The Annals of Applied Statistics
2018, Vol. 12, No. 2, 750–770
https://doi.org/10.1214/18-AOAS1160SF

WHEN SHOULD MODES OF INFERENCE DISAGREE? SOME SIMPLE BUT CHALLENGING EXAMPLES[1]

BY D. A. S. FRASER*, N. REID* AND WEI LIN†

*University of Toronto** *and AidVoice Lab*†

At a recent conference on Bayes, fiducial and frequentist inference, David Cox presented eight illustrative examples, chosen to highlight potential difficulties for the theory of inference. We discuss these examples in light of the efforts of the conference, and related meetings, to study the similarities and differences between the approaches to inference. Emphasis is placed on the goal of finding a distribution for an unknown parameter.

In memory of Steve Fienberg

DF: Steve was an undergraduate student long ago in classes for which I was fortunate to be the instructor; he had that extraordinary enthusiasm for the discipline and life, which makes the classroom scene a joy and a reason in itself. We remained close friends ever since. He was persistent in the pursuit of directions and needs in all areas of our discipline and tireless in bringing together people who could participate and contribute. He will be deeply remembered in his multiple roles.

NR: When I met Steve in recent years, at conferences or committee meetings, he always seemed to be busily tapping away on his iPad, and when he looked up he would say "Annals of Applied Statistics". (Although as it turned out he was editing several journals at the same time.) Steve was an inspiring mentor to me, about editorial work and much else. He showed by example how much fun a professional life can be when you take the effort to meet many people, try to understand what they are working on, keep an open mind about the potential of that work and focus on advancing our field.

1. Introduction. The fourth in a series of workshops on Bayesian, Fiducial, and Frequentist inference took place from May 1 to 3, 2017, at Harvard University; the theme was "foundational thinking in statistics and inference under uncertainty". The series of workshops has been abbreviated to "BFF", meant to connote at the same time "Best Friends Forever". At least part of the motivation for this is the desire to explore the possibility that approaches to inference by the three routes

Received December 2017; revised March 2018.
[1]Supported in part by the Natural Sciences and Engineering Research Council of Canada.
Key words and phrases. Asymptotic theory, confidence distribution, fiducial density, marginalization paradox, noninformative priors.

in the title can all be used to provide something like a distribution for an unknown parameter.

The Bayesian method based on the posterior distribution for a parameter of interest, conditional on the data, is perhaps the most accessible approach to constructing a distribution for a parameter. While "frequentism" is often associated with the approach to inference based on maximizing the power of tests, building on the fundamental lemma of Neyman and Pearson (1933), more recently the emphasis has shifted towards pure frequency properties, and the frequentist label in this trio refers to confidence distribution functions developed from Fisher (1930) and Cox (1958), and reviewed and extended by Xie and Singh (2013) and Schweder and Hjort (2016). The fiducial approach originated with Fisher (1930), was developed in transformation models in Fraser (1961) and more recently has been the focus of generalized fiducial inference in work by Hannig and his collaborators, for example, Hannig (2009) and Hannig et al. (2016). The p-value function or significance function approach of Fraser (1990) that we emphasize below is a frequentist approach, close to the confidence distribution, but taking inspiration from a structural version of Fisher's fiducial argument.

David Cox submitted to the workshop eight seemingly simple statistical examples as challenges for approaches to statistical inference. We list the examples below, and then provide a brief overview of the "B, F, F" approaches to inference. Our view is that the theory of higher order approximation based on the likelihood function addresses some of the issues in the eight examples, by providing both a measure of departure and a calculation of the significance function that is available in wide generality. An outline of the theory is given below, and then applied to the examples.

Although the examples are highly idealized, they isolate issues that can arise in applications where the anomalies might not be so clear-cut, and thus highlight the role of theory in applications. We provide some examples of this in the relevant sections discussing each example.

2. Cox's challenge questions.

A Variables y_1, y_2 are independently normally distributed with means μ_1, μ_2 and common variance 1. The parameter of interest is $\psi = \mu_2/\mu_1$ and the observed data are $y_1 = 0.5$, $y_2 = 10$.

B The same model as A, but with data $y_1 = 0.5$, $y_2 = 0.5$.

C Variables y and θ are independently normally distributed with means θ and 0 and common variance 1. The parameter of interest is θ. The value $y = 6$ is observed.

D Variables y_{i1}, y_{i2} are independently normally distributed with means μ_i and common variance σ^2 for $i = 1, \ldots, n$. The parameter of interest is σ^2.

E Variables y_i, $i = 1, \ldots, n$, are independently normally distributed with means μ_i, and common variance 1. The parameter of interest is $\Sigma\mu_i^2$.

F Variables y_i, $i = 1, \ldots, n$ are independently normally distributed with means $\cos(\theta x_i)$ and common variance 1. For large n, $x_1, \ldots, x_n$ are arbitrarily distributed over $(0, 1)$.

G There are n decimal digits. Is their marginal distribution consistent with those digits being independent and uniformly distributed on the set $\{0, 1, \ldots, 9\}$?

H What is the role of physical randomization?

In these examples, "routine" application of methods, either Bayesian or frequentist, lead to difficulties. Examples A and B highlight anomalies in the construction of confidence intervals; Examples D and E draw attention to the potential for Bayesian, likelihood and fiducial methods to fail in models with large numbers of nuisance parameters. Example C isolates in extreme form disagreement between different sources of information bearing on the problem. This can possibly be resolved in any individual instance, but how do theories of inference deal with this in general? Example F, and to some extent examples A and B, focus attention on the parameterization of the model. In example G the null hypothesis is well-defined, but alternatives are not. Example H draws attention to the divide between inference based on modelling and inference based on the design of the data collection method; the latter can be viewed as nonparametric, although current approaches to nonparametric inference tend to be model-based, even if the models are in some sense infinite-dimensional.

3. BFF. Inference in a Bayesian approach is reported as a posterior distribution for the parameter, conditioned on the observed data. Probabilities for θ are introduced by a prior distribution, so the role of this prior is central. Bayes (1763) used a mathematically convenient prior to suggest probability for θ. Many later writers saw this as very arbitrary, even subjective; see for example, Fisher (1956), Chapter 2. Laplace (1812) was somewhat accepting of the Bayesian approach, and put forward a notion of what would now be called a noninformative prior, meant to express complete lack of prior knowledge about the parameters.

Many, although not all, modern treatments of Bayesian theory for inference accept that inferences based on the posterior distribution from a particular prior should be *calibrated* under the model for the data. That is, the prior should lead to posterior inference that maintains its stated properties under repeated sampling from the model for the response $f(y; \theta)$. Often this calibration is defined by the notion of probability matching of one-sided intervals for scalar parameters, which is extensively developed in Datta and Mukerjee (2004), in which case a posterior region of probability α should be a confidence region at the stated level. There is a lengthy discussion of this requirement of calibration in Berger (2006), Goldstein (2006), Fienberg (2006), and Wasserman (2006) and other papers in the same volume.

In contrast, many modern applications of Bayesian inference do not dwell on the choice of prior or on a need for calibration, but rather make a choice of convenience, possibly in the expectation that this will not influence the result "too

much". Fraser (2011) argues that Bayesian inference obtains justification only through calibration, also called repetition validity or confidence. Rubin (1984) examines calibration in a wider sense, averaging over a range of models for the data, rather than calibrating under the given model.

A major difficulty is that calibration needs to be targeted on the parameter of interest, so the usual approach of marginalizing a multi-dimensional posterior to a series of individual parameters of interest gives posteriors that typically cannot be calibrated. A more general concern is raised in Reid and Cox (2015), who discuss different notions of probability, and question whether probability based on the uncertainty in our information should have the same standing for scientific inference as an empirical probability based on a given sampling mechanism.

