International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2013), 81, 1, 40–41 doi:10.1111/insr.12007

Discussions

David R. Cox

Nuffield College, New Road, Oxford, OX1 1NF, UK
E-mail: david.cox@nuffield.ox.ac.uk

It is a pleasure to comment on a very interesting, meticulous, and wide-ranging account of confidence distributions. Just as the contribution was being completed I was saddened to hear of the death of Professor Singh; our field has lost an enthusiastic and original worker.

It can be hard to trace the history, but the paper of Fisher (1930) is, despite its partly misleading title, perhaps the first to show a formal distribution for a parameter with a clear empirical interpretation not based on inverse probability. The interpretation seems to be that which we now associate with confidence intervals; the quagmire of fiducial probability really only came later when the formal distributions were manipulated like probability distributions. An interesting distinction between Fisherian discussions and those in the spirit of Neyman and Pearson is that Fisher usually begins with a reduction by sufficiency where available, whereas in Neyman–Pearson theory sufficiency typically emerges as the aftermath of an imposed optimality criterion. The former approach seems preferable when applicable, partly for reasons of economy and partly because the conceptual arguments for sufficiency seem at least as or more intuitively compelling and much more general than the optimality criteria of Neyman–Pearson theory. Would there have been any general implications for the present paper had greater weight been placed from the start on sufficiency, including asymptotic sufficiency?

Another general issue about Neyman–Pearson theory raised by the discussion concerns the role of confidence coefficients (and α levels) in that theory. Thus the interpretation of an upper $1-\epsilon$ confidence limit is that we calculate it for each set of data, make the statement that the parameter is less than the limit. We do that again and again and will be wrong only a long-run proportion ϵ of times; no other statement is allowed about individual cases. Each statement is either true or false and all we know is long-run correctness. I recall that Neyman in his verbal presentations strongly emphasized this behaviouristic interpretation. But is that a specification of a hypothetical process that illuminates the operational definition of a confidence limit or is it an instruction on how confidence limits are actually to be used? The latter suggests we have in each application to choose a single specific ϵ appropriate for that application. The former interpretation gives much more flexibility for summarization of evidence, and it was the view that this was usually the appropriate notion that led to the suggestion (Cox, 1958) of the term confidence distribution. Indeed in the conference lecture on which that paper was based, the suggestion that when 95% confidence limits for a normal mean are found then, even if the parameter is outside the calculated range, it will not be too far outside was greeted with some scepticism. For tests the corresponding issue is the use of p-values as contrasted with accept–reject at some single especially appropriate α level. Interestingly in applications Neyman took a quite flexible approach.

Some conditions are, however, needed for the confidence distribution to be an appropriate summary, as is mentioned at the end of the present paper. The ratio of normal means, discussed briefly in the paper, is probably the simplest example. Suppose we observe (X, Y) independently normally distributed with unit variance and unknown means (μ_X, μ_Y) and that interest lies in $\theta = \mu_Y/\mu_X$. Consider two simple examples. First suppose that $x = 0.5$ and $y = 100$. It is clear that the sign of μ_X is not well determined but that μ_Y is relatively close to 100. It follows that the only reasonable inference about θ is that it lies *outside* a suitable interval including zero. Of course precise statements are possible. An even more extreme example is when, say, $x = y = 0.5$. Then no real value of θ can reasonably be dismissed. Thus the confidence region at most levels is the whole real line. In one interpretation this is degenerate; we knew this without the data. A better interpretation is that the analysis provides a strong warning that no value of θ should be considered incompatible with these specific data. This may be an important, if very disappointing, conclusion. In other special cases the appropriate inference may be that the parameter lies within one of a number of disjoint intervals. The issue and its resolution has an interesting theory and history with topical relevance to the use of instrumental variables, for example, in the context of Mendelian randomization.

The general point here is simply that the objective is to provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model) and that this is not always achieved by unqualified specification of a distribution.

References

Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.*, **29,** 357–372.
Fisher, R.A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.*, **26,** 528–535.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 41–42 doi:10.1111/insr.12002

Bradley Efron

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305-4065, USA
E-mail: brad@stat.stanford.edu

An important, perhaps the most important, unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information. Analyses based on uninformative priors, of the type advocated by Laplace and Jeffreys, are now ubiquitous in our journals. Though Bayesian in form, these do not enjoy the scientific justification of genuine Bayesian applications. Confidence distributions can be thought of as a way to ground "objective Bayes" practice in frequentist theory.

Xie and Singh carry out the grounding process with energy and insight. Starting with their attractive Definition 2.1, familiar confidence interval theory is restated in a less familiar confidence distribution format. Connections with Bayesian and fiducial ideas are made along the way, but the main development is entirely frequentistic.

As the authors point out, all of this has something to do with the bootstrap. Let $\{\hat{\theta}_i^*,\ i = 1, 2, \ldots, B\}$ represent B bootstrap replications of $\hat{\theta}$, an estimator of parameter θ (possibly in the presence of nuisance parameters). The α-th empirical quantile of the $\hat{\theta}^*$'s is then the upper endpoint of a first-order accurate-level α confidence interval. In this sense, the bootstrap distribution is an approximate confidence distribution.

The BCa density, Efron & Tibshirani (1998), improves the confidence accuracy by reweighting the B $\hat{\theta}^*$ values. Let $\hat{G}$ be this empirical cdf, and z_0 and a be the *bias correction* and *acceleration* constants in my 1987 paper. Rather than equal weights $1/B$, the BCa density puts weight proportional to

$$\frac{\phi(z_{\theta i}/(1 + az_{\theta i}) - z_0)}{(1 + az_{\theta i})^2 \phi(z_{\theta i} + z_0)} \qquad [z_{\theta i} = \Phi^{-1}\hat{G}(\hat{\theta}_i^*) - z_0]$$

on $\hat{\theta}_i^*$. The reweighted bootstrap distribution then becomes a second-order accurate confidence distribution. Efron (2012) discusses this construction in the context of objective Bayes inference.

I was shocked to learn of Professor Singh's sudden death—a friend and colleague, I have followed his papers with interest since his pioneering work on bootstrap second-order accuracy, and will miss his good-hearted presence, both personally and in the literature.

References

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Ann. Appl. Statist.*, **6**, (4) 1971–1997.
Efron, B. & Tibshirani, R. (1998). The problem of regions. *Ann. Statist.*, **26**, 1687–1718.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 42–48 doi:10.1111/insr.12006

Donald A. S. Fraser

Department of Statistics, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada
E-mail: dfraser@utstat.toronto.edu

1 Introduction

A distribution for a parameter, a distribution of a parameter, a distribution describing a parameter, and variants: the notion is anathema to some, controversial to others, and a field of development for others.

The notion needs a statistical model, a concept central to much of statistics but not all. A statistical model records possible distributions, thus possible probabilities, for variables in an application of interest; and it has a parameter that represents the unknown in the application context, and it assumes that some value of the parameter is the "true" value, the value that corresponds to the actual distribution in the context, all to some reasonable approximation.

Then with observed data, we have the central challenge of statistics: Statistical Inference. For this challenge the discipline of statistics has long maintained the hope that the unknown parameter value can be described also by a probability distribution, as a density, a distribution function, or even a quantile function. And this is the direction followed by Xie and Singh.

The earliest attempt was clearly Bayes (1763): his device was to act as if the context also included a random source for the parameter value, this source acquiring the name: prior distribution. He was severely criticized by many for making such an arbitrary addition, for example, Boole (1854) and Venn (1866), but generally supported by others including the pre-eminent Laplace (1812). See Fisher (1956) and Bernardo & Smith (1994) for two rather different surveys of this area of statistics, once called Inverse Probability.

Fisher (1930, 1935) introduced an alternative to the Inverse Probability procedure, in part to avoid the arbitrariness of an introduced prior, and called it fiducial probability: small modifications suggested by Neyman (1937) then led to confidence as we now know it. Both use the formal inversion of a pivot to get probabilities on the parameter space, for Fisher to get a full distribution while for Neyman to get probabilities now called confidence but just for sets inverse to a pivot set. The Neyman adjustment provided repetition validity for the now widely accepted confidence methodology, and the Neyman diagram became the defining logic.

Xie and Singh propose that we revert to the basic inversion of a pivot, thus producing a distribution on the parameter space as with fiducial; this has much appeal and does provide a rich alternative to the Bayes route to posterior distributions on the parameter space. But both approaches widely overlook the associated risks that have been documented in Neyman (1937), Dawid *et al*. (1973), Fraser (2011), and others. Thus the extension of the name confidence to fiducial distributions should not be without recognizing associated risks that apply generally for distributions for a parameter, Bayes or otherwise.

But also quite generally, concepts for statistical inference have rarely been introduced in any resemblance of definitive form. Typically they arise in an optimistic form, are developed in various directions, often with conflicts among promotors; inverse probability and confidence are a clear instance. This general pattern is often concerned with territory, and seems to contrast with what might be viewed as a pure evolutionary development, seeking the best at each step and fine tuning the risks.

2 Pivot Inversion: Fiducial, Confidence, Structural, and Other Variants

When examining statistical methodology there is merit in looking at very simple examples first, and being sure the methods are coherent and sensible there. Accordingly consider a sample from a Normal(θ, σ_0) model together with observed data that gives $\bar{y}^0$; the distribution function from the obvious variable is $\Phi\{(\bar{y} - \theta)/(\sigma_0/n^{1/2})\}$, where Φ is the standard Normal distribution function.

From a very practical viewpoint, we could record just the statistical position of the data with respect to possible values for the parameter θ; we would then obtain the p-value function $p(\theta; \bar{y}^0) = \Phi\{(\bar{y}^0 - \theta)/(\sigma_0/n^{1/2})\}$. In particular if the parameter value θ_0 were of interest and we had $p(\theta_0; \bar{y}^0) = 15.9\%$, we would know that 15.9% of the distribution indexed by θ_0 was to the left of the data point and 84.1% was to the right of the data value. Thus we would have the "statistical position" of the data value in the distribution labelled by the θ_0 value. Is there more information than this? In various ways the p-value function provides the full statistical story for the particular data value relative to the model.

From the confidence viewpoint (Neyman, 1937), we might choose a statistical position α and seek the corresponding θ value by inverting $\Phi\{(\bar{y} - \theta)/(\sigma_0/n^{1/2})\} = \alpha$, thus obtaining $\hat{\theta}_\beta = \bar{y}^0 - z_\alpha \sigma_0/n^{1/2}$, where $\beta = 1 - \alpha$; accordingly $(-\infty, \hat{\theta}_\beta)$ is a β confidence interval for θ. The change from α to β corresponds to $\bar{y}$ and θ having opposite signs in the distribution function and is typically of no interest in the presence of two-sided symmetric confidence intervals. In particular here, if $\alpha = 15.9\%$ then $z_{15.9\%} = -1$ and $(-\infty, \bar{y}^0 + \sigma_0/n^{1/2})$ is the 84.1% confidence interval for θ. We have deliberately illustrated with a one-sided confidence interval to emphasize that an individual bound of a two-sided confidence interval can often be important in its own right and thus needs its particular repetition meaning!

From the fiducial approach (Fisher, 1935), we can invert the distribution function or more generally invert a pivot. Thus for the example we would have that θ is Normal$(\bar{y}^0, \sigma_0/n^{1/2})$ or in quantile form that $\theta = \bar{y}^0 - \sigma_0 z/n^{1/2}$, where z is a generic standard Normal variable. Fiducial intervals would be obtained then by using quantile intervals for z to obtain intervals for θ, or more generally by directly integrating over intervals of interest to get corresponding probability levels. The fiducial approach was criticized by Bayesians, mostly because providing distributions for a parameter was then viewed as Bayesian territory, and partly on technical grounds (Lindley, 1958) that it did not conform to certain Bayesian rules. In particular, certain ways of choosing an interval of interest, either Bayes or frequency, can lead to repetition frequency that is different from the value used in its calculation, in other words it doesn't do what it claims!

The structural approach (Fraser, 1961, 1979) follows confidence and fiducial closely but restricts attention to transformation-parameter models with an identified error distribution; the restriction avoids various complications that can arise with fiducial and confidence distributions. For the example the structural approach would give the parameter distribution $\theta = \bar{y}^0 - \sigma_0 z/n^{1/2}$, where z is a generic standard Normal variable and thus agrees with confidence and fiducial; it has a stronger repetition interpretation coming from the transformation error model, but also like confidence does depend on the set being chosen on the pivot space rather than on the parameter space.

For the Bayesian approach, Fraser (2011) argues that personal and other prior-type information should be kept separate and not used to do the statistical analyses for one's particular convenience. A completed frequency analysis can of course be accompanied by available personal or prior-type information, thus available for an end user to use or combine as deemed appropriate. It is also argued that an appropriate choice of prior can give approximate and sometimes exact confidence intervals, with repetition validity. For the example, the flat prior $\pi(\theta) = 1$ leads to agreement with the confidence, the fiducial, and the structural result, $\theta = \bar{y}^0 - \sigma_0 z/n^{1/2}$.

For the example, the five approaches lead essentially to the same result, with seemingly just the minor difference, whether intervals are required to come from pivot sets or can be chosen freely. For the second more general route where a distribution is produced for the parameter, risks do exist for all five approaches: this was highlighted in Dawid *et al*. (1973) as coming from marginalization and in Fraser (2011) as coming from parameter curvature. Xie and Singh are thus recommending that we ignore the restriction to confidence sets or equivalent, and free confidence to allow the production of parameter distributions. Certainly distributions are easier to think about, are largely in accord with Fisher's original proposal, and are more in the freedom of the Bayes approach, but they do overlook inherent risks as the preceding references indicate.

3 Promoting Confidence Distribution

Is confidence just probability? In many ways it is, and many users do treat it as probability, even when clearly aware that statistics courses and statistical culture say otherwise! Is a Bayes

posterior value just probability? Well it is called such, just as confidence itself was originally called probability. In either approach, the events underlying the probability statements are events in the past: the θ value came from somewhere or came from the prior when it is part of the given, and its value is realized but inaccessible; and the data value is a realized and known constant. So we are talking about a concealed true value, and we are evaluating it based on antecedent randomness. Xie and Singh are clearly interested in upgrading the probability status of confidence and they have devoted many pages to the task. I extensively support this initiative but with cautions as indicated.

But in doing this, Xie and Singh steer clear of the long-standing stigma associated with fiducial, and this leads them to draw a rather difficult dividing line. Is fiducial really different from confidence when presented in distribution-function form? Or are they making issues with the over-enthusiasm in some of Fisher's arguments for his proposal, where he was the first to seriously confront the arbitrariness in the Bayes (1763) approach, the from-nowhere introduction of a mathematical object called a prior in order to have a bespoke probability at the end of an argument? The amazing thing is the huge following that this Bayes approach engendered, including many elite thinkers: "make up a prior and solve the problem!" A staggering affront to the scientific process. Fisher should be given full credit for his innovative contribution against this background, and saying that he did not get the wording of his promotion in comfortable accord with some present views now seems somewhat irrelevant.

