Journal: BIOMETRIKA

Article doi: asm093

Article title: Studentization and deriving accurate p-values

First Author:

Corr. Author:

AUTHOR QUERIES - TO BE ANSWERED BY THE CORRESPONDING AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer these queries by marking the required corrections at the appropriate point in the text.

A1	Author: The term "whatever" has been replaced by "irrespective of" in the sentence "In their . . . for *panc.*" Please check the edit.	
A2	Author: Please check the usage of "about the true" in the sentence "For the . . . coordinates." Is any word missing after "true" or the sentence is fine as given?	
A3	Author: Please clarify, if reference "Rousseau (2007)" is in press or has already been published?	

Biometrika (2007), **00**, 0, *pp*. 1–16
© 2007 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asm093
Advance Access publication xxxxxx

Studentization and deriving accurate *p*-values

BY D.A.S. FRASER

Department of Statistics, University of Toronto, Toronto, Canada M5S 3G3
dfraser@utstat.toronto.edu

AND JUDITH ROUSSEAU

CEREMADE, University Paris Dauphine, 75016 Paris, France
rousseau@ceremade.dauphine.fr

SUMMARY

We have a statistic for assessing an observed data point relative to a statistical model but find that its distribution function depends on the parameter. To obtain the corresponding p-value, we require the minimally modified statistic that is ancillary; this process is called Studentization. We use recent likelihood theory to develop a maximal third-order ancillary; this gives immediately a candidate Studentized statistic. We show that the corresponding p-value is higher order Un(0, 1), is equivalent to a repeated bootstrap version of the initial statistic and agrees with a special Bayesian modification of the original statistic. More importantly, the modified statistic and p-value are available by Markov chain Monte Carlo simulations and in some cases, by higher order approximation methods. Examples, including the Behrens–Fisher problem, are given to indicate the ease and flexibility of the approach.

Some key words: Ancillary; Bayesian; Behrens–Fisher problem; Bootstrap; Conditioning; Departure measure; Likelihood; p-value; Studentization.

1. INTRODUCTION

Third-order highly accurate p-values are routinely available for assessing scalar parameters in a statistical model having moderate regularity; see, for example, Fraser et al. (1999). We consider here the case where only a null model is available, together with a pragmatically chosen departure measure. We derive third-order p-values by frequentist, Bayesian and bootstrap analyses, and then show them to be equal to the third order: a choice among them would typically depend on the ease of implementation in an application.

Suppose, we have a null model $f(y;\theta)$ for a statistical context and wish to judge the acceptability of the model in the presence of data y^0; the null model may exist on its own, or could be a restriction of a larger embedding model. Also suppose, we have a scalar statistic $t(y)$ that has been proposed as a plausible signed measure of departure of data from the model; the statistic might have arisen pragmatically based on physical properties in the context, or could be a simple departure measure, perhaps of the form estimate-minus-parameter, in an embedding model.

We would naturally expect the departure $t(y)$ to have a distribution that depends on the parameter θ in the null model, and then want to construct a modified departure, $\tilde{t}(y)$ say, that is ancillary, and thus with a θ-free distribution, but still with as much as possible of the structure of the original measure. One would also want to have the corresponding distribution function $H(\tilde{t})$, so as to obtain the observed p-value, $p^0 = H(\tilde{t}^0)$, being the percentage position of the data

with respect to the null model. In this formulation, the only suggestion of possible alternatives to the null is to be found in the choice of the departure statistic $t(y)$. We do not here address this important issue, which may be strongly guided by the physical context. In this paper, we develop p-values for $t(y)$, but based directly on the modified version $\tilde{t}(y)$.

As a simple example, consider a sample purportedly from a normal model with mean μ_0, together with a proposed departure measure $t(y) = \bar{y}$: we might reasonably hope that the indicated p-value would be $p^0 = H_{n-1}(\tilde{t}^0)$, where H_ν is the Student distribution function with ν degrees of freedom and $\tilde{t}^0$ is the observed value of the usual t-statistic; this p-value is, of course, the Student value recording the percentage position of the data with respect to the normal model located at μ_0.

The process of developing a valid $\tilde{t}(y)$ from $t(y)$ is here called general Studentization, a generalization of the Student (1908) conversion of $\bar{y}$ or $\bar{y} - \mu_0$ into the familiar t-statistic with its t_{n-1} distribution function as just described. This general problem has had extensive recent discussion in the literature, particularly the Bayesian literature (Bayarri & Berger, 2000; Robins et al., 2000).

Frequentist theory gives a simple first-order p-value called the plug-in p-value,

$$p^0_{\text{plug}} = G(t^0; \hat{\theta}^0), \tag{1·1}$$

where $G(t;\theta) = \text{pr}\{t(y) < t; \theta\}$ is the distribution function for $t(y)$ and the parameter has been replaced by its observed maximum-likelihood value. This p-value is known to be remarkably unreliable in many contexts (Bayarri & Berger, 2000; Robins et al., 2000).

Bootstrap theory directly addresses (Beran, 1988) the general Studentization problem: one samples from the null model distribution using the observed maximum-likelihood value for the parameter. As such, the bootstrap is calculating the plug-in p-value p^0_{plug} by simulations, and typically centres an initial statistic.

The recent Bayesian literature has developed many p-values for the general Studentization problem, but from a different viewpoint. As mentioned in Bayarri & Berger (2000), 'Bayesians have a natural way to eliminate nuisance parameters: integrate them out.' With a prior $\pi(\theta)$ added to the present context, the Bayesian averaged density for y, called the prior predictive density, is

$$m(y) = c \int f(y;\theta)\pi(\theta)d\theta$$

as appropriately normalized, and gives the prior-based p-value

$$p^0_{\text{prior}} = \text{pr}\{t(y) < t(y^0); m(\cdot)\};$$

see Box (1980). The prior predictive density $m(y)$ can, however, be improper as the normal example using $\pi(\mu, \sigma) = c\sigma^{-1}$ indicates. An alternative Bayesian analogue of the plug-in p-value is the posterior p-value,

$$p^0_{\text{post}} = \int_\Theta \text{pr}\{t(y) \leqslant t^0; \theta\} d\pi(\theta \mid y^0);$$

this also has disadvantages (Bayarri & Berger, 2000; Robbins et al., 2000), but for the normal example is first-order equivalent to the plug-in or bootstrap p-value.

For the normal example, the full plug-in or bootstrap distribution is that of a sample from a normal with mean μ_0 and standard deviation $\hat{\sigma}^0 = \{\sum(y_i - \mu_0)^2/n\}^{1/2}$, and the derived distribution of $\bar{y}$ is correspondingly normal with mean μ_0 and standard deviation $\hat{\sigma}^0/n^{1/2}$; the plug-in p-value is then $p^0 = \Phi\{n^{1/2}(\bar{y}^0 - \mu_0)/\hat{\sigma}^0\}$. This is centred, but underestimates departure from the centre. However, if we examine this p-value as a statistic, as a function of the original data,

we can see that it is one-to-one equivalent to the ordinary Student statistic $\tilde{t} = n^{1/2}(\bar{y} - \mu_0)/s_y$. Thus, if the plug-in or bootstrap approach were to be reapplied to the modified statistic, we would obtain the observed p-value $p^0 = H_{n-1}(\tilde{t}^0)$, up to the accuracy of the sampling simulations used.

In § 3, we use recent likelihood theory (Fraser & Reid, 1995, 2001; Fraser et al., 1999; Fraser, 2003) to develop a definitive p-value. This is obtained from a full $(n - p)$-dimensional third-order ancillary $a(y)$ with density $g(a)$ and requires mild regularity, namely asymptotic properties, p-dimensional continuous parameter and smoothness of the maximum-likelihood statistic $\hat{\theta}(y)$. While the choice of ancillary is not unique, the corresponding distribution $g(a)$ is unique to third order; the device for demonstrating this uses convenient, but non-standard coordinates for the ancillary a. To be specific, the coordinates are taken to be points on the observed maximum-likelihood surface; a fixed choice of maximum-likelihood surface is also a viable option and will be discussed briefly.

