22

Why does statistics have two theories?

Donald A.S. Fraser
Department of Statistical Sciences
University of Toronto, Toronto, ON

The public image of statistics is changing, and recently the changes have been mostly for the better, as we've all seen. But occasional court cases, a few conspicuous failures, and even appeals to personal feelings suggest that careful thought may be in order. Actually, statistics itself has more than one theory, and these approaches can give contradictory answers, with the discipline largely indifferent. Saying "we are just exploring!" or appealing to mysticism can't really be appropriate, no matter the spin. In this paper for the COPSS 50th Anniversary Volume, I would like to examine three current approaches to central theory. As we will see, if continuity that is present in the model is also required for the methods, then the conflicts and contradictions resolve.

22.1 Introduction

L'Aquila and 300 deaths. The earthquake at L'Aquila, Italy on April 5, 2009 had been preceded by many small shocks, and Italy's Civil Protection Department established a committee of seismologists to address the risks of a major earthquake. The committee reported before the event that there was no particularly good reason to think that a major earthquake was coming and the Department's Deputy Head even allowed that the small shocks were reducing the seismic stresses, lowering the chances of a major quake. This gave some reassurance to many who were concerned for their lives; but the earthquake did come and more than 300 died. For some details, see Pielke (2011). Charges were then brought against the seismologists and seven were sentenced to six years in prison for manslaughter, "for falsely reassuring the inhabitants of L'Aquila". Part of the committee's role had been the communication of their findings, statistics being intrinsically involved. See Marshall (2012) and Prats (2012).

Vioxx and 40,000 deaths. The pain killer Vioxx was approved by the US Food and Drug Administration (FDA) in 1999 after a relatively short eight years in the approval process and then withdrawn by the pharmaceutical company Merck in 2004 after an acknowledged excess of cardiovascular thrombotic (CVT) events under Vioxx in a placebo controlled study. But statistical assessments as early as 2000 had indicated the heightened rate of CVT events with the use of Vioxx. Statistician David Madigan of Columbia University rose to the challenge as litigation consultant against Merck, and a five billion dollar penalty against Merck went to the injured and their survivors; some felt this was a bargain for Merck, as the company had made billions in profit from the drug. One estimate from the FDA of the number of deaths attributed to the use of the drug was 40,000. See Abraham (2009).

Challenger and 7 deaths. The space shuttle Challenger had completed nine successful flights but on its tenth take-off on January 28, 1986 disintegrated within the first two minutes. The failure was attributed to the breakdown of an O-ring on a solid rocket booster. The external temperature before the flight was well below the acknowledged tolerance for the O-rings, but the flight was given the go-ahead. The 7 crew members died. See Dalai and Fowlkes (1989) and Bergin (2007).

The preceding events involve data, data analysis, determinations, predictions, presentations, then catastrophic results. Where does responsibility fall? With the various levels of the application of statistics? Or with the statistical discipline itself with its contradicting theories? Or with the attitude of many statisticians. We are just exploring and believe in the tools we use?

Certainly the discipline of statistics has more than one theory and these can give contradictory results, witness frequency-based, Bayes-based, and bootstrap-based methodology; these provide a wealth of choice among the contraindicating methods. Here I would like to briefly overview the multiple theories with a view to showing that if continuity as present in the typical model is also required for the methods, an equivalence emerges among the frequency, the bootstrap, and partially the Bayesian approach to inference.

But also, there is attitude within the discipline that tolerates the contradictions and indeed affects within-discipline valuations of statistics and statisticians. In recent years, an important Canadian grant adjudication process had mathematicians and statisticians evaluating applications from mathematicians and statisticians using standardized criteria but with a panel from mathematics for the mathematicians and a panel from statistics for the statisticians; and it was found that mathematicians rate mathematicians much higher than statisticians rate statisticians, even though it was clear that benefits would be apportioned accordingly. For details, see Léger (2013). The contradictory theory and the contradictory attitude provide a potential for serious challenges for statistics, hopefully not at the level of L'Aquila, Vioxx and Challenger.