Xie and Singh, in promoting the confidence distributions, present three definitions. The first is the classical definition CL and defines the confidence distribution function as the distribution function version of the confidence quantile function $\tilde{\theta}_\beta$, where $(-\infty, \tilde{\theta}_\beta(y^0))$ is a β level confidence interval. But isn't this just what Fisher (1930) did? The second definition (2.1) identifies a confidence distribution function $H(\theta; y)$ as a distribution function in θ for each given y and as a pivot with a uniform distribution for each θ. But this also closely identifies with what Fisher (1930) offered: fiducial but with Fisher's promotion replaced by a claim that the argument is pure frequentist. Of course it is pure frequentist just as Fisher (1930) was pure frequentist, except for Fisher's accompanying claim to having purer probabilities, which was then a direct confrontation to the Bayes aficionados of the time. The third definition (2.2) essentially gives just the quantile equivalent say $\tilde{\theta}_u(y)$ of the distribution function where u is Uniform (0, 1). Aren't confidence quantile and confidence upper bound just different labelling for the same object?

Confidence distributions can have many properties: the distribution function should of course be Uniform(0,1), but also it should inherit continuity when present in the model, should use all available information, and should generally be sensible. These properties aren't really addressed in the authors' proposal. A promotion of confidence distributions should acknowledge these inherent issues and also mention marginal and conditional conflicts as discussed in the literature.

4 Some Details

(i) Signed Likelihood Root. Example 2.5 considers a full exponential model with canonical parameter φ and defines the Signed Likelihood Root r for a scalar θ as $r = \text{sign}(\hat{\theta} - \theta)[2\ell(\hat{\varphi}_\theta)]^{1/2}$. This uses the profile log-likelihood $\ell(\hat{\varphi}_\theta)$ inside a square root and can lead to the square root of a negative number. The usual expression works from the shortfall from the maximum profile and has a change of sign, thus $r = \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\varphi}) - \ell(\hat{\varphi}_\theta)\}]^{1/2}$. The usual expression for q comes in two common versions, one useful for computation and the other for understanding.

(ii) Clean and coherent. In the Introduction the authors refer to their approach as "... clean and coherent". As described in the sections above, the promotional material on fiducial has been deleted but documented risks inherent in fiducial have also been deleted; this is not clean and coherent.

(iii) The bootstrap. The bootstrap as in Section 2.3 provides an approximation to the distributions described by a model, and in doing this the bootstrap can also eliminate the influence of nuisance parameters. It can be applied to statistics or to pivots, with faster effect using suitable pivots. It can be used with least squares, or with maximum likelihood statistics, or with statistical quantities, or anywhere where distributions are wanted. Of course confidence calculations are just one such use but there are many others including of course testing. So there is no particular attachment of the bootstrap to confidence distribution functions other than providing an approximate means of calculation for such.

(iv) Information. The discussion of information in Example 2.4 has disturbing departures from standard usage. The observed Fisher information is usually defined as $-\ell_{\theta\theta}(\hat{\theta})$ in the scalar full parameter case and designated as $i(\hat{\theta})$ or $j(\hat{\theta})$; the subscripts denoting differentiation. The authors use a per-data-value version of information which involves a division by n and they take i or j to be the reciprocal of the usual information in the literature, that is, changing from the variance of the score to the variance of the maximum likelihood value. This provides an uncomfortable connection with the literature.

(v) As an ordinary probability function. The authors claim in Section 3 that a confidence distribution function "can be manipulated as an ordinary probability distribution function". This claim is counter to much material in the literature; see (ii) above; by ignoring properties documented in the literature one gains greater freedom but risks are introduced.

(vi) Modern definition of confidence distribution. As noted earlier the "modern definition of confidence distribution" from Section 3 involves a pivotal quantity with a fixed Uniform (0, 1) distribution which is then inverted. But this is just fiducial probability without the fiducial probability claims. How is this modern? Fiducial leaving out the usual claims, and also leaving out the risks?

(vii) Mathematical coincidence. In Section 3.3, Xie and Singh discuss a property of a location model $f(y-\theta)$, namely that the p-value $p(\hat{\theta}_\alpha)$ is equal to the Bayesian survivor value $s(\hat{\theta}_\alpha)$ using a flat or constant prior $\pi(\theta) = c$, in other words that confidence is equal to Bayes posterior, in this special location structure where Bayes happens to have repetition reliability. They refer to the equality as a "mathematical coincidence"; and also claim that this "hing(es) on the normal assumption". The "coincidence" suggestion is substantially misleading and the claim of hinging on the "normal assumption" is not correct. With a location model the fiducial density is $f(y^0-\theta)$ which is exactly the likelihood $L(\theta; y^0)$. This is not a "mathematical coincidence": it is a direct consequence of the minus sign in "$y-\theta$" which gives the confidence-Bayes equality and does not relate to Normality. A more general claim is that the location property is the background for the approximate repetition validity of the general Bayes approach, see Fraser (2011).

(viii) Optimality. The authors devote an entire Section 5 to optimality for confidence distributions. This considers the stochastic concentration of two confidence distribution functions. But it neglects a large literature on confidence in the conditioning and asymptotics literature, all of which has evolved more or less directly from the confidence literature.

(ix) Vector parameters. In their summary the authors state "It is still an open question how (or whether) one can define a multivariate confidence distribution". This was a crucial

issue in Fisher's fiducial and in part led to confidence (Neyman, 1937); and it has a long history.

(x) *Behrens–Fisher and combining confidence distributions*. With a sample from a Normal (μ, σ^2), the methods discussed in Section 2 lead to the confidence distribution $\mu = \bar{y} - ts_y/n^{1/2}$, where t is a generic Student variable with $n - 1$ degrees of freedom. If one then were interested in the difference $\delta = \mu_2 - \mu_1$ of the means of two Normal populations one could quite reasonably, following Xie and Singh, combine the confidence distributions obtaining $\delta = \bar{y}_2 - \bar{y}_1 - t_2 s_{y_2}/{n_2}^{1/2} + t_1 s_{y_1}/{n_1}^{1/2}$, where t_1 and t_2 are independent Student variables with the appropriate degrees of freedom. This arose with Behrens (1929) and was recommended by Fisher (1935) and was central to the developing stigma that unfortunately became attached to fiducial methodology: the combined distribution did not behave with frequency properties. For some extensive recent simulations, background references, and discussion see Fraser *et al*. (2009). The message: combining confidence distributions is attractive but the resulting distribution may not inherit the initial frequency properties.

5 Discussion

Statistics has a wealth of exploratory procedures and methods: the Bayes where priors are a free choice, the processing-filtering where computing power is dominant, and more. This paper as part of seeking a broader role for confidence uses Fisher's original definition of fiducial, seeks to avoid the stigma of fiducial by renaming it confidence, supplies a new promotion in place of the Fisher arguments, and views the result as just frequentist confidence.

The present comments might be viewed as critical but they are largely concerned with detail, present and absent. From a larger viewpoint, the authors seek to raise the stature of confidence to compete directly with Bayes posterior distributions, a laudable endeavour that avoids the arbitrary mathematical priors in Bayes. Of course Bayes has the property of being approximate confidence (Fraser, 2011), and this arguably provides the sole support for the use of the term probability in the Bayes analysis contest. Xie and Singh are then in effect saying why not switch directly to confidence distributions and enjoy the flexibility of describing the parameter by a distribution. This deserves support, but should not ignore the reality that this would introduce to confidence theory some of the documented risks of the Bayes approach, as a price for the flexibility.

Acknowledgements

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc., London*, **53**, 370–418.

Behrens, W. (1929). Ein beitrag zur fehlerbereichnung bei wenigen beobachtungen. *Landwirtschaftliche Jahresberichte*, **68**, 807–837.

Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. Chichester: Wiley.

Boole, G. (1854). *The Laws of Thought*. New York: Dover.

Dawid, A.P., Stone, M. & Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. R. Stat. Soc. Ser. B*, **35**, 189–233.

Fisher, R. (1930). Inverse probability. *Proc. Cambridge Phil. Soc.*, **26**, 528–535.

Fisher, R. (1935). The fiducial argument in statistical inference. *Ann. Eugenics*, **6**, 391–398.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.
Fraser, D. (1961). The fiducial method and invariance. *Biometrika*, **48**, 261–280.
Fraser, D. (1979). *Inference and Linear Models*. New York: McGraw-Hill.
Fraser, D., Wong, A. & Sun, Y. (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.*, **37**, 161–181.
Fraser, D.A.S. (2011). Is Bayes posterior just quick and dirty confidence? With discussion. *Statist. Sci.*, **26**, 299–316.
Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: V. Courcier.
Lindley, D. (1958). Fiducial distribution and Bayes theorem. *J. R. Stat. Soc. Ser. B*, **20**, 102–107.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc. A*, **237**, 333–380.
Venn, J. (1866). *The Logic off Chance*. London: Macmillan.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 48–52 doi:10.1111/insr.12005

Emanuel Parzen

Department of Statistics, Texas A&M University, College Station, TX, USA
E-mail: eparzen@stat.tamu.edu

This outstanding paper by Professors Xie and Singh is a comprehensive review of history and important applications of confidence distributions. They demonstrate that confidence distributions deserve to be widely taught and practiced as a powerful tool applicable by all researchers concerned with statistical inference and data discovery. My comments will present a confidence quantile interpretation of confidence distributions.

1 Confidence Distributions are Fundamental Methods

We discuss below in Section 6 the connection of confidence distributions to fundamental research by Jerzy Neyman and R. A. Fisher, based on review by Neyman (1941). Their controversy may have been caused by concern over priority for ideas. Today the question should not be about credit for methods, but a framework for tools which are simple and powerful for applications.

In our view the basic definition of a confidence distribution has two parts: it is a conditional distribution of a parameter θ given data $\hat{\theta}$, denoted symbolically as a random variable $\theta \mid \hat{\theta}$; we do not assume that there exists an unconditional distribution for the parameter θ, because the value of the parameter is a non-random number.

In applications our interpretation of a confidence distribution is Bayesian (in the sense of describing our knowledge about a parameter, given data, by a probability distribution). Our computation is frequentist because we do not assume a prior distribution for the parameter. The importance of this objective frequentist approach for the practice of statistical inference is demonstrated by the results of Xie and Singh for combining information about a parameter from several estimators (Section 4 of my discussion presents a confidence quantile version of their recipe).

I note that my discussion is based on my research (called Nonparametric Modeling using Quantiles and Mid-distributions) whose goal is: to unify many cultures of Statistical Science;

Big and Small data science; parametric and non-parametric modelling of univariate, multivariate, high-dimensional variables which can be discrete or continuous.

2 Internal Representations of Probability Distributions

To describe a probability law (and its random variable Y) one usually uses following functions, which we call external representations: Probability mass function $p(y) = p(y; Y)$; Probability density function $f(y) = f(y; Y)$; Distribution function $F(y) = F(y; Y)$; Quantile (inverse distribution) function $Q(u) = F^{-1}(u)$, $0 < u < 1$; Mid-distribution function $F^{\text{mid}}(y) = F(y) - 0.5p(y)$.

We find very useful an additional representation of Y continuous, which we call an internal representation (in the spirit of modelling by stochastic differential equations):

$$Y = g(\theta, W_Y), \tag{1}$$

where θ are unknown parameters, W_Y has known probability distribution, and g is known. Example: Let $Y =$ Gamma (shape $= \nu$, scale $= \theta$), a random variable with distribution Gamma, with unknown θ and known ν. The internal representation of Y is

$$Y = \theta\, W_Y,\ W_Y = \text{Gamma (shape} = \nu, \text{scale} = 1) \text{ has known distribution.} \tag{2}$$

Given random sample $Y_1, \ldots, Y_n$, let V (also denoted $\hat{\theta}$) be estimator of θ. Express sampling distribution of V as internal representation

$$V = h(\theta, W_V). \tag{3}$$

Let V^* denote observed value of V. We regard our knowledge of θ given V^* as a probability distribution, denoted $\theta \mid V^*$, which we call the confidence distribution of θ given sample. We interpret $\theta \mid V^*$ to have properties analogous to a conditional distribution (concept that statisticians apply intuitively, not using the rigourous mathematical definition as a Radon–Nikodym derivative).

3 Confidence Quantiles, Estimating Equations, and Pivots

Our goal of a formula for the confidence distribution $\theta \mid V^*$ is accomplished symbolically by an internal representation $\theta \mid V^* = h^{-1}(V^*, W_V)$, or computationally by its quantile function $Q(P; \theta | V^*)$, called the confidence quantile. A comprehensive outline of the calculus of quantile functions is given in Parzen (2004).

THEOREM 1: *To state an estimating equation for a confidence quantile construct a pivot* $T(\theta, \hat{\theta})$ *with the properties*:

(A) $T(\theta, \hat{\theta}) = Z$, *pivot has same distribution for all* θ;
(B) $T(\theta, \hat{\theta}^*)$ *is increasing in* θ, *for each fixed* $\hat{\theta}^*$.

Then estimating equation for confidence quantile $Q(P; \theta \mid \hat{\theta}^*)$ *is*

$$T[Q(P; \theta \mid \hat{\theta}^*),\ \hat{\theta}^*] = Q(P; Z),\ 0 < P < 1. \tag{4}$$

Often Z is $N(0, 1)$; *then median* $Q(0.5; Z) = 0$. *Median of confidence quantile satisfies*

$$T[Q(.5; \theta \mid \hat{\theta}^*),\ \hat{\theta}^*] = 0. \tag{5}$$

IMPORTANT FACTS: Confidence distribution is a pivot with $Z = U$, Uniform(0,1). General pivot is p-value$(\theta, \hat{\theta}) = \Pr[\hat{\theta} > \hat{\theta}^* \mid \theta]$ when it obeys stochastic order condition that it is an increasing function of θ. Under stochastic order condition, explicit formula for confidence distribution is

$$F(\theta; \theta \mid \hat{\theta}^*) = 1 - F(\hat{\theta}^*; \hat{\theta} \mid \theta). \tag{6}$$

3.1 Graphical Calculation of Confidence Quantile

Solve estimating equation graphically by two plots:

(1) $T = Q(P; Z), 0 < P < 1$;
(2) $T = T(\theta, \hat{\theta}^*)$ as a function of θ.

Given P determine T from (1); then from (2) given T determine θ, which is our desired $Q(P; \theta \mid \hat{\theta}^*)$. Repeat this for selected values of P.

Example 3.2 of Xie and Singh considers statistical inference of parameter θ of model $Y =$ Gamma(ν, θ). Point estimator of θ is $M(Y)$, sample mean. Its sampling distribution has internal representation

$$n\, M(Y) = \theta\, W_M,\ W_M = \text{Gamma}(n\nu, 1). \tag{7}$$

Write the confidence distribution of Xie and Singh (with our ν equal their p_0):

$$P = H_n(\theta_0) = 1 - F(n\, M(Y); \text{Gamma}(n\nu, \theta_0)). \tag{8}$$

Confidence quantile is $\theta_0 = H_n^{-1}(P)$ with explicit formula:

$$\begin{aligned} 1 - P &= F(n\, M(Y); \text{Gamma}(n\nu, \theta_0)) \\ n\, M(Y) &= Q(1 - P; \text{Gamma}(n\nu, \theta_0)) = \theta_0\, Q(1 - P; \text{Gamma}(n\nu, 1)). \end{aligned}$$

Conclude that $Q(P; \theta \mid M(Y)) = \theta_0 = n\, M(Y)/Q(1 - P; \text{Gamma}(n\nu, 1))$, which equals

$$M(Y)\, Q(P; n/\text{Gamma}(n\nu, 1)).$$

Internal representation for confidence quantile:

$$\theta \mid M(Y) = n\, M(Y)/\text{Gamma}(n\nu, 1). \tag{9}$$

4 Combining Estimators Confidence Quantile

In their Section 6, equation (7), Xie and Singh propose a general recipe for combining k independent confidence distributions $H_j(\theta)$ of a parameter θ. Their recipe can be stated in terms of confidence quantiles. Statistician chooses monotonic transformation g_c to be a quantile function $Q_0(u)$. A pivot is defined

$$T(\theta) = \sum_{j=1}^{k} Q_0(H_j(\theta)). \tag{10}$$

Fix P; compute $\theta = Q(P; \theta \mid \text{combinedconfidencedistributions})$ by estimating equation $T(\theta) = Q(P; Z)$, random variable Z defined from k independent Uniform(0,1), U_j by $Z = \sum_j Q_0(U_j)$. The quantile function of Z is a sample quantile computed by simulation. The question of choice of transformation Q_0 should be investigated in each example by comparing several choices for Q_0.

