Statistical Science
2002, Vol. 17, No. 2, 149–172

Setting Confidence Intervals for Bounded Parameters

Mark Mandelkern

Abstract. Setting confidence bounds is an essential part of the reporting of experimental results. Current physics experiments are often done to measure nonnegative parameters that are small and may be zero and to search for small signals in the presence of backgrounds. These are examples of experiments which offer the possibility of yielding a result, recognized a priori to be relatively improbable, of a negative estimate for a quantity known to be positive. The classical Neyman procedure for setting confidence bounds in this situation is arguably unsatisfactory and several alternatives have been recently proposed. We compare methods for setting Gaussian and Poisson confidence intervals for cases in which the parameter to be estimated is bounded. These procedures lead to substantially different intervals when a relatively improbable observation implies a parameter estimate beyond the bound.

Key words and phrases: Confidence bounds, Poisson-with-background, Gaussian-with-boundary.

1. INTRODUCTION

The testing of theoretical models and the estimation of parameters are among the principal tasks of experimental science. Statistical theory has given us methods which work well in most but not all circumstances. We discuss a problem for which statistics does not give a solution that is satisfactory to many scientists. That problem is estimation of a parameter when the data are known a priori to be relatively improbable for all parameter values under consideration. Such a problem arises, for example, when the parameter of interest is bounded and the result of an experiment suggests a parameter estimate beyond the bound.

A crucial element of parameter estimation is quantifying the uncertainty in the estimate. We generally present the uncertainty in the estimate of a parameter by providing an interval based on the data and an associated measure of confidence that the true value lies within that interval. In the frequentist theory, we have the *coverage* $1-\alpha$, the probability that the interval (the random variable) contains the true value. The Bayesian confidence interval (credible interval) is constructed to contain posterior probability $1-\alpha$. Neither construction is unique and for the problem I address here, a number of methods have been proposed. The Bayesian approach requires specification of a prior distribution for the parameter and both approaches require an optimality criterion to uniquely define an interval.

When we do not know a priori whether a particular observation is relatively probable or not, there is a natural and intuitively reasonable choice of frequentist intervals: the central Neyman construction when a two-sided interval is sought, or the one-sided construction when an upper or lower limit is sought.

However, it is sometimes possible to identify an experimental result as relatively improbable for all possible members of a parametric family of distributions. There are two specific cases that appear frequently and lead to difficulty. The first is setting confidence intervals for the mean of a normal pdf $n(X;\mu)$ where the mean is known to be bounded, for example, $\mu \geq 0$, and the variance σ^2 is known. (Here, without loss of generality, $\sigma^2 = 1$.) The Neyman confidence interval in this case is empty when an observation X is sufficiently negative. Any $X < 0$ yields a shorter interval than that obtained for $X = 0$.

This situation occurs frequently in physics, where many fundamental parameters are intrinsically nonneg-

Mark Mandelkern is Professor, Department of Physics and Astronomy, 3158 Frederick Reines Hall, University of California, Irvine, California 92697-4575 (e-mail: markm@uci.edu).

ative. Of particular current interest are the masses of fundamental particles, where the neutrinos are especially important, because whereas we have always believed that neutrinos are massless, there is now indirect experimental evidence that at least two neutrinos have nonzero mass. Major experimental efforts have been mounted to determine whether neutrinos are massless. Measurements of the decay rates of fundamental particles are tests of conservation principles that have long thought to be valid but are now being challenged. Both total decay rates, such as for the proton, and partial decay rates, such as for $K^0 \to e\mu$, a process that violates separate lepton number conservation, are positive parameters that are small and may be exactly zero. The proton is of particular interest since there are strong theoretical reasons to believe that it is unstable yet no decay has been observed after many years of intense experimentation.

The second case leading to difficulty is setting confidence intervals for the parameter μ of the Poisson distribution $p(N; \mu + b)$. Here μ is the unknown nonnegative signal mean and $b > 0$ is a known nonnegative background mean. For an observation N, the usual confidence interval is empty for $N - b$ sufficiently small and any $N < b$ yields a shorter interval than that obtained for the a priori more probable observation $N = b^-$, where b^- is the smallest integer less than or equal to b. Such a circumstance is also common in physics. We frequently search for a small signal above a (unwanted) background. The background may be measured independently or may be estimated. A negative fluctuation of background and/or signal can lead to $N < b$.

An observation of this type causes difficulty only when it appears moderately improbable, since if it is grossly improbable, either the data or model is likely to be so wrong that the observation is not useful. However, a measurement with, for example, a 0.1% or greater probability may simply be a tail (extreme) value rather than indicate a methodological problem. The latter certainly occurs more frequently than one might expect because estimating statistical and systematic uncertainties and incorporating them into the model is one of the most difficult parts of an experiment. We frequently can only approximate the pdfs for random variables, and systematic uncertainties are by definition offsets of which we are ignorant. Backgrounds are not always directly observable and are often calculated. On the other hand, we do not wish to overestimate uncertainties and thereby weaken hypothesis testing and parameter estimation. The rule is to estimate uncertainties one has evidence for, and assume that they are independent unless one knows otherwise. Our best estimates of uncertainties are usually underestimates.

A common reaction of a statistician to a significantly improbable observation may be to advise that either the data are wrong or the model is wrong. While this is certainly true in most cases, the question is what to do. Rejecting the result is not satisfactory, not just because experiments may be unique observations and/or cost millions of dollars, but because rejecting improbable or undesired results is biasing, especially if only some results can be recognized as improbable. We need all the experimental results, each with its uncertainty, to make a globally-best estimate of a constant of nature or test a theoretical model. One can also consider, when obtaining relatively improbable data, exploring alternative models. Physicists are very wary of biasing an analysis by changing the model a posteriori. We have taken great pains to invent and employ blind analysis procedures to avoid even subconscious bias. In most analyses, results for important parameters are typically hidden using a *secret offset* until the analysis is complete. We would require an unbiased procedure to consider alternative models, ones that deal symmetrically with cases in which we can (e.g., because the true value is bounded) or cannot (when it is not) know that the data are improbable.

In the past several years a number of major experimental efforts have produced observations of the types described above. Several well-publicized examples have come in neutrino physics. Neutrinos are uncharged elementary particles with masses that until recently have been thought to be zero. The electron neutrino ν_e was initially postulated by Wolfgang Pauli in order to preserve the conservation of energy and angular momentum in the weak decays of nuclei. That particle was first observed by Clyde Cowan and Frederick Reines in 1956; two additional distinct neutrinos, the muon neutrino ν_μ and tau neutrino ν_τ, have been observed since. Neutrinos have fundamental significance in elementary particle physics and a large number of current experiments are concerned with their properties. Several recent and apparently reliable experiments report results suggesting that at least two neutrinos have nonzero mass. These include the detection of a smaller-than-expected flux of electron neutrinos from nuclear reactions in the sun (Cleveland et al., 1998), the observation of a deficit of muon neutrinos from the decay of particles produced in collisions of cosmic rays with nuclei in the upper atmosphere (Fukuda et al.,

1998, Ahmad et al., 2002, Altmann et al., 2000, Abdurashitov et al., 1999), and the observation by some accelerator neutrino experiments that one type of neutrino appears to turn into a different type (neutrino oscillations) (Athanassopoulos et al., 1998). Thus direct measurements of neutrino masses are of crucial importance.

The most sensitive technique for determining the electron neutrino mass is the measurement of the electron (β) energy spectrum for the decay of tritium, the radioactive isotope of hydrogen. The electron energy spectrum near its upper boundary, where it is most sensitive to the neutrino mass, is given by an expression with the squared neutrino mass as a parameter. The latter is determined from the experimental spectrum by nonlinear regression. Several recent experiments obtained negative values for the fitted value, which is not surprising for a mass either equal to zero or comparable to or smaller than the experimental precision.

The best measurements of the electron neutrino mass are those of the Russian "Troitsk" group, which has reported four results for the squared mass since 1995, (-2.7 eV^2/c^4, 0.5 eV^2/c^4, -3.2 eV^2/c^4, -0.6 eV^2/c^4) for a combined result of $-1.9 \pm 3.4_{\text{stat}} \pm 2.2_{\text{syst}}$ eV^2/c^4 (Lobashev et al., 1999). These results are an example of an a priori relatively improbable observation. The question is how to report the estimate of the neutrino mass and its uncertainty. The authors use a Gaussian model with fixed variance and the Unified method of Feldman and Cousins (1999) to obtain a 95% upper limit of 2.5 eV/c^2. An identical experiment yielding a mass observation of zero would lead to the larger 95% upper limit of 2.8 eV/c^2. Thus an arguably worse measurement produces a smaller upper limit implying better knowledge of the electron neutrino mass.

The Poisson-with-background problem appeared in the Karlsruhe Rutherford Medium Energy Neutrino (KARMEN) experiment, which searched for neutrino oscillations in (among other modes) the appearance reaction $\bar{\nu}_\mu \to \bar{\nu}_e$, by using a beam initially composed of $\bar{\nu}_\mu$ and counting events consisting of a final positron and neutron in coincidence resulting from the interaction of a $\bar{\nu}_e$ with a target proton. There are four background sources that produce events indistinguishable from signal events. Three of these are estimated from the data and the fourth obtained from a simulation. KARMEN (Eitel and Zeitnitz, 1998) reported zero events, estimated the mean background as 2.88 ± 0.13 events and used the Unified method to obtain a 90% upper limit for the signal of 1.07 events for the appearance reaction of interest. If the experiment had observed three events total, for example, three background events and no signal events, it would have reported a 90% upper limit of ~ 4.5 events. It is worth noting that a zero background experiment cannot produce a 90% upper limit smaller than 2.31 events. Thus a relatively improbable measurement from an inferior experiment leads to an upper limit which is smaller than that resulting from a more probable observation, and is also smaller than the best limit a much better experiment can obtain.

While (following the authors of the cited experimental papers), we have quoted the results of using the Unified method, even smaller upper limits follow from the classical Neyman method of setting confidence limits. It is just this pathology that has motivated numerous authors cited here to devise alternative methods.

The following features are considered by physicists to be desirable in a confidence interval construction:

(i) Confidence bounds are determined using a well-defined principle, which is neither arbitrary nor subjective.

(ii) They do not depend upon prior knowledge of the parameter apart from its domain.

(iii) They are equivariant under one-to-one transformations of the parameter.

(iv) They convey an estimate of the experimental uncertainty.

(v) They correspond to a precise statement of probability.

Point (ii) makes confidence bounds for different experimental results independent. Point (v) is usually interpreted as requiring a frequentist construction with constant coverage. However, the Bayesian approach has been advocated by prominent physicists, for example, in Orear (1958, 1982), based on lectures by Enrico Fermi.

The properties of the usual frequentist central and one-sided intervals for unbounded parameters are consistent with these principles and are generally viewed as entirely satisfactory. These are:

(i) For parameter μ, the nonuniqueness in choosing an interval is resolved by choosing the most powerful unbiased test against suitable two-sided or one-sided alternative hypotheses.

(ii) No prior knowledge of μ is required.

(iii) The intervals are equivariant under one-to-one transformations.

(iv) The length of a confidence interval is determined, independent of the particular measurement, by the variance of the parent distribution, so it reflects the experimental uncertainty.

(v) The construction has constant coverage: the probability that a confidence bound *covers* the true value is $1-\alpha$.

On the other hand, the Bayesian construction requires a choice of prior, does not guarantee equivariant intervals, and does not explicitly have constant coverage.

Two desirable features that are relevant to the bounded parameter space and are present in all of the recently proposed methods are:

(vi) The confidence belt transitions from an upper (or lower) limit to a bounded interval. Physicists refer to such constructions as *unified*.

(vii) For any observation the confidence interval is within the bounded parameter space.

I first review the Neyman and Bayesian constructions for the cases of a bounded normal variate and of a Poisson variate with background. I illustrate the difficulties that appear when the data set is a priori relatively improbable. I consider several of the recently proposed constructions for these cases and discuss them with respect to the principles listed above.

2. NORMAL AND POISSON BOUNDS

2.1 Classical and Bayesian Constructions

The construction of *frequentist* confidence limits and intervals was developed by Neyman (1937) from the theory of hypothesis testing. For the measurement of a parameter θ, we have the sampling distribution $f(X;\theta)$ of an estimate X. A *confidence belt* B is a set in the space of (X,θ) pairs defined by the upper boundary function $\theta=\theta_U(X)$ and lower boundary function $\theta=\theta_L(X)$, where for an estimate X the interval $[\theta_L(X),\theta_U(X)]$ is a $1-\alpha$ *confidence interval* for θ. Note that intersecting B with the line $\theta=\theta_0$ gives an *acceptance region* for testing the null hypothesis that $\theta=\theta_0$. The probability that the confidence interval contains the true value of the parameter is the *coverage* and is $1-\alpha$ by construction. This construction can produce a lower limit, upper limit or confidence interval. The defining principle does not fully specify the limit or interval. The optimality requirement suggested by Neyman is to minimize the coverage of false values, which usually minimizes the expected length of the confidence interval (Pratt, 1961).

In the Bayesian theory, the *credible interval* has (posterior) probability $1-\alpha$ and the usual optimality condition is that it contains the largest values of the posterior pdf. The construction does not guarantee constant coverage and requires selecting the prior.