22.2 65 years and what's new

I did my undergraduate studies in mathematics in my home town of Toronto, Ontario. An opportunity to study analysis and algebra in the doctoral program at Princeton University arose in 1947. But then, with a side interest in actuarial things, I soon drifted to the Statistics Group led by Sam Wilks and John Tukey. A prominent theme was Neyman–Pearson theory but a persistent seminar interest focussed on Fisher's writings, particularly those on fiducial inference which had in turn triggered the Neyman (Neyman, 1937) confidence methodology. But also, a paper by Jeffreys (Jeffreys, 1946) kept reemerging in discussions; it offered a default Bayes (Bayes, 1763) approach, often but incorrectly called objective Bayes in present Bayes usage. The striking thing for me at that time was the presence of two theories for statistics that gave contradictory results: If the results were contradictory, then simple logic on theories says that one or the other, or both, are wrong. This latter view, however, was not part of the professional milieu at the time, though there was some puzzlement and vague acceptance of contradictions, as being in the nature of things; and this may even be part of current thinking! "One or the other, or both, could be wrong?" Physics manages to elicit billions in taxpayer money to assess their theories! Where does statistics stand?

With a completed thesis that avoided the frequency-Bayes contradictions, I returned to Canada and accepted a faculty position in the Department of Mathematics at the University of Toronto. The interest in the frequency-Bayes contradictions, however, remained and a conference talk in 1959 and two resulting papers (Fraser, 1961a,b) explored a broad class of statistical models for which the two approaches gave equivalent results: The location model $f(y-\theta)$, of course, and the locally-generated group extensions, the transformation-parameter models. Then an opportunity for a senior faculty position in the Mathematics Department at Princeton arose in 1963, but I was unable to accept. The concerns for the frequency-Bayes contradictions, however, remained!

Now in 2013 with COPSS celebrating its 50th anniversary, we can look about and say "What's new?" And even more we are encouraged to reminisce! There is very active frequency statistics and related data analysis; and there is very active Bayesian statistics; and they still give contradictory answers. So nothing has changed on the frequency-Bayes disconnect: What goes around comes around... Does that apply to statistical theory in the 65 years I have been in the profession? Oh, of course, there have been massive extensions to data exploration, to computer implementation, to simulations, and to algorithmic approaches. Certainly we have Precision, when sought! But what about Accuracy? I mean Accuracy beyond Precision? And what about the frequency-Bayes contradictions in the theory? And even, indeed, the fact that no one seems to care? And then L'Aquila, Vioxx, Challenger, and of course the

contradictory theory? Are perceptions being suppressed? It might wind up in a court, as with L'Aquila, an inappropriate place to address a scientific issue but perhaps not to address a conflict coming from discipline contradictions.

Well yes, something has changed! Now a general feeling in the milieu is acceptance of the frequency-Bayes contradiction: It just doesn't matter, we are just exploring; our models and calculations are just approximations; and we can acquire any Precision we want, even though we may not have used the full information provided by the model, so just run the MCMC longer, even though several million cycles only give two decimal places for some wanted probability or confidence calculation. Or put together an algorithm for processing numbers. Or use personal feelings as in some Bayes methods.

But even for explorations it certainly behooves one to have calibrated tools! And more generally to know with Precision and Accuracy what a model with data implies? Know as a separate issue quite apart from the descriptive Accuracy of the model in a particular context, which of course in itself is an important but separate issue! This Accuracy is rarely addressed! Indeed, as L'Aquila, Vioxx, and Challenger indicate, a concern for Accuracy in the end products of statistics may have an elusive presence in many professional endeavours. An indictment of statistics?

22.3 Where do the probabilities come from?

(i) The starting point. The statistical model $f(y;\theta)$ with data y^0 forms the starting point for the Bayes and often the frequency approach. The Bayesian approach calculates and typically uses just the observed likelihood $L^0(\theta) = f(y^0;\theta)$, omitting other model information as part of a Bayes commitment. The frequency approach uses more than the observed likelihood function: It can use distribution functions and full model calculations, sometimes component model calculations that provide relevant precision information, and more.

(ii) The ingredients for inference. In the model-data context, y^0 is an observed value and is thus a known constant, and θ is an unknown constant. And if a distribution $\pi(\theta)$ is present, assumed, proposed or created, as the source for θ, then a second distribution is on offer concerning the unknown constant. Probabilities are then sought for the unknown constant, in the context of one or two distributional sources: One part of the given and the other objective, subjective, or appended for computational or other reasons. Should these distributions be combined, or be examined separately, or should the added distribution be ignored? No over-riding principle says that distributions of different status or quality should be combined rather than having their consequences judged separately! Familiar Bayes methodology, however, takes the combining as a given, just as the use of only the observed likelihood

function is taken as a given, essentially axioms in the Bayes methodology! For a recent discussion, see Fraser (2011).