5 Regression Parameters Confidence Distribution, Prediction

A regression model $Y = X\beta + e$ has ordinary least squares parameter estimators $\hat{\beta}$ satisfying normal equations

$$X'X\hat{\beta} = X'Y.$$

Write $X'Y = X'X\beta + X'e$. Obtain internal representation for sampling distribution of $\hat{\beta}$:

$$\hat{\beta} = \beta + X^{\#}e, X^{\#} = (X'X)^{-1}X' \text{ generalized inverse of } X. \tag{11}$$

Confidence distribution of β has internal representation

$$\beta \mid \hat{\beta}^* = \hat{\beta}^* - X^{\#}e. \tag{12}$$

These concepts can be extended to Bayesian parameter estimation and confidence distributions for prediction of future observations obeying the regression model. An algorithm is given by Marriott & Spencer (2001) for Bayesian predictive distributions for a linear regression model. Their results can be interpreted: a conjugate prior distribution is equivalent to augmenting the data; their formulas for posterior distributions of parameters and for prediction of future observations can be quickly derived using update formulas for mean and covariance of combined sample from means and covariances of prior and current sample.

6 Confidence Distributions and Fiducial Inference

How to interpret the mathematics of confidence distributions may have been the basis of the controversy between Neyman and Fisher about fiducial inference. We adapt Neyman (1941), equations (10) and (11), about Fisher's reasoning. For statistical inference of mean μ of model $N(\mu, \sigma)$, σ unknown and estimated by S, sample mean $M(X)$, define

$$T(\mu, M(X)) = \sqrt{n}(M(X) - \mu)/S; \tag{13}$$

Pivot, for all μ, has distribution $T(n-1)$, Student distribution $(n-1)$ degrees freedom. Then $P = \Pr[M(X) > M(X)^* \mid \mu] = \Pr[T(n-1) > T(\mu, M(X)^*]$ implies

$$1 - P = F[T(\mu, M(X)^*); T(n-1)],\ T[\mu, M(X)^*] = Q(1-P; T(n-1)). \tag{14}$$

We argue that for fixed P this is an estimating equation for $\mu = Q(P; \mu | M(Y)^*)$ which we solve to obtain

$$\mu \mid M(X)^* = M(X)^* - Q[1-P; T(n-1)]S/\sqrt{n} = M(X)^* + Q[P; T(n-1)]S/\sqrt{n}. \tag{15}$$

7 Conclusion

Confidence distribution and confidence quantiles provide powerful objective methods of statistical inference that deserve to be widely practiced and taught in statistics courses. Their reasoning starts at the same mathematical place as did fiducial inference, but their interpretation is very different. Publications on confidence quantiles by Parzen are Parzen (2008, 2009). Major lectures by Parzen about confidence quantiles include: 2004 Rice University Erich Lehmann symposium lecture "Data Modeling, Quantile/Quartile Functions, Confidence Intervals, Introductory Statistics Reform"; 2005 JSM Noether Prize lecture "All Statistical Methods, Parameter Confidence Quantiles"; 2006 University of Connecticut ASA Distinguished Statisticians Colloquium lecture "Objective Bayesian/ Frequentist Statistics: My Way with Quantiles"; 2008 Texas A&M Parzen Prize Day lecture "United Applicable Statistics,

Confidence Quantiles, Philosophy of Statistical Science, Statistical Education"; 2009 University of Maryland Kedem Celebration lecture "United Applicable Statistics, Mid Distribution, Mid Quantile, Mid P Confidence Intervals Proportion p"; 2012 Interface Computer Science and Statistics lecture "Modeling, Dependence, Classification, United Statistical Science, Many Cultures".

References

Marriott, J. & Spencer, N. (2001). A note on Bayesian prediction from the regression model with informative priors. *Aust. NZ J. Stat.*, **43,** 473–480.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, **32,** 128–150.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statist. Sci.*, **19,** 652–662.

Parzen, E. (2008). United statistics, confidence quantiles, Bayesian statistics. *J. Statist. Plann. Inference,* **137,** 2777–2785.

Parzen, E. (2009). Quantiles, conditional quantiles, confidence quabtiles for p, logodds(p). *Commun. Stat. Theory Methods*, **38,** 3048–3058.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 52–56 doi:10.1111/insr.12003

Christian P. Robert

Université Paris-Dauphine, IUF, and CREST, Paris, France
E-mail: xian@ceremade.dauphine.fr

"We have shown how confidence distributions, as a broad concept, can subsume and be associated to many well-known notions in statistics across different schools of inference." M. Xie and K. Singh

I must first acknowledge I am rather baffled about the overall reason of this review and that this bafflement will necessarily impact the following discussion. Indeed, and this is not truly a coincidence!, I happen (and so do the authors of the review) to have discussed the related paper by Fraser (2011) a few months ago: while I strongly disagreed on the conclusions of this paper, the central point made by Don Fraser was quite clear, namely to show that Bayesian posterior statements were ungrounded. The current paper is mostly missing this type of clear message and it does not convey a true sense of support for using confidence distributions. I find instead that the paper meanders rather aimlessly around the definition of confidence distributions, which are in short dual representations of frequentist confidence sets, and that it never reaches any definitive conclusion about the appeal of relying on those confidence distributions . . . For instance, I had to wait till Section 4 to be introduced to inference based on confidence distributions and I find the description anticlimactic: using confidence distributions to:

(i) construct confidence intervals is hardly surprising, since this is how those distributions are constructed;
(ii) derive point estimators does not show any advance beyond convergence in probability;
(iii) conduct testing of hypotheses simply provides a recovery of the usual p-value both in the

one- and two-sided cases, and again is hardly surprising given the duality between tests and confidence procedures.

Similarly, optimality is defined as to mirror uniformly most powerful unbiased (UMPU) test optimality (Lehmann, 1986). The most fruitful connection witness applications of confidence distributions appear later in Section 7, in particular with the reinterpretation of bootstrap (Section 7.3), although I am far from convinced that "the concept of confidence distribution is much broader" than the one of bootstrap distribution. Furthermore, the very concept of confidence distributions is restricted (at least in the paper) to unidimensional entities, and seems to possess as many avatars as there are ways of constructing confidence intervals. So, by the end of this long review, I do remain skeptical about the innovation (for frequentist theory) brought by adopting the perspective of confidence distributions.

"Clearly a confidence distribution does not have to be a Bayes posterior distribution, as there are numerous ways to derive it." M. Xie and K. Singh

The fundamental difficulty I have with confidence distributions is the same I have with fiducial distributions, namely one of missing a proper target. Some objective Bayes approaches like matching priors (see, e.g., Robert, 2001) are often criticized for having as sole purpose to mimic a frequentist coverage, hence questioning the relevance of going the Bayesian way. It seems to me that this methodology of confidence distributions suffers from the same if reverse drawback: as with Fisher's fiducial distributions, one tries to produce a posterior distribution without following a Bayesian modelling approach, that is, without selecting a reference prior distribution, hence questioning the relevance of *not* going the Bayesian way! As a result, either the constructed confidence distribution corresponds to a valid Bayesian posterior distribution, in which case it is highly preferable to conduct the choice and assessment of this prior on a preliminary and open basis (rather than defaulting to an implicit black-box prior). Or the confidence distribution *does not* correspond to a genuine prior distribution, in which case it is then incoherent in terms of mere probability theory, thus likely to suffer the same woes as most empirical Bayes approaches (like inefficiency and over-fitting). Things somehow turn for the worse when the authors consider a "true" prior distribution $\pi(\theta)$, which may be a confidence distribution resulting from past experiments, and combine it with the confidence distribution associated with the current data, shying away from the genuine likelihood: if nothing else, multiplying two densities of the same random variable together is an impossibility from a probabilistic perspective.

The object of a confidence distribution does thus remain a full mystery for me, as I do not see how to use it with any confidence either as *a Bayesian procedure* or as a *frequentist procedure*. In the former perspective, it does not necessarily correspond to a prior distribution and to perceive the confidence distribution as a way "to assist the development of objective Bayes approaches" is misguided, in that the corresponding "priors" (if any) would then be data-dependent, hence loose the basic coherence of the Bayesian approach. In the latter perspective, having a probability distribution on a fixed parameter θ does not make sense. Except when reinterpreting it as a bootstrap distribution, that is, with a randomness endowed by the observation into $\hat{\theta}$ rather than from the parameter. I must add that the authors of the review do not indicate that the methodology has met with widespread use, beyond their own circle.

In Section 6.2, the fact that expert opinion is available to build prior distributions would sound to me as the most natural way to engage into licit Bayesian activities since the construction of this prior is then validated by the real world. To replace the exact likelihood with a confidence distribution is a way to shoot oneself in the foot, by throwing away a coherent and valid scheme for

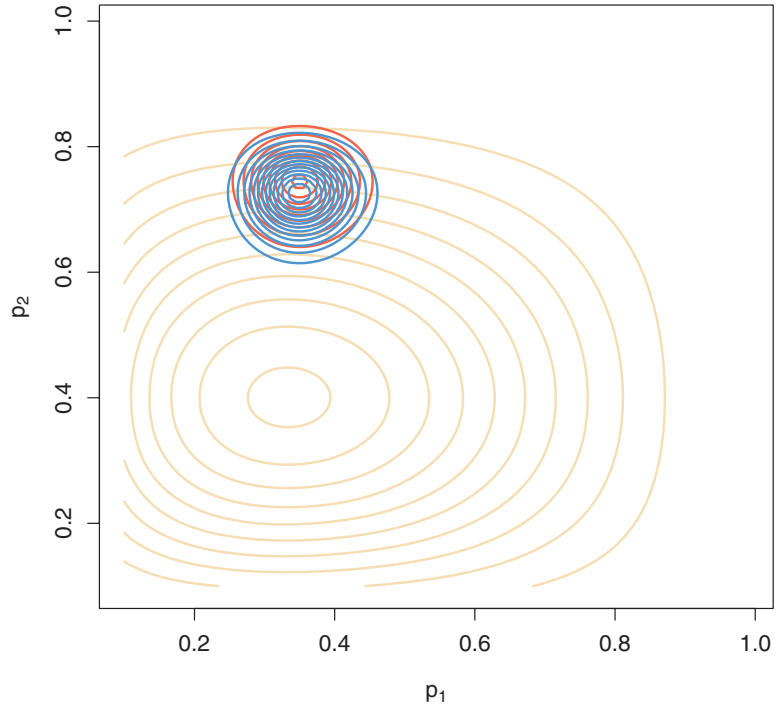

Figure 1. *Contour plots for the prior (light brown), likelihood (red), and posterior (blue) on* (p_1, p_2)*based on 100 binomial observations* $x_i \sim \mathcal{B}(100, p_i)$ *with* $x_1 = 35$ *and* $x_2 = 67$ *and an independent prior,* $p_1 \sim \mathcal{B}e(2, 3)$ *and* $p_2 \sim \mathcal{B}e(3, 4)$.

another one wasting some of the information provided by the data (and contradicting Birnbaum's likelihood principle in addition!).

"The result is a mathematical coincidence, hinged on the normal assumption." M. Xie and K. Singh

The authors seem to consider that having genuine posterior distributions turn into exact confidence distributions cannot have a deeper explanation than being a freak, that is, "a mathematical coincidence". While being a spectator for this kind of exercise, I would think there are deeper reasons for this agreement, first and foremost in connection with the Bayesian interpretation of the best unbiased estimator of Pitman (1938). Furthermore, the work of Welch & Peers (1963) shows that prior distributions can be chosen towards an agreement with the frequentist coverage.

"We can have multiple confidence distributions for the same parameter under any specific setting." M. Xie and K. Singh

As acknowledged by the authors, the notion of confidence distributions suffers from the same taint of *ad-hocquery* as most frequentist (and empirical Bayes, see Robert, 2001) procedures, namely that the confidence distribution can be defined in many ways. Section 5 introduces an ordering of those confidence distributions but it unfortunately is an incomplete ordering, as most frequentist orderings are, and it is thus unlikely that two arbitrary solutions can be ordered according to this principle. The strong connection with UMPU tests—whose own optimality proceeds from an unnatural restriction on testing procedures—reflects this difficulty.

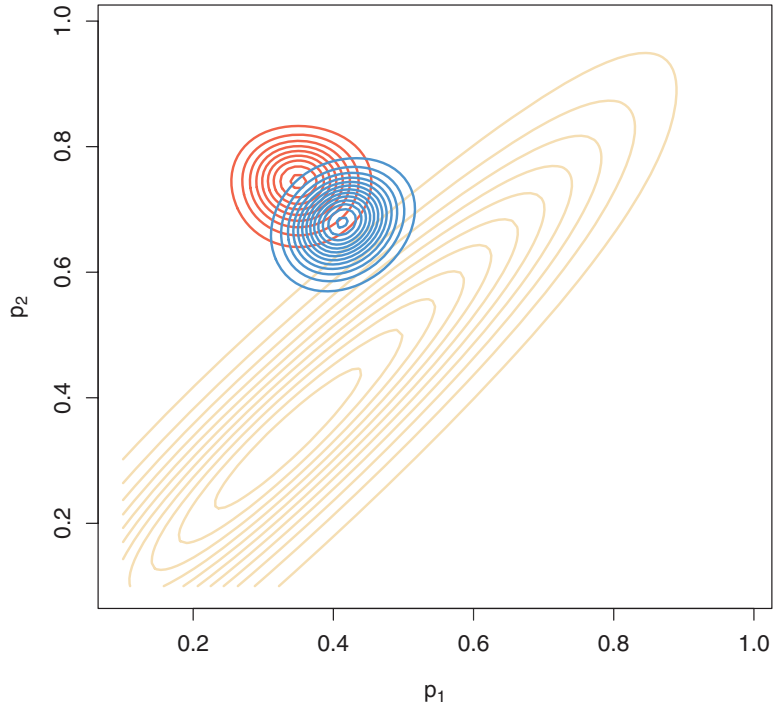

Figure 2. *Same legend as Figure 1 when using a dependent prior with* $p_1 \sim \mathcal{B}e(2, 3)$ *and* $p_2|p_1 \sim \mathcal{N}(p_1, 0.1)$ *restricted to* $(0, 1)$.