Fisher (1930) introduced fiducial inference as a means of establishing a distribution for a parameter that did not require the injection of probability from outside the model. He assumed that a statistic T was available with a distribution depending only on the parameter of interest θ, with a known distribution function $F(\cdot;\theta)$. He further supposed that the distribution F was continuous and stochastically increasing, and argued that fixing T at its observed value t in the equation $p = F(t;\theta)$ established a relationship between p and θ. Then assuming that the relationship was monotone in θ, the fact that p follows a uniform distribution under the model induces a distribution for θ called the fiducial distribution, with density

$$f(\theta|t)\,\mathrm{d}\theta = -\frac{\partial}{\partial\theta}F(t;\theta)\,\mathrm{d}\theta.$$

Fraser (1966) reformulated Fisher's fiducial argument for transformation models. In these models linking the observation, or a statistic, to the parameter, is more direct. For example in a location model, a change from an observed value y to $y+a$ can be offset by a corresponding change in the location parameter θ to $\theta - a$ without changing the basic distribution. Fraser (1966) showed how this could be used to construct structural distributions for parameters in general transformation models. A local location version of that argument leads to

$$f(\theta|t)\,\mathrm{d}\theta = -F_\theta(t;\theta)\,\mathrm{d}\theta = F_t(t;\theta)\left\{\frac{-F_\theta(t;\theta)}{F_t(t;\theta)}\right\}\mathrm{d}\theta = L(\theta)\left(\frac{\mathrm{d}t}{\mathrm{d}\theta}\right)\mathrm{d}\theta, \tag{1}$$

where $L(\theta)$ is the likelihood function and $\mathrm{d}t/\mathrm{d}\theta$ can be described as the effect of θ at a data point t. We have used subscripts as shorthand for partial differentiation, and the derivative of t with respect to θ is for fixed value of the pivot $F(t;\theta)$. The final expression defines a data-dependent prior, discussed in Fraser et al. (2010).

Neyman (1937) restricted the relationship between t and θ to sets C of values for the pivot and constructed what he called confidence sets, replacing the term fiducial. Cox (1958) noted that one-sided confidence bounds could be converted to a *confidence distribution*:

> ... in the common simple cases, where the upper α limit for θ is monotone in α, there seems no reason why we should not work with the confidence distributions for the

unknown parameter. These can either be defined directly, or can be introduced in terms of the set of all confidence intervals at different levels of probability.

More formal notation for this was presented in Efron (1993). Defining the quantile function $\theta_t(\alpha)$ by $\Pr\{\theta \le \theta_t(\alpha)\} = \alpha$ as calculated from the density $f(t;\theta)$, the confidence density is then $\pi_t(\theta) = \mathrm{d}\alpha_t(\theta)/\mathrm{d}\theta$, where $\alpha_t(\theta) = \theta_t^{-1}(\alpha)$.

It is usually the case that the parameter of the statistical model is a vector, but the monotonicity requirement for the construction of fiducial or confidence distributions essentially restricts discussion to a scalar parameter of interest. Although a Bayesian approach seems to have an advantage in this case, because a posterior distribution for a single parameter can be obtained by marginalizing the joint posterior, as noted above the resulting inference is not calibrated unless the prior is targeted on the parameter of interest.

In the fiducial and confidence approaches, it is necessary to find a statistic that measures the parameter of interest, and typically to do this for one scalar parameter at a time. One approach is to use the asymptotic normal distribution of familiar likelihood-based quantities, such as the maximum likelihood estimate or the square root of the log-likelihood ratio statistic. These will give approximate confidence distributions; for recent discussion see Schweder and Hjort (2016), Chapter 3.

Our view is that the theory of higher order approximation based on the likelihood function leads to an essentially unique function of the data measuring a scalar parameter of interest and provides an accurate approximation to its distribution in finite samples. While the accuracy of the approximation is emphasized most often in the literature, the simplification offered by providing a single pivotal quantity may be more important. In the next section we summarize this theory by focussing on the key steps in the argument, and in the following section consider the eight challenge problems. The theory is based on a local location construction which provides approximately ancillary directions and leads to an approximate conditional model. The nuisance parameters are then eliminated by integration, and the resulting distribution is inverted to give confidence bounds and a significance function. In this sense it combines aspects of the fiducial argument as reformulated for transformation models, and the confidence distribution approach. An approach less directly tied to transformation models, but leading to similar expressions, is reviewed in Pierce and Bellio (2017).

4. Likelihood inference and accurate approximation.

4.1. *Local location approximation.* Suppose our model $f(y;\theta)$ for $y = (y_1, \ldots, y_n)$ has parameter $\theta \in \mathbb{R}^p$. Fraser and Reid (1995) obtained a p-dimensional model on the sample space by conditioning, using inference directions $V = (v_1, \ldots, v_p)$ at the observed value y^0. Assuming that a pivotal quantity $z(y;\theta)$ is available, which could be a vector of marginal distribution functions for

independent components, these directions are obtained as

$$V = -\left(\frac{\partial z}{\partial y}\right)^{-1}\left(\frac{\partial z}{\partial \theta}\right)\Big|_{(y^0,\hat\theta^0)} = \frac{\mathrm{d}}{\mathrm{d}\theta}y(z;\theta)\Big|_{(y^0,\hat\theta^0)},$$

where in the second expression z is fixed, and appears in a special context in (1). The inference directions V record how the parameter influences the data in a neighbourhood of the observed data point and associated maximum likelihood estimate. The conditional model in the linear space of these directions can be approximated by an exponential model with p-dimensional canonical parameter $\varphi^{\mathrm{T}}(\theta) = (\partial/\partial V)\ell(\theta; y)|_{y^0}$, called the tangent exponential model, and this model has all the structure needed to provide third-order accurate p-value or significance functions [Fraser and Reid (1995)].

4.2. *Exponential family models.* An exponential model for y has a density of the form

$$f(y;\theta) = \exp\bigl\{\varphi^{\mathrm{T}}(\theta)s(y) - \kappa(\theta)\bigr\}h(y),$$

where s has the same dimension as the canonical parameter φ. The parameter $\varphi(\theta)$ applied to $s(y)$ provides a logarithmic tilt of the base density $h(y)$ and $\kappa(\theta)$ provides the needed norming. In this model inference for θ is clearly based only on the variable $s(y)$, and the distribution of s can be accurately approximated using saddlepoint methods, leading to

$$h(s;\varphi)\,\mathrm{d}s \doteq \frac{e^{k/n}}{(2\pi)^{p/2}}\exp\bigl\{-(\hat\ell - \ell)\bigr\}|\hat\jmath|^{-1/2}\,\mathrm{d}s, \tag{2}$$

where $\hat\ell - \ell = \ell(\hat\varphi) - \ell(\varphi)$ is the log of the likelihood ratio, $\hat\varphi = \varphi(\hat\theta)$ is the maximum likelihood estimate and $\hat\jmath = \jmath_{\varphi\varphi}(\hat\varphi) = -\partial^2\ell(\hat\varphi)/\partial\varphi\,\partial\varphi^{\mathrm{T}}$ is the Fisher information matrix; each of these quantities also depends on s. This approximate density depends on relatively simple and widely used statistical quantities. The notation $\doteq$ is used here and below to indicate the leading term of an asymptotic approximation in n. As described in Section 4.1, it can be used whether or not the original model for y is of exponential family form.