The frequentist p-value is then obtained from the full ancillary density $g(a)$ and its value from observed data is

$$p^0_{\text{anc}} = \text{pr}_{\text{g}}\{t(a) < t^0\} = G_{\text{g}}(t^0; \hat{\theta}^0) = \int_{t(a;\hat{\theta}^0)<t^0} g(a)da, \tag{1·2}$$

where pr_{g} designates probability using the ancillary density $g(a)$, and $G_{\text{g}}(t; \hat{\theta}^0)$ designates the related distribution function for $t(y)$. Interestingly, following Fraser & Reid (1995), we have an explicit and computable expression for p^0_{anc} to third order.

We also obtain, in § 3, a Bayesian ancillary density $\tilde{g}(a)$ by prior averaging, as for $m(y)$, but with the model examined on a region having $\hat{\theta}$ in a small interval $(\pm\delta/2)$ about the observed maximum-likelihood value, or even examined just on the observed maximum-likelihood surface. We find that this modified Bayesian ancillary is equal to the frequentist ancillary to third order, so that $\tilde{g}(a) = g(a)$. This gives a Bayesian p-value p^0_{B} that is equal to the frequentist p-value p^0_{anc} to third order. Section 4 shows that the Bayesian-frequentist ancillary is Un(0, 1) to third order under general regularity and to second order under more relaxed conditions. Section 5 shows that three levels of the plug-in or bootstrap procedure lead to the same third-order p-value.

2. Some background on bootstrap p-values

The difficulties with the direct Bayesian p-values p_{prior} and p_{post} have led to more refined and incisive methods for using prior densities, leading to some preference for two versions designated p_{post} and p_{cpred}. In this direction (Bayarri & Berger, 2000; Robins et al., 2000), a posterior density for θ is derived from some aspect of the data designated D_1,

$$\pi(\theta \mid D_1) = cL(\theta; D_1)\pi(\theta),$$

and then used to eliminate θ from the distribution function G_2 for $t(y)$ derived from some other aspect of the data D_2,

$$p^0 = \int G_2(D_2; \theta)\pi(\theta \mid D_1)d\theta.$$

If the full data y^0 are used in both places, there is a clear conflict in the probability calculations, often described as double use of the data. Some obvious difficulties with the double use of data can be avoided by having D_1 in some sense distinct from D_2. Bayarri & Berger (2000) and Robins et al. (2000) study the case where D_1 is the conditional maximum-likelihood estimator, given the test statistic $t(y)$; for this, Robins et al. (2000) show that p_{cpred} is asymptotically uniform to first order, provided that $t(y)$ is asymptotically normal.

In many settings, however, the preceding conditional maximum-likelihood estimator is extremely difficult to work with, as it can require an explicit expression for the density of $t(y)$, which is often unavailable. Here, following a University of Paris-Dauphine technical report by C.P. Robert and J. Rousseau, we take D_1 to be $\hat{\theta}$ and D_2 to be $y \mid \hat{\theta}$. For this, Robert and Rousseau prove that the resulting p-value p_{cpred} is first-order equivalent to the frequentist p-value $\text{pr}\{t(y) < t \mid \hat{\theta}; \theta\}$, for any statistic $t(y)$. Here, we accept this p_{cpred} as a plausible contender, examine it with other p-values using recent higher order likelihood theory and find that it is third-order equivalent to the frequentist p_{anc} and the direct Bayesian p_{B}, and is thus distributed as Un(0, 1) to third order.

Now consider three repetitions of the plug-in or bootstrap procedure. For this, we let $G_i(t; \theta)$ designate the distribution function for a variable indexed by i, as calculated from the model $f(y; \theta)$. Then, with p_0 designating some initial function $t(y)$ and with the iteration $p_{i+1} = G_i(p_i; \hat{\theta})$, we have that $p_1 = p_{\text{plug}}$ is the plug-in p-value, and that $p_3 = p_{\text{bs}}$ is a proposed three-level bootstrap or plug-in p-value. Section 4 shows that this bootstrap p-value p_{bs} is third-order equivalent to the ancillary p-value p_{anc} under mild asymptotic and regularity conditions.

This paper, in part, thus extends Robins et al. (2000) in several ways: first, by working with a specialized version of p_{cpred}; secondly, by relaxing the hypotheses on the statistic $t(y)$; and thirdly, by obtaining higher order results. The first and second aspects are important, as a test statistic can often be complicated with no available asymptotic distribution; see, for instance, the goodness-of-fit tests described by Robert and Rousseau or in Rousseau (2007). Moreover, a p-value provides a universal scale for a test procedure and can be considered from a Bayesian perspective as a calibration of such test procedures; it is, thus important to be as close as possible to the uniform distribution.

Our results based on large sample likelihood theory for a continuous model with regularity show that general Studentization can be obtained by a frequentist ancillary approach, by a Bayesian ancillary approach or by a three-level bootstrap approach, and that the results are equivalent to third order. We also see, in § 3, that the Bayesian and frequentist p-values are available by direct Markov chain Monte Carlo simulations, while the bootstrap p-values could require double or triple levels and perhaps, not have the same numerical accessibility.

3. THE BAYESIAN AND FREQUENTIST ANCILLARY

Third-order highly accurate p-values initially for exponential models (Lugannani & Rice, 1980) and for exact-ancillary contexts (Barndorff-Nielsen, 1986) are now available for quite general continuous-model contexts (Fraser et al., 1999; Fraser & Reid, 2001), and have been extended to give third-order marginal likelihoods (Fraser, 2003). From this, we use the existence of an approximate ancillary which is second order (Fraser & Reid, 1993, 1995, 2001), but can be upgraded to third order. In addition, for the calculations here, we assume that a particular choice of third-order ancillary has been made; details of such are not needed nor are they readily available. The maximum-likelihood estimator $\hat{\theta}$ of dimension p provides coordinates, given the ancillary and the ancillary has dimension $(n - p)$. An observed maximum-likelihood surface,

$$S_{\hat{\theta}^0} = \{y : \hat{\theta}(y) = \hat{\theta}^0\},$$

will intersect each ancillary contour in a point, and with regularity gives a one-to-one correspondence between ancillary contours and points on $S_{\hat{\theta}^0}$, for fixed $\hat{\theta}^0$.

This allows us to use points on $S^0 = S_{\hat{\theta}^0}$ to index or label the ancillary contours. Thus, if $A(y)$ is a third-order ancillary statistic in familiar form, then our choice of coordinates means that a point y with $A(y) = A$ is projected on to the surface S^0 along its ancillary contour $\{z : A(z) = A\}$ to give

the point $a = a(y)$ on the surface, and a is used in place of A. Accordingly, we take the ancillary variable to be $a(y)$, where a now designates the point on the surface S^0, and correspondingly, we take da to be Euclidean measure on S^0. This rather non-standard choice of coordinates has large benefits, but does need special care. A reason for the special coordinates is that we do not have easy information about what an ancillary contour looks like and as we shall see, we do not need such information.

Thus, in a moderate-deviations neighbourhood of the surface S^0, we have that, a point y is now represented as $(a, \hat{\theta})$, where a is the point on S^0 with the same ancillary value and $\hat{\theta}$ is just $\hat{\theta}(y)$; and we make no direct use of the n-dimensional coordinates on $S_{\hat{\theta}(y)}$.