> *"A counterintuitive 'discrepant posterior phenomenon' that is inherent in a Bayesian approach can be avoided in a confidence distribution-based approach."* M. Xie and K. Singh

The counter-example discussed in Section 6.2 is only relevant in uncovering the approximation due to the confidence distribution-based approach, rather than pointing out an inherent flaw in the Bayesian approach. Indeed, the fact that the posterior distribution is concentrated away from both the prior and the posterior concentrations seems to be (as far as I can infer given the sparse description contained in the paper) due to the use of a *profile likelihood*, which is an imprecise notion throwing away some of the information contained in the data. When checking on a regular Bayesian analysis of this beta-binomial model, I could not spot any discrepancy, using either independent (Figure 1) or dependent (Figure 2) priors. In any case, the more global issue of having partial prior information like marginal priors on proportions p_0 and p_1 does not seem to be such "a challenging question for Bayesian analysis". Indeed, given those two marginals, it is always possible to select one parameterized family of copula distributions and to estimate the parameters of this copula as part of a global Bayesian analysis (Silva & Lopes, 2008).

> *"The review is not intended to re-open the philosophical debate that has lasted more than two hundred years. On the contrary, it is hoped that the article will help bridge the gap between these different statistical procedures."* M. Xie and K. Singh

In conclusion, I fear the authors have not made a proper case in favour of confidence distributions. The notion carries neither consistency nor optimality features of its own, while it fundamentally relies on the choice of another frequentist confidence or p-value procedure.

Worse, the very construction of the confidence distribution as an inversion of the confidence interval, that is, $H_n(\cdot) = \tau_n^{-1}(\cdot)$, reproduces the common and mislead semantic drift from "$(-\infty, \tau_n(\alpha)$ contains the true value θ_0 with probability α" to "θ_0 belongs to the fixed interval $(-\infty, \tau_n(\alpha)$ with probability α". Further, as reflected by the discussion at the end of Section 6, the review reflects deep misunderstandings about Bayesian inference. Indeed, speaking of a "truthful joint prior" or of a "prior that is in agreement with the likelihood evidence" shows that the prior is considered as a mythical (if true) unique entity, rather than as the choice of a reference measure, which is how I do understand priors.

In Memoriam

Between the time I met for the first time with Prof. Singh in Rutgers in early April 2012 and the time I wrote this review, Prof. Singh most sadly passed away. Although I did not know him well, I think he would have appreciated the intellectual challenge raised in this intellectual dispute and responded accordingly. I am quite sorry this opportunity will never occur.

References

Fraser, D. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, **26,** 299–316 (with discussion).
Lehmann, E. (1986). *Testing Statistical Hypotheses*. New York: John Wiley.
Pitman, E. (1938). The estimation of location- and scale-parameters of a continuous population of any given form. *Biometrika*, **30** 390–421.
Robert, C. (2001). *The Bayesian Choice*, 2nd ed. New York: Springer-Verlag.
Silva, R. & Lopes, H. (2008). Copula, marginal distributions and model selection: a Bayesian note. *Statist. Comput.*, **18,** 313–320.
Welch, B. & Peers, H. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B*, **25,** 318–329.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 56–68 doi:10.1111/insr.12004

Tore Schweder[1] and Nils Lid Hjort[2]

[1]*Department of Economics, University of Oslo, Norway*
E-mail: tore.schweder@econ.uio.no
[2]*Department of Mathematics, University of Oslo, Norway*
E-mail: nils@math.uio.no

Min-ge Xie and Kesar Singh are to be congratulated on an excellent job in explaining what confidence distributions (CDs) are and how and why they might be highly useful in statistical work. The authors have also pulled together published work on CDs and related topics in their comprehensive and useful review. We share their optimism regarding the general and so far too modestly explored usefulness of CDs, along with related concepts such as confidence likelihoods, as broadly applicable tools for modern statistics, conceptually and

operationally. These uses include proper frequentist parallels to Bayesian posterior distributions and a sounder methodology for combining different information sources. Below we offer some remarks pertaining to some of the many themes touched on in the article, along with some pointers to further extensions. Further methodological advances and application stories are in our forthcoming book *Confidence, Likelihood, Probability*. Finally we use the opportunity to humbly join ranks with those expressing grief over Kesar's untimely death.

1 Distribution Estimators

"Distribution estimator" is a good term. The goal of a statistical analysis of a set of data is often to interpret the data by selecting an appropriate statistical model within a chosen family of models, and estimate the parameters of primary interest via the selected model, accounting as honestly and as fully as possible for the uncertainties in the estimates, preferably also the uncertainty due to model selection. A distribution *for* such a focus parameter, a confidence distribution (CD), aims at expressing what has been learned regarding the parameter, including what the pattern and amount of uncertainty is, conditional on the model or the family of models used. Such a distribution estimate provides a full inferential result. What more could be asked for?

The most commonly used distribution estimators are Bayesian posterior distributions. Fisher introduced his fiducial distribution as a distribution estimator to overcome problems with Bayesian analysis which then was, and actually often still is based on flat priors to represent non-informativity. Neyman found that fiducial distributions of one-dimensional parameters provide confidence intervals, hence the term "confidence distribution". Xie and Singh say that in the long history of CDs they have been misconstrued as a fiducial concept. Is it really helpful to distance CDs so sharply from fiducial distributions?

A CD with cumulative distribution function $H(x, \theta)$ for a one-dimensional parameter θ, based on the observed data x, has the property $H(X, \theta) \sim U[0, 1]$. This is interpreted in the Neymanian way, that the distribution estimator provides confidence intervals by its quantiles, and also p-values for one-sided hypotheses. This is certainly a purely frequentist interpretation, distinct from Fisher's fiducial interpretation, but as mathematical objects CDs and fiducial distributions are equivalent. The $H(X, \theta)$ is indeed a pivot, and Fisher (1930) found his fiducial distributions by the same pivot. As Fisher, Xie and Singh view CDs for observed data as probability distributions subject to ordinary probability calculus. We will discuss this below. Taken as a pure mathematical object, the CD is subject to ordinary probability calculus. It turns out however that distributions derived from the CD are not in general CDs, not even in the one-dimensional case. Fisher constructed multivariate fiducial distributions by combining one-dimensional fiducial distribution via conditioning. According to Xie and Singh it is an open question how (or whether) multivariate CDs could be defined. Before looking at some illustrative examples, we briefly recap the fiducial debate to recall where Fisher went wrong.

We think that Fisher (1930) saw his one-dimensional fiducial probability as epistemic: "There are two different measures of rational belief appropriate to different cases. Knowing the population we can express our incomplete knowledge of, or expectation of, the sample in terms of probability; knowing the sample we can express our incomplete knowledge of the population in terms of likelihood . . . There are, however, certain cases in which statements in terms of probability can be made with respect to the parameters of the population". These cases are when a pivot exists, and a fiducial distribution is obtained, representing the rational belief.

In papers from 1935, Fisher combined one-dimensional fiducial distributions to multivariate ones, and claimed them to be unique when based on minimal sufficient statistics. From a normal

sample, the empirical variance s^2 and the mean $\bar{x}$ are the statistics. From the chi-squared pivot $(n-1)S^2/\sigma^2$ a fiducial distribution with density $f(\sigma;\widetilde{\sigma})$ is obtained, with $\widetilde{\sigma}=\sigma_{\rm CD}$ the random variable carrying the fiducial distribution for σ (and similar notation for other parameters below). Given $\sigma=\widetilde{\sigma}$, $(\mu-\bar{X})/(\widetilde{\sigma}/\sqrt{n})$ is a normal pivot, yielding the normal fiducial density $f(\mu\,|\,\sigma;\widetilde{\mu}\,|\,\widetilde{\sigma})$. This makes

$$f(\mu,\sigma;\widetilde{\mu},\widetilde{\sigma})=f(\mu\,|\,\sigma;\widetilde{\mu}\,|\,\widetilde{\sigma})f(\sigma;\widetilde{\sigma})$$

the bivariate fiducial probability density for the two unknown parameters. This step-by-step method was found not to yield unique multivariate fiducial distributions. Dempster (1963) found for example a different fiducial distribution by in the first step to find a distribution for $\theta=\mu/\sigma$ and in the second step a conditional fiducial distribution for σ given $\widetilde{\theta}$. Since multivariate fiducial distributions were supposed to be subject to ordinary probability calculus the two distributions should have been equivalent if unique.

Are fiducial distributions ordinary probability distributions, as Fisher claimed, in the sense that fiducial distributions can be transformed by ordinary probability calculus to new fiducial distributions? Pitman (1939) found this to be not generally true. He characterized the functions of the parameters of a location-scale model for which the distribution derived from the joint fiducial distribution indeed are fiducial distributions, but for other functions the machinery fails. Furthermore Stein (1959) showed in the length problem to be considered later that the distribution of $\|\widetilde{\mu}\|$ is badly upwards biased when obtained from the joint normal fiducial distribution for a normal mean vector μ.

Due to these and certain other problems, including Fisher's claim of uniqueness which was found to be untrue, the fiducial method has been de-merited and broadly forgotten during the last 40 years. Xie and Singh say however that fiducial inference provides a systematic way to obtain a CD. We agree, but due to its delicate nature, any distribution obtained by the fiducial argument must be checked, and perhaps modified, before declared to be a CD. Hannig (2009) checks by simulation the distributions he obtains by his generalized fiducial argument, and finds them to be good approximate CDs.

One-dimensional CDs are invariant to monotonic transformations m, say $m(\theta_{\rm CD})\sim\{m(\theta)\}_{\rm CD}$ in suggestive notation. This invariance was emphasized by Fisher (1930) and is indeed important for CDs. As claimed, distributions derived from a CD by ordinary calculus do however not automatically inherit the property of being CDs. As the following example shows, this is the case even in dimension 1. Consider $g(\mu)=\lambda=|\mu|$, where $\widetilde{\mu}\sim{\rm N}(\bar{x},\sigma^2/n)$ is carrying the CD for μ, obtained from a normal sample of known variance. Clearly,

$$G(x,\lambda)=P(-\lambda\le\widetilde{\mu}\le\lambda)=\Phi\left(\frac{\lambda-\bar{x}}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{-\lambda-\bar{x}}{\sigma/\sqrt{n}}\right)$$

is the derived cumulative distribution for λ given x. For any true value $\lambda_0=|\mu_0|$, the $G(X,\lambda_0)$ does not have a $U[0,1]$ distribution and is hence not a CD. When $\lambda_0/(\sigma/\sqrt{n})=1$, for example, its support is $[0,0.683]$ (and only for larger values of this ratio, where it is easier for data to tell us clearly whether μ is positive or negative, does the distribution of $G(X,\lambda_0)$ come close to the uniform). In this example a bona fide CD is available from a pivot based on $|\bar{X}|$, and

$$H(x,\lambda)=\Phi\left(\frac{\lambda-|\bar{x}|}{\sigma/\sqrt{n}}\right)+\Phi\left(\frac{-\lambda-|\bar{x}|}{\sigma/\sqrt{n}}\right)$$

is a distribution function with point mass at $\lambda=0$. Since $H(X,\lambda_0)\sim U[0,1]$, H is indeed a CD.

2 The Length Problem

In their Section 3.3, the authors argue that a posterior distribution can be treated as an approximate CD. This is essentially a consequence of so-called Bernshtein–von Mises theorems, that the posterior distribution of $\sqrt{n}(\theta - \widehat{\theta})$ tends to the same multinormal limit distribution as does $\sqrt{n}(\widehat{\theta} - \theta_0)$, where θ_0 is the reference value governing the generation of data and $\widehat{\theta}$ the maximum likelihood estimator. By the delta method such a result carries over to focus parameters, say $\gamma = g(\theta)$, so the consequent posterior distribution for γ is close to the appropriate $\Phi(\sqrt{n}(\gamma - \widehat{\gamma})/\widehat{\kappa})$, which hence is a valid asymptotic CD. This is actually a sketch of a proof for rather more general results than those outlined in the article's Section 3.3 and Example 2.4, as the reach of the Bernshtein–von Mises theorems is considerable wider than for i.i.d. setups.

Such approximations will typically only work well if the sample size is large compared to the dimension of the parameter vector, however, and it is easy to construct examples where the posterior distribution is far away from proper CDness. For such an illustration, consider the length problem looked at above in dimension 1, but now in dimension p. Let $X \sim \mathrm{N}_p(\mu, I)$ with $\theta = \|\mu\|^2$ being the parameter of interest. The distribution $G(X, \theta)$ obtained from the joint CD for μ gets further away from being a CD the larger the dimension p. This was noted by Stein (1959). Taking the case of $\theta_0 = p$ as an illustration (e.g., with each component $\mu_i = 1$), for $p = 10$, $G(X, \theta_0)$ has support $[0, 0.560]$ and a distribution piling up close to zero, with median equal to 0.0075. For $p = 100$, the support for $G(X, \theta_0)$ is $[0, 0.518]$ and its median is exceedingly close to zero. To be a proper CD, $G(X, \theta_0)$ should have been uniformly distributed on $[0, 1]$! A clean CD is however available. With $\Gamma_p(\cdot, \theta)$ being the distribution function of the non-central chi-square distribution with df $= p$ and parameter of non-centrality θ, $H(x; \theta) = 1 - \Gamma_p(\|x\|^2, \theta)$ is indeed a CD. It has point confidence at $\theta = 0$ but is otherwise continuous. Despite the point mass at zero in each realized $H(x; \theta)$, we have $H(X; \theta) \sim U[0, 1]$.

The joint CD for μ above, which has these unfortunate side effects when one attempts to use it for inference for various focus parameters $\theta = g(\mu)$, is also identical to the Bayesian posterior distribution under the canonical non-informative prior. Thus Bayesians are in dire straits here, for example, with a severe bias in the length problem. There are ways around this, via particular prior constructions that somehow are allowed to let one's prior views on μ be influenced by what one wishes to focus on afterwards. This is conceptually troublesome but leads after considerable efforts to posterior distributions that better match what here must be seen as the frequentist's golden standard; see for example, Berger & Bernardo (1992) for general methodology concerning such focus-driven reference priors and Tibshirani (1989), Datta & Ghosh (1995) for work directly connected to the length problem.

3 CDs Via Profiled Deviances

From these simple examples we learn that CDs are delicate objects. They cannot in general be treated as ordinary probability distributions in the sense that distributions for derived parameters obtained by ordinary probability calculus might not be CDs. Xie and Singh do not provide a definition of CDs in dimensions larger than 1. From what was learned through the fiducial debate, joint CDs should not be sought, we think, since they might easily lead the statistician astray. We will suggest below that ambitions should be lowered to what Xie and Singh call circular CDs.

In the examples, CDs for the one-dimensional derived parameters could be obtained by direct reasoning. This was done by identifying pivots for the parameters in question. Often, pivots are not available. There is thus a need for a generic method. We will argue below that the

profile deviance function might often give rise to CDs (or approximate ones, which then may be modified further).

In smooth models, the asymptotically normal CD based on the maximum likelihood estimator and its Hessian-based standard error is approximated by the integral of the normed profile likelihood (Example 2.4). For data sets where the sample size is moderate to large compared to the parameter length, confidence regions are also routinely obtained from the profile deviance function $D(x,\theta) = D(\theta) = 2\{\ell_{\max} - \ell(\theta)\}$ by the central χ^2 distribution with df $=$ dim$(\theta) = p$. With Γ_p being the cumulative χ^2_p distribution function, the construction is to take the level set $\Gamma_p(D(\theta)) \leq \alpha$ as the confidence region of level α. If $D(X,\theta_0)$ has cumulative distribution F_{θ_0}, $\mathrm{cc}(\theta) = F_\theta(D(\theta))$ could be called the confidence curve for θ. Confidence curves in one dimension are discussed in Section 7.2, and in higher dimensions they are equivalent to CDs in the circular sense. Then by this definition, the confidence curve has its minimum at the maximum likelihood estimate, $\mathrm{cc}(\widehat{\theta}) = 0$.