(iii) The simple location model. Consider the location model $f(y-\theta)$. This is of course rather special in that the error, the variable minus the parameter, has a fixed known distributional shape, free of the parameter. A common added or proposed prior is the flat prior $\pi(\theta) = 1$ representing the translation invariance of the model. As it stands the model almost seems too simple for consideration here; but the reality is that this simple model exists as an embedded approximation in an incredibly broad class of models where continuity of parameter effect is present and should thus have its influence acknowledged.

(iv) Location case: p-value or s-value. The generic version of the p-value from observed data y^0 is

$$p^0(\theta) = \int^{y^0} f(y-\theta)\mathrm{d}y = F^0(\theta),$$

which records just the statistical position of the data relative to the parameter. As such it is just the observed distribution function. This $p(\theta)$ function is uniform on the interval $(0,1)$, which in turn implies that any related confidence bound or confidence interval has validity in the sense that it bounds or embraces the true parameter value with the stated reliability; see Fisher (1930) and Neyman (1937). In parallel, the observed Bayes survivor value is

$$s^0(\theta) = \int_\theta f(y^0-\alpha)\mathrm{d}\alpha.$$

The two different directions of integration correspond to data left of the parameter and parameter right of the data, at least in this stochastically increasing case. The two integrals are mathematically equal as is seen from a routine calculus change of variable in the integration. Thus the Bayes survivor s-value acquires validity here, validity in the sense that it is uniformly distributed on $(0,1)$; and validity also in the sense that a Bayes quantile at a level β will have the confidence property and bound the parameter at the stated level. This validity depends entirely on the equivalence of the integrals and no reference or appeal to conditional probability is involved or invoked. Thus in this location model context, a sample space integration can routinely be replaced by a parameter space integration, a pure calculus formality. And thus in the location model context there is no frequency-Bayes contradiction, just the matter of choosing the prior that yields the translation property which in turn enables the integration change of variable and thus the transfer of the integration from sample space to parameter space.

(v) The simple scalar model. Now consider a stochastically increasing scalar model $f(y;\theta)$ with distribution function $F(y;\theta)$ and some minimum continuity and regularity. The observed p-value is

$$p^0(\theta) = F^0(\theta) = \int^{y^0} F_y(y;\theta)\mathrm{d}y = \int_\theta -F_\theta(y^0;\theta)\mathrm{d}\theta,$$

where the subscripts to F denote partial differentiation with respect to the indicated argument. Each of the integrals records an $F(y, \theta)$ value as an integral of its derivative — the fundamental theorem of calculus — one with respect to θ and the other with respect to y. This is pure computation, entirely without Bayes! And then, quite separately, the Bayes survivor value using a proffered prior $\pi(\theta)$ is

$$s^0(\theta) = \int_\theta \pi(\theta) F_y(y^0; \theta) \mathrm{d}\theta.$$

(vi) Validity of Bayes posterior: Simple scalar model. The second integral for $p^0(\theta)$ and the integral for $s^0(\theta)$ are equal if and only if the integrands are equal. In other words if and only if

$$\pi(\theta) = -\frac{F_\theta(y^0; \theta)}{F_y(y^0; \theta)} = \frac{\partial y(\theta; u)}{\partial \theta}\Big|_{\text{fixed}\, F(y;\theta);y^0}$$

with an appropriate norming constant included. The second equality comes from the total derivative of $u = F(y; \theta)$ set equal to 0, thus determining how a θ-change affects y for fixed probability position. We can also view $v(\theta) = \partial y(\theta; u)/\partial\theta$ for fixed u as being the change in y caused by a change in θ, thus giving at y^0 a differential version of the y, θ analysis in the preceding subsection.

Again, with this simple scalar model analysis, there is no frequency-Bayes contradiction; it is just a matter of getting the prior right. The correct prior does depend on the data point y^0 but this should cause no concern. If the objective of Bayesian analysis is to extract all accessible information from an observed likelihood and if this then requires the tailoring of the prior to the particular data, then this is in accord with that objective. Data dependent priors have been around for a long time; see, e.g., Box and Cox (1964). But of course this data dependence does conflict with a conventional Bayes view that a prior should be available for each model type. The realities of data analysis may not be as simple as Bayes might wish.

(vii) What's the conclusion? With a location model, Bayes and frequency approaches are in full agreement: Bayes gets it right because the Bayes calculation is just a frequency confidence calculation in mild disguise. However, with a non-location model, the Bayes claim with a percentage attached to an interval does require a data-dependent prior. But to reference the conditional probability lemma, relabeled as Bayes lemma, requires that a missing ingredient for the lemma be created, that a density not from the reality being investigated be given *objective* status in order to nominally validate the term probability: This violates mathematics and science.