4.3. *Scalar case.* If the dimension of φ is $p = 1$, (2) can be written

$$h(s;\varphi)\,\mathrm{d}s \doteq \frac{e^{k/n}}{(2\pi)^{1/2}}e^{-r^2/2}\frac{r}{q}\,\mathrm{d}r, \tag{3}$$

where $r^2/2 = \hat\ell - \ell$, $r = \mathrm{sign}(\hat\varphi - \varphi)\{2(\hat\ell - \ell)\}^{1/2}$ and $q = (\hat\varphi - \varphi)|\hat\jmath_{\varphi\varphi}|^{1/2}$ is the Wald departure. The distribution function evaluated at the observed value s^0 is the p-value or significance function:

$$p(\varphi) = H\bigl(s^0;\varphi\bigr) \doteq \int_{-\infty}^{r^0}\frac{e^{k/n}}{(2\pi)^{1/2}}e^{-r^2/2}\frac{r}{q}\,\mathrm{d}r \doteq \Phi\left(r^0 - \frac{1}{r^0}\log\frac{r^0}{q^0}\right) = \Phi\bigl(r^*\bigr). \tag{4}$$

This records the percentile position of s^0 relative to a parameter value φ. The approximation in (4) has relative error $O(n^{-3/2})$ when the distribution of s is continuous, and $O(n^{-1})$ when it is discrete [Barndorff-Nielsen (1991)].

4.4. *Scalar parameter of interest.* When φ is a vector, and we have a scalar parameter of interest $\psi(\varphi)$, we need a distribution for a scalar variable that measures ψ. For this we are led to the marginal density of an ancillary recorded as a function of s on $L_\psi^0 = \{s : \hat{\lambda}_\psi = \hat{\lambda}_\psi^0\}$, where $\lambda = \lambda(\varphi)$ is a complementing nuisance parameter. This integral on the ancillary contour can be approximated by Laplace's method, leading to

$$h(s;\psi) \doteq \frac{e^{k/n}}{(2\pi)^{1/2}} e^{-r_\psi^2/2} |\hat{j}|^{-1/2} \big|_{J(\lambda\lambda)}(\tilde{\varphi})\big|^{1/2}, \qquad s \in L_\psi^0, \tag{5}$$

where $\tilde{\varphi} = \hat{\varphi}_\psi$ is the constrained maximum likelihood estimate, $r_\psi = \operatorname{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\varphi}) - \ell(\tilde{\varphi})\}]^{1/2}$ and $|_{J(\lambda\lambda)}(\tilde{\varphi})| = |_{J\lambda\lambda}(\tilde{\varphi})||\tilde{\varphi}_\lambda^{\mathrm{T}}\tilde{\varphi}_\lambda|$ is the nuisance information for λ rescaled to φ; see for example Fraser, Reid and Wu (1999). The integration is along an ancillary contour for given ψ, and the adjustment factor involving the nuisance information determinants is independent of the contour. Essentially the nuisance parameter is eliminated by marginalizing over the distribution that describes that parameter, for a fixed value of ψ [Fraser (2016)].

The p-value function $p(\psi)$ for the scalar $\psi(\varphi)$ is the observed distribution function from (5). This is available to third order by using the step from (3) to (4) but replacing q by an adjusted Wald departure

$$Q = \operatorname{sign}(\hat{\psi} - \psi)|\hat{\chi} - \tilde{\chi}|\{|\tilde{j}_{(\lambda\lambda)}|/|\hat{j}_{\varphi\varphi}|\}^{-1/2}, \tag{6}$$

where $\chi = \chi(\varphi)$ is a local linear approximation to the parameter of interest on the φ scale [Brazzale, Davison and Reid (2007), Chapter 8.5]. The associated significance function is

$$p(\psi) = H^0(s;\psi) \doteq \Phi\left(r - \frac{1}{r}\log\frac{r}{Q}\right) = \Phi(r^*), \tag{7}$$

where we have integrated along the line L_ψ^0 described above. As with (4), the approximation has relative error $O(n^{-3/2})$ in continuous models and $O(n^{-1})$ in discrete models.

The integration from (3) to (4), or from (5) to (7), relies on an asymptotic expansion of r in terms of q, or Q, of the form $r = q + aq/n^{1/2} + bq^2/n + O(n^{-3/2})$, and in particular requires that r is a monotone function of q or Q, at least to $O(n^{-1/2})$.

The density (5) also gives the marginal log-likelihood function for ψ provided we use a reparametrization for which the observed Fisher information is rotationally symmetric, in order to have a common reference distribution on different L_ψ^0 lines. This marginal log-likelihood function is

$$\ell_m(\psi) = \ell(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|\tilde{j}_{(\lambda\lambda)}|, \tag{8}$$

assuming that the observed Fisher information $\hat{j}_{\varphi\varphi} = I$; see Fraser (2003).

4.5. *Vector parameter of interest*. The density approximation (5) applies for scalar or vector parameters of interest, but the p-value calculation (7) requires a one-dimensional integration. Davison et al. (2014) and Fraser, Reid and Sartori (2016) show how the exponential family model can be used to construct a directional test and a directional p-value.

5. Cox's challenge questions.

5.1. *A*, *B*: *Ratio of normal means*. This is an abstraction of a problem treated in Bliss (1935a, 1935b) in connection with probit modelling of biological assays. The dose at which a 50% response is expected, the ED50, is estimated by a ratio of independent normal variables: in Bliss's case the estimates of the intercept and slope of the regression. Fieller (1954) noted that the confidence region for the ED50 given in Bliss (1935b) could apply more generally, described this as the fiducial distribution and noted that the same distribution arose by inverting a pivotal quantity. Fieller (1954) and Bliss (1935a, 1935b) assumed the variances were estimated from the data, and used the t distribution as the reference, but in the present example the variances are known, and the relevant pivotal quantity is $w = (y_2 - \psi y_1)/(1 + \psi^2)^{1/2}$, which has a standard normal distribution. In personal communication David Cox wrote that it was Fisher who gave the solution in Bliss (1935b); in the acknowledgments Bliss writes: "I am indebted especially to Prof. R.A. Fisher, without whose help it could not have been written". Fieller (1954) emphasized that pivotal inversion of w in the usual way could result in a confidence interval that was the whole real line, or confidence sets of the form $(-\infty, a] \cup (b, \infty)$, which he called exclusive. Example A leads to an exclusive set.

The pivotal quantity w has an unusual feature: it measures the departure from ψ in an increasing direction if $y_1 > 0$ and in a decreasing direction if $y_1 < 0$, and these contradictory measures of information are averaged in using the normal distribution for w. Thus the fiducial or confidence distribution $\Phi(w)$ is not monotone in $\psi \in \mathbb{R}$; see for example Schweder and Hjort (2016), Section 4.6. As David Cox has pointed out in personal communication, if the observed value of y_1, and hence the maximum likelihood estimate of μ_1 is close to zero, then under the model there is appreciable chance that it is less than zero, and the sign of ψ is not well determined.