A CD for a one-dimensional focus parameter $\gamma = g(\theta)$ generates a family of equitailed confidence intervals, viz. $[H^{-1}(\varepsilon), H^{-1}(1-\varepsilon)]$ for $\varepsilon \in [0, \frac{1}{2}]$, with the same probability of missing the true value on either side. These endpoints are naturally captured by the confidence curve $\mathrm{cc}(\gamma) = |1 - 2H(\gamma)|$, as solving $\mathrm{cc}(\gamma) = \alpha$ provides the level α equitailed confidence interval. Often such confidence curves are reached directly, for example, via the deviance profile recipes, after which we if required may translate to the CD scale, with $H(\gamma)$ equal to $\frac{1}{2}\{1 + \mathrm{cc}(\gamma)\}$ for $\gamma \geq \gamma^*$ and $\frac{1}{2}\{1 - \mathrm{cc}(\gamma)\}$ for $\gamma \leq \gamma^*$, with $\gamma^* = H^{-1}(\frac{1}{2})$ the associated median confidence estimator.

The maximum likelihood estimator is asymptotically unbiased (both in the mean and in the median sense). For the present paragraph assume for simplicity of presentation that θ is one-dimensional, though generalizations to the general case of focus parameters $\gamma = g(\theta)$ in bigger models may be worked out. For small to moderate sample size, let $b(\theta) = \mathrm{med}(\widehat{\theta})$ be its median function. One may show that for $D(b(\theta_0)) \sim F(\cdot;\theta_0)$ the bias corrected confidence curve $\mathrm{cc}_{\mathrm{bc}}(\theta) = F(D(b(\theta));\theta)$ is approximating equitailedness at order $n^{-3/2}$ in a large class of one-parametric models.

Coming back to the general case of a one-dimensional focus γ as a function of a model parameter vector θ, we suggest for reasons sketched above that a CD should be obtained from the probability transformed profile deviance

$$D(\gamma) = 2\{\ell_{\mathrm{prof,max}} - \ell_{\mathrm{prof}}(\gamma)\},$$

where $\ell_{\mathrm{prof}}(\gamma) = \max\{\ell(\theta) : g(\theta) = \gamma\}$ is the profiled log-likelihood function in question. Intervals generated from $D(\gamma) \leq \Gamma_1^{-1}(\alpha)$ are generally known to give more accurate results than by using the symmetric first-order normal approximations for the maximum likelihood estimator. The likelihood could also be reduced by conditioning, if possible, or by integration. The median bias function for the maximum estimator $\widehat{\gamma}$ should be estimated along with the distribution of $D(b(\gamma))$, usually by simulation. If this distribution depends markedly on the nuisance parameters relative to γ, a good CD is not obtained. This work should be carried out for each single parameter of interpretive importance.

Setting appropriate confidence intervals and hence a full CD is tricky when issues of model selection are taken as part of the procedure. This remains true also when a clear limit distribution can be described, as for intervals following the Akaike Information Criterion (AIC) or the Focused Information Criterion (FIC) methods; this may involve non-linear mixtures of correlated and non-central χ^2_1 variables, see Claeskens & Hjort (2008). Procedures via bootstrapping and acceleration and bias corrections may be suggested, as in Schweder & Hjort (2002), but further exploration is required.

4 Ratio of Normal Means

In Example 3.1, Xie and Singh give a bivariate normal distribution for two normal means (μ_1, μ_2). This appears to be a bivariate CD carried by the CD-random vector $(\widetilde{\mu}_1, \widetilde{\mu}_2) \sim \mathrm{N}_2((\bar{x}_1, \bar{x}_2), S)$, but should perhaps be seen as a fiducial distribution. Here S is a known diagonal covariance matrix. A distribution for $\delta = \mu_1/\mu_2$ is by ordinary marginalization given by that of $\widetilde{\delta} = \widetilde{\mu}_1/\widetilde{\mu}_2$. The authors give $H_r(\delta) = \Phi((\delta - \widehat{\delta})/s_r)$ as an asymptotic CD, where s_r is the standard error of $\widehat{\delta} = \bar{x}_1/\bar{x}_2$ obtained by the delta method, disregarding the structure $\mu_1 = \delta\mu_2$. They note that this distribution is not quite a CD, and the challenge is to find a clearer CD for small to moderate sample sizes.

This is the so-called Fieller–Cressy problem, with an associated separate literature presenting solutions from various perspectives; see, for example, Raftery & Schweder (1993). The CD approach we propose above via profiled deviance leads to a clear solution. For simplicity of presentation let us simply take $X_1 \sim \mathrm{N}(\mu_1, \sigma^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma^2)$ with known σ. The profile log-likelihood (apart from constants) is

$$\ell_{\mathrm{prof}}(\delta) = -\tfrac{1}{2}(1/\sigma^2) \min_{(\mu_1,\mu_2):\mu_1/\mu_2=\delta} \{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2\}.$$

Some algebra, minimizing the quadratic under the constraint $\mu_1 = \delta\mu_2$, leads to the deviance statistic

$$D(\delta) = \frac{1}{\sigma^2} \frac{(x_1 - \delta x_2)^2}{1 + \delta^2}.$$

Under the true δ_0, this $D(\delta_0)$ has the exact χ^2_1 distribution. The consequent exactly correct confidence regions take the form

$$\{\delta : D(\delta) \le \Gamma_1^{-1}(\alpha)\} = \{\delta : \Gamma_1(D(\delta)) \le \alpha\}.$$

These are illustrated in Figure 1. They are finite intervals for small levels α; half-infinite intervals for higher levels; unions of two half-infinite intervals for yet higher α; and finally equal to the full real line for the highest levels where the data do not support any sharper conclusions. With more data the σ becomes smaller and the solution produces genuine intervals for a higher range of α.

The solution here, now arrived at via the general profile deviance recipe, is equivalent to the Fieller–Cressy method; see Raftery & Schweder (1993). It can easily be generalized to the case of unknown σ and be modified for other models where the focus is on a ratio of two parameters. We also stress that the fact that the CD here is improper should not be seen as a drawback and hence suggest extending the authors' Definition 2.1 to such cases. When the amount of information is weak about a parameter, as here when x_2/σ is close to zero, it is appropriate to lose some amount of confidence to infinity.

5 Performance and Optimality

The article touches on issues of performance in Section 5. In addition to the optimality theorems of Schweder & Hjort (2002), briefly discussed there, one may also delve into investigations related to reductions by invariance, by sufficiency (Rao–Blackwell results for CDs), construction of nearly optimal CDs via approximations to exponential models, etc. Here we shall merely make a few points; further theory and applications are in Schweder & Hjort (2013).

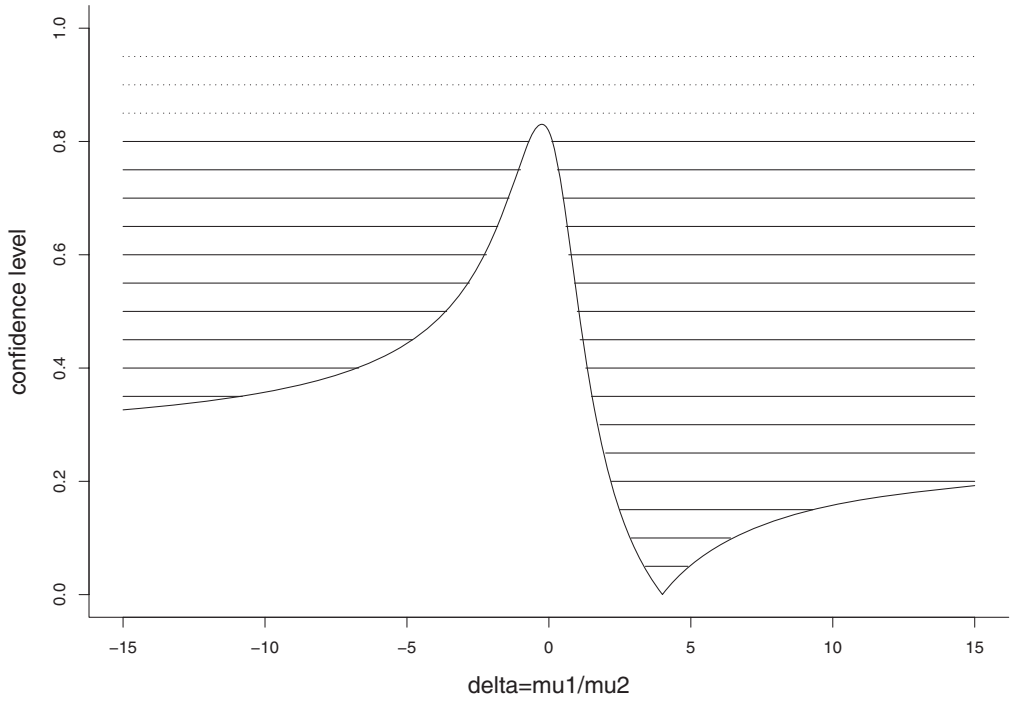

Figure 1. *The probability transformed profile deviance curve* $\Gamma_1(D(\delta))$ *for the focus parameter* $\delta = \mu_1/\mu_2$*, based on* $X_1 \sim \mathrm{N}(\mu_1, 1)$ *and* $X_2 \sim \mathrm{N}(\mu_2, 1)$ *observed to be 1.333 and 0.333; the maximum likelihood estimate* $\widehat{\delta} = 4.003$ *is where the curve hits zero. The horizontal lines indicate confidence regions (intervals or union of intervals) corresponding to vigintic confidence levels from 0.05 to 0.80. The data do not give support to any regions smaller than the full real line when the levels are higher than 83.1%.*

First, it ought to be realized that the list of exponential models for which CDs with uniformly best performance can be constructed is a large one, not merely comprising the traditional textbook examples but extending to generalized linear models, Ising and Potts models for images, Strauss-type models for point patterns, etc. This is of some importance in that inference for some of these complex models is typically hard in the first place, and where the literature sometimes advocates the use of, for example, asymptotic normality, perhaps even via mathematically complicated theorems (with long proofs). Our point then is that inference can be carried out via CDs in an optimal fashion and without necessarily caring about approximate normality or indeed about approximation results at all. Such operations will typically involve certain computational challenges, though, for example, for setting up required simulation schemes for handling complex conditional distributions.

Secondly we wish to add a bit of further insight to the article's brief mention of mean squared error calculus for CDs (more general loss functions may also be worked with). The quadratic loss function version, for a CD of type $H(x;\gamma)$ for some focus parameter $\gamma = g(\theta)$, may then be expressed as

$$\int (\gamma - \gamma_0)^2 \,\mathrm{d}H(x;\gamma) = \mathrm{Var}\,\gamma_{\mathrm{CD}} + (\bar{\gamma} - \gamma_0)^2,$$

computed under a reference parameter value θ_0 with consequent $\gamma_0 = g(\theta_0)$. Here γ_{CD} is a random variable drawn from the (random) CD and $\bar{\gamma} = \mathrm{E}\,\gamma_{\mathrm{CD}}$ its (random) mean. The risk, that

is, the expected loss under reference value θ_0, is therefore

$$R(\theta_0, H) = \mathrm{E}_{\theta_0} \operatorname{Var} \gamma_{\mathrm{CD}} + \mathrm{E}_{\theta_0}(\bar{\gamma} - \gamma_0)^2.$$

To check these formulae in an informative setting let us consider the case where the CD exactly or approximately takes the familiar form $H(x;\gamma) = \Phi(\sqrt{n}(\gamma - \widehat{\gamma})/\widehat{\kappa})$, perhaps from first-order large-sample approximations, and typically involving an estimator $\widehat{\gamma}$ with estimated standard deviation $\widehat{\kappa}/\sqrt{n}$. This CD over γ values has mean $\widehat{\gamma}$ and variance $\widehat{\kappa}^2/n$. Hence the confidence loss is $\widehat{\kappa}^2/n + (\widehat{\gamma} - \gamma_0)^2$ and the confidence risk becomes

$$R(\theta_0, C) = \mathrm{E}_{\theta_0}\widehat{\kappa}^2/n + \mathrm{E}_{\theta_0}(\widehat{\gamma} - \gamma_0)^2.$$

This quantity again is for large samples close to $2\,\kappa^2/n$, where $\kappa = \kappa(\theta_0)$ is the standard deviation parameter of the implied limit distribution of $\sqrt{n}(\widehat{\gamma} - \gamma_0)$. Not surprisingly the CDs with best performance are those with smallest limiting standard deviation, which we see enter the risk equally in two places. That these two contributions to the risk are essentially equal, from the dispersion of the CD and the variability of the estimator, is no coincident; the dispersion of the CD is actually by construction reflecting this variability.

6 CD Meta-Analysis of Related Two-by-Two Tables

We agree with the importance the authors place on the role and potential further usefulness of CD methodology for meta-analysis and other forms of combining information sources (Section 6). Branches of modern statistics will need to finesse such tools further, both practically and conceptually. How can data-hunting schemes be made more efficient and relevant? How can one utilize "cheap but soft and flimsy" data from a Google crawling operation counting occurrences of a few key words (perhaps from masses of Twitter and Facebook messages) with "expensive but harder" types of data, to predict swings in an influenza epidemic, or the start of a social revolt in a society, in real time? We suggest tools associated with CDs for focus parameters and ways to combine pieces of information via estimated confidence likelihoods may be important in such endeavours; for some illustrations and further discussion, see Schweder & Hjort (1996, 2002, 2013).

We shall not pursue these grander aims here but wish to contribute to an ongoing debate concerning an application Xie and Singh point to in their Section 6. It concerns attempted and partly conflicting meta-analyses of 48 two-by-two tables. These relate to alleged increased health risks for users of a certain antidiabetic drug used to treat type 2 diabetes mellitus (working as an insulin sensitizer by making certain fat cells more responsive to insulin). The manufacturing pharmaceutical company in question is over the course of 2012 alone the subject of more than 13,000 lawsuits. For more on the background, in addition to other pointers by Xie and Singh, see Nissen & Wolski (2007). Part of the statistical debate is how to handle the "null tables"; 8 of the 48 tables are blessed with zero deaths and zero myocardial infarction (MI) events.