2.1.1 *Normal variate with nonnegative mean.* For a normal variate X, distributed as $n(X;\mu,\sigma_0)$, where $-\infty<\mu<\infty$, the $1-\alpha$ upper limit for a single measurement X is $X+f\sigma_0$. The $1-\alpha$ confidence belt for μ is $[X-g\sigma_0, X+g\sigma_0]$. For $1-\alpha=0.6827(0.9)$, $f=0.475(1.28)$ and $g=1(1.64)$.

The treatment for $\mu\geq 0$ is discussed, for example, by Cox and Hinkley (1974, Section 7.2). The upper limit $\max(0, X+f\sigma_0)$ is shown in Figure 1, and is zero for $X\leq -f\sigma_0$. This construction has coverage $1-\alpha$ for all $\mu>0$. For $\mu=0$, it overcovers, with coverage 1.0, since the upper limit is ≥ 0 for any X. The confidence interval $[\max(0, X-g\sigma_0), \max(0, X+g\sigma_0)]$, also shown in Figure 1, contains only zero for $X\leq -g\sigma_0$. The construction has coverage $1-\alpha$ for $\mu>0$ and is conservative for $\mu=0$, with coverage $1-\alpha/2$. It gives a unified description, since for small X, the lower limit is 0, making the interval effectively an upper limit.

The Neyman confidence interval for $X<0$ underestimates the uncertainty in μ. The worst cases are $X\leq -f\sigma_0$ (for upper limits) and $X\leq -g\sigma_0$ (for the confidence belt), where the interval contains only 0. Except at $\mu=0$, the most powerful test for μ is insensitive to the absence of alternatives $\mu'<0$. Although there are fewer alternative means $\mu'<\mu$ than for the unbounded case, the hypothesis test is unchanged. In other words, a negative measurement is known a priori to be less than maximally probable, yet the hypothesis test does not make use of this additional information. It seems reasonable to reduce the power of the test against smaller alternatives in favor of increased

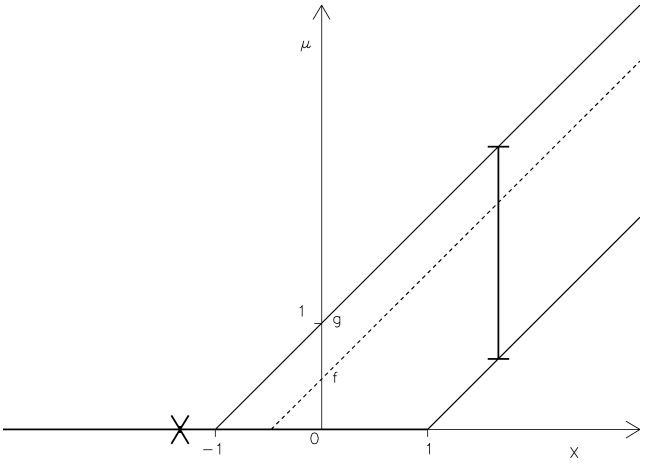

FIG. 1. *The solid lines give the* 68.27% *confidence belt and the dashed line the* 68.27% *upper limit contour, both for* $n(X;\mu,1)$ *with nonnegative* μ *in the classical Neyman construction.* $g=1.0$ *and* $f=0.475$. *For* $X\leq g(f)$, *the confidence interval is empty.*

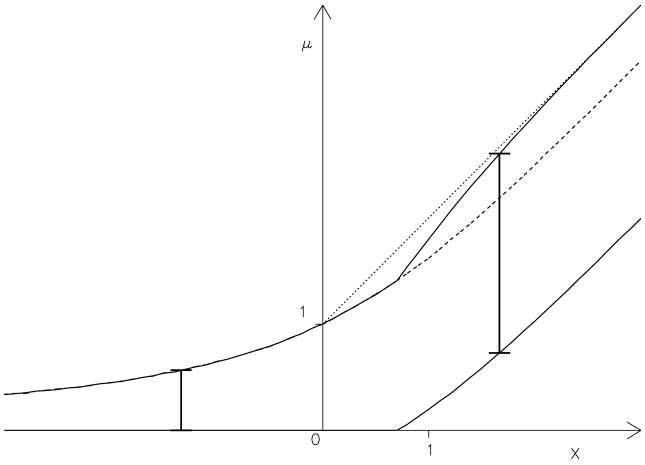

FIG. 2. *The solid curves give the* 68.27% *confidence belt and the dashed curve the* 68.27% *upper limit contour, both for* $n(X;\mu,1)$ *with nonnegative* μ *in the Bayesian construction (uniform prior). The modification of Roe and Woodroofe* (2001) *is shown as a dotted line.*

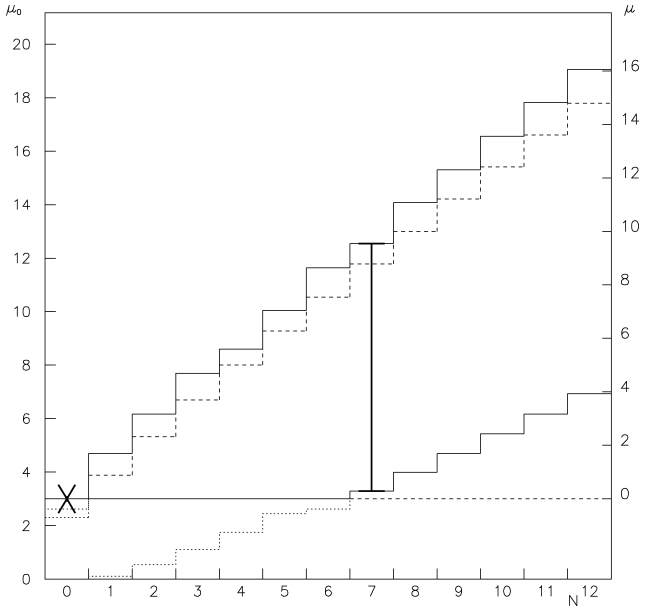

FIG. 3. *The solid contours give the* 90% *confidence belt and the dashed contour the* 90% *upper limits, both for* $p(N;\mu+3)$ *in the classical Neyman construction. For* $N=0$*, the confidence interval is empty.* $\mu_0=\mu+b$.

power against larger alternatives for small μ, which makes the upper bound less restrictive. This is the strategy of the methods discussed below. I note that for $X \le g\sigma_0$ the $1-\alpha$ explicit upper limit, derived from the most powerful test against smaller alternatives, is $\max(0, X+f\sigma_0)$ while the $1-\alpha$ confidence interval, derived from the (less powerful) unbiased test, has the larger upper bound $\max(0, X+g\sigma_0)$.

For an unbounded normal variate, the Bayesian $1-\alpha$ upper limit for a uniform prior is identical to the Neyman $1-\alpha$ upper limit, that is, $X+f\sigma_0$. For the bounded mean $\mu \ge 0$, the Bayesian upper limit is $X+\sigma_0\Phi^{-1}[1-\alpha\Phi(\frac{X}{\sigma_0})]$, where Φ is the standard normal distribution function. It gradually decreases for $X<0$ and is never zero. The Bayesian central interval for the bounded mean (uniform prior, largest posterior pdfs) is $[\max(X-d,0), X+d]$, where $d=\sigma_0\Phi^{-1}[1-\alpha\Phi(\frac{X}{\sigma_0})]$ for $-\infty < X \le x_0$ and $d=\sigma_0\Phi^{-1}[\frac{1}{2}+\frac{1}{2}\Phi(\frac{X}{\sigma_0})]$ for $x_0 < X < \infty$, where $x_0=\sigma_0\Phi^{-1}(\frac{1}{1+\alpha})$. For $1-\alpha=0.6827$ ($x_0=0.7034\sigma_0$) these bounds are shown in Figure 2. The explicit upper limit coincides with the confidence belt upper bound for $X \le x_0$.

2.1.2 *Poisson variate with background.* The confidence intervals for the Poisson distribution $p(N;\mu_0)$ can only have approximately constant coverage because of discreteness. For a Poisson variate N distributed as $p(N;\mu+b)$, where b is known and nonnegative and μ is nonnegative, Figure 3 shows the classical Neyman 90% upper limit contour and 90% confidence belt, both for $b=3$. The latter is unified in that for a small observation N the lower bound is zero so we effectively have an upper limit. For both constructions, that upper limit is zero for $N-b$ sufficiently small, in analogy to the Gaussian-with-boundary problem. The explicit upper limit contour is everywhere distinct from the confidence belt upper bound.

The Bayesian 90% upper limit contour and 90% confidence belt (uniform prior, largest posterior probabilities) for $b=3$ are given in Figure 4. The upper bound for $N=0$ is 2.31 independent of b for either construction, an intuitively attractive result that motivated the construction of Roe and Woodroofe (1999) and is discussed below. The explicit upper limits coincide with the confidence belt upper bound for $N<7$.

2.2 Proposed Constructions

The freedom in choosing an optimality condition has been exploited by several authors who have proposed confidence interval constructions that take into account the relative improbability of small measurements. The resulting constructions are unified and give larger upper limits for $X<0$ (or $N<b$) than the classical Neyman construction in an effort to make them useful measures of confidence.

2.2.1 *Unified approach.* For the Gaussian-with-boundary problem, Feldman and Cousins (1998) apply

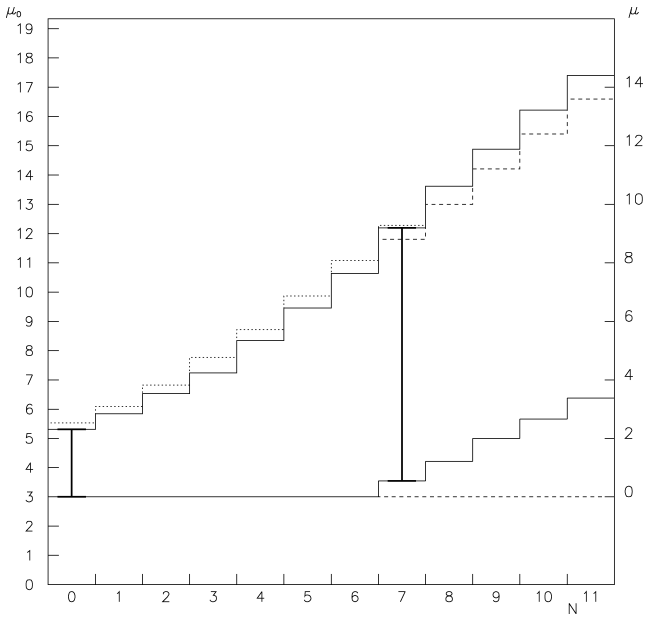

FIG. 4. *The solid contours give the* 90% *confidence belt and the dashed contour the* 90% *upper limits, both for* $p(N;\mu+3)$ *in the Bayesian construction* (*uniform prior*). *The modification of Roe and Woodroofe* (2001) *is shown as a dotted line.*

the (generalized) likelihood ratio test (LRT); for each μ the acceptance region is obtained from

$$R = \frac{n(X;\mu,\sigma_0)}{\max_{\mu'\geq 0} n(X;\mu',\sigma_0)} \geq c, \tag{1}$$

subject to the coverage requirement. The LRT has recently been discussed by Perlman and Wu (1999), who emphasize that optimality (maximum power) is not necessarily the best criterion for a hypothesis test. Let μ_{best} be that value of $\mu' \geq 0$ that maximizes $n(X;\mu',\sigma_0)$ for the observation X. For $X \geq 0$, $\mu_{\text{best}} = X$. However, for $X < 0$, $\mu_{\text{best}} = 0$, giving larger values for R than for the unconstrained case. Therefore for small μ, where acceptance regions contain $X < 0$, they are shifted to the left, giving reduced power against smaller alternatives μ'. The corresponding test is biased. The construction has constant coverage for all μ and the resulting confidence intervals always contain some $\mu > 0$ as shown in Figure 5. However, for $X_0 < 0$, the interval is short, and like that of the Neyman construction, underestimates the uncertainty in μ.

The same principle is applied to the Poisson-with-background case. The construction, shown in Figure 6, yields intervals that always contain some $\mu > 0$, but are short for $N < b$, showing the same behavior as the Gaussian case. For $N = 0$, the interval significantly depends upon b.

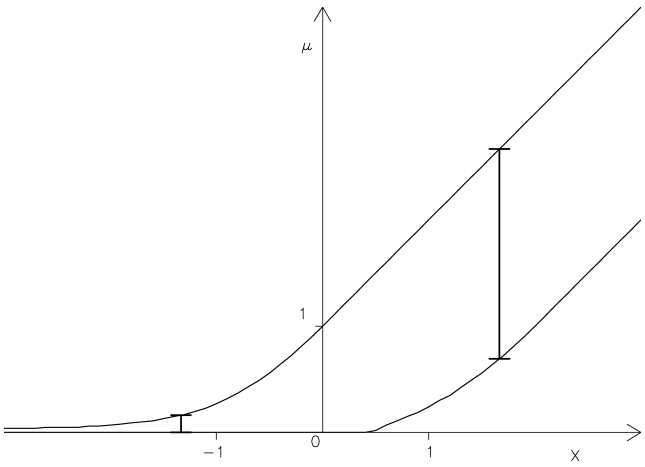

FIG. 5. *The* 68.27% *confidence belt for* $n(N;\mu,1)$ *with nonnegative* μ *in the Unified approach.*

A construction called the New Ordering approach is proposed by Giunti (1999), in which μ_{ref} replaces μ_{best} in the expression for R, where $\mu_{\text{ref}} \geq 1$ is the Bayesian expectation value for the measurement N. That $\mu_{\text{ref}} \geq \mu_{\text{best}}$ increases the shift to the left of the acceptance region, thus further diminishing the power of the test against $\mu' < \mu$. This Bayesian-motivated frequentist construction gives less restrictive upper bounds than the classical Neyman and Unified methods, as shown in Figure 6.

