instead the probability of covering false values does not avoid this difficulty, however. In the first place, the two approaches are related, as shown in this paper. More fundamentally, small chance of covering false values is also desirable when the true value is covered but not necessarily otherwise. A natural way to avoid the difficulty would be to consider both expected length and the probability of covering false values conditional on the true value being covered. The basic relation between the two approaches still holds, that is:

$$\begin{aligned} E\{\mu(R(X)) \mid R(X) \ni \tau_0\} &= \int P\{R(X) \ni \tau \mid R(X) \ni \tau_0\} d\mu(\tau) \\ &= \int P\{X \in A(\tau) \mid X \in A(\tau_0)\} d\mu(\tau) \end{aligned} \tag{42}$$

and, again, the omission of τ_0 from the range of integration does not change the value of the integrals.

I have not pursued the idea of conditioning further. The immediate reason is that its possible relevance only recently occurred to me, as one result of some questions raised in correspondence with the editor to whom I wish to express my gratitude. Furthermore, there is in my opinion a more important consideration than either expected length or the probability of covering false values, namely, to what extent the confidence level has a meaning when a confidence statement is regarded as a statement about the unknown parameter τ for the X observed. My real objection to the interval (15) is not that it is a poor confidence procedure according to some criterion involving the probability of covering false values, but that one would not ordinarily have the same confidence in the statement that τ is in the interval (15) when $|X|$ is small and the interval is short as when $|X|$ is large and the interval is long; in particular for $X>\xi_\alpha$, as X increases it seems to become increasingly likely that the true τ is in the interval. This is true despite the fact that, before X is observed, the probability of covering the true value is exactly $1-\alpha$ no matter what the true value (including $\tau=0$ if a trivial change is made). Just because a statement about X and τ has probability $1-\alpha$ for each τ and hence has confidence level $1-\alpha$, it does not automatically follow that there is any useful sense in which confidence $1-\alpha$ is attributable to the statement for each X. Though some further discussion of this important problem is possible within the frequency theory of probability, along the line of [2] and [9], for instance, I doubt that the frequency theory alone can really solve it.

Though the weighting function used to select a confidence procedure in Section 4 and in the examples might be regarded as a prior distribution, the procedure obtained is generally not a Bayes procedure corresponding to this prior. This has already been pointed out in Sections 7 and 9 in connection with the second and third examples. In the first example (the normal mean, Section 5), the weight function put all weight at 0; taking this as a prior distribution, the posterior distribution is also concentrated at 0, which does not correspond to (15). The situation is reminiscent of a method of testing hypotheses sometimes suggested, namely, to use as the test statistic the posterior odds against the null hypothesis (computed under some prior), judging it not as odds but by the usual criteria of testing theory.

The position in orthodox confidence interval theory of choosing a weight function in Section 4 seems to me like that of choosing a statistic in testing hypotheses or choosing a design for an experiment. That is, we may use our previous information and hunches as we see fit in choosing an experimental design, a test statistic, or a confidence procedure, provided we make the choice before (or as if before) taking the observations, and provided our procedure has the stated confidence or significance level. What are we to say, then, about an experimenter who thinks in advance his mean will be near 0 and accordingly decides to use a procedure near (15) instead of the usual confidence procedure for a normal mean?

REFERENCES

[1] Blyth, Colin R. and Hutchinson, David W. "Table of Neyman-shortest unbiased confidence intervals for the binomial parameter," *Biometrika*, 47 (1960), 381–91.

[2] Buehler, Robert J. "Some validity criteria for statistical inferences," *Annals of Mathematical Statistics*, 30 (1959), 845–63.

[3] Clopper, C. J. and Pearson, E. S. "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, 26 (1934), 404–13.

[4] Crow, Edwin L. "Confidence intervals for a proportion," *Biometrika*, 43 (1956), 423–35.

[5] Lehmann, E. L. *Testing Statistical Hypotheses.* New York: John Wiley & Sons, Inc., 1959.

[6] Lindley, D. V., East, D. A., and Hamilton, P. A. "Tables for making inferences about the variance of a normal distribution," *Biometrika,* 47 (1960), 433–7; Introduction by D. V. Lindley.

[7] Sterne, Theodore E. "Some remarks on confidence or fiducial limits," *Biometrika,* 41 (1954), 275–8.

[8] Tate, R. F. and Klett, G. W. "Optimal confidence intervals for the variance of a normal distribution," *Journal of the American Statistical Association,* 54 (1959), 674–82.

[9] Wallace, David L. "Conditional confidence level properties," *Annals of Mathematical Statistics,* 30 (1959) 864–76.