Now following Fraser & Reid (1995) and with $\ell(\theta; y)$ designating log-density, we use the Jacobian of the change of variable for a point y on S^0 and obtain the probability differential

$$f(y; \hat{\theta}^0)dy = \exp\{\ell(\hat{\theta}^0; a)\} \mid \ell_{\theta;y}(\hat{\theta}^0; a) \mid^{-1} \mid j_{\theta\theta}(\hat{\theta}^0; a) \mid da d\hat{\theta}, \tag{3·1}$$

where $\mid \ell_{\theta;y} \mid^{-1} \mid j_{\theta\theta} \mid$ is the Jacobian. We next divide by the conditional density value $f(\hat{\theta}^0 \mid a; \hat{\theta}^0)$, expressed as $e^{c/n}(2\pi)^{-p/2} \mid j_{\theta\theta}(\hat{\theta}; y) \mid^{1/2}$ by Barndorff–Nielsen's p^* formula. This gives the marginal density for the ancillary to third order:

$$g(a)da = \frac{(2\pi)^{p/2}}{e^{c/n}} \exp\{\ell(\hat{\theta}^0; a)\} \mid \ell_{\theta;y}(\hat{\theta}^0; a) \mid^{-1} \mid j_{\theta\theta}(\hat{\theta}^0; a) \mid^{1/2} da, \tag{3·2}$$

which is an $O(n^{-1/2})$ adjustment to the expression in (3·1). Although a third-order ancillary does not have uniqueness to third order, we do have from the above, that the corresponding density $g(a)$ has such uniqueness (Fraser & Reid, 1995, 2001), up to the labelling of the ancillary partition sets. In other words, if $A(y)$ is an ancillary variable, then

$$\text{pr}\{A(y) \in B\} = \int_{A(a,\hat{\theta}^0)\in B} g(a)da\{1 + O(n^{-3/2})\}.$$

We also have (Fraser & Reid, 1995) that the ancillary $a(y)$ recorded with n-dimensional coordinates can be represented equivalently to third order by a finite number of coordinates; this facilitates the use of Laplace integration techniques.

The simple normal example illustrates some aspects of this, and for notational ease, we take the null value μ_0 to be zero. The observed maximum-likelihood surface has $\hat{\sigma} = \hat{\sigma}^0$ and is the sphere with Σy_i^2 equal to the observed sum of squares; an obvious ancillary is the unit direction $y/\mid y \mid$, but there are many others, such as $O(\mid y \mid)y/\mid y\mid$, where $O(s)$ is some rotation matrix chosen as a smooth function of radial distance s. If, however, we project probability along ancillary contours to the observed maximum-likelihood sphere, we obtain a unique distribution, which here is uniform on a sphere. This unique distribution on a maximum-likelihood surface is a general likelihood result and is the basis for the third-order p-values and the marginal likelihoods.

While the distribution of the ancillary as recorded on any chosen cross section S^0 is unique to third order, subject, of course, to the coordinate labelling, there still can be various ancillaries as noted for the simple normal example. Thus, when we write $a(y)$, we are implicitly assuming a particular choice of ancillary and thus, a particular linking of points from one maximum-likelihood surface to another. This can raise certain technical issues and can lead to different parameter inference statements, not of interest here. We do note, however, that with independent scalar coordinates and continuity in the parameter-to-variable relationship, the inference issue does not arise, and that for independent vector variables, the inference statements can depend on how the parameter is related to the variables, as given typically by coordinate distribution functions or other sensible pivotal quantities.

Now, consider a Bayesian ancillary density: we average the model density with respect to the prior and then normalize the resulting expression over a small neighbourhood of the observed maximum-likelihood surface S^0. For this, let $f(\hat{\theta};\theta)$ denote the marginal density of $\hat{\theta}$, given θ; then the proposed Bayesian ancillary density is

$$\tilde{g}(a \mid \hat{\theta^0}) = \frac{\int_\Theta f(a,\hat{\theta}^0;\theta)\pi(\theta)d\theta}{\int_\Theta f(\hat{\theta}^0;\theta)\pi(\theta)d\theta}$$

$$= \frac{\exp\{\ell(\hat{\theta}^0;a)\} \mid \ell_{\theta:a}(\hat{\theta}^0;a) \mid^{-1} \mid \hat{j}_{\theta\theta}(\hat{\theta}^0;a) \mid \int_\Theta \exp\{\ell(\theta;a)-\ell(\hat{\theta}^0;a)\}\pi(\theta)d\theta}{\int_{S_{\hat{\theta}^0}} \exp\{\ell(\hat{\theta}^0;a)\} \mid \ell_{\theta:a}(\hat{\theta}^0;a) \mid^{-1} \mid \hat{j}_{\theta\theta}(\hat{\theta}^0;a) \mid \int_\Theta \exp\{\ell(\theta;a)-\ell(\hat{\theta}^0;a)\}\pi(\theta)d\theta da}.$$

The Bayesian integration over the p-dimensional parameter space can be accomplished by the usual Laplace integration and for example, the denominator integration would take the form

$$\int_\Theta \exp\{\ell(\theta;a)-\ell(\hat{\theta}^0;a)\}\pi(\theta)d\theta = (2\pi)^{p/2} \mid j_{\theta\theta}(\hat{\theta}^0;a) \mid^{-1/2} \pi(\hat{\theta}^0)\exp\{H(\hat{\theta}^0,a)/n\}$$

to third order. For this, $\exp H(\hat{\theta}^0,a)/n$ is the fraction of the Laplace normal integral that reproduces the initial integral and it has no $O(n^{-1/2})$ component, as such, terms cancel in the usual Laplace manner. As noted earlier in this section, the ancillary a can be represented by a finite number of coordinates with other coordinates distribution free of θ to the third order. It follows then, that when $H(\hat{\theta}^0,a)/n$, already of second order, is further expanded in terms of these simplified coordinates, there will be no dependence on them to the next order, which is the third order that we are working to.

Thus, the H terms in the numerator and denominator of the expression for $\tilde{g}(a \mid \hat{\theta^0})$ cancel and we obtain,

$$\tilde{g}(a \mid \hat{\theta^0}) = \frac{\exp\{\ell(\hat{\theta}^0;a)\} \mid \ell_{\theta:a}(\hat{\theta}^0;a) \mid^{-1} \mid \hat{j}_{\theta\theta}(\hat{\theta}^0;a) \mid^{1/2}}{\int_{S_{\hat{\theta}^0}} \exp\{\ell(\hat{\theta}^0;a)\} \mid \ell_{\theta:a}(\hat{\theta}^0;a) \mid^{-1} \mid \hat{j}_{\theta\theta}(\hat{\theta}^0;a) \mid^{1/2} da}$$

$$= g(a)\cdot$$

We, thus, have the significant result that Bayesian averaging of probability in an interval region about the observed maximum-likelihood surface generates a distribution $\tilde{g}(a)$ on that surface which is equal to the ancillary distribution recorded on that surface. Alternatively, from a somewhat different perspective, we can use the prior $\pi(\theta)$ to integrate the density function of y for fixed $\hat{\theta}^0$, leading to $g(a)da$ to third order. Thus, in effect, we have without loss interchanged the order, in which we do a $\hat{\theta}$-sectioning and a θ-marginalization. Either way, we obtain the distribution $g(a) = \tilde{g}(a)$, which is the marginal distribution of the ancillary statistic a.

Now consider the Bayesian p-value proposed by Robert and Rousseau in their report. The posterior distribution from the marginal for $\hat{\theta}$ at the observed $\hat{\theta}^0$,

$$\pi(\theta \mid \hat{\theta}^0) = cf(\hat{\theta}^0;\theta)\pi(\theta),$$

is combined with the conditional distribution for $y \mid \hat{\theta}^0$, producing $\pi(\theta)f(y^0;\theta)$; this is then averaged over θ, which, as we have noted, just gives $c\tilde{g}(a)$. We, thus, have that the proposed modified Bayesian p_{cpred},

$$p^0_{\text{cpred}} = \int_\Theta \text{pr}\{t(y) < t^0 \mid \hat{\theta}^0;\theta\}\pi(\theta \mid \hat{\theta}^0)d\theta = \int_{t(a,\hat{\theta}^0)<t^0} \tilde{g}(a \mid \hat{\theta}^0)da,$$

is equal to the ancillary and the direct Bayesian p-values, to third order.

From this, we have the intriguing result that the present p-values can be obtained by Markov chain Monte Carlo sampling: we use Markov chain Monte Carlo to obtain a sample sequence $\{\theta^t : t = 1, \ldots, T\}$ from $\pi(\theta) f(\hat{\theta}^0; \theta)$; then for each θ^t , we obtain a sample sequence $\{y^{tl} : l = 1, \ldots, L\}$ from the density of y for given $\hat{\theta}^0$, $l = 1, \ldots, L$; and then for each of these y^{tl}, we calculate $t(y^{tl})$. This gives the Bayesian p-value approximation,

$$\hat{p} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{L} \sum_{l=1}^{L} \mathbb{I}_{t(y^{tl}) < t^0}.$$

A similar algorithm can be used to obtain p_{anc} as it has a similar form.