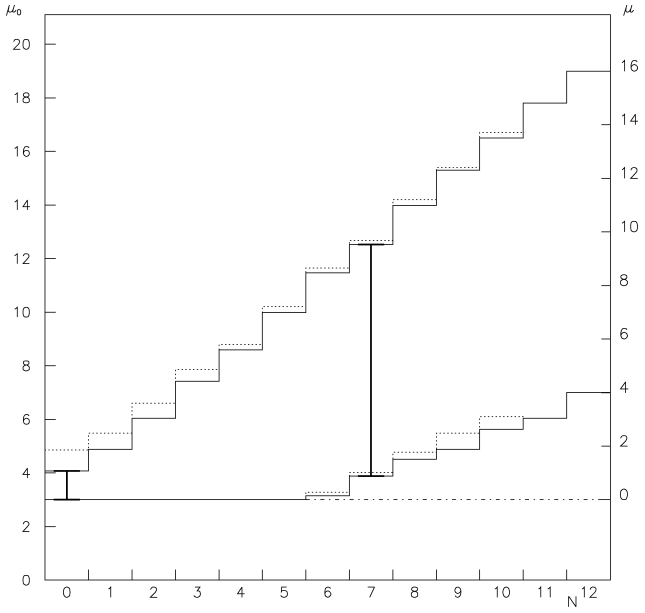

FIG. 6. *The* 90% *confidence belt for* $p(N;\mu+3)$ *in the Unified* (*solid*) *and New Ordering* (*dotted*) *constructions, the latter for* $N \leq 10$.

2.2.2 *Maximum Likelihood Estimator approach.* Mandelkern and Schultz (2000a) argue that X (or $N-b$) is an unsuitable estimate for the nonnegative parameter μ. They obtain an estimator μ^* from a likelihood function which incorporates the constraint $\mu \geq 0$. For the Gaussian-with-boundary problem we have

$$L(\mu) = \prod_{i=1}^{N} n(X_i; \mu, \sigma_0)\theta(\mu), \tag{2}$$

where $\theta(\mu) = 1$ for $\mu \geq 0$ and 0 otherwise. Maximizing L for the case of a single measurement X leads to $\mu^* = \max(X, 0)$ and the pdf,

$$P(\mu^*; \mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(\mu^* - \mu)^2}{2\sigma_0^2}\right] + \delta(\mu^*)\Phi(-\mu/\sigma_0), \tag{3}$$

which is normal for $\mu^* > 0$ plus a Dirac delta function at $\mu^* = 0$ multiplied by the remaining probability. A frequentist confidence belt, shown in Figure 7, is obtained using Neyman optimality, by applying the most powerful test against two-sided alternatives, in this case an unbiased test, to $P(\mu^*; \mu)$. For $\mu \leq g\sigma_0$ the test has no power against $\mu' < \mu$, resulting in an upper bound for $X < 0$ that is equal to that for $X = 0$.

The Poisson-with-background case is treated in exactly the same way, leading to the confidence belt shown in Figure 8. Unlike the other methods described here, this construction does not produce more restrictive (improved) upper limits for (negative) measurements that are known a priori to be relatively improbable. I thank Carlo Giunti (INFN Torino) for informing

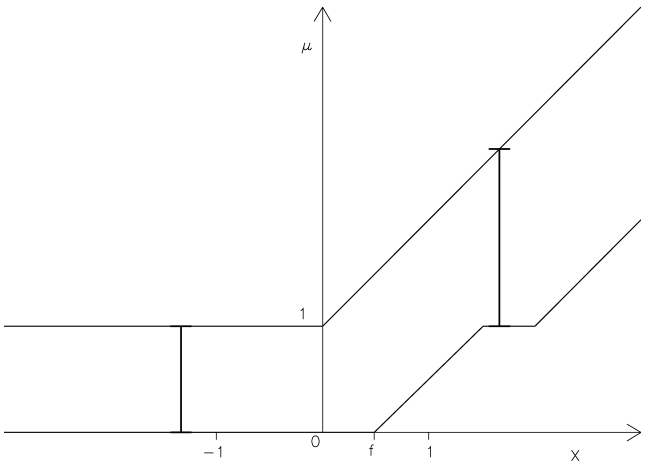

FIG. 7. *The* 68.27% *confidence belt for* $N(X; \mu, 1)$ *with nonnegative* μ *in the Maximum Likelihood Estimator approach.* $f = 0.475$.

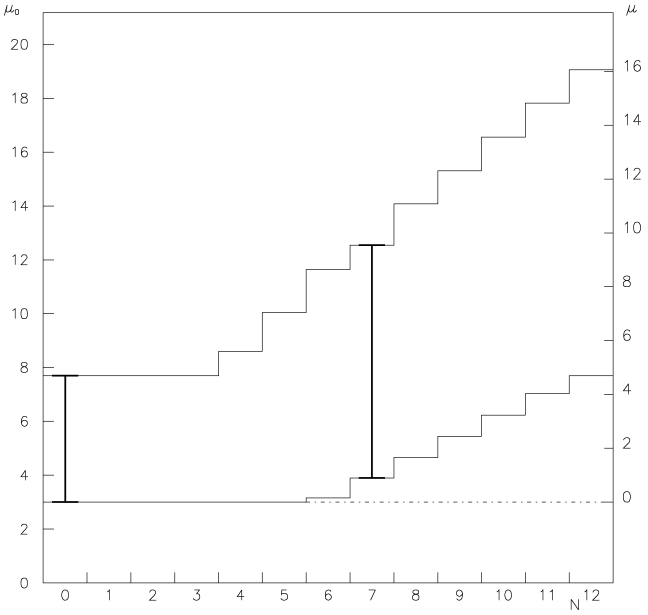

FIG. 8. *The* 90% *confidence belt for* $p(N; \mu+3)$ *in the Maximum Likelihood Estimator approach.*

me of the earlier work of S. Ciampolillo (1998), who constructed the same confidence intervals using a different rationale for obtaining the estimator.

The authors conclude that, for a particular experiment, the observation $X < 0$ gives no more information regarding μ than the observation $X = 0$. For an analysis of several experiments, one would compute an overall mean from all of the data, again leading to a nonnegative overall estimator μ^*. Thus if the overall mean were negative, the corresponding upper bound would be the same as for zero overall mean.

2.2.3 *Conditional Probability approach.* Roe and Woodroofe (1999) observe, for the Poisson-with-background problem, that since the observation $N = 0$ implies that zero signal (and zero background) is seen, the resulting estimate for μ is zero, independent of b. They argue that the confidence interval for μ for $N = 0$ must also be independent of b, as in the Bayesian construction, which effectively conditions on the observation. Noting that $N = N_s + N_b$ implies $N_b \leq N$, they consider the probability for N' conditioned on $N_b \leq N$,

$$q(N'; \mu, b, N_b \leq N) = \begin{cases} \dfrac{p(N'; \mu + b)}{P_b(N)}, & \text{if } N' \leq N, \\ \displaystyle\sum_{j=0}^{N} \frac{p(j; b)p(N' - j; \mu)}{P_b(N)}, & \text{if } N' > N, \end{cases} \tag{4}$$

where $P_b(N)$ is the cumulative Poisson probability for mean b.

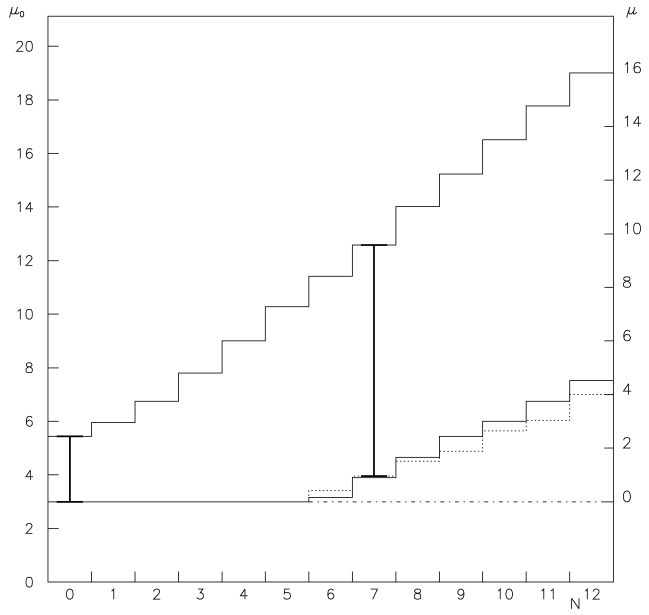

FIG. 9. *The* 90% *confidence belt for* $p(N;\mu+3)$ *in the Conditional Probability approach. The modification of Mandelkern and Schultz* (2000b) *is shown as a dotted line. The solid* (*dotted*) *contours give the limits before* (*after*) *adjustment for coverage. The unmodified construction undercovers for those* μ *where the solid lower limit contour is to the left of the dotted contour and vice versa.*

For each $N > N_b$ a *conditional confidence belt* is constructed. The acceptance region is obtained from

$$\frac{q(N';\mu,b,N_b\le N)}{\max_{\mu'\ge 0} q(N';\mu',b,N_b\le N)} \ge c, \tag{5}$$

subject to $\sum_{N'} q(N';\mu,b,N_b\le N)\ge 1-\alpha$. For the set of hypothetical experiments $[N'=N_s+N_b,\ N_b\le N]$, the construction is said to have *conditional coverage*.

An overall confidence belt is constructed, given in Figure 9 for $b=3$ and $\alpha=0.1$, with the vertical interval for N taken equal to that for $N'=N$ from the conditional confidence belt for $N>N_b$. Since $q(N';\mu,b,N=0)=p(N';\mu)$, the conditional confidence belt for $N=0$ is independent of b, thus the overall confidence interval for $N=0$ is independent of b with upper bound 2.44. We have the usual intervals for $N\gg b$ since $q(N';\mu,b,N_b\le N)\to p(N';\mu+b)$.

Because of conditioning on the observation, the overall confidence belt does not have conventional coverage. The undercoverage for $b=3$ is minimal (coverage ~ 0.87 at $\mu\sim 0.4$), and in any case coverage cannot be exact for a discrete variate. However, undercoverage is more severe for greater b; for $b=10$, the minimum coverage ~ 0.78.

As described by Mandelkern and Schultz (2000b), and also shown in Figure 9, the construction can be modified to have conventional coverage without affecting its attractive features, namely the confidence limits for $N\le b$, by retaining the left-hand boundary and adjusting the right-hand boundary so that for all μ the horizontal intervals contain probability greater than or equal to $1-\alpha$. This ad hoc modification corresponds to a particular choice of hypothesis test of significance α for all μ.

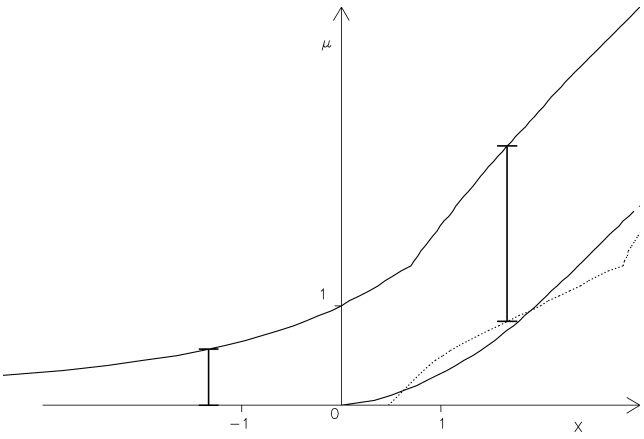

FIG. 10. *The* 68.27% *confidence belt for* $n(X;\mu,1)$ *with nonnegative* μ *in the Conditional Probability approach. The modification of Mandelkern and Schultz* (2000b) *is shown as a dotted line. The unmodified construction undercovers for those* μ *where the solid lower limit contour is to the left of the dotted contour and vice versa.*

Cousins (2000) constructs intervals from a conditional pdf for the Gaussian-with-boundary problem. The confidence belt has generally poor coverage properties but asymptotically approaches the classical confidence belt for large μ, as shown in Figure 10, where I also give the contour modified to have coverage as described by Mandelkern and Schultz (2000b). The contour for $X\le 0$ is identical to the Bayesian contour shown in Figure 2.