Asymptotic calculations are not needed here, but the requirement of monotonicity in the theory, combined with Figure 1, suggests a reparametrization to the angle $\alpha = \tan^{-1}(\psi)$; the radial distance ρ gives a convenient complementary nuisance parameter. Parametrizing the model with the angle α as the parameter of interest provides a way to enforce the continuity and monotonicity required by the theory outlined in Section 4, at the expense of changing the parametrization. Although in principle α takes values on the circle in $\mathbb{R}^2$, the meaningful range for α is $\hat{\alpha} \pm \pi/2$, as outside this range we again lose monotonicity and continuity; see Figures 2

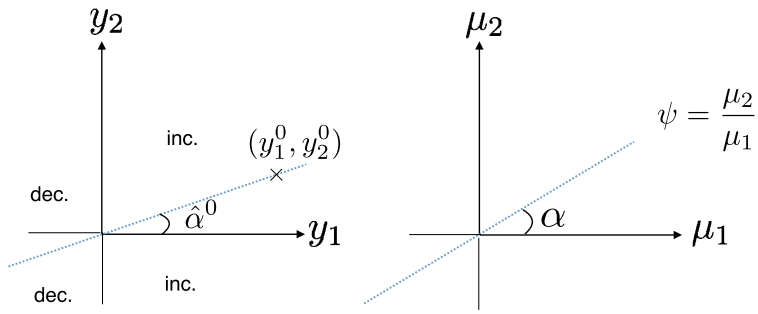

FIG. 1. *The dotted lines correspond to points with a given slope ψ. On the sample space (left), w measures departure relative to increasing ψ when $y_1 > 0$ and decreasing ψ otherwise. The parameter space (right), shows that fixing the parameter of interest together with a direction corresponding to increasing parameter value leads naturally to the angle α as an appropriate parametrization.*

and 3. A referee has noted that the parameter space is data-dependent, which is unusual, but seems unavoidable here. The line L_α^0 in the sample space, where the nuisance parameter is fixed at its constrained estimate, goes through the observed value (y_1^0, y_2^0) and is in the direction $\alpha + \pi/2$ corresponding to increasing α on the sample space. The distribution on this line is standard normal.

In Figure 2 we centre the graph at the maximum likelihood value; the pivotal statistic is $w = -y_1 \sin\alpha + y_2 \cos\alpha$, the p-value function for α is $\Phi(w)$ and the likelihood function is $\exp(-w^2/2)$. This pivotal is algebraically the same as the Fieller pivotal, but in effect is used conditionally and handles the directional effect of ψ on the data point.

For the data of example A the p-value function is monotone in α, has range $(0, 1)$ and provides a set of nested confidence intervals for α in the usual way. For the data of example B both y_1 and y_2 are close to 0, and the p-value function and the likelihood function take values over a much smaller interval; see Figure 3.

The intervals for α obtained by this route can if desired be converted back to the scale of the parameter of interest ψ, and the plot of the significance function on this scale is consistent with the solution using the pivotal w, but the evident discontinuity in the plot draws attention to the unusual nature of the inference problem. In example A, the upper limit for α transforms to a negative limit for ψ, and vice versa.

This example suggests that any theory of distributions for parameters needs to pay close attention to the parametrization of the model. Schweder and Hjort [(2016), Section 4.6] use confidence densities that are not required to be monotone, so that exclusive intervals, for example, are included in the approach. Xie and Singh (2013) treat this example in an independent sampling context by finding the approximate distribution of $\bar{x}_1/\bar{x}_2$, in their notation, but don't address the difficulty highlighted by this example.

A Bayesian approach to this problem can avoid "difficult" confidence sets by using informative prior distributions that downweight values of μ_1 near zero, but

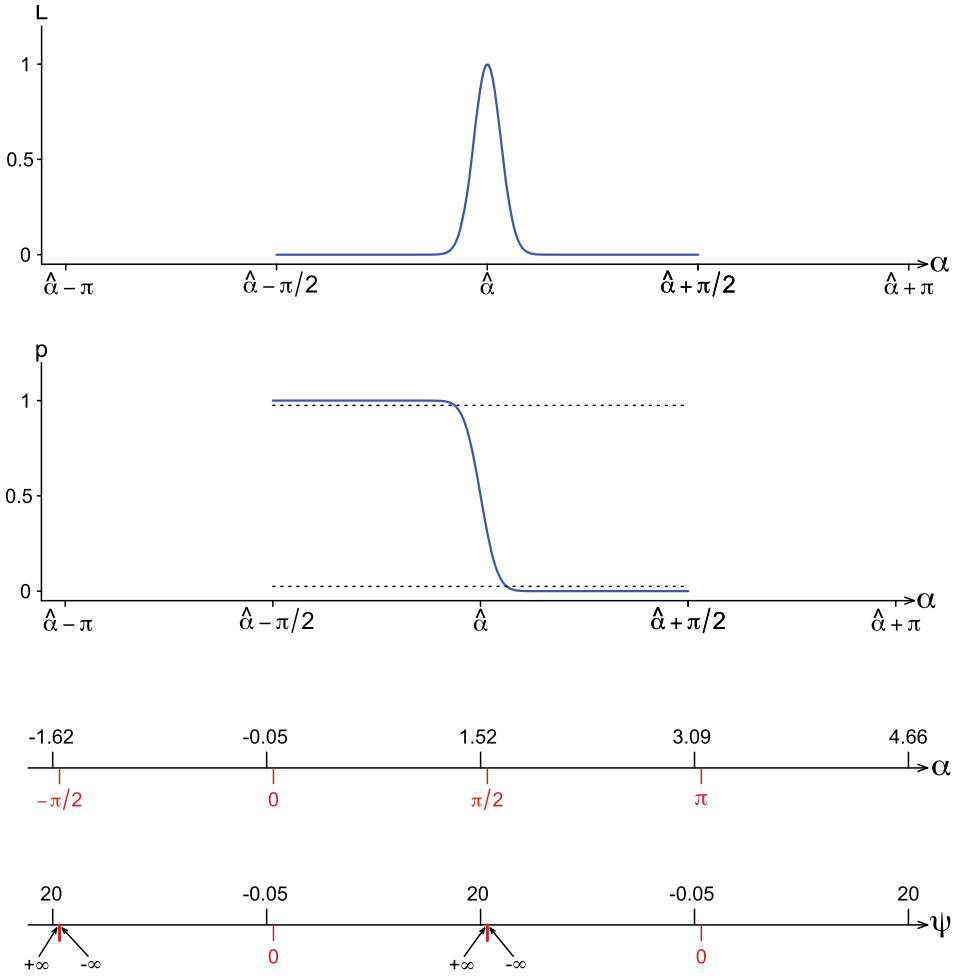

FIG. 2. *The significance function and likelihood function for example, A:* $y_1 = 0.5$, $y_2 = 10$. *The maximum likelihood estimate is* $\hat{\alpha} = 1.52$ *and we show both the* α *and* ψ *scale for the parameter. There are two discontinuities in* ψ *shown, at* $\alpha = \pm\pi/2$, *where* y_1 *changes from positive to negative. The discontinuity at* $\pi/2$ *leads to what Fieller* (1954) *called exclusive confidence regions for* ψ *at confidence level* 0.95. *Horizontal dotted lines are drawn at* $p = 0.975$ *and* $p = 0.025$.

this seems an artificial solution. Ghosh (2011) discusses probability matching priors for this model; as with the approximation of Xie and Singh (2013) this requires an asymptotic setting for validity, and does not draw attention to the need for monotonicity or to the discontinuity noted above.

Estimation of a ratio arises in more complex examples of regression calibration, as well as other contexts; for example, the estimation of density ratios is discussed in the context of the analysis of photometric and spectroscopic data in astronomy in Izbicki, Lee and Freeman (2017), Section 4.