4. The Bayesian-frequentist p-value is asymptotically uniform

4·1. *The effective statistic*

In § 1, for a statistic $t(y)$, we used a distribution on an observed maximum-likelihood surface to calculate a p-value; however, the distribution was a marginal distribution projected on to that surface and the statistic was examined only on that surface. We now define a modified statistic $\tilde{t}(y)$ that links from one possible maximum-likelihood surface to another such surface; the linkage is by using contours of $t(y)$ having the same p-value. For this, consider a particular surface S_0, corresponding to some maximum-likelihood value $\hat{\theta}_0$, and on S_0 , take $\tilde{t}(y)$ to be the value of the given statistic $t(y)$, and on any other maximum-likelihood surface at point y, do the p-value calculation for that surface and take $\tilde{t}(y)$ to be the value of $t(\cdot)$ on S_0 that has the same p-value; thus,

$$\tilde{t}(y) = G_g^{-1}[G_g\{t(y); \hat{\theta}(y)\}; \hat{\theta}_0].$$

We have immediately, that $\tilde{t}(y)$ agrees with $t(y)$ on S_0, but may differ everywhere else; that on any maximum-likelihood surface, its partition agrees with that of $t(y)$; and that on the full space, its partition is independent of the choice of initial surface S_0.

However, from the construction, we have that $\tilde{t}(y)$ satisfies

$$\text{pr}\{\tilde{t}(y) \leqslant \tilde{t}_0; \theta\} = \text{pr}\{p_{\text{anc}}(t, \hat{\theta}) \leqslant p_{\text{anc}}(t_0, \hat{\theta}_0); \theta\}$$

for any t_0. Thus, if we prove that

$$\text{pr}\{\tilde{t}(y) \leqslant \tilde{t}_0; \theta\} = p_{\text{anc}}(t_0, \hat{\theta}_0) + O_P(n^{-3/2}),$$

we obtain the required third-order uniformity.

Towards the proof to come of the uniformity property, we construct an intermediate statistic $\bar{t}(y)$. From § 1, we have that there exist third-order ancillaries and we assumed a particular choice of such an ancillary, designating it as $a(y)$, now to be based on the present reference surface S_0; we can then replace y by the alternative coordinates $(a, \hat{\theta})$. The intermediate statistic $\bar{t}(y)$ is taken to be equal to $t(y)$ on S_0 and otherwise to be constant on contours of the chosen ancillary. We, thus, define $\bar{t}$ by lifting from S_0 in accordance with the chosen ancillary: $\bar{t}(a, \hat{\theta}) = \bar{t}(a, \hat{\theta}^0)$. Thus,

$$\text{pr}\{\bar{t}(a, \hat{\theta}) < t; \theta\} = \text{pr}_g\{\bar{t}(a, \hat{\theta}_0) < t; \hat{\theta}_0\} = p_{\text{anc}}(t; \hat{\theta}_0), \tag{4·1}$$

which is distributed as Un(0, 1). This intermediate statistic is very clearly dependent on our chosen reference surface, but it is third-order Un(0, 1) and leads us to the proof that the same holds for $\tilde{t}(y)$.

For convenience, later, we let $T_t = \{(a, \hat{\theta}); \bar{t}(a; \hat{\theta}) < t\}$ be the related cylinder set and let T_{t^0} be the observed cylinder set. We now need more detail about the asymptotic distribution of $\bar{t}(y)$ and for this, we use an embedding model.

4·2. *An embedding model*

Let $f(y; \theta, \gamma)$ be a model with an additional scalar parameter γ that could be obtained from the initial model by exponential tilting with factor $e^{\gamma t(y)}$; the relative density is trivially defined, but the normalizing constant could be unavailable. With this augmented model, the ancillary will drop dimension by unity from $(n - p)$ to $(n - p - 1)$, giving a modified ancillary $d(y)$. As a complement to $d(y)$ to designate the reduced-dimensional space, we could use $t(y)$, but rather choose $\bar{t}(y)$ which conveniently coincides with $t(y)$ on S_0. As $\bar{t}(y)$ is a smooth statistic, we have that a is also a smooth one-to-one function of $(\bar{t}, d)$.

Also, from likelihood theory (Fraser & Reid, 1993, 2001), we have that the conditional model, given an ancillary depends on third order on just a finite number k of characteristics of the ancillary; accordingly, for the analysis, we condition on the surplus characteristics and then let y and $(\hat{\theta}, \bar{t}, d)$ have fixed dimensions k and $(p, 1, k - p - 1)$, respectively.

Consider the simple normal example with $\mu_0 = 0$: the augmented model by tilting with respect to $t(y) = \bar{y}$ is just that of a sample from the $N(\mu, \sigma^2)$; and the statistic $d(y)$ corresponds to the location-scale standardized residual and under normality has no effect on the conditional model. We have then, that $\bar{t}(y) = \bar{y}/s$, so that the initial y can be replaced by $(\hat{\theta}, \bar{t}, d)$ with the effective $d(y)$ void and thus, of dimension zero.

4·3. *Example*

Consider the regression model $y = X\beta + \sigma z$, where z is distributed as $N(0, I)$ in $\mathbb{R}^n$, I is the identity matrix in $\mathbb{R}^n$ and X is the design matrix with full column rank r. Let $t(y) = x'_{r+1}y$ be a suggested test statistic, with x_{r+1} linearly independent of X and thus, not in the span $\mathcal{L}(X)$ of the vectors X. The maximum-likelihood value is then given by $(\hat{\beta}, \hat{\sigma}) = (b, s/n^{1/2})$, where b is the least-squares estimator and $s^2 = \sum_i (y_i - \hat{y}_i)^2$ is the sum of squares of residuals; let $\hat{z}$ be the unit residual as standardized by the length s. The observed maximum-likelihood surface is $S_{\hat{\theta}^0} = \{Xb^0 + s^0\hat{z}; \hat{z} \in \mathcal{S}_0\} = Xb^0 + s^0\mathcal{S}_0$, where $\mathcal{S}_0$ is taken to be the unit sphere in the $(n - r)$-dimensional space $\mathcal{L}^{\perp}(X)$ orthogonal to the span of X. From normal distributional symmetry, we have then, that the distribution of the ancillary as projected on to the maximum-likelihood surface is uniform with respect to surface volume on the sphere and correspondingly uniform relative to surface volume on $\mathcal{S}_0$. Any contour $\{y; x'_{r+1}y = t\}$ of the test statistic $t(y)$ that intersects $S_{\hat{\theta}^0}$ will do so in a sphere of one fewer dimension and divide the initial sphere into two caps. Let p^0 be the surface volume of the cap corresponding to $\{t(y) < t^0\}$ as a proportion of the surface volume of the full sphere. The modified t-statistic $\bar{t}(y)$ can be expressed as $\bar{t}(y) = \tilde{x}'_{r+1}(Xb + s\hat{z})$, where $\tilde{x}_{r+1}$ is the orthogonal projection of x_{r+1} on to $\mathcal{L}^{\perp}(X)$. The set T^0 can then be expressed as

$$T^0 = \{y : \tilde{x}'_{r+1}y/s < \tilde{x}'_{r+1}y^0/s^0\},$$

and $\tilde{t}(y) = \bar{t}(y) = \tilde{x}'_{r+1}y/s$, which is equivalent to the usual Student statistic for testing regression on x_{r+1} after eliminating regression on X.

More generally, if $t(y)$ can be written to second order as a smooth function of certain derivatives of the loglikelihood and not asymptotically equivalent to a function of $\hat{\theta}$, then the assumptions are satisfied.

