Statistical Science
2004, Vol. 19, No. 2, 333–369
DOI 10.1214/088342304000000323
© Institute of Mathematical Statistics, 2004

Ancillaries and Conditional Inference

D. A. S. Fraser

Abstract. Sufficiency has long been regarded as the primary reduction procedure to simplify a statistical model, and the assessment of the procedure involves an implicit global repeated sampling principle. By contrast, conditional procedures are almost as old and yet appear only occasionally in the central statistical literature. Recent likelihood theory examines the form of a general large sample statistical model and finds that certain natural conditional procedures provide, in wide generality, the definitive reduction from the initial variable to a variable of the same dimension as the parameter, a variable that can be viewed as directly measuring the parameter. We begin with a discussion of two intriguing examples from the literature that compare conditional and global inference methods, and come quite extraordinarily to opposite assessments concerning the appropriateness and validity of the two approaches. We then take two simple normal examples, with and without known scaling, and progressively replace the restrictive normal location assumption by more general distributional assumptions. We find that sufficiency typically becomes inapplicable and that conditional procedures from large sample likelihood theory produce the definitive reduction for the analysis. We then examine the vector parameter case and find that the elimination of nuisance parameters requires a marginalization step, not the commonly proffered conditional calculation that is based on exponential model structure. Some general conditioning and modelling criteria are then introduced. This is followed by a survey of common ancillary examples, which are then assessed for conformity to the criteria. In turn, this leads to a discussion of the place for the global repeated sampling principle in statistical inference. It is argued that the principle in conjunction with various optimality criteria has been a primary factor in the long-standing attachment to the sufficiency approach and in the related neglect of the conditioning procedures based directly on available evidence.

Key words and phrases: Ancillaries, conditional inference, inference directions, likelihood, sensitivity directions, pivotal.

1. INTRODUCTION

Sufficiency has a long and firmly established presence in statistical inference; it provides a major simplification for many familiar statistical models and often gives a variable with a simple relationship to the parameter. The assessment of this reduction of the statistical problem is done implicitly in terms of repeated performances of the full investigation under study; call this the global repeated sampling principle.

Certain conditional methods have almost as long a history in statistical theory, but rather strangely are discussed and used extremely rarely. In Section 2 we examine two important early papers (Welch, 1939; Cox, 1958) that discuss conditional inference and quite extraordinarily come to opposite views on the merits of conditioning. Note, however, that the two papers differ in their orientation toward statistics, the first be-

D. A. S. Fraser is Professor, Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3 (e-mail: dfraser@utstat.toronto.edu).

ing decision theoretic and the second being inferential. The conditional approach examined in the second paper does violate, however, the global repeated sampling principle, because the model used for statistical inference refers just to repeated performances of the measurement instrument that actually gave the observed data.

In Sections 3 and 4 we examine two simple normal measurement contexts and find of course that sufficiency produces the essential variables for forming tests and confidence procedures. In each of these sections we then progressively replace the normality and location relationship by alternative conditions concerning the distribution form and the continuity in the parameter–variable relationship. We find that sufficiency is no longer available and that definitive conditioning procedures from likelihood asymptotics give the appropriate variable with a simple relationship to the parameter. We also find that if these procedures are applied to the initial location normal cases, they duplicate the results from sufficiency. We are thus led to the view that sufficiency and the global repeated sampling principle together have been a major delaying factor to recognition of the conditional approach. These two sections also include an overview of the methods provided by recent likelihood theory; these methods in wide generality produce highly accurate p values and highly accurate likelihoods for component parameters of interest. The methods are assessed in terms of just the measurement processes that gave the actual data; accordingly the methods do not conform to the global repeated sampling principle.

In Section 5 we examine criteria for the use of conditioning and for the construction of statistical models for purposes of statistical inference. In Section 6 we survey some traditional ancillary examples and how these relate to the criteria in Section 5. Then in Section 7 we consider the role of global repeated sampling assessments and how these assessments interact with familiar optimization criteria.