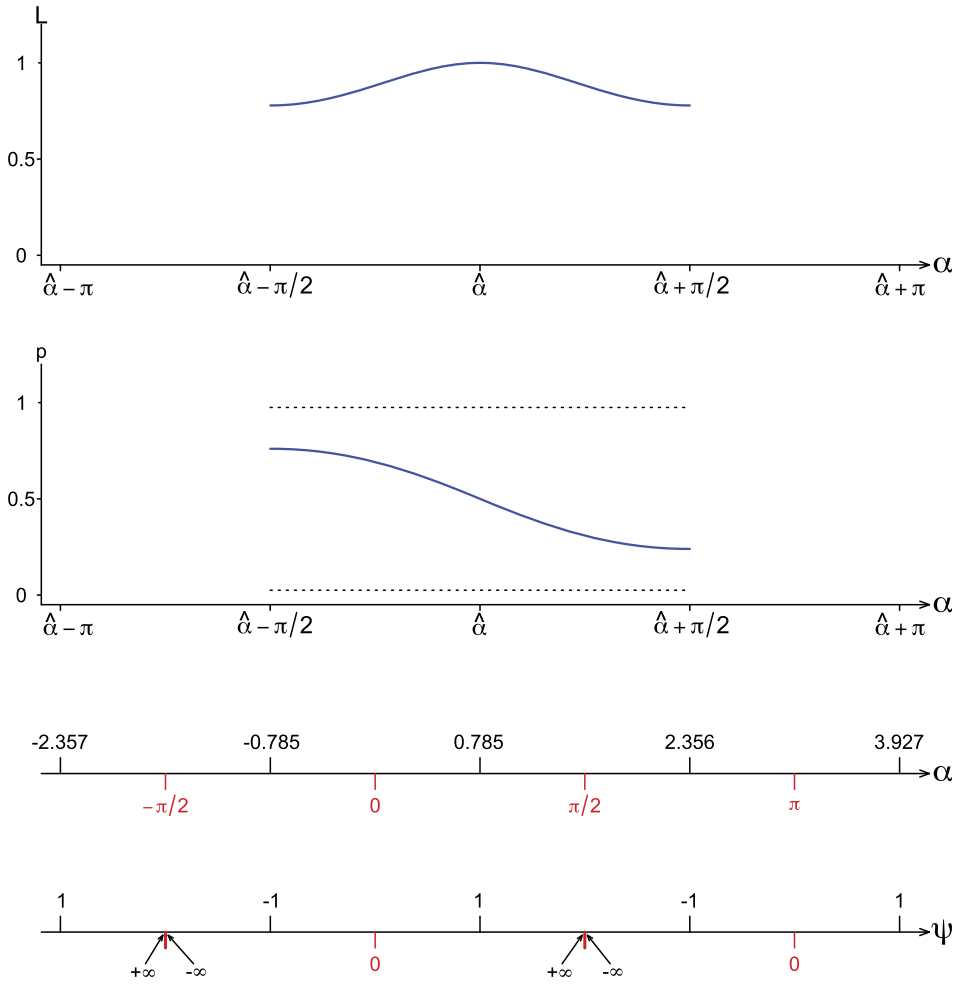

FIG. 3. *The significance function and likelihood function for example, B,* $(y_1 = 0.5, y_2 = 0.5)$. *The maximum likelihood estimate is* $\hat{\alpha} = 0.785$ *and we show both the* α *and* ψ *scale for the parameter. There are discontinuities in* ψ, *at* $\alpha = \pm\pi/2$, *where* y_1 *changes from positive to negative. Confidence intervals for* α *are only available for a restricted range of confidence levels,* $(0.24, 0.76)$. *For other confidence levels all values of* α *and all values of* ψ *are consistent with the data. Horizontal dotted lines are drawn at* $p = 0.975$ *and* $p = 0.025$.

5.2. *C*: *Prior-data conflict*. If we assume that the normal distribution for θ is a genuine prior in the sense of Efron (2013), for example, based on prior observation of a random process, then the prior is more accurately described as part of the model, and standard probability calculus enables derivation of the conditional distribution for θ given $y = y^0$, which is $N(y^0/2, 1/2)$. From Figure 4 however we see that this distribution is neither compatible with the initial distribution for θ nor with the observed value from the model for y, and it seems appropriate to report

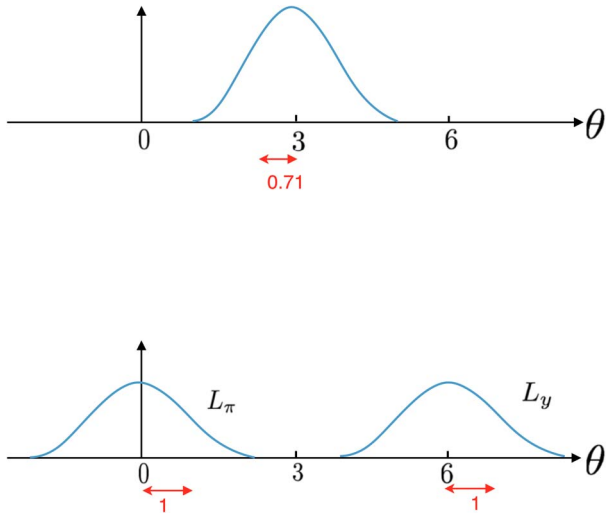

FIG. 4. *Conditional likelihood function for* θ, *given* $y = 6$ (*top*). *Individual likelihood functions* (*bottom*) *show the incompatibility of this with both the prior and the model.*

the individual density functions shown there. The separation of the two models suggests that one or other assumption is incorrect. From a practical viewpoint, this seems uncontroversial.

For a more formal approach, Box (1980) suggested using the marginal distribution for y which is $N(0, 1.41)$; an observation of 6 conveys the inadequacy of the modelling. The marginal distribution for y is also called the prior predictive distribution; the posterior predictive distribution or some modification of it is often recommended. Two recent examples implementing posterior predictive checks are Simoiu, Corbett-Davies and Goel (2017) and Hartmann et al. (2017). A formal theory for assessing conflict between the prior and the data is summarized in Evans (2015), Chapter 5. In the application studied in Keele and Quinn (2017), an informative prior for the parameter of interest was required, and sensitivity of the inference to that prior is carefully considered.

A referee questioned what the fiducial approach might suggest in this example. This question highlights one difficulty, that "the fiducial approach" is not very well defined. If the information about θ is ignored, then the fiducial distribution for θ is $N(y^0, 1)$. If both the $N(0, 1)$ model for θ and the $N(\theta, 1)$ model for y are considered to be on equal footing, but with θ unobserved, then there seems to be no basis for the construction of a fiducial density.

In this example the theory of Section 4 is not directly helpful, because all the calculations outlined there are valid under the assumption that the model for y, given θ is correct. However the general principle of using diagnostic plots or calculations to check the assumptions continues to apply, no matter how the inference is conducted. In a setting where there is an asymptotic theory in n, one might expect that

a large discrepancy between the conventional normal approximation and a higher order approximation casts doubt on the validity of the model. We are not aware however of any work that has succeeded in demonstrating a general result along these lines. Typically even the first order approximation is dependent on the correctness of the model, although exceptions to this are discussed in Ogden (2017), where conditions on approximation likelihoods are given that ensure standard first order theory continues to be valid.

5.3. *D*: *Neyman–Scott problem.* This is an exponential model, with n nuisance parameters and a scalar parameter of interest, $\psi = \sigma^2$. The canonical parameter is

$$\varphi = \left(\frac{\mu_1}{\sigma^2}, \ldots, \frac{\mu_n}{\sigma^2}, \frac{1}{\sigma^2}\right),$$

and the canonical variable is

$$\left\{\bar{y}_1, \ldots, \bar{y}_n, \sum_{i=1}^{n}(y_{i1}^2 + y_{i2}^2)\right\}.$$

The saddlepoint density approximation (5) leads to an expression equivalent to deriving inference from the marginal density of $s^2 = \Sigma_i(y_{i1} - y_{i2})^2$, which is $2\sigma^2\chi_n^2$. The method of marginalizing along an ancillary contour for fixed σ sketched in Section 4.3 is equivalent in this example to eliminating the nuisance parameters μ_i by marginalizing over the distribution of $\bar{y}_i$. The saddlepoint approximation is not needed as the exact distribution for s^2 is readily available, but if used, the information correction term in (5) gives the correct degrees of freedom adjustment for inference about σ^2 [Cox and Reid (1987)]. Use of the usual likelihood methods based on the profile log-likelihood function in this example leads to an inconsistent estimator of σ^2. Although in principle the higher order theory does not apply to the setting with increasing numbers of nuisance parameters, the methodology that it provides for eliminating nuisance parameters does lead to the accepted solution. It is a feature of the approximations that the elimination of nuisance parameters in examples where there is an exact conditional or marginal density for the parameter of interest is a close approximation to this exact density, even with increasing numbers of parameters; see for example Sartori (2003).