2.2.4 *Mixed Bayesian–Frequentist approach.* Roe and Woodroofe (2001) suggest a second method, in which they start with the Bayesian confidence belts shown in Figures 2 and 4 and make a "conservative ad hoc modification," also shown in the figures, to improve the conventional coverage properties. The latter consists of a small increase in the upper limits by applying the one-sided limits for a smaller significance $\alpha'<\alpha$. For the Gaussian-with-boundary problem, they choose $\alpha'=\alpha/2$ and obtain a 68.27% confidence belt that overcovers ($>95\%$) for $\mu<1.0$. The choice $\alpha'=0.08$ yields a 90% Poisson-with-background belt for $b=3$ that overcovers ($>97\%$) for $\mu<5.0$. The Poisson upper limit for $N=0$ is independent of b for

TABLE 1
The results of different methods for obtaining 68.27% upper limits and confidence intervals, in units of σ_0, for the unknown mean of a Gaussian of variance σ_0^2

	Classical		Bayesian		Unified		Max Lik		Cond(mod)		Bayes-Freq	
X	lower	upper	lower	upper	lower	upper	lower	upper	lower	upper	lower	upper
−3.0	0	0	0	0.34	0	0.04	0	1.00	0	0.30	0	0.34
−2.0	0	0	0	0.45	0	0.07	0	1.00	0	0.48	0	0.45
−1.0	0	0	0	0.65	0	0.27	0	1.00	0	0.64	0	0.65
0.0	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00	0	1.00
1.0	0	2.00	0.20	1.80	0.24	2.00	0.53	2.00	0.48	1.81	0.20	2.00
2.0	1.00	3.00	1.03	2.97	1.00	3.00	1.00	3.00	1.00	2.94	1.00	3.00
3.0	2.00	4.00	2.00	4.00	2.00	4.00	2.00	4.00	1.75	3.96	2.00	4.00

fixed α' but the "appropriate choice" for α', and thus the $N = 0$ upper limit, depends "weakly" on b.

2.3 Summary

Results for the methods discussed above are given in Tables 1 and 2, which contain the 68.27% bounds for the Gaussian-with-boundary problem and 90% bounds for the Poisson-with-background problem. These methods can be compared from the frequentist point of view. All have at least approximately constant coverage; in fact the desire for conventional coverage is an important concern for all of the authors. They are all unified in that they automatically produce upper limits when a small observation is made. For a nonnegative observation (I refer to Gaussian observations $X < 0$ and Poisson observations $N < b$ as negative), all give nearly the same result. The confidence belt is determined by a hypothesis test against two-sided alternatives for all possible values of the parameter. The classical Neyman construction is derived from an unbiased test. However it produces upper limits that are very small or zero for modestly negative observations. These do not convey an estimate of the experimental uncertainty. The Unified (LRT) construction follows from a weaker test against smaller alternatives but still produces limits that are overly restrictive for negative observations and the authors warn against interpreting them as measures of uncertainty. The test producing the Maximum Likelihood Estimator construction is least powerful against smaller alternatives for small μ and gives a conservative upper limit, no more restrictive than that obtained when zero is observed. The authors argue that a negative measurement is no more informative than a nonnegative measurement and should not be rewarded by a smaller upper limit. The Conditional Probability construction and Mixed Bayesian–Frequentist construction correspond to tests of intermediate power and, for a negative observation, yield upper bounds (for a non-

TABLE 2
Results of different methods for obtaining 90% confidence intervals for an unknown nonnegative Poisson signal μ in the presence of a Poisson background with known mean $b = 3.0$

	Classical		Bayesian		Unified		Max Lik		Cond(mod)		Bayes-Freq	
N	lower	upper	lower	upper	lower	upper	lower	upper	lower	upper	lower	upper
0	0	0	0	2.31	0	1.08	0	4.69	0	2.44	0	2.53
1	0	1.70	0	2.84	0	1.88	0	4.69	0	2.95	0	3.09
2	0	3.17	0	3.55	0	3.04	0	4.69	0	3.75	0	3.82
3	0	4.69	0	4.35	0	4.42	0	4.69	0	4.80	0	4.71
4	0	5.60	0	5.33	0	5.60	0	5.60	0	6.01	0	5.74
5	0	7.04	0	6.44	0	6.99	0	7.04	0	7.28	0	6.85
6	0	8.64	0	7.63	0.15	8.47	0.16	8.64	0.16	8.42	0	8.07
7	0.29	9.54	0.55	9.21	0.89	9.53	0.90	9.54	0.90	9.58	0.55	9.29
8	0.99	11.08	1.21	10.62	1.51	11.00	1.66	11.08	1.66	11.02	1.21	10.62
9	1.70	12.30	1.90	11.91	1.88	12.30	2.44	12.30	2.44	12.23	1.90	11.91
10	5.43	13.55	2.64	13.24	2.63	13.50	3.23	13.55	3.00	13.51	2.64	13.24

negative parameter) that are modestly more restrictive than when zero is observed.

3. DISCUSSION

The experimenter is placed in a quandary when the data set obtained from an experiment is a priori relatively improbable, such as a negative estimate for an intrinsically positive parameter. If the experimental result is extremely improbable, it is plausible to discard the measurement as systematically flawed or the model as incorrect. However, a procedure that discards improbable measurements a posteriori or calls for repeating an experiment depending upon its results clearly biases the final answer. In many cases it is impossible or impractical to repeat an experiment, such as results from astronomical events or from gigantic experiments which require years of operation and/or cost millions of dollars. In any case the systematic errors are likely to be different. The experimenter must report the result in such a way that it conveys information about the parameter, including a measure of statistical uncertainty or confidence.

The classical method for reporting confidence, the Neyman confidence interval, works well for unbounded parameters, but is unsatisfactory to many scientists when the parameter is bounded. The models in which recognizably improbable data are possible and this problem appears are estimation of a bounded normal mean and estimation of a Poisson mean which is the sum of an unknown signal and a known nonnegative background. I have reviewed the classical Neyman and Bayesian constructions of confidence bounds and recently proposed modifications for these models. None of the constructions seems satisfactory. A fully justified principle has not been articulated for any of the methods; some have explicitly ad hoc features. While some methods produce very stringent upper limits for only modestly improbable measurements, another gives much more conservative limits. The reaction of a scientist may be to question whether any method is meaningful if there are so many different ways of setting confidence intervals.

I conclude by asking several questions:

- Is the frequentist approach that which best satisfies the needs of scientists for reporting experimental results?
- Is the requirement for constant coverage important?
- Is the subjectiveness of the Bayesian approach overstated?
- Since Neyman optimality does not seem appropriate for choosing among frequentist constructions for this problem, is there a suitable alternative criterion?
- Are there more flexible models that avoid the difficulties described here, that for example take into account the intrinsically incomplete knowledge of experimental uncertainties? Perhaps a technique other than confidence bounds (or credible intervals) may be useful.
- Finally, independent of the method for reporting uncertainty, is it reasonable to obtain a more restrictive measure of confidence for a priori improbable data than for the most probable data? In other words does a fortuitously improbable observation give improved knowledge of an unknown parameter?

ACKNOWLEDGMENTS

I thank Professor Jonas Schultz for helpful discussions and assistance. I am also grateful for the assistance of Professor Leon Gleser and the useful comments of the *Statistical Science* reviewers.

REFERENCES

ABDURASHITOV, J. N., et al. (1999). Measurement of the solar neutrino capture rate by SAGE and implications for neutrino oscillations in vacuum. *Phys. Rev. Lett.* **83** 4686–4689.

AHMAD, Q. R., et al. (2002). Direct evidence for neutrino flavor transformation from neutral–current interactions in the Sudbury Neutrino Observatory. *Phys. Rev. Lett.* **89** 011301.

ALTMANN, M. et al. (2000). GNO solar neutrino observations: Results for GNO I. *Phys. Lett. B* **490** 16–26.

ATHANASSOPOULOS, C. et al. (1998). Results on $\nu_\mu \to \nu_e$ neutrino oscillations from the LSND experiment. *Phys. Rev. Lett.* **81** 1774–1777.

CIAMPOLILLO, S. (1998). Small signal with background: Objective confidence intervals and regions for physical parameters from the principle of maximum likelihood. *Il Nuovo Cimento* **111A** 1415–1430.

CLEVELAND, B. T., et al. (1998). Measurement of the solar electron neutrino flux with the Homestake chlorine detector. *Astrophysical J.* **496** 505–526.

COUSINS, R. D. (2000). Comment on "Improved probability method for estimating signal in the presence of background," by B. P. Roe and M. B. Woodroofe. *Phys. Rev. D* **62** 098301.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

EITEL, K. and ZEITNITZ B. (1998) The search for neutrino oscillations $\bar{\nu}_\mu \to \bar{\nu}_e$ with KARMEN. Available at arXiv/hep-ex/9809007.

FELDMAN, G. J. and COUSINS, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57** 3873–3889.

FUKUDA, Y. et al. (1998). Evidence for oscillations of atmospheric neutrinos. *Phys. Rev. Lett.* **81** 1562–1567.

FUKUDA, Y. et al. (2001). Solar ^{8}B and hep neutrino measurements from 1258 days of Super-Kamiokande data. *Phys. Rev. Lett.* **86** 5651–5655. Constraints on neutrino oscillations using 1258 days of Super-Kamiokande solar neutrino data. *Phys. Rev. Lett.* **86** 5656–5660.

GIUNTI, C. (1999). New ordering principle for the classical statistical analysis of Poisson processes with background. *Phys. Rev. D* **59** 053001.

LOBASHEV, V. M., et al. (1999). Direct search for mass of neutrino and anomaly in the tritium beta spectrum. *Phys. Lett. B* **460** 227–235.

MANDELKERN, M. and SCHULTZ, J. (2000a). The statistical analysis of Gaussian and Poisson signals near physical boundaries. *J. Math. Phys.* **41** 5701–5709.

MANDELKERN, M. and SCHULTZ, J. (2000b). Coverage of confidence intervals based on conditional probability. *J. High Energy Phys.* (**11**) 036.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. Roy. Soc. London Ser. A* **236** 333–380.

OREAR, J. (1958, 1982). Notes on statistics for physicists. Lab. for Nuclear Studies, Cornell University, CLNS 82/511. Revised version of UCRL-8417 distributed in 1958 by the Lawrence Berkeley Laboratory.

PERLMAN, M. D. and WU, L. (1999). The emperor's new tests (with discussion). *Statist. Sci.* **14** 355–381.

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

ROE, B. P. and WOODROOFE, M. B. (1999). Improved probability method for estimating signal in the presence of background. *Phys. Rev. D* **60** 053009.

ROE, B. P. and WOODROOFE, M. B. (2001). Setting confidence belts. *Phys. Rev. D* **63** 013009.

Comment

George Casella

1. INTRODUCTION

Professor Mandelkern is to be congratulated for so clearly motivating and explaining a fascinating problem in statistics. Boundedness of any kind is always a plague to statistical procedures, as our usual assumptions tend to break down at a boundary.

This problem is interesting for two reasons. First, there is the belief that the model is correct ("... a measurement with, for example, a 0.1% or greater probability may simply be an outlier ...") and second, there is the reality of frequent observance of data that have small probability under the model, as "models in which recognizably improbable data are possible" seem to produce a lot of improbable data.

My own feelings about the paper and the problem have come full circle. When I first read the paper my immediate reaction was that the model was incorrect, and some hard thinking was required by the physicists. After some discussion, thought, and re-reading of the paper, I started to think that the fault was maybe not with the model, but with the statistics. But there seemed to be no basis for faulting the statistics, so I am back to faulting the model.

George Casella is Professor and Chair, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545 (e-mail: casella@stat.ufl.edu).

2. DOING THE RIGHT THING?

There is no question that the frequentist confidence intervals are doing the right thing or, at least, doing what they promise to do. That is, if $X \sim n(\mu, \sigma^2)$, then $X \pm 2\sigma$ will cover μ 95% of the time. If μ is restricted to $[0, \infty)$, then $X \pm 2\sigma \cap [0, \infty)$ will do the same thing. The fact that negative values of X are less probable is taken into account in the location of the interval.

When an observation is "unlikely" under the hypothesized model, we either can believe that we observed an unusual occurence, or we can believe that the model is incorrect, and there is a true model under which the observation is likely. As Professor Mandelkern wants to believe that the model is correct, very negative observed values are then extremely strong evidence that the true mean is zero. The fact that the resulting intervals collapse to a point just reflects the strength of the evidence.

3. A PERFECT SETUP

There is not much more that the frequentist interval can do—it is frequentist and doesn't know how to effectively use prior information. What we then have is a perfect set-up for a Bayesian solution. Although Professor Mandelkern is concerned with the subjectivity of the Bayes inference, we can avoid that by using a *Bayesian solution* to construct the interval and a *frequentist evaluation* to evaluate it.

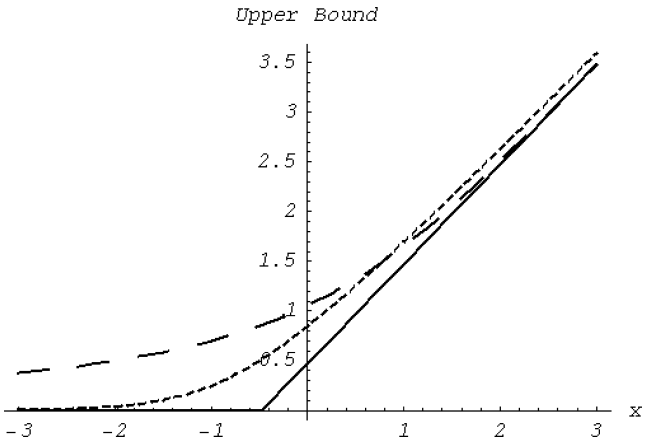

FIG. 1. *One-sided* 68% *confidence intervals from the "usual" frequentist method* (*solid*); *the frequentist-calibrated Bayes solution of* (1) (*long dashes*) *and the alternative frequentist model of* (2) (*short dashes*).