Consider a collection of two-by-two tables in the form of paired binomial experiments,

$$Y_{i,0} \sim \mathrm{Bin}(m_{i,0}, p_{i,0}) \quad \text{and} \quad Y_{i,1} \sim \mathrm{Bin}(m_{i,1}, p_{i,1}),$$

in biostatistical applications typically parameterized via the logistic transformation

$$p_{i,0} = \exp(\theta_i)/\{1 + \exp(\theta_i)\} \quad \text{and} \quad p_{i,1} = \exp(\theta_i + \psi_i)/\{1 + \exp(\theta_i + \psi_i)\}.$$

In the application described above, $Y_{i,1}$ is the number of deaths or MI events in the drug group and $Y_{i,0}$ the corresponding number of deaths or MI events in the control group. The typical

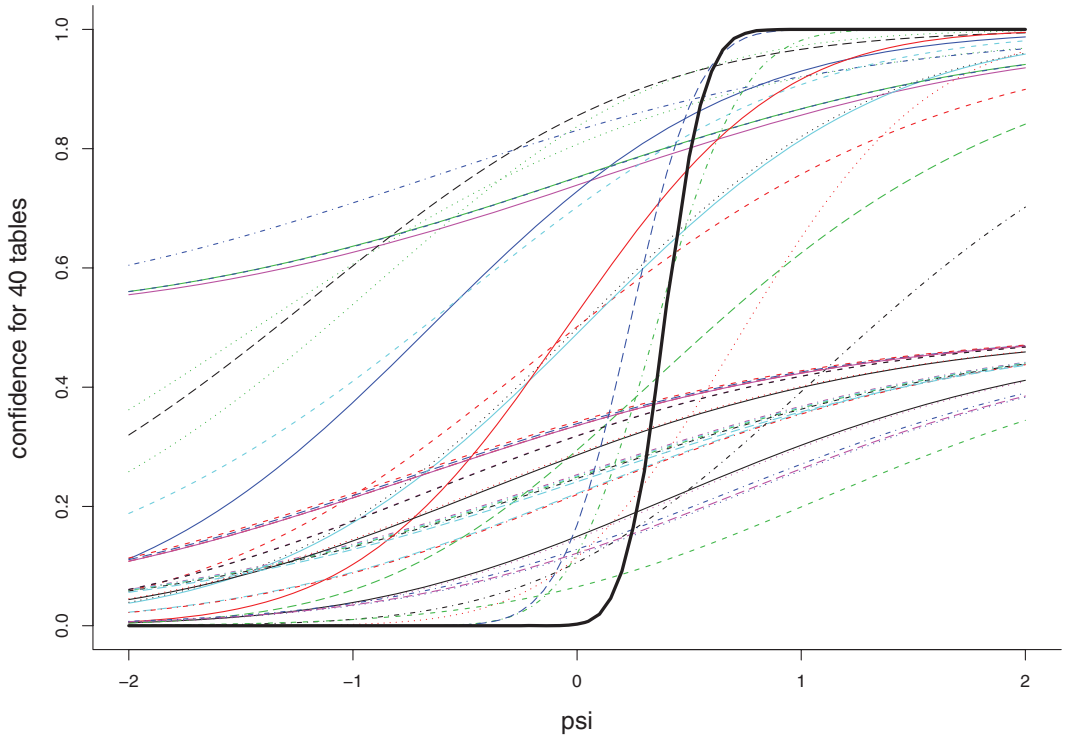

Figure 2. *Confidence distributions for each individual log-odds difference ψ_i, see Section 6. This is one way of representing the relevant part of the statistical information in the relevant 40 tables. The fatter line is the optimal overall confidence distribution computed via meta-analysis.*

meta-model here is to take the log-odds difference ψ constant across groups. The natural CD-driven meta-analysis approach is then (a) to compute, display, and compare the CD for ψ for each of the two by two tables, followed by constructing an overall CD for the common ψ using all the information.

The log-likelihood function is found to be

$$\ell_n = \sum_{i=1}^{n} \big[y_{i,1}\psi + z_i\theta_i - m_{i,0}\log\{1+\exp(\theta_i)\} - m_{i,1}\log\{1+\exp(\theta_i+\psi)\}\big]$$

over the $n = 48$ tables, where $z_i = y_{i,0} + y_{i,1}$. We may use the i-th component here to form the optimal CD for ψ based on the i-th pair of tables, namely

$$H_i^*(\psi) = P_\psi\{Y_{i,1} > y_{i,1,\mathrm{obs}} \mid z_{i,\mathrm{obs}}\} + \tfrac{1}{2}P_\psi\{Y_{i,1} = y_{i,1,\mathrm{obs}} \mid z_{i,\mathrm{obs}}\},$$

employing the usual half-correction for discreteness and using "obs" to indicate observed value. The conditional distribution in question is of the excentric hypergeometric type, depending as per general exponential class theory only on ψ and not θ_i; indeed $f(y_{i,1} \mid z_i)$ is proportional to $\binom{m_{i,0}}{z_i - y_{i,1}}\binom{m_{i,1}}{y_{i,1}}\exp(\psi y_{i,1})$ for $y_{i,1} = 0, \ldots, z_i$. Secondly the optimality result of Schweder & Hjort (2002) applies also to the grander task of constructing a CD for ψ using data from all tables (in effect focussing on one parameter of the 49-parameter model and treating the other 48 as nuisance parameters),

$$H^*(\psi) = P_\psi\{T > t_{\mathrm{obs}} \mid z_{1,\mathrm{obs}}, \ldots, z_{n,\mathrm{obs}}\} + \tfrac{1}{2}P_\psi\{T = t_{\mathrm{obs}} \mid z_{1,\mathrm{obs}}, \ldots, z_{n,\mathrm{obs}}\},$$

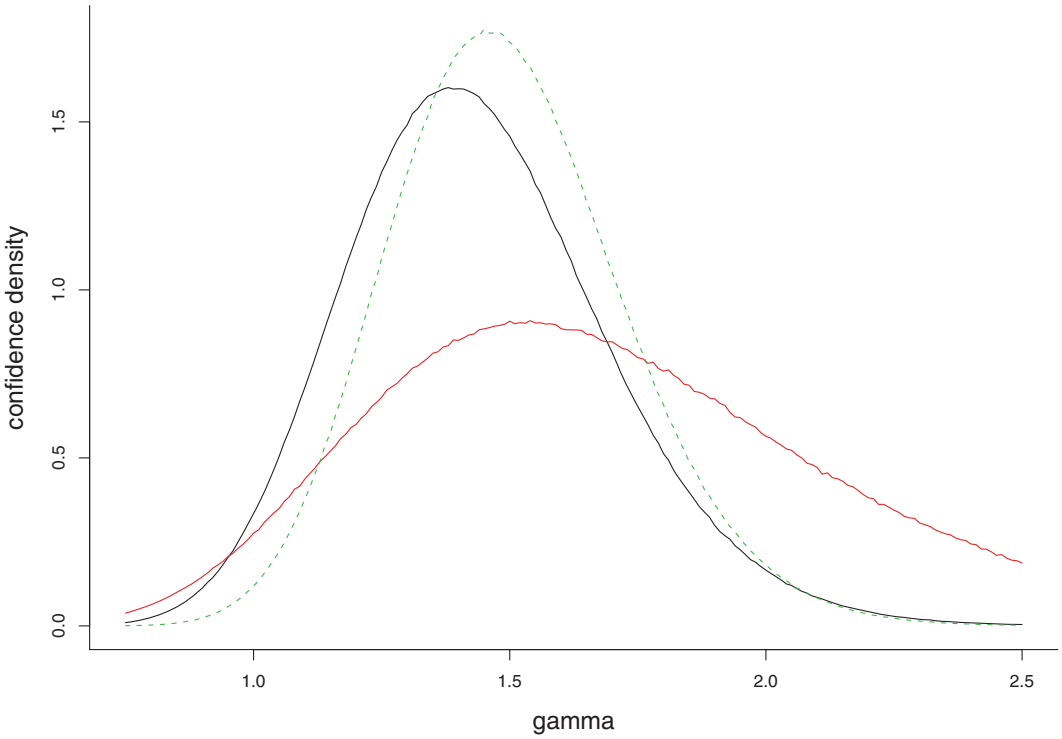

Figure 3. *Three confidence densities $h^*(\gamma)$ associated with the optimal confidence distributions $H^*(\gamma)$, for the risk proportionality parameter of the Poisson models when used for meta-analysis of the data described in Section 6. These are for MI only (full line, ML point estimate 1.421); for cardiovascular disease related death only (full line, ML point estimate 1.659); and for the combined group MI + Death (dotted line, ML estimate 1.482). It is apparent that the Roziglitazone drug increases the risk for MI and cardiovascular disease related death with more than 40%.*

where $T = \sum_{i=1}^{n} Y_{i,1}$. The conditional distribution of T is complicated but may be evaluated through simulation of each $Y_{i,1} \mid z_{i,\text{obs}}$.

As mentioned earlier there is a still ongoing debate in the literature concerning the "null tables" where both $y_{i,0}$ and $y_{i,1}$ are zero, with proposals ranging from leaving them out of discussion and *ad hoc* modifications of large-sample approximations to empirical Bayesian solutions. We suggest that the above provides a principled solution to the problem, without *ad hoc* arguments: The optimal CD is that of H^*, and for the 8 out of 48 tables for which $z_{i,\text{obs}} = 0$, the $Y_{i,1}$ is simply zero too; in other words, only the 40 tables where there is at least one death really contribute to the conditional distribution of U. Figure 2 displays each individual CD for ψ along with the optimal overall CD, clearly indicating that the drug increases the log-odds risk (with point estimate 0.405, 95% interval [0.122, 0.680], and p-value 0.003 for $H_0 : \psi \leq 0$, all read off from the CD). Only the 40 tables that matter are represented in the figure however, as per the comment just made; also, such a null table carries no confidence information about ψ.

Data such as these may also fruitfully be analysed using Poisson models, since luckily the patients we are in the "high n, low p" domain of the binomial model. A natural model here takes $Y_{i,0} \sim \text{Pois}(e_{i,0}\lambda_i)$ and $Y_{i,1} \sim \text{Pois}(e_{i,1}\lambda_i\gamma)$, where $e_{i,0}$ and $e_{i,1}$ are exposure factors (e.g., proportional to sample size, and possibly involving other covariates thought to influence any differences between pairs of tables). The model has $n + 1 = 49$ parameters, with γ the crucial

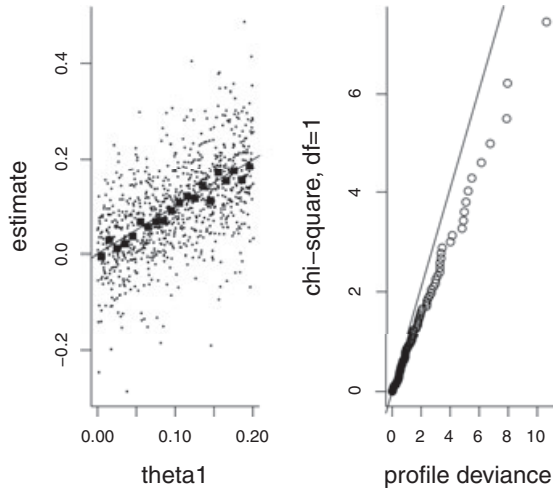

Figure 4. *A scatter of simulated* $\widehat{\theta}_1$ *by* θ_1, *with binned medians as black squares nearly following the diagonal; and a qq plot of* $D(\theta_0)$ *against the* χ^2_1 *distribution.*

focus parameter reporting on the extent to which the drug in question increases the risk of death or an MI event. Figure 3 displays three confidence densities, say $h^*(\gamma)$ (derivatives of CDs $H^*(\gamma)$), again optimally constructed via the appropriate conditional distribution, corresponding to death, to MI, and to the combined event death or MI. These carry the same flavour as a Bayesian's posteriors, but are crucially different in construction and interpretation in that no Bayesian prior camels are swallowed. They report in an optimally informative fashion on what we are interested in, using the data and nothing but the data.

7 A CD-Driven Re-Analysis of Sims's Bayesian Story

Together with Thomas Sargent, Chistopher A. Sims was awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel (alias the Nobel Prize in Economics) for 2011. He used the occasion of his Nobel lecture to promote Bayesian methods in econometrics (Sims, 2012). To show that CD methods also work in more complex models, we re-analyse the data Sims used and shall compare his Bayesian result for a key parameter with the CD obtained from the profile likelihood.

Sims considered the following macro-economic data. For year t let C_t be consumption, I_t investment, Y_t total income, and G_t Government spending. The following model was estimated using annual, chain-indexed, real GDP component data for the United States, $t = 1929, \ldots, 1940$:

$$\begin{aligned} C_t &= \beta_0 + \beta_1 Y_t + \sigma_C Z_{1,t}, \\ I_t &= \theta_0 + \theta_1 (C_t - C_{t-1}) + \sigma_I Z_{2,t}, \\ Y_t &= C_t + I_t + G_t, \\ G_t &= \gamma_0 + \gamma_1 G_{t-1} + \sigma_G Z_{3,t}. \end{aligned}$$

Here the $Z_{i,t}$ are taken as i.i.d. and standard normal. The multiplier θ_1 is of special interest. It cannot be negative according to Sims. He therefore assumes a flat prior on the six regression

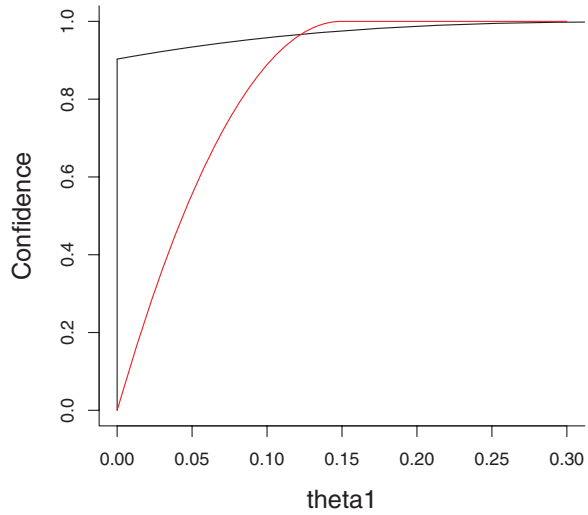

Figure 5. *The cumulative confidence distribution for θ_1 (with point mass at zero) together with an approximation to the cumulative posterior distribution obtained by Sims (2012).*

coefficients of the model but restricted to the positive halfline for θ_1, along with a prior that is flat in $1/\sigma^2$ for the three variance terms. From this he obtains a posterior distribution for the interest parameter that is nearly triangular on [0, 0.17] with mode at $\theta_1 = 0$.

The left plot of Figure 4 indicates that the unrestricted maximum likelihood estimator $\widehat{\theta}_1$ is free of any bias in the median. The qq plot to the right in the figure is similar to the other qq plots we looked at for various values of the parameters, and indicates that the profile deviance evaluated at the assumed value θ_1^0, $D(\theta_1^0)$, is nearly χ_1^2 distributed. Assuming this distribution, the unrestricted confidence curve $\text{cc}(\theta_1) = \Gamma_1(D(\theta_1))$ is obtained (not shown here). This is then converted to a CD, following our general recipe mentioned above. Setting $C(\theta_1) = 0$ for $\theta_1 \leq 0$, corresponding to Sims's prior constraint, we reach our CD; see Figure 5 for Sims's cumulative credibility and our cumulative confidence. We should be 90.03% confident that $\theta_1 = 0$, that is, investment was insensitive to changes in consumption in the pre-war period!

References

Berger, J.O. & Bernardo, J. (1992). On the development of reference priors [with discussion]. In *Bayesian Statistics 4*, Eds. J. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, pp. 35–60. Oxford: Oxford University Press.

Claeskens, G. & Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Datta, G.S. & Ghosh, M. (1995). Some remarks on noninformative priors. *J. Am. Stat. Assoc.*, **90,** 1357–1363.

Dempster, A. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, **34,** 884–891.

Fisher, R.A. (1930). Inverse probability. *Proc. Cambr. Philos. Soc.*, **26,** 528–535.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, **19,** 491–544.

Nissen, S.E. & Wolski, K. (2007). Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New Engl. J. Med.*, **356,** 2457–2471.

Pitman, E.J.G. (1939). The estimation of location and scale parameters of a continuous population of any given form. *Biometrika*, **30,** 391–421.

Raftery, A.E. & Schweder, T. (1993). Inference about the ratio of two parameters, with application to whale censusing. *Am. Stat.*, **47,** 259–264.