This example is also discussed in Schweder and Hjort [(2016), Section 4.4], where the pivotal quantity $\Sigma_i(y_{i1} - y_{i2})^2/(2\sigma^2)$ is deduced from the form of the profile log-likelihood function. With this choice of pivotal quantity the confidence density is the same as the fiducial density. As noted above any joint fiducial distribution for the vector parameter $(\mu_1, \ldots, \mu_n, \sigma)$ cannot be marginalized to a valid marginal fiducial distribution. A historical survey of this discussion is provided in Schweder and Hjort (2016), Chapter 6. A Bayesian approach to this example is discussed in Ghosh [(2011), Example 3], where it is shown that the improper prior for location-scale models $\mathrm{d}\mu_1 \cdots \mathrm{d}\mu_n \, \mathrm{d}\sigma/\sigma$ leads again to the same solution

based on the χ^2_n distribution, but the posterior based on Jeffreys' prior leads to inconsistent inference for σ^2. Ghosh (2011) motivates the location-scale prior as a two-group reference prior.

5.4. *E*: *Curved parameter of interest*. This example is due to Stein (1959), and is a simple illustration that using a flat prior for μ gives a posterior marginal distribution for the parameter of interest $\psi = \|\mu\|^2$ that is far from calibrated; see for example, Cox and Hinkley (1974), page 383. It has also been discussed as an example where marginalizing a joint fiducial density fails.

The model is an exponential family with canonical parameter μ and variable y. Following the development of Section 4.4, the density measuring ψ is on the line in the sample space where the nuisance parameter estimate is constant, $\hat{\lambda}_\psi = \hat{\lambda}^0_\psi$. The easiest version of the nuisance parameter for the calculations is $\lambda = \mu/\|\mu\|$, so that $\mu_i = \psi^{1/2}\lambda_i$, $\hat{\lambda}_{i,\psi} = \psi^{1/2} y_i/\|y\|$, and the line L^0_ψ goes through y^0 and $\psi^{1/2} y^0/\|y^0\|$. Integration to this line is marginalization over the sphere through y^0 with radius ψ. The inference summaries $p(\psi)$ and $\ell(\psi)$ are then given by (7) and (8). In a slightly different setting, where $y_i \sim N(0, 1/n)$, $i = 1, \ldots, k$, Reid and Sun (2010) show that the normal approximation to r^*_ψ is very close to the exact distribution of $n\|y\|^2$, which is a noncentral χ^2_k distribution with noncentrality parameter $n\psi^2$, even for very small values of n and large values of k. In this version of the problem this exact distribution can be recovered in a Bayesian argument by using a matching prior for ψ; the prior $\pi(\mu) \propto \|\mu\|^{-(k-1)}$ is both a probability matching prior and a reference prior [Datta and Ghosh (1995)]. This prior will not be a matching prior for any other function of μ; it is necessary to target the prior on the parameter of interest.

Nearly all applications of Bayesian models in complex applied settings involve marginalization of a joint posterior to construct inference about one or more parameters of interest. While the Stein example may be an extreme case, it seems plausible that some version of this could apply in other settings, unless the prior is effectively swamped by the data. Tak et al. (2017) explicitly check the frequentist coverage of their posterior intervals in simulations, but this does seem to be the exception.

5.5. *F*: *Cosine regression*. This model has a scalar parameter and an n-dimensional variable y and the log-likelihood function can be multi-modal. To reduce from the dimension of the data to that of the parameter using the approximate conditioning described in Section 4.1, we compute the inference direction v,

$$\begin{aligned} v &= \left\{ \frac{\partial}{\partial\theta}\cos(\theta x_1), \ldots, \frac{\partial}{\partial\theta}\cos(\theta x_n) \right\}^{\mathrm{T}} \Big|_{(y^0, \hat{\theta}^0)} \\ &= -\{x_1 \sin(\hat{\theta}^0 x_1), \ldots, x_n \sin(\hat{\theta}^0 x_n)\}^{\mathrm{T}}. \end{aligned}$$

The tangent exponential model is normal with variance one and mean

$$\mu(\theta) = \big(v^{\mathrm{T}}/\|v\|\big)\big\{\cos(\theta x_1), \ldots, \cos(\theta x_n)\big\}.$$

This has likelihood function $L(\mu) = \phi\{(v^{\mathrm{T}}/\|v\|)y - \mu\}$ and significance function

$$p(\mu) = \Phi\big\{\big(v^{\mathrm{T}}/\|v\|\big)y - \mu\big\};$$

μ is the canonical parameter for this exponential family.

An example of a log-likelihood function with equally spaced values of x on $(0, 1)$ is illustrated in Figure 5 (left). Removing the interval constraint, here by letting $x_i = i$, gives an oscillating log-likelihood function supporting disjoint intervals for values of θ, as described in Cox and Hinkley (1974), Chapter 7.5. Cox [(2006), Example 7.4] considers a related time-series regression problem, showing that the usual asymptotic theory for likelihood inference does not apply to the estimation of θ in the context of Figure 5 (right). This example has some similarities to the ratio problem of examples A and B in that inference for the given

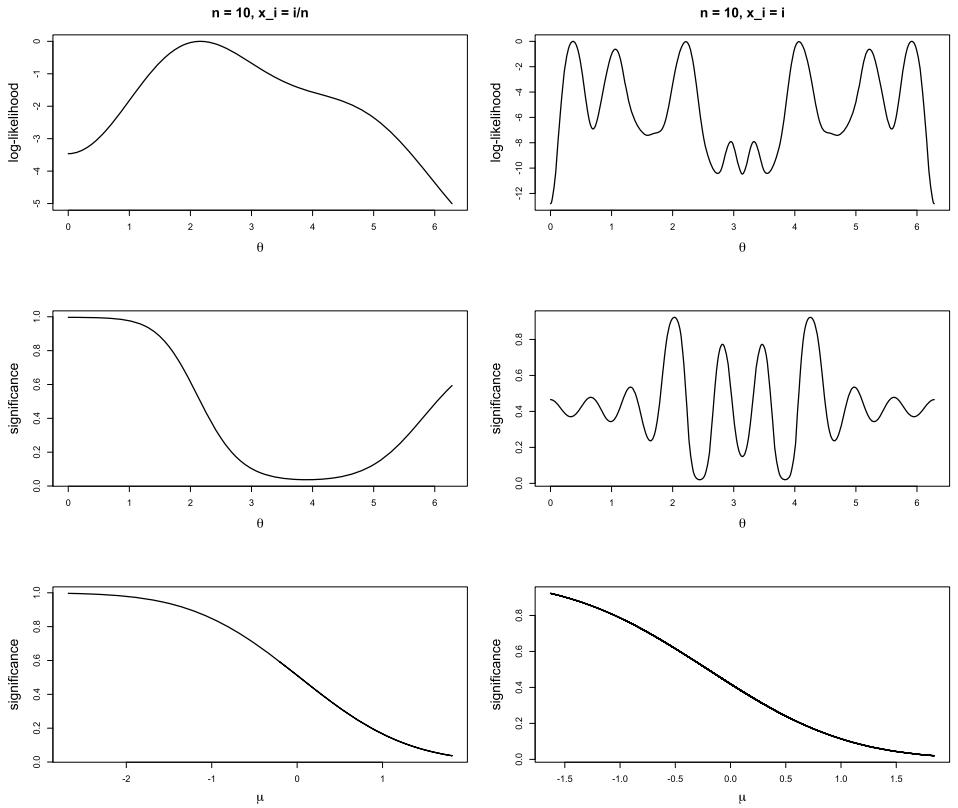

FIG. 5. *Cosine regression* (*example F*). *Plots of a log-likelihood function* (*top*) *and significance function against* θ (*middle*) *and against* μ (*bottom*) *for* $n = 10$, $\theta_0 = 1$, *with* $x_i = i/n$ (*left*); $x_i = i$ (*right*).

parameter of interest does not lead to a series of nested confidence intervals, but there is a parametrization in which this can be attained. However in this case the parametrization seems more difficult to interpret.