22.4 Inference for regular models: Frequency

(i) Normal, exponential, and regular models. Much of contemporary inference theory is organized around Normal statistical models with side concerns for departures from Normality, thus neglecting more general structures. Recent likelihood methods show, however, that statistical inference is easy and direct for exponential models and more generally for regular models using an appropriate exponential approximation. Accordingly, let us briefly overview inference for exponential models.

(ii) Exponential statistical model. The exponential family of models is widely useful both for model building and for model-data analysis. The full exponential model with canonical parameter φ and canonical variable $u(y)$ both of dimension p is $f(y;\varphi) = \exp\{\varphi' u(y) + k(\varphi)\}h(y)$. Let y^0 with $u^0 = u(y^0)$ be observed data for which statistical inference is wanted. For most purposes we can work with the model in terms of the canonical statistic u:

$$g(u;\varphi) = \exp\{\ell^0(\varphi) + (\varphi - \hat\varphi^0)'(u - u^0)\}g(u),$$

where $\ell^0(\varphi) = a + \ln f(y^0;\varphi)$ is the observed log-likelihood function with the usual arbitrary constant chosen conveniently to subtract the maximum log-likelihood $\ln f(y^0;\hat\varphi^0)$, using $\hat\varphi^0$ as the observed maximum likelihood value. This representative $\ell^0(\varphi)$ has value 0 at $\hat\varphi^0$, and $-\ell^0(\varphi)$ relative to $\hat\varphi^0$ is the cumulant generating function of $u - u^0$, and $g(u)$ is a probability density function. The saddle point then gives a third-order inversion of the cumulant generating function $-\ell^0(\varphi)$ leading to the third-order rewrite

$$g(u;\varphi) = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\{-r^2(\varphi;u)/2\}|\jmath_{\varphi\varphi}(\hat\varphi)|^{-1/2},$$

where $\hat\varphi = \hat\varphi(u)$ is the maximum likelihood value for the tilted likelihood

$$\ell(\varphi;u) = \ell^0(\varphi) + \varphi'(u - u^0),$$

$r^2(\varphi;u)/2 = \ell(\hat\varphi;u) - \ell(\varphi;u)$ is the related log-likelihood ratio quantity,

$$\jmath_{\varphi\varphi}(\hat\varphi) = \frac{\partial}{\partial\varphi\partial\varphi'}\,\ell(\varphi;u)|_{\hat\varphi(u)}$$

is the information matrix at u, and finally k/n is constant to third order. The density approximation $g(u;\varphi_0)$ gives an essentially unique third-order null distribution (Fraser and Reid, 2013) for testing the parameter value $\varphi = \varphi_0$.

Then if the parameter φ is scalar, we can use standard r^*-technology to calculate the p-value $p(\varphi_0)$ for assessing $\varphi = \varphi_0$; see, e.g., Brazzale et al. (2007). For a vector φ, a directed r^* departure is available; see, e.g., Davison et al. (2013). Thus p-values are widely available with high third-order accuracy,

all with uniqueness coming from the continuity of the parameter's effect on the variable involved; see in part Fraser et al. (2010b).

(iii) Testing component parameters. Now consider more generally a component parameter $\psi(\varphi)$ of dimension d with $d < p$. If ψ is linear in φ, then a rotation of coordinates lets us write $\varphi = (\chi, \lambda)$ with χ equivalent to ψ and with say (s, t) as the corresponding canonical coordinates. Statistical inference is available from the d-dimensional conditional distribution on the profile line or plane $L^0 = \{(s, t^0)\}$ with parameter χ. This uses in an essential way the profile likelihood ratio

$$r^2(\chi; s)/2 = \ell^{\mathrm{P}}(\hat{\chi}; s) - \ell^{\mathrm{P}}(\chi; s) = \ell(s, t^0; \hat{\chi}, \hat{\lambda}) - \ell(s, t^0; \chi, \hat{\lambda}_\chi)$$

and related saddle point, but does need a norming constant dependent on χ.