4·4. *Asymptotic uniformity*

We have a statistic $\tilde{t}(y)$ and wish to show that the corresponding probability integral transformation is distributed as Un(0, 1) for each possible θ value. For this, we choose an arbitrary θ value θ_0, and use notation based on $\hat{\theta}_0 = \theta_0$ and the corresponding surface S_0, and then for convenience, take $\theta_0 = 0$; we have, of course, that the probability integral transformation of the resulting $\bar{t}(y)$ is distributed as Un(0, 1). We, thus, need to show just that the probability integral transformation of $\tilde{t}(y)$ under θ_0 is distributed as Un(0, 1); that is, to show that, to third order,

$$\Delta = \text{pr}\{\tilde{t}(y) < t; \theta_0\} - \text{pr}\{\bar{t}(y) < t; \theta_0\}.$$

For ease of notation, we work with scalar $\hat{\theta}$ and d, but the calculations extend to the vector case. We assume that $t(y)$ and $(\hat{\theta}, \bar{t}, d)$ are regular with an asymptotic normal distribution and expansions, as discussed in Cakmak et al. (1994, 1998); and we assume that $G_g(t; \theta)$ is continuously differentiable in (t, θ) and has positive density. For notational ease, we assume that the scaling has been adjusted and write $(\hat{\theta}, \bar{t}, d)$ again for the standardized variables, now asymptotically standard normal. Asymptotic independence between $\hat{\theta}$ and $(\bar{t}, d)$ follows from the ancillarity of a, and that between $\bar{t}$ and d, from the ancillarity of d in the embedding model.

We examine the probability difference Δ for a typical value $t = t_0$. Our region of interest $\{\tilde{t}(\hat{\theta}, \bar{t}, d) < t_0\}$ has boundary set $\{\tilde{t}(\hat{\theta}, \bar{t}, d) = t_0\}$, as defined implicitly. We solve and express $\bar{t}$ explicitly as a function of $(\hat{\theta}, d)$, obtaining $\bar{t} = t_0 + b(\hat{\theta}, d)$. In this form, the difference from the contour of $\bar{t}(y) = t_0$ to the contour $\tilde{t}(y) = t_0$ is described by the displacement $b(\hat{\theta}, d)$, and we can then write the probability difference Δ as

$$\Delta = \int_d \int_{\hat{\theta}} \left\{ \int_{t_0}^{t_0 + b(\hat{\theta}, d)} f(\hat{\theta}, \bar{t}, d; \hat{\theta}_0) d\bar{t} \right\} d\hat{\theta} dd, \tag{4·2}$$

where the inner integral gives a positive or negative contribution according to the sign of $b(\hat{\theta}, d)$. In the Appendix, we generate an asymptotic expansion for the boundary $b(\hat{\theta}, d)$ and then evaluate the integral (4·2). We find that $\Delta = 0$ to third order, and thus, that $\tilde{t}(y)$ is ancillary and p_{anc} is uniform to third order.

4·5. *Asymptotic uniformity under weaker conditions*

This final subsection outlines how uniformity to a lower order may be obtained under weaker assumptions. In their report, Robert and Rousseau prove under relaxed conditions, that the special p_{cpred} is asymptotically uniform to first order, irrespective of the statistic $t(y)$; thus, the same holds for p_{anc}. This robustness property with respect to the test statistic is of wide interest as the test statistic may often be too complicated to yield anything easily concerning its limiting distribution.

We can obtain second-order uniformity for p_{cpred} or p_{anc} under somewhat stronger conditions: we do not require asymptotic normality of the test statistic $t(y)$ as in Robins et al. (2000), but do require familiar regularity conditions on the model as in Bhattacharya & Ghosh (1978) and Bickel & Ghosh (1990); this provides Laplace and Edgeworth expansions for the posterior and for the maximum-likelihood estimator. We have then the following theorem, where we let $t(y)$ now be the standardized version and let Z_2 be the re-centred and renormalized vector formed from the components of the matrix of second derivatives of the loglikelihood. We denote by Z the vector formed of the components of Z_2 that are linearly independent, so that $(Z, \hat{\theta})$ has positive definite asymptotic covariance matrix.

THEOREM 1. *Given standard regularity conditions on the model* $f(y; \theta)$, *so that* $(t, u, Z) = \{t(y), \sqrt{n}(\hat{\theta} - \theta), Z\}$ *converges to a distribution with density h, then* p_{cpred} *being distributed as*

Un(0, 1) *to the second order is equivalent to*

$$\int h(t,u,Z)\mathbb{I}_{t<G^{-1}(p\,|\,u)}\left[u'Z_2u-\mathrm{tr}\{i^{-1}(\theta_0)Z_2\}\right]dtdudZ=0, \tag{4·3}$$

for all $p\in[0,1]$*, where* $G(t\,|\,u)$ *is the asymptotic conditional distribution function of* $t(y)$*, given* u*, and* $i(\theta)$ *is the Fisher information matrix.*

The proof is outlined in the Appendix.

Condition (4·3) is satisfied, in particular, when the limiting distribution is Gaussian; it is also satisfied as soon as t is asymptotically independent of $\hat\theta$, even though the limiting distribution might not be Gaussian. A third-order result could also be obtained following the preceding route, but would involve tedious calculations that might not be as enlightening as the Studentization approach in the previous section.

5. THE BOOTSTRAP p-VALUE

5·1. *Overview*

In this section, we show under moderate regularity that the bootstrap applied to a statistic $t(y)$ produces a new statistic whose distributional dependence on the parameter θ is reduced by order $n^{-1/2}$ and in three stages converts an initial statistic into the Bayesian-frequentist p-value, which is third-order ancillary. Again, we let $G_i(t;\theta)$ denote the distribution function of an ith statistic $t_i(y)$ and let $p_{i+1}=G_i(p_i;\hat\theta)$ be the plug-in modification of p_i; we show that p_3 is the Bayesian-frequentist p-value.

5·2. *Alternative coordinates*

We assume the conditions in § 4 leading to the third-order asymptotic uniformity established in § 4.4, and for ease of exposition, we work with the scalar parameter case. Also, we find it convenient to examine the model variable in the alternative coordinates $(\hat\theta,\tilde t,d)$, rather than the initial coordinates. The asymptotics that drive the theory are the asymptotics of n appropriate to this variable, not the N of any resampling used for implementation.

At the core of the bootstrap are two issues, namely the scalar statistic used at a particular iteration and the sampling distribution obtained at that stage to order points statistically in relation to the original data value point; for the latter, the sampling distribution is obtained by replacing θ by the observed maximum-likelihood estimator $\hat\theta^0$. Working with the sampling distribution for $(\hat\theta,\tilde t,d)$, we have that $(\tilde t,d)$ is ancillary and its bootstrap distribution is unchanged in an iteration. Thus, the deviation between the bootstrap distribution and the true distribution at each iteration is entirely in the conditional distribution of $\hat\theta$. Thus, when θ is replaced by $\hat\theta$, the resulting composite distribution is inflated; it is, in effect, a double compound of itself, and when standardized, would again approximate the initial standard normal, but perhaps be smoother. However, the statistic used to order sampled values relative to the original data point changes radically; we examine this next.

For the initial step in the bootstrap with resampling from a current $\hat\theta^0$ value, we have assumed that the variables are standardized, so that $(\hat\theta,\tilde t\,)$ is first-order standard normal with some detail in the Appendix and that $t(y)$ has multiple correlation ρ with $\hat\theta$, where $|\,\rho\,|<1$. As $t(\hat\theta^0,\tilde t\,)=\tilde t$, we are able to re-express $t(y)$ as $\tilde t(y)+\{\rho/(1-\rho^2)^{1/2}\}\hat\theta$ to first order.

5·3. *First-level bootstrap*

Consider the data point y^0. From the methods in the Appendix, we have that $\tilde t=\bar t+O_P(n^{-1/2})$, so that the bootstrap distribution of $(\hat\theta,\tilde t\,)$ is distributed as $N(0,I)$ to first order. From the

correlation ρ above between t and $\hat{\theta}$, it follows that $t(y)$ is distributed as $N(0, \gamma^2)$ to first order, where $\gamma^2 = 1 + \rho^2/(1-\rho^2) = (1-\rho^2)^{-1}$, and thus, that the observed p-value based on the bootstrap sample $(\hat{\theta}, \tilde{t})$ is

$$p_1 = \mathrm{pr}\{t(y) < t; \hat{\theta}_0\} = \Phi\{(1-\rho^2)^{1/2}\tilde{t}\} + O_P(n^{-1/2}). \tag{5·1}$$

As $\tilde{t}$ is distributed as $N(0, 1)$, it follows that p_1 is first-order conservative, unless $\rho = 0$. This gives the result in Robins et al. (2000), as $\rho = 0$ is equivalent to $t(y)$ having asymptotic mean independent of θ. Thus, the first-order bootstrap is distributed as Un(0, 1), if and only if $t(y)$ and $\hat{\theta}$ are asymptotically independent. Towards bootstrap results to the next order, we now use an asymptotic statistic $t_1(y)$ equivalent to p_1, which has the form $t_1(y) = \tilde{t}(y) + O_P(n^{-1/2})$, and obtain $p_1(y^0) = \mathrm{pr}\{t_1(y) \leqslant t_1(y^0); \theta_0\}$.