2. TWO MEASUREMENT INSTRUMENTS

As part of a general discussion of statistical inference, Cox (1958) considered two measurement instruments, both unbiased and normal, but with different variances; the context includes an equally likely random choice of which instrument to use to make a single measurement on a parameter θ. The example was also discussed by Cox and Hinkley (1974, page 96) and Casella and Berger (2002), but despite its importance seems not to appear in most texts on statistics. A somewhat related example was considered earlier by Welch (1939).

Cox initially considered the appropriate sample space for statistical inference, but then developed it in terms of conditioning on an ancillary statistic (Fisher, 1925, 1934, 1935). A statistic is *ancillary* if it has a fixed distribution, that is, if its distribution is free of the parameter in the problem. A related notion of *reference set* was introduced by Fisher (1961).

Cox noted that the indicator variable, say a, for the choice of measurement instrument has a fixed distribution with probability $1/2$ at $a = 1$ or 2 according as the first or second instrument is used; a is thus ancillary. The Fisher conditionality approach is to condition on the observed value of the ancillary a and thus to use the normal model that corresponds to the instrument that actually made the measurement. From a practical perspective this seems very natural, and some related theory is developed in Section 5.

Cox (1958) and Cox and Hinkley (1974) considered the two measurement instruments example numerically in terms of the testing of a point null hypothesis. We recast this in terms of confidence intervals.

EXAMPLE 2.1. For the two measurement instruments we assume that the standard deviations are $100\sigma_0$ and σ_0, respectively. A 95% confidence interval based on the measurement instrument actually used has the form

$$\begin{aligned} &(y \pm 196\sigma_0) \quad \text{if } a = 1, \\ &(y \pm 1.96\sigma_0) \quad \text{if } a = 2. \end{aligned} \tag{2.1}$$

Suppose now that we consider the problem in terms of ordinary confidence methods and then invoke some optimality criterion such as minimizing the average length of the confidence interval. We might then prefer the 95% confidence interval

$$\begin{aligned} &(y \pm 164\sigma_0) \quad \text{if } a = 1, \\ &(y \pm 5\sigma_0) \quad \text{if } a = 2. \end{aligned} \tag{2.2}$$

We can see that this has 90% conditional confidence if $a = 1$ and has almost certain conditional confidence if $a = 2$; and we then see that this averages and does give the desired 95% overall confidence. The first interval (2.1) has average length $197.96\sigma_0$ and the second interval (2.2) has a substantially shorter average length $169\sigma_0$. The second interval (2.2) acquires this shorter average length within the overall 95% confidence by presenting a slightly longer interval in the

precise measurement case $a = 2$ and a very much shorter interval in the imprecise measurement case $a = 1$. A similar argument in the hypothesis testing context shows that the overall power of a size α test analogous to (2.1) can be increased by allowing a slight decrease in power in the precise measurement case with a large increase in power in the imprecise case. The raw message for applications from this optimality approach is, "Get your minimum length or maximum power where it is cheap in terms of contribution to confidence level or test size." Here, we are viewing this in terms of a random choice of measurement instrument, but we could also view it in a larger context, say that of a major consultant who advertised that his or her 95% intervals are shorter on average. His or her policy might be to give the clients with more accurate measuring instruments longer intervals and give the clients with less precise instruments shorter intervals. He or she thus maintains the overall confidence level at 95%, but is able to provide shorter confidence intervals on average than some other confidence interval provider who might feel constrained to restrict the coverage probability at 95% for each instrument used. This would perhaps not be done overtly, but is presented here because of its patent violation of good sense and because the phenomenon as just described is intrinsically embedded in almost all applications when an optimality approach is used. The next example will clearly display this strange trade-off.

Let us consider the two measurement instruments example in Welch (1939). For this we have two measurements y_1 and y_2 of θ with independent errors that are uniform $(-1/2, 1/2)$; there is nothing special in the choice of a uniform distribution other than simplicity and its clear departure from normality in the form of having very short tails.