But more generally when the interest parameter ψ is non-linear and thus curved in the initial φ parameterization, the conditional approach just described is effectively unavailable and a marginal approach coming from recent likelihood asymptotics is needed. This involves integrating out over a nuisance parameter variable, and gives to third order the marginal distribution for an ancillary variable under $\psi = \psi_0$, viz.

$$f_{\mathrm{m}}(s; \psi_0) = \frac{e^{k/n}}{(2\pi)^{d/2}} \exp(\tilde{\ell} - \hat{\ell}) \times |\jmath_{\varphi\varphi}\{\hat{\varphi}(s, t^0)\}|^{-1/2} |\jmath_{(\lambda\lambda)}(\psi_0, \hat{\lambda}_{\psi_0}; s, t^0)|^{1/2}, \quad (22.1)$$

on $L^0 = \{(s, t^0)\}$ using rotated coordinates (χ, λ) and (s, t) having $\chi = \chi_0$ first derivative equivalent to $\psi = \psi_0$ at $\hat{\varphi}^0_{\psi_0}$. Here $\hat{\ell} - \tilde{\ell}$ is the log-likelihood ratio at (s, t^0) for the tested value ψ_0, and the nuisance information uses λ with given $\psi = \psi_0$ and λ derivatives for fixed $\psi = \psi_0$ then rescaled in terms of the φ parameterization at $\hat{\varphi}(s, t^0)$ as indicated by the parentheses and described in Brazzale et al. (2007), Fraser and Reid (1993) or Davison et al. (2013). This distribution is essentially unique if continuity of parameter effect is respected; and it is simple, involving only the log-likelihood ratio for ψ_0 and information determinants. In the linear parameter case where the conditional approach is available, this agrees with that conditional result; but here with curvature where no easily accessible conditional approach is available the present marginal approach is the reference standard. My only purpose here is to report on the availability of these unique null distributions and on the availability of p-values, for both linear and curved parameters; for details see, e.g., Fraser and Reid (2013).

(iv) Regular statistical model. Now consider a statistical model $f(y; \theta)$ with continuity in parameter effect and general regularity. For such models we can find, quite widely, a quantile representation $y = y(\theta, u)$ as discussed briefly for a simple case earlier. Such is widely used for simulations and is routinely and definitively available in cases where the model has independent scalar coordinates. Let $V(\theta, y) = \partial y(\theta; u)/\partial\theta$ be the $n \times p$ matrix giving the vectors

that record the effect on y of change in the parameter coordinates $\theta_1, \ldots, \theta_p$; and let $V = V(\hat{\theta}^0, y^0) = \hat{V}^0$ be the observed matrix. Then V records tangents to an intrinsic ancillary contour, say $a(y) = a(y^0)$, that passes through the observed y^0. Thus V represents directions in which the data can be viewed as measuring the parameter, and $\mathcal{L}V$ gives the tangent space to the ancillary at the observed data, with V having somewhat the role of a design matrix. For development details, see Fraser and Reid (1995).

From ancillarity it follows that likelihood conditionally is equal to the full likelihood $L^0(\theta)$, to an order one higher than that of the ancillary used. And it also follows that the sample space gradient of the log-likelihood in the directions V along the ancillary contour gives the canonical parameter, viz.

$$\varphi(\theta) = \frac{\partial}{\partial V}\, \ell(\theta; y)\Big|_{y^0},$$

whenever the conditional model is exponential, or gives the canonical parameter of an approximating exponential model otherwise. In either case, $\ell^0(\theta)$ with the preceding $\varphi(\theta)$ provides third order statistical inference for scalar parameters using the saddle point expression and the above technology. And this statistical inference is uniquely determined provided the continuity in the model is required for the inference (Fraser and Rousseau, 2008). For further discussion and details, see Fraser et al. (2010a) and Fraser and Reid (1995).

22.5 Inference for regular models: Bootstrap

Consider a regular statistical model and the exponential approximation as discussed in the preceding section, and suppose we are interested in testing a scalar parameter $\psi(\varphi) = \psi_0$ with observed data y^0. The bootstrap distribution is $f(y; \psi_0, \hat{\lambda}^0_{\psi_0})$, as used in Fraser and Rousseau (2008) from a log-model perspective and then in DiCiccio and Young (2008) for the exponential model case with linear interest parameter.

The ancillary density in the preceding section is third-order free of the nuisance parameter λ. Thus the bootstrap distribution $f(y; \psi_0, \hat{\lambda}^0_{\psi_0})$ provides full third-order sampling for this ancillary, equivalent to that from the true sampling $f(y; \psi_0, \lambda)$, just the use of a different λ value when the distribution is free of λ.