5·4. *Iterated bootstrap*

For the effect of a second bootstrap iteration, we expand t_1 in terms of $\hat{\theta}$ about the true, which is zero in the standardized coordinates **A2**

$$t_1(y) = \tilde{t}(y) + c_1(\tilde{t})\hat{\theta}n^{-1/2} + O_P(n^{-1}). \tag{5·2}$$

To assess the second-order effect of the last term, we need to use only the first-order standard normal distribution for $\hat{\theta}$; thus,

$$p_2 = \mathrm{pr}\{t_1(y) < t; \theta_0\} = \mathrm{pr}\{\tilde{t}(y) < t\} + O_P(n^{-1}).$$

To obtain results to the next order, we work with an asymptotic statistic equivalent to p_2, which from the preceding discussion has the form $t_2(y) = \tilde{t}(y) + O_P(n^{-1})$. We then expand t_2 as before, in terms of $\hat{\theta}$ about the true value, which is zero:

$$t_2(y) = \tilde{t} + c_2(\tilde{t})\hat{\theta}n^{-1}. \tag{5·3}$$

For the last term, we again need only the first-order standard normal distribution for $\hat{\theta}$; thus, to the third order,

$$p_3 = \mathrm{pr}\{t_2(y) < t; \theta_0\} = \mathrm{pr}\{\tilde{t}(y) < t\} + O_P(n^{-3/2}) = p_{\mathrm{anc}}(t; \hat{\theta}) + O_P(n^{-3/2}).$$

Note that, these three iterations are needed, in general, to reach $\tilde{t}$ as the next-order term in each iteration has a component of the form $(\hat{\theta} - \hat{\theta}^0)^2$ which does not disappear under the standard normal averaging.

6. The Behrens–Fisher problem

The Behrens–Fisher problem (Behrens, 1929; Fisher, 1935) has two normal distributions and interest focuses on the difference of the means: let $(y_{11}, \ldots, y_{1m})$ be a sample from the $N(\mu_1, \sigma_1^2)$, $(y_{21}, \ldots, y_{2n})$ be a sample from the $N(\mu_2, \sigma_2^2)$ and let $\delta = \mu_1 - \mu_2$ be the parameter of interest; also, let $(\bar{y}_1, \bar{y}_2)$ designate the sample means and (s_1^2, s_2^2) designate the sample variances. Despite its apparent simplicity, this problem has challenged statistical theory since its origins. Ghosh & Kim (2001) give some general background and propose a second-order default Bayesian prior as an improvement on previously available priors; simulations are given to compare with other methods. We examine this problem using the present Studentization method and report on simulations made available to us by colleagues Augustine Wong and Ye Sun of York University.

Table 1. *Behrens–Fisher problem. Coverage frequencies for the left and right bounds of the central 90% confidence intervals using Jeffreys prior, Ghosh–Kim prior, signed likelihood root and third-order Studentization procedure, with simulation limits about target*

	Left	Right
Target	5%	95%
Signed likelihood root	13·2%	86·9%
Jeffreys	0·7%	99·1%
Ghosh–Kim	1·7%	97·9%
Studentization	4·2%	95·8%
95% simulation limits	(4·9%, 5·1%)	(94·9%, 95·1%)

We consider the assessment of a particular value of δ, and for convenience and without loss of generality (Fraser, 2004), we use $\delta = 0$ and work with the sufficiency variables $(\bar{y}_1, \bar{y}_2, s_1^2, s_2^2)$. For fixed σ's, the maximum-likelihood estimator for μ is the reciprocal-variance weighted average of the sample means, and the maximum-likelihood values for the σ^2's are available by simple iteration. Thus, the maximum-likelihood surface is, in fact, just a curve in the 4-space of the sufficient statistic. We want the p-value determined from the ancillary distribution on this curve, and a convenient statistic for describing the points on the curve is $\bar{y}_1 - \bar{y}_2$. While a numerical integration or a Markov chain Monte Carlo simulation would be interesting, the p-value from this ancillary distribution is directly available from third-order likelihood approximations (Fraser et al., 1999), using an embedding model which conveniently is just the original normal model.

From the Bayesian viewpoint, various priors are available. Jeffreys (1961) proposed the composite of the right-invariant priors for the component normals:

$$\sigma_1^{-1}\sigma_2^{-1} d\mu_1 d\mu_2 d\sigma_1 d\sigma_2.$$

Ghosh & Kim (2001) proposed a second-order prior,

$$\sigma_1^{-3}\sigma_2^{-3}(\sigma_1^2/m + \sigma_2^2/n) d\mu_1 d\mu_2 d\sigma_1 d\sigma_2,$$

which appears as a type of weighted combination of the component right-invariant priors.

For a frequentist comparison short of the present ancillary approach, we used the signed likelihood-root quantity, and tested relative to its usual first-order standard normal approximation.

Simulations have been performed for various parameter combinations. We report here on just an extreme case, the smallest sample size case with $m = n = 2$, and with $\sigma_1^2/m = 2$, $\sigma_2^2/n = 1$ and $\delta = 2$; the simulation size was $N = 100\,000$. For frequentist cases, we calculated the central 90% intervals and checked for coverage of the true value, up to the left and up to the right interval bound; for the Bayesian cases, we report the results from Ghosh & Kim (2001), but we did record quite similar results ourselves. The corresponding coverage proportions are recorded in Table 1 for each of the methods.

This is an extreme case with the smallest possible sample sizes, and the individual sample t-statistics have Cauchy distributions. For this extreme case, the coverage probabilities based on third-order Studentization do just miss the 95% simulation limits under the nominal. Nonetheless, we find that third-order Studentization provides a large improvement over the signed likelihood root, the Jeffreys and the Ghosh and Kim confidence and posterior intervals. Coverage results for larger sample sizes were generally excellent.

7. More examples and discussion

For any full exponential family and for any scalar statistic $t(y)$ that is not a function of the maximum-likelihood statistic, p_{cpred} is equal to the conditional p-value as a consequence of the sufficiency and is thus, exactly uniformly distributed. This covers Example 2.2 in Bayarri & Berger (2000) for a sample from the scale exponential, using the statistic $t(y) = \min_i y_i$; it also covers the example in Gelman et al. (1995, p. 166) for a sample from the $N(\mu, \sigma^2)$ and the same minimum value statistic. Also, for the discrete case, following Davison et al. (2006), we have similar results. Consider, for example, a goodness-of-fit chi-squared test against a smooth parametric family $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$, together with a test statistic $t(y) = \sum_{j=1}^k \{N_j - np_j(\hat\theta)\}^2/\{np_j(\hat\theta)\}$ using a fixed number k of bins and N_j observations in the jth bin. Simple Taylor expansion around the true value θ implies that

$$t(y) = \sum_{j=1}^{k} \frac{[\sqrt{n}\{N_j/n - p_j(\theta)\} - \sqrt{n}(\hat\theta - \theta)p_j'(\theta)]^2}{p_j(\theta)} + O_P(n^{-1/2})$$

$$= \sum_{j=1}^{k-1} w_j^2 \left\{ \frac{1}{p_j(\theta)} + \frac{1}{p_k(\theta)} \right\} + \frac{1}{p_k(\theta)} \sum_{j \neq l} w_j w_l + O_P(n^{-1/2}),$$

where $w_j = \sqrt{n}\{N_j/n - p_j(\theta)\} - \sqrt{n}(\hat\theta - \theta)p_j'(\theta)$. Asymptotically and conditionally on $u = \sqrt{n}(\hat\theta - \theta)$, the vector $w = (w_1, \ldots, w_{k-1})$ is distributed as $N(0, \Omega)$, for some covariance matrix Ω independent of u, as soon as

$$i(\theta) = \int \frac{f_\theta(y_1; \theta)^2}{f(y_1; \theta)} dy_1 > \sum_{j=1}^{k} \frac{p_j'(\theta)^2}{p_j(\theta)},$$

where the subscript on the density denotes differentiation; otherwise, the distribution is degenerate. Therefore, t is asymptotically independent of u and p_{cpred} is second-order uniform. The discreteness of the count variables N_j inhibits the use of higher order expansions for the distribution of $t(y)$, so that third-order uniformity is not available; however, for some second-order calculations, see Davison et al. (2006).