The "uniform prior Bayes" is not a good choice. Since the improper prior puts infinite mass outside of every compact set, the resulting interval has an upper endpoint that is pulled too far from zero. Alternatively, one can hypothesize an exponential or gamma prior on μ. As a simple example, for $\sigma = 1$ suppose that we take an exponential (5) prior on μ: that is, we have the model

$$(1) \quad X \mid \mu \sim N(\mu, 1), \quad \mu \sim \tfrac{1}{5}e^{-\mu/5}, \quad \mu > 0.$$

Calculation of the posterior distribution, $\pi(\mu \mid x)$ is straightforward, and the upper endpoint of an interval $[0, \mu_B(x)]$ can be obtained as the solution to the integral equation $\int_0^{\mu_B(x)} \pi(\mu \mid x)\, d\mu = 1 - \alpha_B$, where α_B is a specified value.

To eliminate the subjectivity in the Bayes solution, we now choose α_B so that the frequentist coverage probability of $[0, \mu_B(x)]$ is at the desired level. For example, to attain coverage probability 0.68, we set $1 - \alpha_B = 0.75$, and get the interval given in Figure 1.

4. ENHANCING THE FREQUENTIST SOLUTION

Although it may be reasonable to assume that $X \sim n(\mu, \sigma^2)$, we should model a bit further. The previous section showed that a simple Bayesian solution, with a frequentist calibration, will produce reasonable intervals. But what can be done from a pure frequentist view? First, it seems reasonable to attempt to further model the error structure (especially near zero), and a first thought is a random effects model,

$$X \mid \mu, a \sim N(\mu + a, \sigma^2), \quad a \sim N(0, \sigma_a^2)$$
$$\Longrightarrow \quad X \mid \mu \sim N(\mu, \sigma_a^2 + \sigma^2).$$

While this model lessens the "point-mass" problem, it does not alleviate it, merely resulting in a straight line confidence bound that hits the axis at a smaller value of x.

As frequentists, we can go a bit further with modeling the error variance and suppose that it is function of μ with the necessary properties,

$$(2) \quad X \mid \mu \sim N\big(\mu, \sigma^2 h(\mu)\big), \quad h(0) = \infty,\ h(\infty) = 1.$$

One function that satisfies these properties is $h(\mu) = [1 + \log(1 + \mu^{-1})]^2$, with the resulting one-sided confidence interval given in Figure 1. As can be seen, it provides a reasonable upper bound, not collapsing to zero, being in between the usual frequentist interval and the frequentist-calibrated Bayesian interval.

5. IT REALLY MUST BE THE MODEL

At this point it seems clear that it really must be the model, or at least we must come to the realization that the model of first choice is inadequate to explain the situation. Although it is not necessary to totally abandon the simplicity of assuming $X \sim n(\mu, \sigma^2)$, $\mu > 0$, this assumption is only part of what is needed to accurately model the entire process.

When there is strong prior information, it seems that we must use this information, in the model, in a strong way. The frequentist paradigm, while a preferred one for evaluation, is not the place for building models. That exercise better fits under the Bayesian umbrella, where hierarchical models like (1) can often provide us with both understanding and accurate representation of a process.

It is important to differentiate between a model (or a solution), and the mechanism that is used to evaluate the worth of the model (or solution). The hierarchical Bayes model (1) and the frequentist variance model (2), come from different places. However, each can produce intervals that maintain a specified level of frequentist coverage and avoid the "paradox" of collapsing to a point.

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-99-71586.

Comment

Leon Jay Gleser

INTRODUCTION

I cannot resist commenting on Professor Mandelkern's most interesting and timely paper. My own research has concerned some of the issues that bother him, particularly the meaning and measurement of uncertainty and the role of measures of uncertainty in the combining of information.

FREQUENTIST CONFIDENCE INTERVALS AND RECOGNIZABLE SUBSETS

Frequentist confidence intervals are often described in textbooks as conveying both a point estimator of a parameter (usually the midpoint of the interval) and an indication of one's uncertainty in that estimate after the data is drawn (the minimal probability of coverage, the interval length). However, the probability of coverage of a confidence interval is a predata measure of uncertainty; its role as a postdata measure of uncertainty depends upon assertions such as "If the probability that the estimator is within d units of the parameter is 0.95, then the uncertainty that the parameter is within d units of the estimator is also 0.95." Such an assertion is not always true, as illustrated by Professor Mandelkern's example of estimating a mean known to be positive. Although the standard Neyman 95% confidence interval has minimum (predata) probability of 0.95 of covering the true value of the mean, if the sample mean is observed to be two standard deviations or more to the left of zero, then the (postdata) conditional probability that the interval will contain the true value of the mean is zero.

The subset of samples having the property that the sample mean is two standard deviations to the left of zero would have been called a "recognizable subset" by Fisher (1956). A classic example of such a recognizable subset in the case of confidence intervals is the set of samples in which a one-sided test of hypothesis concerning the mean is rejected (or accepted); see Brown (1967). Examples such as these led Scheffé (1977) to advocate using confidence regions to construct tests of hypothesis, rather than vice versa, as in Neyman's construction of confidence intervals. Buehler (1959), and later Robinson (1979), introduced the notion of *conditionally admissible* tests and confidence intervals—those procedures whose frequentist control of error (coverage probability, level of significance) was not adversely affected by the realization that a given data set belonged to a recognizable subset of samples. A good survey of this literature can be found in Casella (1987). Although many of the classic normal-theory procedures were found to be conditionally admissible, Professor Mandelkern's example shows that the classic Neyman confidence interval for the mean is not conditionally admissible in the case of estimating a positive mean. Extension of this result to other cases of bounded parameters is obvious. In short, once something about the data is known, it is possible for the frequentist properties of the confidence interval to change; *the predata measure of risk is not necessarily the correct postdata measure of uncertainty*.

The Neyman–Pearson theory of statistical inference and its generalization to statistical decision theory by Wald were intended to guide choice of a statistical design and corresponding analysis of the data. Necessarily, this choice (because it involves the design) is made predata. Choice of a design–analysis combination is viewed in the same way as a physicist might view the choice of a measuring device. Such a device is chosen to have a required level of accuracy; once the device is chosen, its measurements are treated as if they are error free. Similarly, a design–analysis combination is chosen to control certain probabilities of false decision (or risks of false decision), but once the data is gathered and a conclusion reached, further assessment of uncertainty is not considered. (The decision is made, and its risks have been described.) It is unfortunate that Neyman's confidence intervals and Fisher's fiducial intervals are often confused. Fisher meant his fiducial intervals to describe postdata uncertainty, whereas Neyman probably did not. (Indeed, Neyman called his intervals "interval estimators" and talked of "confidence."

Leon Jay Gleser is Professor, Department of Statistics, University of Pittsburgh, 2732 Cathedral of Learning, Pittsburgh, Pennsylvania 15260 (e-mail: ljg@stat.pitt.edu).

On the other hand, Fisher certainly thought of Neyman's approach as a competitor to his own, and I don't find any evidence that Neyman disabused him of this idea.)

One frequentist response to the problems which recognizable subsets and conditional inadmissibility raise is to try to establish a theory of conditional frequentist inference, trying to replace Fisher's device of conditioning on ancillary statistics with something that applies more generally (see, e.g., Fraser and Reid, 1995, 2001). Another approach, implicit in the bootstrap approach to confidence intervals and made explicit by Kiefer (1977), is to try to estimate from the data the risk of the procedure under the true distribution that governs the data. The latter approach, however, is subject to the same conditional inadmissibility criticism as the unconditional frequentist approach.

LIKELIHOOD AS A MEASURE OF UNCERTAINTY

The Likelihood Principle (see Casella and Berger, 2002) proposes the likelihood function as a measure of evidence about the parameters of the model used to describe the data. What does the likelihood function tell us in the two problems discussed by Professor Mandelkern?

In each case, as the data become less and less likely under the family of distributions assumed by the model, the narrowing of the interval corresponds to an increasing concentration of the likelihood at the boundary value 0 of the parameter. Consider, for example, the problem of estimating a positive mean assuming normality. Following Professor Mandelkern in assuming that the sample mean X is normally distributed with mean μ and known standard error d, we can write the likelihood function for μ as

$$L(\mu) = f(X \mid \mu)/f(X \mid 0) = \left[\left(\mu X - (1/2)\mu^2\right)/d^2\right],$$

where $f(x \mid \mu)$ is the density function of X when the population mean equals μ. (Recall that the likelihood function is defined up to a constant of proportionality that may depend on X.)

It is easy to see that when $X < 0$ and μ is positive, $L(\mu)$ is strictly decreasing in μ, and further that the rate of decrease increases the more negative that X becomes. Thus, the likelihood function of μ becomes more and more concentrated about 0 the more negative X becomes, eventually becoming totally concentrated (spiked) at 0. Consequently, if one is willing to base inference on the likelihood, more and more unusual X-values under the model lead to tighter and tighter inferences about μ, in agreement with the evidence (uncertainty) indicated by the classic Neyman confidence interval for μ. It follows that any confidence intervals that keep a constant width as X becomes more negative, as some of the physicists seem to desire, are indicating not necessarily what the data shows through the model and likelihood, but rather desiderata imposed external to the statistical model.

Notice that basing inference about μ solely on the model does not permit the data to challenge the model. Consequently, there is no way for the likelihood to reflect uncertainty about the model. It is possible, because of the way I have defined $L(\mu)$, to test each possible value of μ against the alternative that X is normally distributed with zero mean and standard deviation d. In this case, sufficiently negative values of X suggest that μ equals 0; less negative values of X favor either 0 or values close to 0. In many experiments of the sort that Professor Mandelkern describes, an inference that the parameter is 0 suggests either that no quantity was measured (X reflects only background noise) or else that measurement was insufficiently accurate to detect the value of the parameter.

In more complex parametric models than the ones presented by Professor Mandelkern, it is impractical to present the likelihood of a vector parameter as a measure of uncertainty. Statisticians seem to disagree how to construct a measure of evidence or uncertainty for a single parameter based on a multiparameter likelihood. Two suggestions often advanced are (1) to report the marginal (integrated) likelihood of a parameter, and (2) to report a pseudo-likelihood (with other parameters replaced by their modal estimators). Other arguments and issues arise with the use of likelihood functions. In some problems, likelihood functions are difficult to compute. Nonetheless, a likelihood function of a parameter is one measure of uncertainty for that parameter that directly summarizes what the data has to say about the value of that parameter *through the model*. Consequently, a likelihood function for a parameter can be communicated without ambiguity from one scientist to another, and can be combined by multiplication with other likelihoods to provide information across studies about the value of that parameter.

SUMMARY

Statistical measures of uncertainty will have their limitations, and thus it is necessary to keep clearly in

mind what information various ways of presenting uncertainty provide. Of particular importance is the distinction between predata probabilities of error for statistical design–analysis procedures and postdata measures of uncertainty for the values of the parameters (particularly if given as probabilities). Also, it should be kept in mind that inference procedures designed to work in the context of a given model, particularly likelihood-based procedures, cannot test that model directly. Checks of the model require separate inference procedures. In the problems Professor Mandelkern describes, checks of the model are likely to require meta-analytic examination of the results of similar experiments; for example, does the problem of estimators having values outside of the parameter space occur more often than the model would predict? If so, some thought should be given to generalizing the model to account for this extra variability.

Comment

Larry Wasserman

UNCERTAINTY ABOUT WHAT?

Strictly speaking, the Neyman interval is correct in that it has the advertised coverage. The perception that the confidence intervals are inadequate arises partly because it fails Professor Mandelkern's property (iv) that the interval should "convey an estimate of experimental uncertainty." The question is: what is experimental uncertainty?

One type of experimental uncertainty is the risk of the estimator. Consider n observations from a regular model with unrestricted scalar parameter θ. The usual interval is $I = [\hat{\theta} - \hat{se} z_{\alpha/2},\ \hat{\theta} + \hat{se}\, z_{\alpha/2}]$ where $\hat{se}$ is the estimated standard error and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard Normal. This interval not only has correct coverage (asymptotically) but the length of the interval is proportional to the square root of the risk (mean squared error) of the point estimator. This is because the MLE $\hat{\theta}$ has bias $O(n^{-1})$ and standard deviation $O(n^{-1/2})$. The length of the interval will shrink to 0 if and only if the risk goes to 0.

In the Normal problem with nonnegative mean μ, the confidence interval fails to represent *estimator* uncertainty. For example, when X is very negative, the length of the Neyman interval is 0, which clearly underestimates the risk. The risk (with $\sigma = 1$) of the MLE is

$$R(\mu) = \mu^2[1 - \Phi(\mu)] - \mu\phi(\mu) + \Phi(\mu)$$

and hence $1/2 \le R(\mu) \le 1$. We could supplement the confidence interval with an estimate of risk. For example, the confidence interval for μ leads immediately to a confidence interval for $R(\mu)$ which degenerates to $1/2$ when X is highly negative. At least this makes clear that there is still experimental uncertainty. To avoid the dangers of misinterpretation, we could even demand that the length of the interval never be allowed to shrink below some function of the estimated risk. We will then pay a price in conservativeness of the interval but the payoff is that users will not misinterpret the degenerate interval to imply that their is no uncertainty.