Consider the profile line L^0 through the data point y^0. In developing the ancillary density (22.1), we made use of the presence of ancillary contours cross-sectional to the line L^0. Now suppose we have a d-dimensional quantity $t(y, \psi)$ that provides likelihood centred and scaled departure for ψ, e.g., a signed likelihood root as in Barndorff-Nielsen and Cox (1994) or a Wald quantity, thus providing the choice in DiCiccio and Young (2008). If $t(y)$ is a function of the ancillary, say $a(y)$, then one bootstrap cycle gives third order,

a case of direct sampling; otherwise the conditional distribution of $y|a$ also becomes involved and with the likelihood based $t(y)$ gives third order inference as in the third cycle of Fraser and Rousseau (2008).

This means that the bootstrap and the usual higher-order calculations are third-order equivalent in some generality, and in reverse that the bootstrap calculations for a likelihood centred and scaled quantity can be viewed as consistent with standard higher-order calculations, although clearly this was not part of the bootstrap design. This equivalence was presented for the linear interest parameter case in an exponential model in DiCiccio and Young (2008), and we now have that it holds widely for regular models with linear or curved interest parameters. For a general regular model, the higher order routinely gives conditioning on full-model ancillary directions while the bootstrap averages over this conditioning.

22.6 Inference for regular models: Bayes

(i) Jeffreys prior. The discussion earlier shows that Bayes validity in general requires data-dependent priors. For the scalar exponential model, however, it was shown by Welch and Peers (1963) that the root information prior of Jeffreys (1946), viz.

$$\pi(\theta) = j_{\theta\theta}^{1/2},$$

provides full second-order validity, and is presented as a globally defined prior and indeed is not data-dependent. The Welch–Peers presentation does use expected information, but with exponential models the observed and expected informations are equivalent. Are such results then available for the vector exponential model?

For the vector regression-scale model, Jeffreys subsequently noted that his root information prior (Jeffreys, 1946) was unsatisfactory and proposed an effective alternative for that model. And for more general contexts, Bernardo (1979) proposed reference posteriors and thus reference priors, based on maximizing the Kullback–Leibler distance between prior and posterior. These priors have some wide acceptance, but can also miss available information.

(ii) The Bayes objective: Likelihood based inference. Another way of viewing Bayesian analysis is as a procedure to extract maximum information from an observed likelihood function $L^0(\theta)$. This suggests asymptotic analysis and Taylor expansion about the observed maximum likelihood value $\hat{\theta}^0$. For this we assume a p-dimensional exponential model $g(u;\varphi)$ as expressed in terms of its canonical parameter φ and its canonical variable u, either as the given model or as the higher-order approximation mentioned earlier. There are also some presentation advantages in using versions of the parameter and of the

variable that give an observed information matrix $j^0_{\varphi\varphi} = I$ equal to the identity matrix.

(iii) Insightful local coordinates. Now consider the form of the log-model in the neighborhood of the observed data $(u^0, \hat{\varphi}^0)$. And let $\mathbf{e}$ be a p-dimensional unit vector that provides a direction from $\hat{\varphi}^0$ or from u^0. The conditional statistical model along the line $u^0 + \mathcal{L}\mathbf{e}$ is available from exponential model theory and is just a scalar exponential model with scalar canonical parameter ρ, where $\varphi = \hat{\varphi}^0 + \rho\mathbf{e}$ is given by polar coordinates. Likelihood theory also shows that the conditional distribution is second-order equivalent to the marginal distribution for assessing ρ. The related prior $j^{1/2}_{\rho\rho}\mathrm{d}\rho$ for ρ would use $\lambda = \hat{\lambda}^0$, where λ is the canonical parameter complementing ρ.

(iv) The differential prior. Now suppose the preceding prior $j^{1/2}_{\rho\rho}\mathrm{d}\rho$ is used on each line $\hat{\varphi}^0 + \mathcal{L}\mathbf{e}$. This composite prior on the full parameter space can be called the differential prior and provides crucial information for Bayes inference. But as such it is of course subject to the well-known limitation on distributions for parameters, both confidence and Bayes; they give incorrect results for curved parameters unless the pivot or prior is targeted on the curved parameter of interest; for details, see, e.g., Dawid et al. (1973) and Fraser (2011).

(v) Location model: Why not use the location property? The appropriate prior for ρ would lead to a constant-information parameterization, which would provide a location relationship near the observed $(y^0, \hat{\varphi}^0)$. As such the p-value for a linear parameter would have a reflected Bayes survivor s-value, thus leading to second order. Such is not a full location model property, just a location property near the data point, but this is all that is needed for the reflected transfer of probability from the sample space to the parameter space, thereby enabling a second-order Bayes calculation.