In a Bayesian approach a $U(0, 2\pi)$ prior on θ may well be viewed as non-informative; under this prior the posterior distribution and the log-likelihood function have the same behaviour, and the construction of posterior quantiles for θ would lead to the same anomalies as standard likelihood inference. An informative prior for θ could in principle smooth out the oscillations in the likelihood, at the expense of injecting information into the solution. To construct a confidence distribution, or a fiducial distribution, requires specification of a pivotal quantity by some means; the most direct pivotal quantity is the residual sum of squares, but this is equivalent to the log-likelihood function and will lead to the same difficulties. A very similar log-likelihood function arises in a much more complex model described in Tak et al. (2017), where the profile log-likelihood is used for the parameter of interest, which is a time delay in a trigonometric regression.

5.6. *G*: *Random numbers*. The decimal digits are $(x_1, \ldots, x_n)$ and we define $y_0, \ldots, y_9$ to be the number of $0's, \ldots, 9's$. The hypothesis to be assessed is that the marginal distribution of the $y_0, \ldots, y_9$ is consistent with the digit vector $(x_1, \ldots, x_n)$ having a uniform distribution on $S = \{0, 1, \ldots, 9\}^n$.

Under the uniform distribution the vector $y = (y_0, \ldots, y_9)$ has a multinomial distribution $(n; 1/10, \ldots, 1/10)$ with expected value $(4.5, \ldots, 4.5)$. We embed this in an exponential model based on the Poisson distribution with canonical parameter $\varphi = (\log p_0, \ldots, \log p_9)$; conditioning on $\Sigma y_i = n$ recovers the multinomial distribution. The parameter of interest in this model is a vector, so we consider a directional test as developed in Davison et al. (2014) and Fraser, Reid and Sartori (2016).

The starting point is the saddlepoint approximation of (5)

$$h(s; \varphi)\,\mathrm{d}s \doteq \frac{e^{k/n}}{(2\pi)^{10/2}} e^{\ell - \hat{\ell}} |\hat{\jmath}|^{-1/2}\,\mathrm{d}s,$$

with $\ell(\varphi; y) = \{\varphi_1 y_0 - \log\varphi_0, \ldots, \varphi_9 y_9 - \log\varphi_9\}$; the domain for the model is $S^* = (-0.5, 9.5)^{10} \cap \{\sum y_i = n\}$. As in Davison et al. (2014) we measure the discrepancy of y from its expected value conditional on the direction of departure. We compute the p-value using the squared length $s^2 = \sum(y_i - 4.5)^2$ and integrate from the origin along the vector $(y_0 - 4.5, \ldots, y_9 - 4.5)^{\mathrm{T}}$. The boundary point is where some coordinate attains the maximum possible absolute value 4.5. Denoting the squared length of this boundary point by $(s^b)^2$ the directional p-value is

$$p^0 = \int_0^{s^0} s^9 e^{\ell - \hat{\ell}} |\hat{\jmath}_{\varphi\varphi}|^{-1/2}\,\mathrm{d}s \Big/ \int_0^{s^b} s^9 e^{\ell - \hat{\ell}} |\hat{\jmath}_{\varphi\varphi}|^{-1/2}\,\mathrm{d}s.$$

Further calculation shows that the p-value is the same as that obtained using the chi-squared test of fit of the uniform multinomial distribution. Although it is based

on the squared length of the vector, any monotone departure measure in the same vector direction will lead to the same p-value.

Specific types of departure from the uniform are not addressed by the directional approach. It is likely that careful construction of pivotal quantities tailored to a particular departure could be used for the development of a confidence distribution or a fiducial probability, but there seems to be no driving principle that could guide this.

5.7. *H*: *The role of physical randomization*. We interpret physical randomization as referring to the random allocation of treatments to experimental units in a designed experiment, or to the selection of units for measurement in a sample survey. Randomization introduces crucial symmetries that may be exploited in a parametric model, strengthening the model assumptions, but the model-based approach discussed here is different from the randomization analysis of a designed experiment or a sample survey. In simple cases where there is an identified parameter, for example an additive treatment effect, it may be possible to construct a set of confidence bounds from the randomization distribution, as suggested for example, in Schweder and Hjort [(2016), Section 11.6], and this can provide a confidence distribution for that parameter.

6. Discussion. As David Cox has stated, the questions he posed are simplified in order to bring out some essential issues, and the issue is not so much whether an answer can be obtained, as whether theories of inference can address these issues. We have emphasized the approach to significance or p-value functions developed from higher order approximations, as this seems to give a principled approach to both to the elimination of nuisance parameters, and to determining a quantity, r_ψ^*, which is pivotal to a high order of approximation and can be used directly to assess individual parameters in light of the data.

Several examples closer to the practice of statistics are discussed in Brazzale, Davison and Reid (2007) and Fraser, Wong and Sun (2009). As one example, the latter paper treats the transformed regression model of Box and Cox (1964), with independent observations $y_i^\lambda = \alpha + \beta x_i + \sigma z_i$, where z_i is assumed to follow a $N(0, 1)$ distribution. The inference directions form an $n \times 4$ matrix with ith row

$$V_i = \left(\frac{\partial}{\partial\theta}\right)(\alpha + \beta x_i + \sigma z_i)^{1/\lambda}\Big|_{(y,\hat\theta)} = \frac{y_i}{\hat\lambda y_i^{\hat\lambda}}\big(1, x_i, \hat z_i, -y_i^{\hat\lambda}\log y_i\big), \tag{9}$$

where $\theta = (\alpha, \beta, \sigma, \lambda)$ and $\hat z_i = \hat\sigma^{-1}(y_i^{\hat\lambda} - \hat\alpha - \hat\beta x_i)$. Differentiating the log-likelihood in the directions V gives the canonical parameter

$$\varphi^{\mathrm{T}} = \Sigma_i\big\{-\lambda y_i^{\lambda-1}\big(y_i^\lambda - \alpha - \beta x_i\big)/\sigma^2 + (\lambda - 1)y_i^{-1}\big\}V_i. \tag{10}$$

Numerical work presented in Fraser, Wong and Sun [(2009), Section 6] shows that the p-value function computed using (7) is very accurate for various parameters in the model.

The development of the approximate p-value function, and particularly the requirement of monotonicity, highlights the role of the parametrization of the model. Examples A, B and F feature non-monotonicity that raises issues of interpretation. The importance of a theory concerning parametrization of statistical models has been emphasized by McCullagh (2002), but there seems to be little else on this point in the literature. Perhaps the theory of significance functions can be used as a diagnostic to alert the user to potential problems with a given model parametrization.