Another type of experimental uncertainty is whether the model is correct. Professor Mandelkern is correct that posthoc outlier rejection and posthoc model changes lead to bias. But it would be straightforward to modify the model to include the realistic possibility of model violations and this will lead to more intuitive confidence intervals. A simple example of such a model is the contaminated Normal $(1 - \varepsilon)N(\mu, \sigma) + \varepsilon N(\mu, K\sigma)$ where ε is the probability of an outlier and $K > 1$. With only one observation (and continuing to take σ known), (μ, ε, K) are not identifiable and the confidence interval for μ is $[0, \infty)$. One could argue that this is the correct expression of our uncertainty. This shows how strongly the inferences depend on model assumptions. Or perhaps one might simply fix K and ε at reasonable values since this is better than taking $\varepsilon = 0$. With more observations, one could treat all the parameters as unknown.

Larry Wasserman is Professor, Department of Statistics, Carnegie Mellon University, Baker Hall 228A, Pittsburgh, Pennsylvania (e-mail: larry@stat.cmu.edu).

Comment

David A. van Dyk

1. A PRIORI UNLIKELY DATA OR MODEL MISSPECIFICATION?

The seemingly poor properties of standard confidence intervals given a priori unlikely data described by Professor Mandelkern have received much attention in physics. I am delighted that the author has solicited the advice of the statistical community through this publication and that the editors of *Statistical Science* have given me the opportunity to comment.

It seems to me that the basic difficulty is summarized well in the final question of Mandelkern's discussion, namely, "Is it reasonable to obtain a more restrictive measure of confidence for a priori unlikely data than for the most probable data." To answer this question, we consider the Poisson case with $N \sim \text{Poisson}(\mu + b)$, where b is *assumed* to be known from background calibration. Figure 1 illustrates the sampling distribution of the 95% confidence interval for μ when $\mu = 1.25$ and $b = 2.88$. The simulation values are taken from the description of the KARMEN 2 experiment given in the article and in Roe and Woodroofe (1999). The confidence intervals were computed using the frequentist method of Garwood (1936) for $\mu + b$ and subtracting off b. In Figure 1 the horizontal range of each rectangle corresponds to the confidence interval for the given observed value of N and the height of each rectangle corresponds to the sampling probability of the confidence interval; the dashed vertical line indicates the supposed value of $\mu = 1.25$. That the confidence interval grows longer as N increases is readily apparent in Figure 1. Thus, unlikely values of N *that are small* can result in highly restrictive measures of confidence, that is, narrow intervals. Of course, this is wholly dependent on the choice of scale; the corresponding intervals for $\log(\mu)$ have finite length only for $N \geq 8$. Even on the original scale, this property is not surprising; smaller values of N make smaller values of $\mu + b$ and the correspondingly smaller Poisson variability more credible. Although the situation is intensified by the known background intensity, since $\mu + b$ is bounded below not by zero but by b, the confidence intervals remain a reasonable frequentist summary *under the model*. The reason these frequentist intervals are so short when $N = 0$ is that *under the model and given b* only very small values of μ make $N = 0$ at all likely.

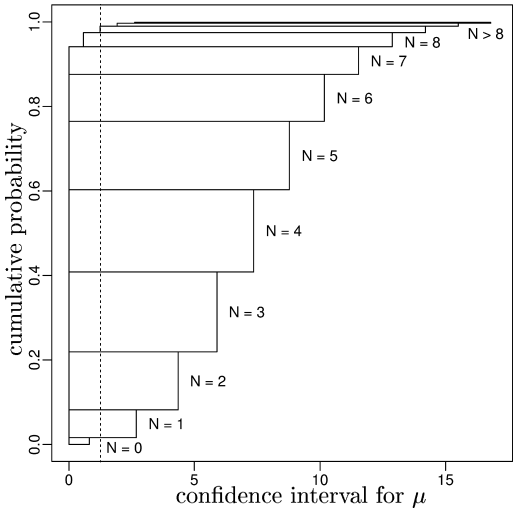

FIG. 1. *The sampling distribution of the standard* 95% *Poisson confidence interval for* μ *with* $b = 2.88$ *and* $\mu = 1.25$. *The horizontal width of each rectangle is the confidence interval for the corresponding value of* N; *the height of each rectangle indicates the sampling probability for the interval. The figure illustrates that if the model is correctly specified, very short intervals should be rare.*

I emphasize that it is unquestionably reasonable that smaller values of N result in shorter frequentist intervals *but only if the model is a plausible representation of the data generating mechanism*. The italicized caveat is critical. *For any probability calculations* (*frequentist or Bayesian*) *to be meaningful and relevant the statistical model must adequately represent the data*. In theory, this means that if the experiment were repeated many times, the resulting counts would follow a Poisson distribution with intensity $\mu + b$ for some $\mu \geq 0$. Of course, models should be viewed as tools that offer a parsimonious summary of the relevant aspects of the data, rather than a complete and full description. Thus, model selection is inherently a subjective art: it is dependent not only on the characteristics of the data and data collection process but also the aims and intentions of the scientist. Nonetheless, to be useful a model must

David A. van Dyk is Associate Professor, Department of Statistics, Harvard University, one Oxford Street, Cambridge, Massachusetts 02138 (e-mail: vandyk@stat.harvard.edu).

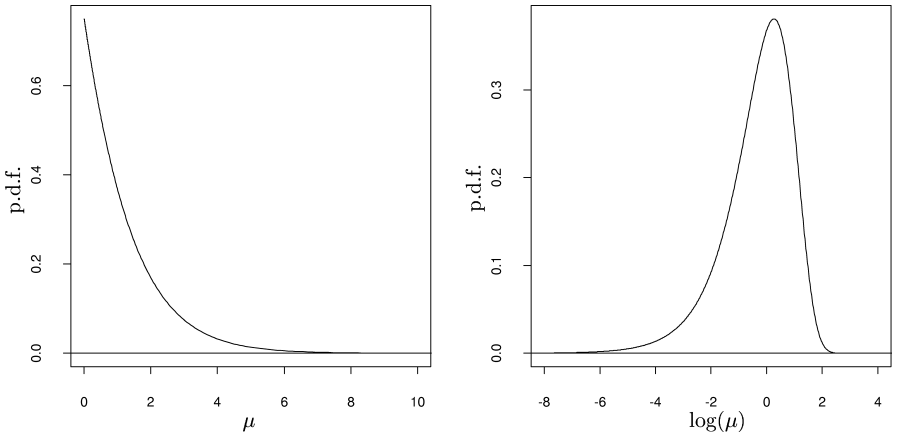

FIG. 2. *The posterior distribution of* μ (*first panel*) *and* $\log(\mu)$ (*second panel*) *under a point mass prior for* β ($b = 3$) *and with* $N = 1$. *The figure illustrates the effect of the symmetrizing log transformation.*

offer a credible summary of the character and variability of the scientifically interesting aspects of the data.

A second observation that can be drawn from Figure 1 is that very short confidence intervals are quite uncommon. Indeed, frequentist confidence intervals are not designed to behave well for particular realizations of the data, but rather are designed to have predictable coverage *in repeated realizations*; if one is interested in conditioning on the particular realization of the data, in principle Bayesian methods are better suited. Indeed, the interval in Figure 1 resulting from $N = 0$, $(0.00, 0.081)$ has a sampling probability of about 1.6%; if the probability of such a short interval is considered too high, a higher confidence coefficient should be used. Of course, unlikely events do occur. But if they occur often, one might begin to question how their likelihood is being quantified. In particular, one would expect that such unsatisfactory intervals would be quite rare in physics experiments. This, however, does not seem to be the case. Instead there are a variety of proposed statistical quick fixes and even capacity-crowd workshops devoted to the topic at CERN and Fermi Lab, all presumably motivated by the common occurrence of unsatisfactory intervals. (The Workshop on "Confidence Limits" was held at CERN January 17–18, 2000; see cern.web.cern.ch/CERN/Divisions/EP/Events/CLW/Welcome.html. The Workshop on Confidence Limits was held at Fermi Lab March 27–28, 2000; see conferences.fnal.gov/c12k/.) I wonder if anyone has undertaken a systematic investigation of how frequently major physics experiments result in unsatisfactory intervals. Such an investigation is clearly mandated.

Since retaining an inadequate model can have unpredictable consequences for the resulting statistical inference, careful model checking is unavoidable. Although the methodology of model selection, checking, and diagnosis is among the most controversial and ill-defined topics of statistical science, in this case the situation seems clear cut. If a confidence interval is empty (e.g., as with $N = 1$ in Figure 2 of the paper) the observed data is unlikely, as measured by the confidence coefficient, *for any value of the parameter*. Put another way, we can reject the null hypothesis that $\mu = \mu_0$ for any $\mu_0 \geq 0$. By any measure, the model does not offer an adequate representation of the scientifically most interesting aspect of the data. This difficulty cannot be addressed by reformulating the procedure for computing the confidence interval under the same model. Thus, the basic notion of developing new, creative, or ad hoc formulations of interval estimates under the same model is misguided in this situation.

Mandelkern correctly points out that discarding data or changing the model a posteriori can bias the final answer. As we shall see, however, retaining an inadequate model is not the path to unbiased inference! Rather than worrying about the biases that are introduced by model checking, the science would be better served by learning about the form of an adequate model that can be used in future experimentation and analysis.

2. RESPECIFICATION OF THE MODEL

From my distant vantage point it is impossible to propose a model that might be more suitable to the data. Thus, my goal is this section is not to propose a specific solution (indeed, there is surely no all-purpose solution), but rather to illustrate the construction of highly structured models and how they can be used for statistical inference. A more detailed and specific

example from my own work in high energy astrophysics, which uses Poisson models and accounts for background contamination, blurring, absorption and stochastic censoring of counts can be found in van Dyk, Connors, Kashyap and Siemiginowska (2001), Protassov, van Dyk, Connors, Kashyap and Siemiginowska (2002) and van Dyk and Hans (2002).

For illustration, I propose to generalize the Poisson model in two ways. First by allowing for stochastic censoring of the data; based on the problems outlined in the paper it seems plausible that some instruments do not detect as many events as some might hope. Second, I do not assume that the background intensity, b, is a known constant. Indeed, it is reported with error bars for the KARMEN experiment and Mandelkern reports that b is "measured independently" or "estimated," presumably with error. Thus, I propose,

$$N \sim \text{Poisson}\{\alpha(\mu + \beta)\}, \tag{1}$$

where α is the proportion of events that are recorded (e.g., not absorbed or otherwise missed), β is the background intensity, and μ is the source intensity. Of course, not all three parameters in (1) are jointly identifiable. This is not a reason to fix $\alpha = 1$ and $\beta = b$ but rather a reason to aim to design experiments that can identify the parameters, for example, by obtaining additional counts due only to background,

$$N_B \sim \text{Poisson}(\beta), \tag{2}$$

or by producing M events and observing how many are detected. Undoubtedly, some such instrumental calibration is already done—what is important here is that the uncertainty involved in calibration must be accounted for in the final analysis.

In the remainder of this section, for simplicity we fix $\alpha = 1$, treat μ as the parameter of interest, and treat β as a nuisance parameter. We discuss Bayesian and frequentist intervals for μ under (1) and investigate the consequences of the model misspecification of fixing $\beta = b$ when really the data is generated under (1).

In a Bayesian analysis we can replace (2) with a prior distribution for β. This need not be and indeed should not be a subjective prior distribution. Rather data or simulations can be used to construct the prior distribution; for example, with the KARMEN experiment the prior specification can reflect such information as $b = 2.88 \pm 0.13$. In this case, we specify a conjugate gamma prior distribution with shape and scale parameters ξ_β and ψ_β, respectively; that is, $\beta \sim \gamma(\xi_\beta, \psi_\beta)$. Likewise, we specify a prior distribution for μ, $\mu \sim \gamma(\xi_\mu, \psi_\mu)$, but this prior distribution is ordinarily uninformative; for example, for a flat prior on μ we set $\xi_\mu = 1$ and $\psi_\mu = +\infty$. The highly skewed character of the resulting *marginal* posterior distribution for μ,

$$p(\mu \mid N) = \int_0^\infty p(\mu, \beta \mid N)\, d\beta, \tag{3}$$

is evident in the first panel of Figure 2, which plots the posterior distribution resulting from $N = 1$ and a point mass prior for β; that is, β is fixed at $b = 3$. Point estimates are computed using the posterior mean, but only after a transformation which aims to symmetrize the distribution, in this case the log transformation; see the second panel of Figure 2. Equal tailed interval estimates are invariant to transformation and should correspond closely to the shortest interval under a symmetrizing transformation, at least for unimodal distributions. Alternatively, highest posterior density intervals or upper bounds can be computed. The effect of the prior specification (i.e., error in b) is illustrated in Figure 3, which varies ψ_β but fixes $\xi_\beta = 3/\psi_\beta$ and thus fixes the prior mean of β at 3. A point mass prior distribution, which fixes β at $b = 3$ corresponds to $\psi_\beta = 0$; as ψ_β increases the intervals grow wider.