(vi) Second-order for scalar parameters? But there is more. The conditional distribution for a linear parameter does provide third order inference and it does use the full likelihood but that full likelihood needs an adjustment for the conditioning (Fraser and Reid, 2013). It follows that even a linear parameter in an exponential model needs targeting for Bayes inference, and a local or global prior cannot generally yield second-order inference for linear parameters, let alone for the curved parameters as in Dawid et al. (1973) and Fraser (2013).

22.7 The frequency-Bayes contradiction

So is there a frequency-Bayes contradiction? Or a frequency-bootstrap-Bayes contradiction? Not if one respects the continuity widely present in regular statistical models and then requires the continuity to be respected for the frequency calculations and for the choice of Bayes prior.

Frequency theory of course routinely leaves open the choice of pivotal quantity which provides the basis for tests, confidence bounds, and related intervals and distributions. And Bayes theory leaves open the choice of the prior for extracting information from the likelihood function. And the bootstrap needs a tactical choice of initial statistic to succeed in one bootstrap cycle. Thus on the surface there is a lot of apparent arbitrariness in the usual inference procedures, with a consequent potential for serious contradictions. In the frequency approach, however, this arbitrariness essentially disappears if continuity of parameter effect in the model is respected, and then required in the inference calculations; see Fraser et al. (2010b) and the discussion in earlier sections. And for the Bayes approach above, the arbitrariness can disappear if the need for data dependence is acknowledged and the locally based differential prior is used to examine sample space probability on the parameter space. This extracts information from the likelihood function to the second order, but just for linear parameters (Fraser, 2013).

The frequency and the bootstrap approaches can succeed without arbitrariness to third order. The Bayes approach can succeed to second order provided the parameter is linear, otherwise the prior needs to target the particular interest parameter. And if distributions are used to describe unknown parameter values, the frequency joins the Bayes in being restricted to linear parameters unless there is targeting; see Dawid et al. (1973) and Fraser (2011).

22.8 Discussion

(i) Scalar case. We began with the simple scalar location case, feeling that clarity should be present at that transparent level if sensible inference was to be available more generally. And we found at point (ii) that there were no Bayes-frequency contradictions in the location model case so long as model continuity was respected and the Bayes s-value was obtained from the location based prior. Then at point (v) in the general scalar case, we saw that the p-value retains its interpretation as the statistical position of the data and has full repetition validity, but the Bayes requires a prior determined by the form of the model and is typically data dependent. For the scalar model case this is a radical limitation on the Bayes approach; in other words inverting the distribution function as pivot works immediately for the frequency approach whereas inverting the likelihood using the conditional probability lemma as a tool requires the prior to reflect the location property, at least locally. For the scalar model context, this represents a full vindication of Fisher (1930), subject to the Neyman (1937) restriction that probabilities be attached only to the inverses of pivot sets.

(ii) Vector case. Most models however involve more than just a scalar parameter. So what about the frequency-Bayes disconnect away from the very

simple scalar case? The Bayes method arose from an unusual original example (Bayes, 1763), where at the analysis stage the parameter was retroactively viewed as generated randomly by a physical process, indeed an earlier performance of the process under study. Thus a real frequency-based prior was introduced hypothetically and became the progenitor for the present Bayes procedure. In due course a prior then evolved as a means for exploring, for inserting feelings, or for technical reasons to achieve analysis when direct methods seemed unavailable. But do we have to make up a prior to avoid admitting that direct methods of analysis were not in obvious abundance?

(iii) Distributions for parameters? Fisher presented the fiducial distribution in Fisher (1930, 1935) and in various subsequent papers. He was criticized from the frequency viewpoint because his proposal left certain things arbitrary and thus not in a fully developed form as expected by the mathematics community at that time: Welcome to statistics as a developing discipline! And he was criticized sharply from the Bayes (Lindley, 1958) because Fisher proposed distributions for a parameter and such were firmly viewed as Bayes territory. We now have substantial grounds that the exact route to a distribution for a parameter is the Fisher route, and that Bayes becomes an approximation to the Fisher confidence and can even attain second-order validity (Fraser, 2011) but requires targeting even for linear parameters.