Acknowledgement

The authors express very deep appreciation to Nancy Reid for initiating the present Bayesian-frequentist-bootstrap analysis and for continuing with many fruitful discussions of the material, during and after a research visit to CEREMADE at the University of Paris. They thank Professors Augustine Wong and Ye Sun for making available the Behrens–Fisher simulation results, and they thank the referee who suggested the Behrens–Fisher example. They offer special thanks to the editor, an associate editor and a referee for penetrating and helpful comments that led to a substantially improved and more focused manuscript. The Natural Sciences and Engineering Research Council of Canada has provided support for this research.

Appendix

Technical details

Third-order uniformity. We prove that $\Delta = 0$ for $\theta = \theta_0$. Following from § 4·4, we use standardized notation and have a standard normal asymptotic distribution for $(\hat\theta, \bar t, d)$, and of course, marginally for $(\bar t, d)$. We also have a statistic $\tilde t(\hat\theta, \bar t, d)$ that coincides with $\bar t$ when $\hat\theta = 0$, and satisfies marginal distribution

conditions for other $\hat\theta$ values. In addition, we want to show that $\bar t = t_0$ and $\tilde t = t_0$ define a difference set expressed by $b(\hat\theta, d)$ that has probability content 0 to third order.

We expand $b(\hat\theta, d)$ in a Taylor series around $(0, 0)$ to order $O(n^{-3/2})$ and make use of $b(0, d) = 0$:

$$b(\hat\theta, d) = (a_{10} + a_{11}d/n^{1/2} + a_{12}d^2/2n)\hat\theta + (a_{20}/n^{1/2} + a_{21}d/n)\hat\theta^2/2 + (a_{30}/n)\hat\theta^3/6\cdot \tag{A1}$$

The definition of the boundary on a surface $S_{\hat\theta}$ gives

$$\int_d \left\{ \int_{t^0}^{t^0 + b(\hat\theta, d)} g(\bar t, d) d\bar t \right\} dd = O(n^{-3/2}), \tag{A2}$$

and we have the very strong result, that this holds for each $\hat\theta$ in moderate deviations.

The complicating feature for analyzing the difference set is the presence in (A1) of the cross term $a_{11}d\hat\theta/n^{1/2}$, which describes how $\tilde t$ has a twist relative to $\bar t$. Our approach is to apply a transformation to the space of $(\hat\theta, \bar t, d)$ to remove the twist in the initial statistic $t(y)$; of course, the transformation does not alter the distribution of $\tilde t(y)$ and, as we will see, does not alter the marginal distribution $g(\bar t, d)$. To be specific, at each $\hat\theta$ level, a rotation is applied to $(\bar t, d)$ to make the contours of the initial statistic $t(y)$ second-order parallel to the d axis at $d = 0$; the contours of $\tilde t(y)$ being just $t(y)$ contours then have the same property. We then examine the difference set after the transformation.

An $O(n^{-1/2})$ rotation of the distribution of $(\bar t, d) \mid \hat\theta$ does not, to third order, affect the fourth derivative terms, but does alter to $O(n^{-1})$, the third derivative terms by an amount proportional to the rotation, as shown below. When this is averaged over $\hat\theta$, we have that $g(\bar t, d)$ is unchanged. We have, however, that the structure of $\tilde t(y)$ is now more easily examined as it is second-order parallel to the d axis direction at $d = 0$.

We again expand the boundary, but now in the new notation:

$$b(\hat\theta, d) = (a_{10} + a_{11}d/n + a_{12}d^2/2n)\hat\theta + (a_{20}/n^{1/2} + a_{21}d/n)\hat\theta^2/2 + (a_{30}/n)\hat\theta^3/6. \tag{A3}$$

Then, using the definition (A2) for the boundary, we have that the boundary (A1) can be written as

$$b(\hat\theta, d) = \{a_{11}d/n + a_{12}(d^2 - 1)/2n\}\hat\theta + (a_{21}d)\hat\theta^2/2n + a_{30}\hat\theta^3/6n\cdot$$

Then substituting in (4·2), expanding the inner upper limit of integration and using the symmetry of the density function gives zero to the third order, thus establishing the third-order uniformity.

Effect of rotations. In two dimensions, a rotation through an angle $a/n^{1/2}$ involves a cosine written as $1 - a/n$ and a sine written as $a/n^{1/2}$, all to third order. Then, for the effect on the log of a standardized density, there is no effect on the quadratic term, the cubic term is unchanged to the second order, but acquires an $O(n^{-1})$ adjustment, and the fourth-order terms are unchanged; the angle $a/n^{1/2}$ appears linearly in the second-order adjustment.

Second-order uniformity: Connection to conditional p-value. The proof is straightforward and from (3·1), we can write

$$f(y \mid \hat\theta^0; \theta) = \frac{\mathbb{1}_{S_{\hat\theta^0}} \exp \ell(\theta; y) \mid \ell_{\theta;y} \mid^{-1} \mid \hat j \mid}{f(\hat\theta^0; \theta)},$$

where $f(\hat\theta; \theta)$ is the marginal density of $\hat\theta$. From (1·2) and using the ratio relative to a true density-labelled θ_0 density, we obtain

$$\begin{aligned} p^0_{\text{cpred}} &= \int_\Theta \text{pr}\{t(y) < t^0 \mid \hat\theta^0; \theta\}\pi(\theta \mid \hat\theta^0) d\theta \\ &= \int_{S_{\hat\theta^0}} f(y \mid \hat\theta^0; \theta_0) \mathbb{1}_{t(y) < t^0} f(\hat\theta^0; \theta_0) \frac{\int_\Theta e^{\ell(\theta;y) - \ell(\theta_0;y)} \pi(\theta) d\theta}{\int_\Theta f(\hat\theta^0; \theta)\pi(\theta) d\theta} da\cdot \end{aligned} \tag{A4}$$

Under the conditions in § 3, we use an Edgeworth expansion (Bhattacharya & Ghosh, 1978) for the density of the maximum-likelihood estimator which is uniform in θ, together with the usual Laplace

expansions as in § 3 and obtain

$$\int_{\Theta} f(\hat{\theta}^0;\theta)\pi(\theta)d\theta = 1 + O_P(n^{-1}).$$

For the numerator integral in (A4), we use a Laplace expansion of the integral, together with an Edgeworth expansion of the density $f(\hat{\theta};\theta_0)$ and obtain

$$p^0_{\text{cpred}} = \int_{S_{\hat{\theta}^0}} f(y \mid \hat{\theta}^0;\theta_0)\mathbb{I}_{t(y)<t^0}A(y)dy_c + R_n n^{-1}, \tag{A5}$$

with the adjustment factor $A(y)$ given as

$$A(y) = \left\{1 + H_1(u_0)/n^{1/2}\right\}\left\{1 + \frac{1}{2}\left(u_0'\{\hat{j} - i(\theta_0)\}u_0 - \text{tr}[i(\theta_0)^{-1}\{\hat{j} - i(\theta_0)\}]\right)\right\},$$

where $u^{1/2}(\theta_0 - \hat{\theta})$, $u_0 = n^{1/2}(\theta_0 - \hat{\theta}^0)$, $H_1(u_0)$ is an odd polynomial function, and $R_n = O_P(1)$. In these calculations, we make no assumption concerning the behaviour of $t(y)$, but we do invoke the usual regularity conditions on the likelihood function. The equation (A4) shows that p_{cpred} is first-order equivalent to the conditional p-value: $\text{pr}\{t(y) < t^0 \mid \hat{\theta}^0;\theta_0\} = p_{\hat{\theta}^0;\theta_0}(t^0)$ for any $t(y)$, and is thus uniformly distributed to first order.