Frequentist regions for (μ, β) can also be computed. In this case, however, one generally incorporates information regarding β through data, for example, as in (2) rather than via a prior distribution. A joint confidence region (with confidence coefficient $1 - \alpha$) can be com-

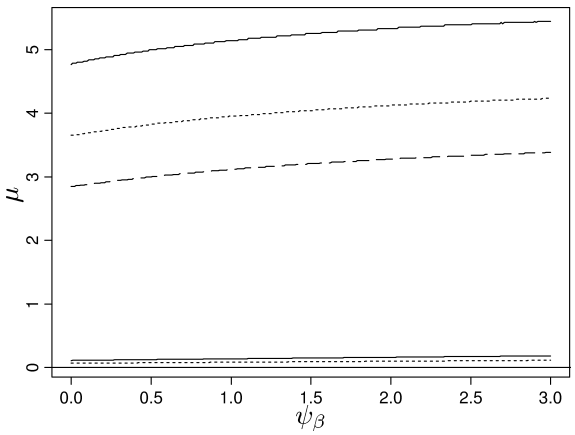

FIG. 3. *The effect of the error in b on the* 90% *posterior interval for* μ. *The figure illustrates how the confidence intervals for* μ *grow wider as the error in b increases, measured here via the prior parameter,* ψ_β. *The solid lines correspond to the upper and lower limits of the highest posterior density interval under the log transformation of* μ; *the dotted lines corresponds to the upper and lower limits of the equal tailed interval*; *and the dashed line is an upper limit.*

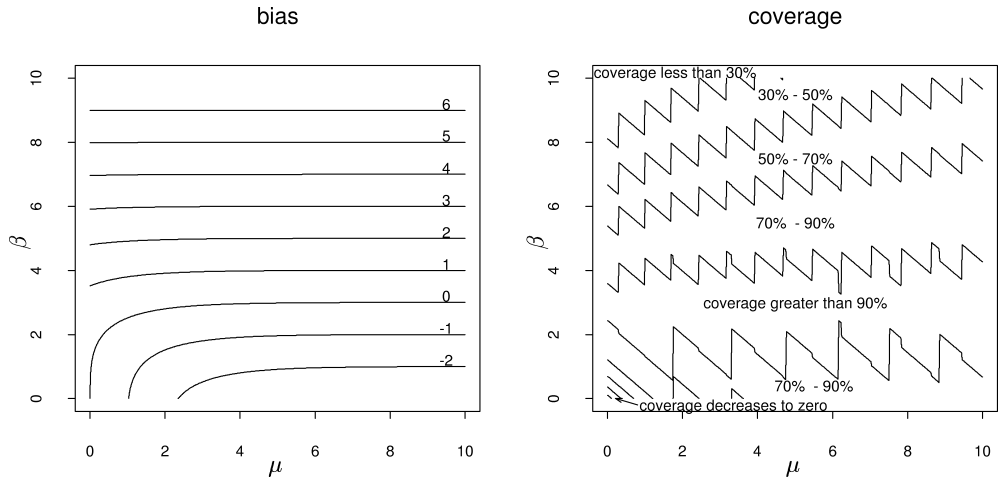

FIG. 4. *Bias of the maximum likelihood estimate* (*first panel*) *and under coverage of the nominal* 90% *interval* (*second panel*) *caused by model misspecification. The figures assume the data are generated according to model* (1) *with various values of* μ *and* β (*and* $\alpha = 1$), *but is fit with* β *fixed at* $b = 3$.

puted as

$$\{(\mu, \beta) : (N, N_B) \in R(\mu, \beta)\}, \tag{4}$$

where for each $\mu \geq 0$ and $\beta \geq 0$, $R(\mu, \beta)$ is a set of values of (N, N_B) such that $\Pr\{(N, N_B) \in R(\mu, \beta) \mid \mu, \beta\} \geq 1 - \alpha$. Such regions are often constructed as acceptance regions for a particular α-level hypothesis test, perhaps with attention paid to the power of the test. Constructing a frequentist "marginal" interval for β is both more subjective and analytically complicated than for the Bayesian marginal interval. Ideally, we condition on a sufficient statistic for the nuisance parameter β (Neyman, 1937), but such a statistic is not always forthcoming.

We conclude by illustrating the effect of model misspecification, by computing the bias of the maximum likelihood estimate and the coverage of the standard frequentist interval of Garwood (1936). Both the estimate and the interval are computed with β fixed at $b = 3$, but the data is generated under (1) with various values of μ and β (and $\alpha = 1$); the results appear in Figure 4. Although the bias induced by this simple model misspecification is clear, we emphasize that this is only an illustration of the perils of model misspecification. In the current situation, the error in b may be small and the effects correspondingly small. Nonetheless, frequent a priori unlikely data and empty confidence intervals are strong evidence of model misspecification. Unfortunately, the biases resulting from ignoring the misspecification are not easily quantified.

3. ARE BAYESIAN METHODS TOO SUBJECTIVE?

The subjective nature of specifying a prior distribution, as required with Bayesian methods, has been repeatedly pointed out. Here Mandelkern's first desirable feature for confidence intervals explicitly forbids basing intervals on arbitrary or subjective "principles." Of course, the principles behind Bayesian methods, that is, the principles of probability calculus, are anything but arbitrary and subjective. Indeed, the principles behind other methods may be far more subjective, especially in the presence of nuisance parameters. When given a choice, basing a frequentist interval on a more powerful test is preferred, but not at the expense of the conditionality principle, for example, conditioning on ancillary statistics. Of course, ancillary statistics and the corresponding intervals may not be unique. Even without nuisance parameters there may be no clear optimal interval; witness the variety of methods outlined in Section 2 of the paper. On the other hand, given the model (including the prior specification) the posterior distribution of the parameters of interest is uniquely defined by probability calculus.

This leaves three seemingly subjective tasks in computing a Bayesian interval: reducing the inference to an interval, selecting the likelihood, and selecting the prior distribution. The first task is not unique to Bayesian methods and there are of course guiding principles; highest posterior density intervals result in the shortest interval for a given parameterization and equal tailed (or other percentile based) intervals are invariant to one-to-one monotone transformations. Nonetheless,

the real problem stems from a desire to construct an interval to summarize the posterior distribution. The posterior distribution itself is invariant to transformations and is a much more informative summary of the statistical inference. It should be preferred over any particular Bayesian interval.

The second task, specifying a model for the sampling distribution (or likelihood), is truly subjective. In any given analysis some models are clearly inappropriate, but there always remain models among which the data are unable to distinguish. In some cases we make a parsimonious choice and in others the choice has little effect on the final analysis. In any case, specification of the sampling distribution is a subjective task common to all statistical analyses. The choice is critical, sometimes highly influential, and thus should be approached with care and checked when possible against the data, rather than holding to an arbitrary initial proposal.

I save the seemingly most potent criticism for last. Indeed in her discussion of Bayesian methods as a potential solution to the difficulties encountered by frequentist methods in the presence of nuisance parameters, Reid pointed to the necessary specification of a "prior [distribution] for a high-dimensional nuisance parameter" as justification for her conclusion that "the fact that the Bayesian approach is logically consistent strikes me as somewhat irrelevant" (Reid, 1995, see also McCullagh, 1995). Here, however, these concerns do not seem to apply. In particular, the prior distributions for nuisance parameters are neither subjective nor uninformative; they are based on calibration data and merely enable the inference to reflect uncertainty in the calibration variables. The parameter of interest is of low dimension, dimension one in the current model formulation, where $p(\mu) \propto 1$ is an obvious choice. Even with higher dimensional parameters, hierarchical models or hierarchical prior specifications serve to mitigate Reid's concern. The sensitivity of the final analysis to the choice of prior distribution as well as the frequency properties of the resulting intervals can be explored. Indeed, in this case, a prior distribution seems neither difficult to specify nor subjective, at least not when compared with the subjective nature of the principles underlying the alternatives.

ACKNOWLEDGMENT

The author thanks Tom Loredo for pointing out several references and the recent relevant workshops at CERN and Fermi Lab. Funding was partially provided by NSF Grant DMS-01-04129 and by NASA Contract NAS8-39073 (CXC).

Comment

Michael Woodroofe and Tonglin Zhang

We thank Professor Mandelkern for his informative review of statistical problems that have been plaguing physicists and his attempts to address them. We have some minor quibbles with the "desirable features," some brief comments on the Bayesian and unified methods with known b and σ^2, and more extensive comments on treating σ^2 as an estimated parameter instead of a known one.

Quibbles. In (i), statisticians have been searching for a general method that is neither arbitrary or subjective and makes intuitive sense for a long time now without any general consensus on what that method is. In (ii), there is certainly a need for a method that does not require prior information; but using prior information should not be precluded when it exists. Also, requiring equivariance under one-to-one transformations, as in (iii), rules out many intuitive optimality criteria.

Known b and σ^2. The unified method was developed explicitly to deal with problems of a restricted parameter space. It clearly provides an improvement over the Neyman intervals and has attracted a wide following among physicists. We agree with Mandelkern, however, that it can produce unbelievably short intervals. The Bayesian intervals are not especially short in the Poisson case, as is clear from Mandelkern's Figure 4. In the extreme case $N = 0$, the length of the Bayesian interval is $\log(1/\alpha)$, and this is the right answer in the absence of prior information. To elaborate,

Michael Woodroofe is Professor and Tonglin Zhang is a graduate student Department of Statistics, University of Michigan, 4082 Frieze Building, Ann Arbor, Michigan 48109 (e-mail: michael@umich.edu).

suppose that $N = B + S$, where $B \sim \text{Poisson}(b)$ and $S \sim \text{Poisson}(\theta)$ are independent, b is known, and θ is unknown. If $N = 0$, then $B = 0$ and $S = 0$, and common sense dictates that the confidence interval for θ should be the same as if we had observed (only) $S = 0$. When $S = 0$ is observed the Bayesian and Neyman upper credible/confidence bounds are both $\log(1/\alpha)$, and the unified bound is only slightly larger. So we do not believe that the Bayesian intervals are counterintuitive in the Poisson case.

In the normal case, the Bayesian credible interval shrinks to $\{0\}$ as $x \to -\infty$; that is, letting $u(x) = x + d(x)$ denote the upper credible limit,

$$\lim_{x \to -\infty} u(x) = 0. \tag{1}$$

This may appear counterintuitive, for reasons given in the paper, but it is consistent with the solution to the Poisson case (which we maintain is the right solution). For if b is large, then $X = 2[\sqrt{N} - \sqrt{b}]$ is approximately normal with mean $\mu = 2[\sqrt{b+\theta} - \sqrt{b}]$ and unit variance; and if $N = 0$, then upper credible bounds for θ and μ are $\log(1/\alpha)$ and

$$2\left[\sqrt{b + \log\left(\frac{1}{\alpha}\right)} - \sqrt{b}\right] < \frac{1}{\sqrt{b}} \log\left(\frac{1}{\alpha}\right) = \frac{2}{|x|} \log\left(\frac{1}{\alpha}\right).$$

So (1) does not seem unreasonable when applied to $X = 2[\sqrt{N} - \sqrt{b}]$. To the extent that (1) appears counterintuitive, it does so because large values of $-x$ cast doubt on the model.

Estimated nuisance parameters. Reassessing the model can introduce biases, as Mandelkern says, but it is necessary sometimes and does not always introduce severe biases. In the present context, the values of σ^2 and b were assumed known. This is almost certainly an oversimplification (recalling that b was given by $b = 2.88 \pm 0.13$ in an example). We show below how treating σ^2 as an estimated, rather than known, parameter in the normal case leads to important differences in the nature of the confidence bounds for negative x. Thus, suppose that there are independent data $S^2 \sim \sigma^2 \chi_r^2 / r$ and $X \sim \text{Normal}(\mu, \sigma^2)$, where $0 \le \mu < \infty$ and $0 < \sigma^2 < \infty$ are both unknown, but $r \ge 1$ is known. Then σ^2 is unbiasedly estimated by S^2, and the likelihood function is

$$L(\mu, \sigma^2 \mid x, s^2) \propto \frac{1}{\sigma^{r+1}} \exp\left[-\frac{rs^2 + (x-\theta)^2}{2\sigma^2}\right].$$

This simple change in the model has a profound effect on the nature of the Bayesian credible intervals for large values of $-x/s$.

Estimated σ^2: The Bayesian view. In the enlarged model, credible intervals for μ may be obtained from the (improper) prior $d\mu\, d\sigma^2/\sigma^2$ over $0 \le \mu < \infty$, $0 < \sigma^2 < \infty$. After some routine calculation, the (marginal) posterior distribution of μ is

$$g(\mu \mid x, s) = \frac{1}{H_r(t)} h_r\left(\frac{\mu - x}{s}\right),$$

where $t = x/s$ and h_r and H_r are the density and distribution function of the t-distribution on r degrees of freedom. Equivalently, the posterior distribution of $(\mu - x)/s$, given $X = x$ and $S = s$, is a t-distribution with r degrees of freedom, conditioned to exceed $-t$. There is then a complete analogue with the results of Roe and Woodroofe (2001). Letting $t_0 = H_r^{-1}[1/(1+\alpha)]$, a level $1-\alpha$ Bayesian credible interval for μ has the form $[s\ell(t), su(t)] = [\max(0, x - sd), x + sd]$, where $d = H_r^{-1}[1 - \alpha H_r(t)]$ if $t \le t_0$ and $d = H_r^{-1}[\frac{1}{2} + \frac{1}{2}(1-\alpha)H_r(t)]$ if $t > t_0$. Further, a level $1-\alpha$ credible interval has (frequentist) coverage probability at least $(1-\alpha)/(1+\alpha)$, even without any ad hoc modification, and the latter bound is conservative. See Zhang and Woodroofe (2001) for the derivations.