Schweder, T. & Hjort, N.L. (1996). Bayesian synthesis or likelihood synthesis: what does Borel's paradox say? *Rep. Int. Whal. Commn.*, **46,** 475–479.

Schweder, T. & Hjort, N.L. (2002). Likelihood and confidence. *Scand. J. Stat.*, **29,** 309–322.

Schweder, T. & Hjort, N.L. (2013). *Confidence, Likelihood, Probability.* Cambridge: Cambridge University Press.
Sims, C.A. (2012). Statistical modeling of monetary policy and its effects. Lecture upon acceptance of the Nobel Prize in Economics. *Am. Econ. Rev.*, **102,** 1187–1205.
Stein, C. (1959). An example of wild discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, **30,** 877–880.
Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, **76,** 604–608.

[*Received November 2012, accepted November 2012*]

International Statistical Review (2013), 81, 1, 68–77 doi:10.1111/insr.12001

Rejoinder

Min-ge Xie

Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854
E-mail: mxie@stat.rutgers.edu

1 Introduction

Sincere thanks to Professors David R. Cox, Brad Efron, Donald A.S. Fraser, Nils L. Hjort, Emanuel Parzen, Christian P. Robert, and Tore Schweder for their discussions. I am grateful for their insightful and scholarly contributions, which will no doubt clarify many important issues pertaining to the topic of confidence distributions and general statistical inference as a whole. While we were waiting and looking forward to receiving the discussions, Professor Kesar Singh, my co-author of the article and a personal friend, sadly had a massive heart attack and passed away much too young and too sudden. I have been left alone to carry out this rejoinder and our research. Professor Singh was a great thinker and a dear colleague. He will be deeply missed in the statistics community and also in my personal life.

In this rejoinder, I will begin by answering a few common questions from the discussants. In particular, I will provide in Section 2 a pragmatic view of statistical inference shared by Professor Singh and myself. We have subscribed to this pragmatic view in our approach to confidence distributions. And it is our hope that this view could provide a potential conciliation point for the Bayesian-fiducial-frequentist controversies of the past. In Section 3, I will respond to a question raised by Professors Fraser, Schweder, and Hjort on whether it is helpful to distance confidence distributions so sharply from fiducial distributions. In Section 4, I will discuss a subtle difference in perception between a confidence interval and a confidence distribution, drawing from a discussion by Professor Cox. Finally, in Section 5, I will offer our response to each discussant.

2 A Pragmatic View of Statistical Inference and a Frequentist Measure

Statistical approaches have evolved in many forms to tackle more and more complicated data problems in every corner of our life. Although there are many different statistical data-analysis

approaches of varying depth and complexity, they can be loosely described in two grossly oversimplified steps: A probability structure (model) is assumed to describe the uncertainty of the variables in the sample data; then, data analysis is performed to learn more about the assumed model and to make inference and statistical conclusions. In a frequentist approach, the data analysis is performed conditional solely on the assumed model structure. In the frequentist approach, the parameters in the model are unknown fixed quantities. In a Bayesian approach, in addition to the assumed probability model on the data, a prior probability structure for the model parameters is also assumed, and both the model and prior assumptions are utilized in data analysis. In the Bayesian approach, the parameters in the model are unknown random quantities. The frequentist and Bayesian approaches represent two major groups of research and paradigms. They have co-existed for a long time but with protracted battles, vociferously fought, from both camps. An interesting tidbit is that R. A. Fisher "began life a Bayesian" and Bayesian theory (then referred to as "the theory of inverse probability") was "an integral part of the subject" before Fisher's introduction of fiducial inference (c.f., Zabell, 1992). Fisher was among the first to question the "subjective and arbitrary" Bayesian approach because it "depended upon an arbitrary assumption [referring to the prior]." He proposed an "objective alternative," known as fiducial inference, to "define a distribution for parameters of interest that captures all of information that data contain about these parameters . . . without assuming a prior distribution." (Hannig, 2009). The now well-accepted concept of confidence interval by Neyman (1937) came later after Fisher's introduction of his fiducial inference. In fact, Neyman described his development as "an alternative description and development of Fisher's fiducial probability" and Fisher referred to it as "a generalization of the fiducial argument" (Zabell, 1992). But the relationship between Fisher and Neyman became "completely broken down" soon after. Based on what we have read, their feud appeared to have hampered, to a large degree, any potential, mutually cordial and consistent development of fiducial and confidence theory.

The Bayesian-frequentist-fiducial debates reflect "different attitudes to the process of doing science" (Efron, 2005). As an applied science providing analytic tools for other scientific subjects, a key element of statistics (statistical science) seems to be that a statistical approach should provide a sensible result to help advance scientific development. Perhaps, we should, as advocated by Kass (2011) and many others, "move beyond the frequentist-Bayesian controversies of the past" and abandon any attempt to take "as its goal exclusive ownership of inference" which is "doomed to failure." Instead of arguing which attitude or defining logic is superior, we from a pragmatic viewpoint could perhaps judge each approach based on its end result and focus more on whether or not the result from the approach is sensible. As pointed out by Professor Cox (in a private communication), regardless of which method we use, Bayesian, frequentist or fiducial, in the end the ultimate judgment is that "a statistical result should make sense." If we are conscientious about our effort to provide sensible results and conclusions, it is our opinion that we provide a good service to society and make a contribution to a better understanding of science.

Different statistical paradigms may apply different rules or criteria to judge whether or not the end result is sensible. To many, the frequentist repetitive interpretation of the coverage rate of a confidence interval is a good criterion that can be used to measure whether an inference conclusion is a sensible one. It is simple, widely accepted and easy to explain in layman's terms. In the review article, we have subscribed to this frequentist principle and have concentrated on whether or not an approach can provide intervals with sensible frequentist coverage. Throughout the article, we have tried to articulate the theme that *any approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of confidence distributions, as long as it can be used to build confidence intervals of all levels, exactly or asymptotically*.

Under this development, a fiducial or a Bayesian approach can be viewed as just one of many procedures that can potentially provide confidence distributions ("distribution estimators" with proper frequentist coverage rates). The role that a fiducial or a Bayesian approach plays in the distribution estimation to provide a confidence distribution is similar to the role that an MLE procedure or a set of estimating equations plays in point estimation to provide a consistent estimator for a parameter of interest. Of course, the consideration of the coverage rate is just one criterion. The need to consider additional requirements, for instance, some optimality criteria, may arise depending on the specific context of each application. Nevertheless, we think measuring a performance by the frequentist coverage criterion, which underlines our development of the confidence distribution concept, is at least a good starting point.

Many may prefer to use a Bayesian criterion to judge whether a result is sensible. If confidence or fiducial distributions can be unified under a Bayesian framework, as long as the approaches provide sensible solutions that meet the demands of science and applications, then we see no flaws in such an approach.

3 Is It Helpful to Distance Confidence Distributions So Sharply from Fiducial Distributions?

Professors Schweder and Hjort state that as mathematical objects, confidence distributions and fiducial distributions are equivalent. This is especially true when we use the same pivot. Professor Fraser also questions whether our treatment of confidence distributions draws a rather difficult dividing line between confidence and fiducial. Some of the answers may possibly be explained by our education and our journey towards doing research on the topic.

The first time that I learned about fiducial distributions was during a half-hour seminar by Professor Xiru Chen when I was a senior undergraduate student in the University of Science and Technology of China. At that time, it was very difficult for me to understand how a fixed number (parameter) could have a distribution and we were warned that the fiducial idea was very controversial. I had never seriously encountered fiducial inference again until I started to work on confidence distributions. Professor Singh's first encounter with the fiducial idea was the one-page introduction in Rao (1973), which also suggested that the development is "controversial" with "several objections." He would probably have also avoided the fiducial concept completely, if he had not started to do research on confidence distributions. Our experience is very typical for many statisticians of our generation. We are either minimally or never exposed to the fiducial concept. Even with a minimal exposure, we are immediately warned that it is controversial and advised to stay away from the concept. This is just an unfortunate truth, since we agree with Professor Fraser that the fiducial idea is one of Fisher's major innovative contributions involving deep insight and wisdom. We suspect that the widespread negative exposure of fiducial inference is partly caused by the fact that a fiducial distribution has never been fully defined and "Fisher's own thoughts on fiducial inference underwent a substantial evolution over time" (Zabell, 1992). Also, Fisher was very passionate about his fiducial inference idea and he attached, in our view, requirements that are too constraining. Many of the so-called fiducial paradoxes are directly related to the strong requirements imposed by Fisher. In addition, his disputes with Pearson, Neyman, and others, and his furious rejection of Neyman's interpretation of fiducial distributions (i.e. confidence distributions), did not help the development either.

Our research on confidence distributions started with our effort to better understand bootstrap distributions conceptually, and not merely treating them as computing objects. This lead to our attempt to expand the estimation concept from a single point or an interval to a sample-dependent

distribution function. The paper by Efron (1998) helped shape our development and choice of the term confidence distribution. It was never our intention to re-open the long-lasting philosophical debate in the history of statistics. To us, the frequentist interpretation of probability coverage rate is simple, easy to explain and widely accepted by the majority statisticians and scientists in other fields. Under this interpretation, we can clearly define a confidence distribution in the framework of estimation. Professor Fraser asks whether we are making issues with the over-enthusiasm in some of Fisher's arguments. Compared with Fisher's proposal of fiducial distributions, it is true that the key difference of the confidence distributions studied in our article is that they are freed from some of those restrictive constrains set forth by Fisher. We also avoid Fisher's interpretation that randomness can be transferred from sample data to parameters (which works nicely in location and scale parameter models, but not necessarily all), by treating a confidence distribution as a "distribution estimator" instead of an inherent distribution of the target parameter. We choose to take this route, not because we are trying to make issues with some of the fiducial arguments but to follow our objective (in Professor Cox's statement) "to provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)." Some examples of confidence distributions discussed in our review article may not pass the test of being a bona fide fiducial distribution judged by Fisher's original set of rules. The frequentist development allows us to sidestep several controversies and contradictions in the classical fiducial inference. Based on our experience to date, the method seems to provide reasonable answers in many practical problems.

Although we think that a confidence distribution is a clean and coherent frequentist concept (similar to a point estimator), we do not believe that it can be developed into a new fiducial theory or a new philosophy (either frequentist or Bayesian) that can solve all statistical paradoxes. Professor Cox points out that our goal to provide simple and interpretable summaries "is not always achieved by unqualified specification of a distribution." Indeed, a sensible confidence distribution may not always exist. Sometime, even when we can find a confidence distribution, it may not provide a sensible solution as judged by optimality or other criteria. From the viewpoint of estimation, these issues are not different from those in point estimations, since we also encounter situations and examples in which a sensible point estimator does not exist or is hard to find. While we think that the confidence distribution is a useful tool for statistical inference, we thank and concur with Professors Cox, Fraser, Schweder, and Hjort for highlighting the risks of over extrapolation and interpretation of the concept.

4 Confidence Interval Versus Confidence Distribution

In his discussion, Professor Cox explains his suggestion of using the term confidence distribution in his 1958 paper. The key, from our interpretation, seems to be that a confidence distribution can have "more flexibility" for a summarization of evidence than a confidence interval of a fixed significance coefficient (α level). As described by Professor Cox, one often has a sense that "when 95% confidence limits of a normal mean are found then, even if the parameter is outside the calculated range, it will not be too far outside." This cannot be captured based on the definition of a 95% confidence interval, but can be clearly displayed by a confidence distribution. Cox (1958) highlighted the distinction between inference and decision. Professor Cox has also discussed the role of significance coefficient. Although a confidence interval is an inference procedure, the choice of significance coefficient involves an operational decision. A confidence distribution does not involve any choice of a significant coefficient or any operational decision. As an inference tool, it can provide a fuller picture to summarize all evidence for the target parameters.

We share the same view with Professor Fraser and several other discussants that we need to raise the stature of confidence (intervals) to a distributional inference. We also agree with Professors Schweder and Hjort that the usefulness of distributional inference under frequentist paradigm is "too modestly explored," a sentiment also shared by Professor Parzen in his discussion.

5 Response to the Discussants

5.1 David R. Cox

Professor Cox's discussion provides a precise and clear portrait of a confidence distribution, its relationship with Fisher's original fiducial idea, its connection to Neyman's original introduction of a confidence interval, and also an interesting distinction between Fisherian discussions and those in the spirit of Neyman and Pearson. Professor Cox points out that the objective of a confidence distribution is to "provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)." He also stresses that "some conditions are, however, needed for the confidence distribution to be an appropriate summary." We fully embrace these messages and are very grateful for Professor Cox's contribution, insight and wisdom.

Professor Cox points out the interesting distinction that Fisher usually begins with a sufficient statistic where available, whereas Neyman–Pearson theory starts with sample data and sufficiency is the aftermath of an imposed optimality criterion. He asks whether there would have been any general implications for our paper if greater weight had been placed from the start on sufficiency, including asymptotic sufficiency. We have no definite answer for this question. So far, our work on confidence distributions is in the spirit of Neyman and Pearson and starts with sample data. Some recent fiducial developments under the framework of generalized fiducial inference (see, e.g. Hannig, 2009) have also started with sample data. Since a meaningful reduction by sufficiency is not always available, starting from sample data allows the developments applicable to broader settings. But, as stated by Professor Cox, when applicable, the sufficiency concept is more general than the optimality criteria of Neyman–Pearson theory. Although we have not carefully investigated the issue, we speculate that in some (but not all) situations when reduction by sufficiency is applicable, starting from sample data instead of a sufficient statistic may cause us to pay some price in terms of optimality. It would be certainly interesting to start our CD development with sufficiency (or asymptotic sufficiency) and investigate optimality issues.

5.2 Bradley Efron

Professor Efron states that "an important, perhaps the most important, unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information." He points out that the development of confidence distributions can potentially be a way to practice "objective Bayes." He also provides an additional bootstrap example leading to a higher order (second order) confidence distribution. The points raised by Professor Efron are interesting and worth further pursuing. We are very grateful to Professor Efron for his consistent support, encouragement, and suggestions relating to our developments on this subject.

As mentioned previously, our research started with an investigation to understand a bootstrap distribution as a "distribution estimator" of a parameter of interest. Viewing a bootstrap distribution function as a frequentist distribution estimator for inference instead of merely a

computation procedure has helped shape our view on confidence distributions. In our article, we have also briefly mentioned the potential connection of confidence distributions to objective Bayesian approaches. In addition, we have reviewed and briefly explored the connections of a bootstrap distribution to a confidence distribution and a CD random variable. We think that the connections can perhaps also be further extended to some computing methods developed under a fiducial framework or based on fiducial ideas.

5.3 Donald A.S. Fraser

Professor Fraser's discussion provides an overview of distributional inference along with historical information and his own experiences. The scholarly discussion is very valuable and can help us better understand the classical fiducial development and frequentist inference. We share many common views with Professor Fraser regarding distributional inference. In particular, like Professor Fraser, we also believe that Fisher's innovative idea of fiducial inference should have been better received by, and more broadly introduced to, the general statistical community. We also embrace Professor Fraser's strong warnings against overzealous statements on distributional inference that overlook potential inherent risks. We have never believed or hinted that treating a confidence distribution as a distribution estimator can solve all problems in statistical inference. Instead, we discussed the limitation of the confidence distribution approach in our discussion section. We view these difficulties as essentially similar to those problems encountered in point estimations.