The examples D and E involving several nuisance parameters highlight the failure of likelihood or Bayesian methods that are not specifically targeted on the parameter of interest. In these problems the asymptotic theory outlined here seems satisfactory; it effectively eliminates the nuisance parameters as part of the development of a measure for the parameter of interest, and for example E, shows what the Bayesian prior will need to be in order that the resulting inference be calibrated in a frequency sense. The necessity for targeting parameters of interest in the development of noninformative, or reference, priors, is well-known in the theoretical literature but seems to have had less impact on applications.

Asymptotic theory is often argued to be of limited importance for practical applications either because the sample size is sufficiently large that refined approximations are not seen to be necessary, or because the models are sufficiently tentative that achieving very precise inference under the model is not very important for the application at hand. For the former, it does often seem to be the case that the number of parameters, or the model complexity, increases with the sample size, and examples D and E are reminders of the difficulties that this can create. When there are several parameters of interest the construction of confidence distributions or significance functions treats each parameter separately. In some settings the vector approach outlined in Section 4.5 and illustrated in example G may be useful. Posterior inference for each parameter of interest in turn is readily implemented in a Bayesian approach, but will only be calibrated if the prior is adapted for each calculation [Fraser et al. (2016)].

If the schools of thought are indeed to be best friends forever they can of course reasonably share a commitment to finding ultimate resolutions. In a Bayesian approach this seems to require a commitment to ensuring, or assessing, calibration of the resulting inference, at least to some order of approximation. The confidence distribution approach takes as its starting point essentially a nested set of upper or lower confidence bounds, but is somewhat agnostic about what method is used for this. Our approach to p-value or significance functions emphasizes the role of higher order asymptotics in determining this starting point. In very complex models for applications of the type that feature in this journal, it must be said that a Bayesian approach often seems conceptually simpler, if computationally demanding. The simple but challenging examples serve as a useful caution to reliance on simple solutions.

Acknowledgments. We are grateful to David Cox for his challenge questions and for his many thoughtful comments on draft versions of this manuscript. We thank Ruobin Gong, Xiao-Li Meng, and Madeleine Straubel for their organization of the Fourth BFF conference. We also thank the reviewers for their prompt and helpful comments.

REFERENCES

BARNDORFF-NIELSEN, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78** 557–563. MR1130923

BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* **53** 370–418.

BERGER, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402. MR2221271

BLISS, C. (1935a). The calculation of dosage-mortality curves. *Ann. Appl. Biol.* **22** 134–167.

BLISS, C. (1935b). The comparison of dosage-mortality data. *Ann. Appl. Biol.* **22** 307–333.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BRAZZALE, A. R., DAVISON, A. C. and REID, N. (2007). *Applied Asymptotics*: *Case Studies in Small-Sample Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2342742

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Stat.* **29** 357–372. MR0094890

COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge Univ. Press, Cambridge. MR2278763

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. MR0893334

DATTA, G. S. and GHOSH, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90** 1357–1363. MR1379478

DATTA, G. S. and MUKERJEE, R. (2004). *Probability Matching Priors*: *Higher Order Asymptotics*. *Lecture Notes in Statistics* **178**. Springer, New York. MR2053794

DAVISON, A. C., FRASER, D. A. S., REID, N. and SARTORI, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *J. Amer. Statist. Assoc.* **109** 302–314. MR3180565

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. MR1225211

EFRON, B. (2013). Bayes' theorem in the 21st century. *Science* **340** 1177–1178. MR3087705

EVANS, M. (2015). *Measuring Statistical Evidence Using Relative Belief*. *Monographs on Statistics and Applied Probability* **144**. CRC Press, Boca Raton, FL. MR3616661

FIELLER, E. C. (1954). Symposium on interval estimation: Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B* **16** 175–185. MR0093076

FIENBERG, S. E. (2006). Does it make sense to be an "objective Bayesian"? (comment on articles by Berger and by Goldstein). *Bayesian Anal.* **1** 429–432. MR2221275

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

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

FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280. MR0133910
FRASER, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53** 1–9. MR0196840
FRASER, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77** 65–76. MR1049409
FRASER, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90** 327–339. MR1986650
FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001
FRASER, D. A. S. (2016). The p-value function: The core concept of modern statistical inference. *Ann. Rev. Stat. Appl.* **4** 1–14.
FRASER, D. A. S. and REID, N. (1995). Ancillaries and third order significance. *Util. Math.* **47** 33–53. MR1330888
FRASER, D. A. S., REID, N. and SARTORI, N. (2016). Accurate directional inference for vector parameters. *Biometrika* **103** 625–639. MR3551788
FRASER, D. A. S., REID, N. and WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264. MR1705367
FRASER, D. A. S., WONG, A. and SUN, Y. (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.* **37** 161–181. MR2531825
FRASER, D. A. S., REID, N., MARRAS, E. and YI, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 631–654. MR2758239
FRASER, D. A. S., BÉDARD, M., WONG, A., LIN, W. and FRASER, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.* **31** 578–590. MR3598740
GHOSH, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.* **26** 187–202. MR2858380
GOLDSTEIN, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Anal.* **1** 403–420. MR2221272
HANNIG, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19** 491–544. MR2514173
HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954
HARTMANN, M., HOSACK, G. R., HILLARY, R. M. and VANHATALO, J. (2017). Gaussian process framework for temporal dependence and discrepancy functions in Ricker-type population growth models. *Ann. Appl. Stat.* **11** 1375–1402. MR3709563
IZBICKI, R., LEE, A. B. and FREEMAN, P. E. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *Ann. Appl. Stat.* **11** 698–724. MR3693543
KEELE, L. and QUINN, K. M. (2017). Bayesian sensitivity analysis for causal effects from 2×2 tables in the presence of unmeasured confounding with application to presidential campaign visits. *Ann. Appl. Stat.* **11** 1974–1997. MR3743285
LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.
MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320
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.
NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. A* **231** 289–337.
OGDEN, H. E. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika* **104** 153–164. MR3626485
PIERCE, D. A. and BELLIO, R. (2017). Modern likelihood-frequentist inference. *Int. Stat. Rev.* **85** 519–541. MR3723615
REID, N. and COX, D. R. (2015). On some principles of statistical inference. *Int. Stat. Rev.* **83** 293–308. MR3377082

REID, N. and SUN, Y. (2010). Assessing sensitivity to priors using higher order approximations. *Comm. Statist. Theory Methods* **39** 1373–1386. MR2753513

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SARTORI, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90** 533–549. MR2006833

SCHWEDER, T. and HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics* **41**. Cambridge Univ. Press, New York. MR3558738

SIMOIU, C., CORBETT-DAVIES, S. and GOEL, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* **11** 1193–1216. MR3709557

STEIN, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Stat.* **30** 877–880. MR0125680

TAK, H., MANDEL, K., VAN DYK, D. A., KASHYAP, V. L., MENG, X.-L. and SIEMIGINOWSKA, A. (2017). Bayesian estimates of astronomical time delays between gravitationally lensed stochastic light curves. *Ann. Appl. Stat.* **11** 1309–1348. MR3709561

WASSERMAN, L. (2006). Frequentist Bayes is objective (comment on articles by Berger and by Goldstein). *Bayesian Anal.* **1** 451–456. MR2221280

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496

D. A. S. FRASER
N. REID
DEPARTMENT OF STATISTICAL SCIENCES
UNIVERSITY OF TORONTO
TORONTO, ONTARIO M5S 3G3
CANADA
E-MAIL: dasfraser@gmail.com
reid@utstat.utoronto.ca

WEI LIN
AIDVOICE LAB
TORONTO, ONTARIO M4Y2W4
CANADA
E-MAIL: wei.lin@mail.utoronto.ca