Graphs of the $\ell(t)$ and $u(t)$ are included in Figure 1 for selected α and r. Comparing the latter figure with Mandelkern's Figure 2 shows that including an unknown σ^2 in the model changes the nature of the upper credible limit qualitatively for large values of $-x$. In the enlarged model the upper confidence limit is *decreasing* in x for fixed s when $-x$ is large and even approaches ∞ as $x \to -\infty$. The explanation for this behavior is that if $-x$ is large, then the posterior distribution of σ^2 can be quite diffuse, even if

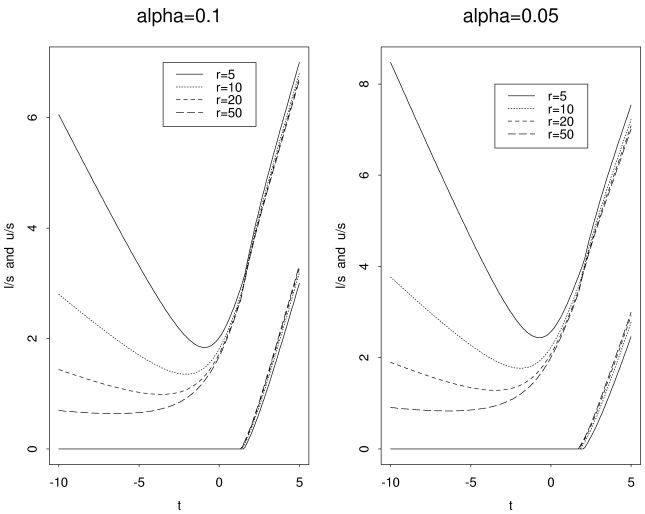

FIG. 1. *Bayesian confidence limits when $s = 1$.*

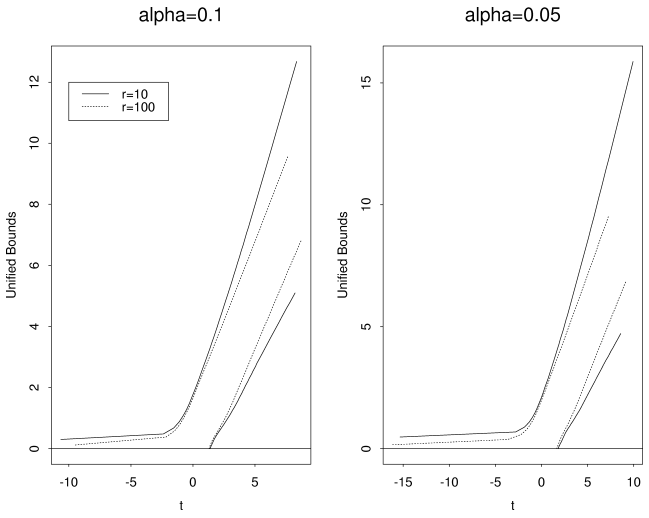

FIG. 2. *Unified confidence limits for* $\delta = \mu/\sigma$.

$s = 1$. Comparing Figure 1 with Mandelkern's Figure 3 shows that the Bayesian approach with estimated σ^2 discounts large values of $-x$ to an even greater extent that the maximum likelihood approach (with known σ^2).

The unified approach. Including an estimated σ^2 in the model causes some problems for the unified approach, which is naturally suited to one parameter families or to constructing confidence regions for the vector of all parameters in a multiparameter model. There are three possible ways to proceed. It is straightforward to construct confidence regions for (μ, σ^2), but then projections on the μ axis will lead to very long intervals. It is natural to reduce by invariance. The unified method can be applied to the distributions of $T = X/S$ to construct confidence intervals for μ/σ. The family of noncentral t-distributions is hard to use in this way, however. A simpler approach is to use the likelihood ratio statistic for composite hypotheses, $H_\delta : \mu/\sigma = \delta$. That is, letting

$$R_\delta(t) = \frac{\sup_{\mu=\delta\sigma} L(\mu, \sigma^2 \mid x, s^2)}{\sup_{\mu,\sigma\geq 0} L(\mu, \sigma^2 \mid x, s^2)},$$

where $L(\mu, \sigma^2 \mid x, s^2)$ is the likelihood function, the unified confidence intervals for δ are $\{\delta : R_\delta(t) \geq c_\delta\}$, where c_δ are determined by $P_\delta[R_\delta(T) \geq c_\delta] = 1 - \alpha$. Here $R_\delta(t)$ depends only on t by scale invariance, and $T = X/S$ has the noncentral t-distribution with r degrees of freedom and noncentrality parameter δ. After some calculation,

$$R_\delta(t) = \left[\frac{r + t_-^2}{r+1}\right]^{(r+1)/2} \psi_\delta(t)^{r+1} \cdot \exp\left(-\frac{1}{2}\delta^2 + \frac{1}{2}\delta t \psi_\delta(t)\right),$$

where t_-^2 is the square of the negative part of t and

$$\psi_\delta(t) = \frac{\sqrt{\delta^2 t^2 + 4(r+1)(t^2+r)} + \delta t}{2(t^2+r)}.$$

If $\delta = 0$, then $\psi_0(t) = \sqrt{(r+1)/(r+t^2)}$, $R_0(t) = 1$ for $-\infty < t \leq 0$, and $R_0(t) = [r/(r+t^2)]^{(r+1)/2}$ is decreasing in $0 \leq t < \infty$. For $\delta > 0$, let $\tau_\delta =$

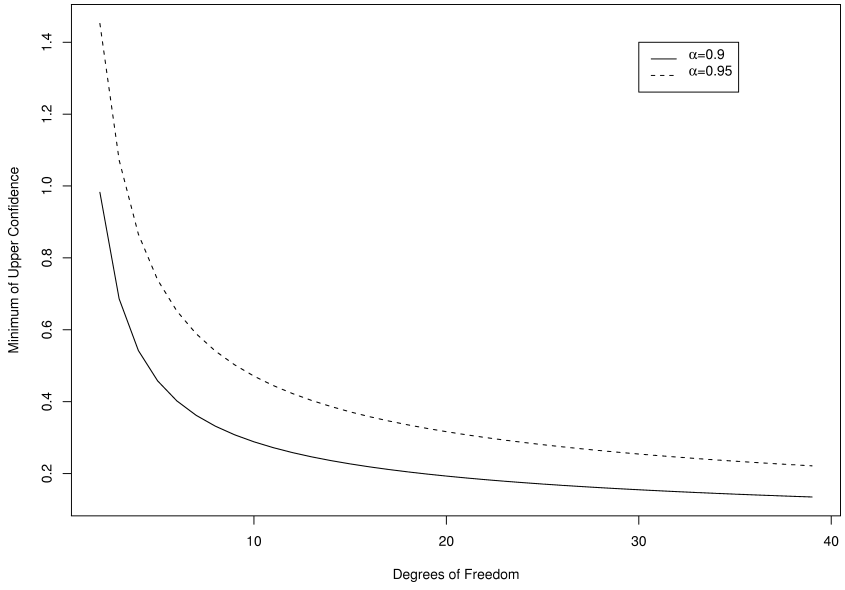

FIG. 3. *The minimum unified upper confidence limit for* δ.

$\delta\sqrt{r/(r+1)}$. Then, differentiation shows that $R_\delta(t)$ is increasing in $-\infty < t \le \tau_\delta$ and decreasing in $\tau_\delta \le t < \infty$. Further, $\lim_{t\to\infty} R_\delta(t) = 0$, and

$$\lim_{t\to-\infty} R_\delta(t) = \left[\frac{\sqrt{\delta^2+4(r+1)}-\delta}{2\sqrt{r+1}}\right]^{r+1} \cdot \exp\left[-\frac{\delta^2+\delta\sqrt{\delta^2+4(r+1)}}{4}\right]$$

by direct calculation. So $\{t : R_\delta(t) \ge c_\delta\} = \{t : a_\delta \le t \le b_\delta\}$, where $-\infty \le a_\delta < b_\delta < \infty$ are determined by $P_\delta[a_\delta \le T \le b_\delta] = 1-\alpha$, and $R_\delta(a_\delta) \ge R_\delta(b_\delta)$ with equality if $a_\delta > -\infty$. It is straightforward to compute a_δ and b_δ numerically and to solve the equation $a_u = t$ and $b_\ell = t$ for $u(t)$ and $\ell(t)$, which then serve as the upper and lower boundaries of a level $1-\alpha$ confidence interval for δ. Graphs of $\ell(t)$ and $u(t)$ are included in Figure 2 for selected α and r.

In this case, the upper boundary is increasing in t and has a positive limit, δ_0 say, as $t \to -\infty$. Letting $H_{r,\delta}$ denote the noncentral t-distribution with r degrees of freedom and noncentrality parameter δ, δ_0 is the solution to the equation $R_\delta(-\infty) = R_\delta[H^{-1}_{r,\delta}(1-\alpha)]$ and is graphed as a function of r in Figure 3 for selected values of α. So shorter intervals are again obtained for large values of $-t$, but they do not shrink to 0 as $t \to -\infty$. In any case a small value of δ can arise from a small μ, a large σ, or both.

Rejoinder

Mark Mandelkern

I think I can speak for physicists in appreciating the interest of the statistical community in the problem of confidence intervals for bounded parameters. The variety of comments by five distinguished mathematical statisticians suggests that our community has not overlooked a satisfactory procedure that has previously been published. The comments have two main themes: (1) that a Bayesian solution may be most suitable, perhaps with frequentist modification to rationalize the coverage properties as suggested by Professor Casella in his comment and previously by Mandelkern and Schultz (2000a, b) and Roe and Woodroofe (2001); (2) that enlargement or respecification of the model, even a posteriori, may be appropriate. A number of distinct suggestions have been made in the latter regard.

While the procedure used to compute a confidence interval is usually discussed in the original experimental work, it is rarely carried forward to subsequent experimental papers, reviews and theoretical analyses. For this reason it is important for intervals to be evaluated in a consistent and uniform way. Consistency is certainly more important than are the absolute values obtained. It would be particularly valuable if statisticians working in this area would propose a procedure for computing confidence intervals, which could then be adopted as a standard by the experimental physics community.

Finally, it may be most appropriate to, at least in ambiguous cases, give up the notion of characterizing experimental uncertainty with a confidence interval and instead, as suggested by Professor Gleser in his comment, to present the likelihood function for this purpose. It is interesting that Enrico Fermi, who introduced the likelihood method to physicists (Orear, 1958, 1982), suggested that likelihood functions for different experiments be multiplied for overall estimation of parameters.

ADDITIONAL REFERENCES

BROWN, L. (1967). The conditional level of Student's t-test. *Ann. Math. Statist.* **38** 1068–1071.

BUEHLER, R. J. (1959). Some validity criteria for statistical inferences. *Ann. Math. Statist.* **30** 845–863.

CASELLA, G. (1987). Conditionally acceptable recentered set estimators. *Ann. Statist.* **15** 1363–1371.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

FRASER, D. A. S. and REID, N. (1995). Ancillaries and third order significance. *Util. Math.* **47** 33–53.

FRASER, D. A. S. and REID, N. (2001). Ancillary information for statistical inference. *Empirical Bayes and Likelihood Inference*. *Lecture Notes in Statist.* **148** 185–207. Springer, New York.

GARWOOD, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika* **28** 437–442.

KIEFER, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72** 789–827.

MCCULLAGH, P. (1995). Comment on "The roles of conditioning in inference," by N. Reid. *Statist. Sci.* **10** 177–l79.

PROTASSOV, R., VAN DYK, D. A., CONNORS, A., KASHYAP, V. and SIEMIGINOWSKA, A. (2002). Statistics, handle with care: Detecting multiple model components with the likelihood ratio test. *Astrophysical J.* **571** 545–559.

REID, N. (1995). The roles of conditioning in inference (with discussion). *Statist. Sci.* **10** 138–199.

ROBINSON, G. K. (1979). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742–755.

SCHEFFÉ, H. (1977). A note on a reformulation of the *S*-method of multiple comparison. *J. Amer. Statist. Assoc.* **72** 143–144.

VAN DYK, D. A. and HANS, C. M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory. In *Spatial Cluster Modelling* (D. Denison and A. Lawson, eds.) 175–198. CRC Press, London.

VAN DYK, D. A., CONNORS, A., KASHYAP, V. and SIEMIGINOWSKA, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophysical J.* **548** 224–243.

ZHANG, T. and WOODROOFE M. (2001). Credible and confidence sets for restricted parameter spaces. *J. Statist. Plann. Inference*. To appear.