Professor Fraser appears to prefer keeping Fisher's fiducial inference in its entirety, while we are adopting a viewpoint more in the spirit of Neyman and Pearson. One key difference, which we have stressed in our article, is that we do not view a confidence distribution as an inherent distribution of a fixed target parameter. Rather, we treat the confidence distribution approach as a sample-dependent function used to estimate the parameter of interest. Although Fisher proposed an elegant interpretation of transferring randomness from data to parameter after observing data in location/scale family problems, we still feel it is difficult to understand and explain to the general public that a fixed number (parameter) has a distribution. On the contrary, our experience tells us that the Neymanian interpretation is easier to explain and communicate. This interpretation also frees us from some of the constraints Fisher over-enthusiastically imposed (such as uniqueness, optimality and manipulation of distributions, etc.) and allows us to sidestep many so-called fiducial "paradoxes." We agree with Professor Parzen who states—"Today the question should not be about credit for methods, but a framework for tools which are simple and powerful for applications." It is our hope that the fiducial idea and confidence distribution developments will not just be a research topic for elite statisticians but also be an inference tool that can be understood and used by general public and practitioners of statistics. We thank and appreciate Professor Fraser for his engaging discussion and insightful comments.

5.4 Emanuel Parzen

Professor Parzen provides an interesting interpretation of confidence distributions using confidence quantiles, with which we were not previously familiar (and guilty of negligence of the related literature in our article). In his discussion, Professor Parzen introduces confidence quantiles using estimating equations based on pivots. The definition (equation) is similar to the classical fiducial equation and also so-called generalized fiducial inference (but where estimating equations are introduced based on individual data; see, e.g. a review article by Hannig, 2009). Professor Parzen's discussion provides an interesting look and connection to the CD random variable in our article. When the structure and solution exist, the approach provides

a systematic and easy way to obtain confidence quantiles and also a confidence distribution. One question that I have on Professor Parzen's discussion is whether the h function (internal representation) always exists or, if not, under which general conditions does it exist. I think it is worthwhile to further explore the connections between the existing developments of confidence quantiles and confidence distributions, in particular under the scope of distributional-inference-based computing, including fiducial, CD-based and bootstrap computational methods. Finally, we really appreciate and thank Professor Parzen for his constant support and for raising the interesting connection.

5.5 Christian P. Robert

Professor Robert raises many objections to our article and non-Bayesian inferences. From his discussion, it is not difficult to tell that Professor Robert is very passionate about Bayesian inference. I appreciate the passion and conviction exhibited in Professor Robert's discussion, even though I am surprised by many of his comments. Professor Robert's objections can probably categorized into two groups: 1) the statements that are invalid in Professor Robert's view of Bayesian philosophy and principles and 2) the statements that are misinterpreted and misunderstood by Professor Robert. Items in category 1 include those concerning the sacred status of priors and the accusation of being "ad-hocquery" for allowing multiple estimators in frequentist practices, among others. To avoid prolonging the existing patriotic debate between different philosophical points of view, which our review article has set out to overcome to begin with, I will not address category 1 and will simply let the reader be the judge. Among the objections related to category 2, quite many of them seem to be simple misunderstandings and misinterpretations of our wording. Some of the others, aside from our disagreements on some philosophical issues, contains technical statements made by Professor Robert which are not quite accurate. In particular, I would like to reply to Professor Robert's discussion on the "discrepant posterior phenomenon," for which I believe has broad implications and practical value.

The discrepant posterior phenomenon was reported in Xie *et al.* (2012) as a cautionary tale and also as a reflection as how we at times may overly emphasize a methodology to provide a solution without paying enough attention to the undesirable consequences that could be brought about by the methodology. Professor Robert uses two simulated numerical examples to show that the discrepant posterior phenomenon does not exist in a regular Bayesian analysis. He further suggests a copula procedure that can always produce a global prior from marginal priors, and thus a Bayesian solution to the real data problem addressed in Xie *et al.* (2012). But, in fact, this is exactly a blind spot in a regular practice that was highlighted in Xie *et al.* (2012)—we are at times too easily satisfied to have a methodology to provide a solution without further examining its validity in specific applications. Also, very often, we have been overly reliant on simulation results under specific settings for proving a concept. Indeed, in Professor Robert's Figure 2, which is used to "prove" that the discrepant posterior phenomenon does not exist, we can clearly see the undesirable discrepant posterior phenomenon. Although it is not an issue with the marginal distributions of $p_1 - p_0$, the undesirable discrepant posterior phenomenon exists on the marginal distributions of p_1: there is a triangle formed by the centres of the three (blue, light brown, and red) contours and the mode of the marginal posterior of p_1 (around 0.4) is bigger than both the mode of the marginal prior ($1/3 \approx 0.33$) and the MLE ($35/100 = 0.35$) directly obtained from the data. Although the phenomenon is relatively mild, it clearly presents in Professor Robert's Figure 2.

A much simpler example (see Figure 1 below) to demonstrate the discrepant posterior phenomenon is perhaps to use bivariate normal data with an informative bivariate normal

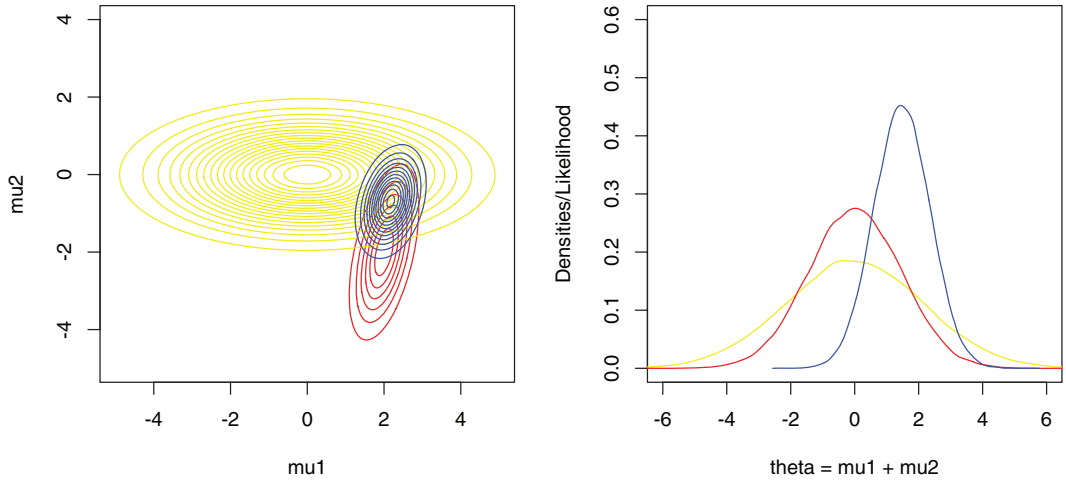

Figure 1. *The setup for the figure is: a sample of* $n = 3$ *points with a sample mean* $(\bar{x}_1, \bar{x}_2) = (2, -2)$ *is observed from a bivariate normal* $N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_1\sigma_2\rho & \sigma_2^2\end{pmatrix}\right)$; *for simplicity, the variance matrix is assumed to be known with* $(\sigma_1^2, \sigma_2^2, \rho) = (0.64, 4, 0.5)$. *The parameter of interest is* $\theta = \mu_1 + \mu_2$, *which has a marginal prior* $\pi(\theta) \sim N(0, 4.64)$. *In additional, we also know the marginal prior of* $\pi(\mu_1) \sim N(0, 4)$. (a) *Contour plots of the independent joint prior* [*in yellow*] *of* $(\mu_1, \mu_2) = \pi(\mu_1)\pi(\mu_2)$ *with* $\pi(\mu_1) \sim N(0, 4)$ and $\pi(\mu_2) \sim N(0, 0.64)$, *the likelihood function* [*in red*] *based on the observed sample mean* $(\bar{x}_1, \bar{x}_2)$, *and posterior* [*in blue*]; (b) *density plots of the marginal prior of* ~~μ_1~~ [*in yellow*], *the density function of* $N(0, 2.08)$ [*in red*], *and the marginal posterior of* ~~μ~~ [*in blue*].

prior. (Professor David Draper first pointed out that the discrepant posterior phenomenon exists even in bivariate normal examples in a follow-up discussion of my talk in his department at University of California at Santa Cruz. In the bivariate normal example, Professor Draper and his colleague worked out, mathematically, a set of explicit conditions (parameter settings) under which a discrepant posterior phenomenon occurs.) The bivariate normal example is cleaner mathematically than the beta-prior example introduced in Xie *et al.* (2012) and, unlike the beta-prior example, it can also be examined under the context of the marginalization paradox (Dawid *et al.*, 1973). (Note that in the beta-prior example of Xie *et al.* (2012), it is not possible to have the marginal model $f(data|\theta)$, and thus it is not covered by the regular narrative of the marginalization paradox; see Section 5 of Xie *et al.* (2012) for more discussions.) The marginalization paradox showed that Bayes formula can produce different (marginal) Bayesian posteriors for the same parameter, depending on whether we use a full model with all parameters or we use only its corresponding marginal models. The standard Bayesian inference sidesteps this paradox by advocating the use of the full model. But there is a further message by the discrepant posterior phenomenon: the full joint modelling approach advocated by the standard Bayesian inference may not provide us desirable solutions in some situations. The question raised in the real data example of Xie *et al.* (2012) is not whether we have a method to produce a numerical result. Rather, we ask whether we have a sensible methodology to provide a coherent solution when skewed distributions are involved in a multivariate setting.

This discrepant posterior phenomenon may have broad implications in the general practice of Bayesian analysis. For instance, many researchers have routinely drawn conclusions solely based on marginal posterior distributions without checking the validity of such conclusions. The discrepant posterior phenomenon suggests that further care is needed. It also raises a general question about using informative priors, regardless of whether or not we treat a prior as a reference measure. In particular, we need to ask: should we still incorporate an

informative prior distribution with the likelihood function even if they are clearly mismatch? The general discussion of the discrepant posterior phenomenon is also of relevance to some current research topics in Bayesian statistics, specifically in the attempts to quantify the impact of prior distribution on posterior inference as well as the work on calibrating a prior distribution using partial sample data; see, e.g. O'Hagan (1995) and Berger & Pericchi (2004).

5.6 Tore Schweder and Nils L. Hjort

Schweder & Hjort (2002) and other publications of confidence distributions by Professors Schweder and Hjort have used the same interpretation of confidence distributions in the spirit of Newman and Pearson as we do. We thank Professors Schweder and Hjort for their insightful discussion, which touches a broad range of topics with depth and insight. Their discussion, from the classical inference prospective, including their account of the intertangled relationship between fiducial and confidence distributions and their examples of confidence distributions in several non-trivial settings, provides a rich source of knowledge for understanding better confidence distributions and statistical inference as a whole. The discussion reflects their understanding and passion for the development of confidence distributions. We share their sentiment that distributional inference, especially confidence distributions and confidence likelihoods, has so far only been "too moderately explored" and that confidence distributions and confidence likelihoods should be used more "as broadly applicable tools for modern statistics."

We would like to elaborate the discussion of "ordinary probability calculation" and clarify our position on this particular issue. In Section 2 of their discussion, Professors Schweder and Hjort bring out a nice example in the classical literature on fiducial inference to show that distributions derived from fiducial distributions (or confidence distributions) are not, in general, confidence distributions. The same message was conveyed, though less directly, in the latter half of Section 3.1 in our review article. The key is that we view a confidence distribution as simply an object (a function that meets the requirements of being a distribution function) to estimate the parameter of interest, rather than as an inherent distribution of the parameter. There are three layers of messages that we would like to elaborate. First, the requirements of being a distribution function in the CD definition follow the ordinary definition of a cumulative distribution function (i.e. non-decreasing, from 0 to 1, etc.). Thus, in its mathematical form and as a pure mathematical object, a confidence distribution is an ordinary distribution function. This consideration is our statement in Section 3.1 of our review article that "like a bootstrap distribution, a confidence distribution is an ordinary probability distribution function for each given sample." Second, unlike the classical fiducial interpretation, the confidence distribution is *not* an inherent distribution of the parameter of interest. Thus, as discussed by Professors Schweder and Hjort that distributions derived from a manipulation of confidence distributions by ordinary probability calculation (treating them as distributions of parameters) generally do not give us confidence distributions (because they do not guaranteed a proper coverage rate of their corresponding intervals). We have the same message in our Example 3.1. This being said for the general case, we have an additional (the third layer of) message in Section 3.1 of our review article. That is, in some special situations such as monotonic or smooth transformations, the manipulation of confidence distributions actually does lead to either an exact or asymptotic confidence distribution for the transformed parameter. Thus, a manipulation of confidence distributions using ordinary probability calculation may sometime be used as a systematic procedure to obtain a confidence distribution. Since there is no guarantee in general (in viewing of the message in layer 2), we agree with Professors Schweder and Hjort that one must check each time that the sample-dependent function obtained by a manipulation satisfies the requirements of being a confidence distribution or not.

Finally, our review article has stated that a fiducial approach (and also, to some extent, a Bayesian approach) provides a systematic way to obtain confidence distributions. This statement is only in a general sense and it is in the same spirit as the statement that an MLE is often consistent. Due to its delicate nature, as described by Professors Schweder and Hjort, one must check any distribution obtained by a fiducial (or a Bayesian) argument before declaring it to be a confidence distribution. Our article and Table 1 in Singh and Xie (2011) have used an analogy of MLEs versus consistent estimators to compare the relationship between fiducial and confidence distributions. The fiducial and Bayesian methods are like our "MLEs" which provide concrete ways to obtain sample-dependent distribution functions for inference. Just as an MLE does not automatically guarantee a consistent point estimator, a fiducial distribution or a Bayesian posterior does not always have the frequentist coverage property and does not have to be a confidence distribution. But, often a fiducial distribution and, to some extent, a Bayesian posterior can be viewed as a confidence distribution, since under some regularity conditions, the sample-dependent distribution function from a fiducial or Bayesian argument satisfies the frequentist coverage requirement, exactly or asymptotically.

Acknowledgements

This research is partly supported by research grants from NSF DMS1107012, DMS0915139, and SES0851521.

References

Berger, J.O. & Pericchi, L. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.
Cox, D.R. (1958). Some problems with statistical inference. *Ann. Math. Stat.*, **29**, 357–372.
Dawid, A.P., Stone, M. & Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. R. Stat. Soc. Ser. B*, **35**, 189–233.
Efron, B. (1998). R.A. Fisher in the 21st Century. *Stat. Sci.*, **13**, 95–122.
Efron, B. (2005). Bayesian, frequentists, and scientists. *J. Amer. Statist. Assoc.*, **100**, 1–5.
Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, **19**, 491–544.
Kass, R. (2011). Statistical inference: the big picture. *Stat. Sci.*, **26**, 1–9.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc.* A, **237**, 333–380.
O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. R. Stat. Soc. Ser. B*, **57**, 99–138.
Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley & Sons.
Schweder, T. & Hjort, N.L. (2002). Confidence and likelihood. *Scand. J. Stat.*, **29**, 309–332.
Singh, K. & Xie, M. (2011). Discussions on Professor Fraser's article on "Is Bayes posterior just quick and dirty confidence?" *Stat. Sci.*, **26**, 319–321.
Xie, M., Liu, R.Y., Damaraju, C.V. & Olson, W.H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. Available at http://www.imstat.org/aoas/next_issue.html.
Zabell, S.L. (1992). R.A. Fisher and fiducial argument. *Stat. Sci.*, **7**, 369–387.

[*Received November 2012, accepted November 2012*]