Second-order uniformity: Discrepancy from conditional p-value. We now examine the discrepancy between p_{cpred} and $p_{\hat{\theta};\theta_0}(t^0)$:

$$\Delta_c = \text{pr}\left(p_{\text{cpred}} < p\right) - \text{pr}\left\{p_{\hat{\theta};\theta_0}(t^0) < p\right\}. \tag{A6}$$

Consider the expression (A4) and the form of the adjustment factor $A(y)$ and let $z_2 = n^{1/2}\{j(\theta_0) - i(\theta_0)\}$. Then, $\hat{j} - i(\theta_0) = z_2 n^{-1/2} - u_0'\mu_3(\theta_0)n^{-1/2} + O_P(n^{-1})$, where $u_0'\mu_3(\theta_0)$ is the $p \times p$ matrix whose (a, b) component is $\sum_{r=1}^{p} u_r^0 E_{\theta_0}\{D_{abr}\log f(X;\theta_0)\}$ and D_{abr} designates the third derivative with respect to the parameter coordinates a, b, r. Let

$$W(t^0, u^0) = \int_{S_{\hat{\theta}^0}} f(y \mid \hat{\theta}^0, \theta_0)\mathbb{I}_{t(y)<t^0}[u_0'z_2u^0 - \text{tr}\{i(\theta_0)^{-1}z_2\}]/2dy_c.$$

The calculations at (A4) then show that,

$$p_{\text{cpred}} = p_{\hat{\theta}^0;\theta_0}(t^0)\left\{1 + H_2(u_0)n^{-1/2}\right\} + W(t^0, u_0)n^{-1/2} + O_P(n^{-1}),$$

where H_2 is an odd polynomial function.

We now compare $\text{pr}(p_{\text{cpred}} < p;\theta_0)$ with $p = \text{pr}(p_{\hat{\theta}^0;\theta_0}(t^0) < p;\theta_0)$. For this, we assume that some standardized version of $t(y)$, $t_s(y)$ say, has, as n goes to infinity a limiting conditional density, given $\hat{\theta}^0$ which is positive under the θ_0 distribution. We denote $t_s(y^0)$ by t_s^0 and let $G(\cdot \mid \hat{\theta}, \theta_0)$ be the distribution function of $t_s(y)$, given $\hat{\theta}$ and θ_0. We also let E_{θ_0} designate the expectation taken with respect to $f(y;\theta_0)$. We assume that the transformation from $t(y)$ to $t_s(y)$ is monotone increasing, in other words, that $\{t(y) < t^0\} \cap S_{\hat{\theta}^0} = \{t_s(y) < t_s^0\} \cap S_{\hat{\theta}^0}$. In the following expression, the probabilities are calculated under $f(y;\theta_0)$ and we use the simpler notation $G^{-1}(p \mid u)$ instead of $G^{-1}(p \mid u, \theta_0)$, and work up to order $O(n^{-1})$:

$$\begin{aligned}
\Delta_c &= \text{pr}(p_{\text{cpred}} < p;\theta_0) - \text{pr}(p_{\hat{\theta},\theta_0} < p;\theta_0)\\
&= \text{pr}(G^{-1}(p \mid u) < t_s < G^{-1}[p\{1 - H_2(u)n^{-1/2}\} - W_n\{G^{-1}(p \mid u), u\}n^{-1/2} \mid u])\\
&\quad - \text{pr}(G^{-1}(p \mid u) \geqslant t_s^0 > G^{-1}[p\{1 - H_2(u)n^{-1/2}\} - W_n\{G^{-1}(p \mid u), u\}n^{-1/2} \mid u])\\
&= -E_{\theta_0}[H_2(u)n^{-1/2} - W_n\{G^{-1}(p \mid u), u\}n^{-1/2}].
\end{aligned}$$

As H_2 is an odd polynomial function in u, we find that,

$$\Delta_c = (1/2)n^{-1/2}\int f_n(t, u, z_2)\mathbb{I}_{t<G^{-1}(p \mid u)}[u'z_2u - \text{tr}\{i^{-1}(\theta_0)z_2\}]dtdudz_2,$$

where $f_n(t, z_2, u)$ is the joint density of $\{t_s(y), z_2, n^{1/2}(\theta_0 - \hat{\theta}\}$. If $f_n(t, z_2, u)$ converges almost surely to a density function $f(t, u, z_2)$, then $\Delta_c = 0$, if and only if

$$\int f(t, u, z_2) \mathbb{I}_{t<G^{-1}(p\,|\,u)}[u' z_2 u - \mathrm{tr}\{i^{-1}(\theta_0) z_2\}] dt du dz_2 = 0. \tag{A7}$$

This completes the proof of Theorem 1.

REFERENCES

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardised signed log likelihood ratio. *Biometrika* **73**, 307–22.

BAYARRI, M. J. & BERGER, J. O. (2000). *p*-values for composite null models. *J. Amer. Statist. Assoc.* **95**, 1127–42.

BEHRENS, W. V. (1929). Ein Beitrag zur Fehlerbereichnung bei wenigen Beobachtungen. *Landwirt. Jahr.* **68**, 807–37.

BERAN, R. J. (1988). Pre-pivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.* **83**, 687–97.

BHATTACHARYA, R. N. & GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–51.

BICKEL, P. J. & GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—A Bayesian argument. *Ann. Statist.* **18**, 1070–90.

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

CAKMAK, S., FRASER, D. A. S., MCDUNNOUGH, P., REID, N. & YUAN, X. (1998). Likelihood centered asymptotic model: Exponential and location model versions. *J. Statist. Plann. Inference* **66**, 211–22.

CAKMAK, S., FRASER, D. A. S. & REID, N. (1994). Multivariate asymptotic model: Exponential and location approximations. *Utilitas Math.* **46**, 21–31.

DAVISON, A. C., FRASER, D. A. S. & REID, N. (2006). Improved likelihood inference for discrete data. *J. R. Stat. Soc. B* **68**, 495–508.

FISHER, R. A. (1935). The logic of inductive inference. *J. R. Stat. Soc.* **98**, 39–54.

FRASER, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–39.

FRASER, D. A. S. (2004). Ancillaries and conditional inference (with Discussion). *Statist. Sci.* **19**, 333–69.

FRASER, D. A. S. & REID, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. *Statist. Sinica* **3**, 67–82.

FRASER, D. A. S. & REID, N. (1995). Ancillaries and third order significance. *Utilitas Math.* **47**, 330–53.

FRASER, D. A. S. & REID, N. (2001). Ancillary information for statistical inference. In *Empirical Bayes and Likelihood Inference*, Ed. S. E. Ahmed and N. Reid, pp. 185–207. New York: Springer-Verlag.

FRASER, D. A. S., REID, N. & WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–64.

GELMAN, A., CARLIN, J. B., STERN, H. & RUBIN, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

GHOSH, M. & KIM, Y.-H. (2001). The Behrens–Fisher problem revisited: A Bayesian-frequentist synthesis. *Canad. J. Statist.* **29**, 5–17.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.

LUGANNANI, R. & RICE, S. O. (1980). Saddlepoint approximation for the distribution of the sums of independent random variables. *Adv. in Appl. Probab.* **12**, 457–90.

ROBINS, J. M., VAN DER VAART, A. & VENTURA, V. (2000). The asymptotic distribution of *p*-values in composite null models. *J. Amer. Statist. Assoc.* **95**, 1143–56.

ROUSSEAU, J. (2007). Approximating interval hypotheses : *p*-values and Bayes factor. In *Bayesian Statistics* 8, Ed. J. M. BERNARDO, J. O. BERGER, A. P. DAWID and A. F. M. SMITH. Oxford: Oxford University Press.

STUDENT (1908). The probable error of a mean. *Biometrika* **6**, 1–25.

[*Received January* 2005. *Revised August* 2007]