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.

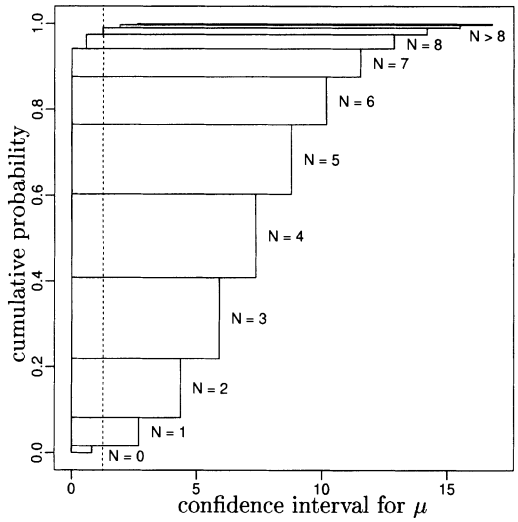

FIG. 1. *The sampling distribution of the standard 95% Poisson confidence interval for μ with $b = 2.88$ and $\mu = 1.25$. The horizontal width of each rectangle is the confidence interval for the corresponding value of N; the height of each rectangle indicates the sampling probability for the interval. 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).

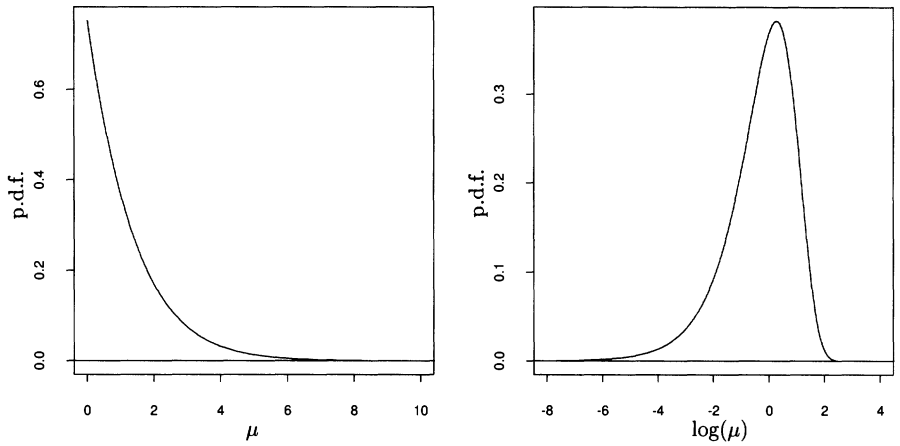

FIG. 2. *The posterior distribution of* μ *(first panel) and* $\log(\mu)$ *(second panel) under a point mass prior for* β $(b = 3)$ *and with* $N = 1$. *The figure illustrates the effect of the symmetrizing log transformation.*

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-

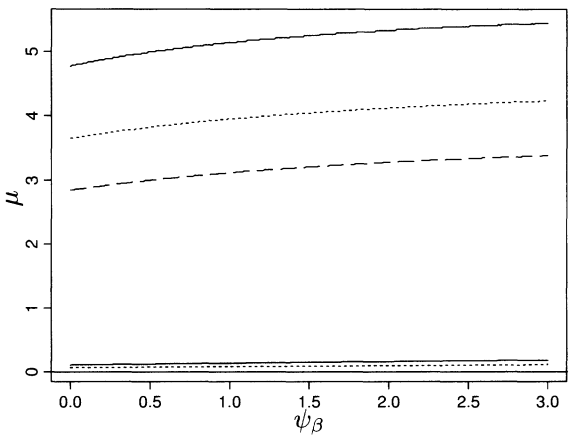

FIG. 3. *The effect of the error in b on the* 90% *posterior interval for* μ. *The figure illustrates how the confidence intervals for* μ *grow wider as the error in b increases, measured here via the prior parameter,* ψ_β. *The solid lines correspond to the upper and lower limits of the highest posterior 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.*

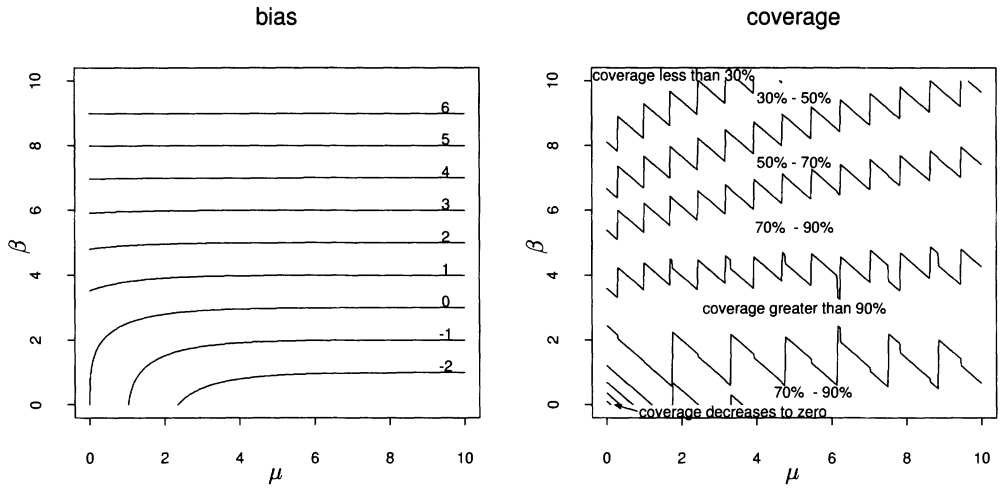

FIG. 4. *Bias of the maximum likelihood estimate (first panel) and under coverage of the nominal 90% interval (second panel) caused by model misspecification. The figures assume the data are generated according to model* (1) *with various values of* μ *and* β *(and* $\alpha = 1$*), but is fit with* β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