But the root problem is that a distribution for a vector parameter is inherently invalid beyond first order (Fraser, 2011). Certainly in some generality with a linear parameter the routine frequency and routine Bayes can agree. But if parameter curvature is allowed then the frequency p-value and the Bayes s-value change in *opposite directions*: The p-value retains its validity, having the uniform distribution on the interval $(0, 1)$ property, while the Bayes loses this property and associated validity, yet chooses to retain the label "probability" by discipline commitment, as used from early on. In all the Bayes cases the events receiving probabilities are events in the past, and the prior probability input to the conditional probability lemma is widely there for expediency: The lemma does not create real probabilities from hypothetical probabilities except when there is location equivalence.

(iv) Overview. Most inference contradictions disappear if continuity present in the model is required for the inference calculations. Higher order frequency and bootstrap are consistent to third order for scalar parameters. Bayes agrees but just for location parameters and then to first order for other parameters, and for this Bayes does need a prior that reflects or approximates the location relationship between variable and parameter. Some recent preliminary reports are available at `http://www.utstat.toronto.edu/dfraser/documents/` as `260-V3.pdf` and `265-V3.pdf`.

Acknowledgment

This research was funded in part by the Natural Sciences and Engineering Research Council of Canada, by the Senior Scholars Funding at York University, and by the Department of Statistical Sciences at the University of Toronto. Thanks to C. Genest and A. Wang for help in preparing the manuscript.

References

Abraham, C. (2009). Vioxx took deadly toll study. *Globe and Mail* `http://www.theglobeandmail.com/life/study-finds-vioxx-took-deadly-toll/article4114560/`

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, London*, 53:370–418.

Bergin, C. (2007). Remembering the mistakes of Challenger. `nasaspaceflight.com`.

Bernardo, J.M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society, Series B*, 41:113–147.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26:211–252.

Brazzale, A.R., Davison, A.C., and Reid, N.M. (2007). *Applied Asymptotics*. Cambridge University Press, Cambridge, UK.

Dalai, S. and Fowlkes, B. (1989). Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association*, 84:945–957.

Davison, A.C., Fraser, D.A.S., Reid, N.M., and Sartori, N. (2013). *Accurate directional inference for vector parameters*. Submitted for publication.

Dawid, A.P., Stone, M., and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society, Series B*, 35:189–233.

DiCiccio, T.J. and Young, G.A. (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika*, 95:497–504.

Fisher, R.A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

Fisher, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398.

Fraser, A.M., Fraser, D.A.S., and Fraser, M.J. (2010a). Parameter curvature revisited and the Bayesian frequentist divergence. *Statistical Research: Efron Volume*, 44:335–346.

Fraser, A.M., Fraser, D.A.S., and Staicu, A.M. (2010b). Second order ancillary: A differential view with continuity. *Bernoulli*, 16:1208–1223.

Fraser, D.A.S. (1961a). The fiducial method and invariance. *Biometrika*, 48:261–280.

Fraser, D.A.S. (1961b). On fiducial inference. *The Annals of Mathematical Statistics*, 32:661–676.

Fraser, D.A.S. (2011). Is Bayes posterior just quick and dirty confidence? (with discussion). *Statistical Science*, 26:299–316.

Fraser, D.A.S. (2013). *Can Bayes inference be second-order for scalar parameters*? Submitted for publication.

Fraser, D.A.S. and Reid, N.M. (2013). *Assessing a parameter of interest: Higher-order methods and the bootstrap*. Submitted for publication.

Fraser, D.A.S. and Reid, N.M. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statistica Sinica*, 3:67–82.

Fraser, D.A.S. and Reid, N.M. (1995). Ancillaries and third order significance. *Utilitas Mathematica*, 47:33–53.

Fraser, D.A.S. and Rousseau, J. (2008). Studentization and deriving accurate p-values. *Biometrika*, 95:1–16.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society, Series A*, 186:453–461.

Léger, C. (2013). The Statistical Society of Canada (SSC) response to the NSERC consultation on the evaluation of the Discovery Grants Program. *SSC Liaison*, 27(2):12–21.

Lindley, D.V. (1958). Fiducial distribution and Bayes theorem. *Journal of the Royal Statistical Society, Series B*, 20:102–107.

Marshall, M. (2012). Seismologists found guilty of manslaughter. *New Scientist*, October 22, 2012.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Series A*, 237:333–380.

Pielke, R. (2011). Lessons of the L'Aquila lawsuit. *Bridges* 31, `http://www.bbc.co.uk/news/world-europe-20025626`.

Prats, J. (2012). The L'Aquila earthquake. *Significance*, 9:13–16.

Welch, B.L. and Peers, H.W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *Journal of the Royal Statistical Society, Series B*, 25:318–329.