EXAMPLE 2.2. The variable (y_1, y_2) has a uniform density equal to 1 on the unit square $(\theta - 1/2, \theta + 1/2) \times (\theta - 1/2, \theta + 1/2)$. If we take $z_1 = \overline{y}$ and $z_2 = (y_2 - y_1)/2$ we see easily that z_2 has the triangular density

$$p(z_2) = 2(1 - 2|z_2|)$$

on the interval $(-1/2, +1/2)$ and that $z_1|z_2$ has the uniform density

$$p(z_1|z_2) = (1 - R)^{-1}$$

on the interval $\{\theta \pm (1 - R)/2\}$, where $R = 2|z_2|$ is the sample range for (y_1, y_2). Obviously z_2 is ancillary, and clearly, it is describing the physical nature of the sample, the within-sample characteristics typically presented by residuals. Its analog in more general contexts is called a *configuration statistic*. A β-level confidence interval conditional on the ancillary R is then given as

$$\{\overline{y} \pm \beta(1 - R)/2\}; \tag{2.3}$$

the $\beta = 75\%$ acceptance region for testing a value θ corresponding to (2.3) is recorded in Figure 1(a).

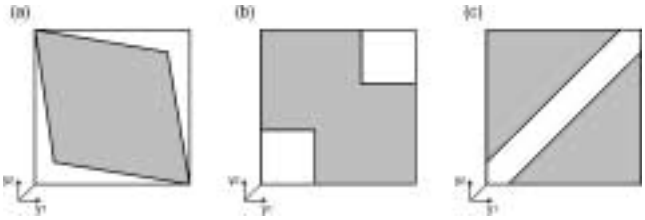

FIG. 1. *Acceptance region in* $(\theta \pm 1/2)' \times (\theta \pm 1/2)$: (a) *conditional*; (b) *max power*; (c) *min length*.

A likelihood ratio argument can be used to obtain the most powerful (often called, rather inappropriately, most accurate) unbiased or symmetric β-level interval:

$$\begin{aligned} &\left\{\overline{y} \pm \frac{1 - R}{2}\right\} && \text{if } R > \left(\frac{1 - \beta}{2}\right)^{1/2}, \\ &\left[\overline{y} \pm \left\{\frac{1 + R}{2} - \left(\frac{1 - \beta}{2}\right)^{1/2}\right\}\right] \\ &&& \text{if } R \le \left(\frac{1 - \beta}{2}\right)^{1/2}. \end{aligned} \tag{2.4}$$

This interval gives the full range of possible θ values for large R. The $\beta = 75\%$ acceptance region for testing a value θ that corresponds to (2.4) is recorded in Figure 1(b). Similarly a length-to-density ratio argument can be used to obtain the shortest on average symmetric $\beta = 75\%$ confidence interval, which has the form

$$\begin{aligned} &\left(\overline{y} \pm \frac{1 - R}{2}\right) && \text{if } R > (1 - \sqrt{\beta}), \\ &\varnothing && \text{if } R \le (1 - \sqrt{\beta}). \end{aligned} \tag{2.5}$$

This confidence interval is either the full range of possible θ values or the empty set; the acceptance region that corresponds to this confidence interval (2.5) is recorded in Figure 1(c). Again we see that we can reduce average length or gain power by removing the requirement that the confidence level be controlled conditionally. Also we note that the two optimality criteria lead to quite different confidence intervals, both with rather extreme properties. In particular, the most

powerful 75% interval is the full range of possible values some of the time (and then always covers θ); the minimum average length 75% interval is the empty set 25% of the time (and then never covers θ). These are certainly extraordinary and unpleasant properties that hopefully would not easily be explained away to a client.

Cox (1958) offered general support for the conditionality approach from Fisher (1961). Welch (1939) invoked optimality conditions and argued against conditionality using a similar example. Similar opposite viewpoints can be found in Fraser and McDunnough (1980) and Brown (1990). The viewpoint from Fisher and Cox and supported here is that anomalies such as these argue in fact against the appropriateness of the optimality approach applied on a global or repeated sampling basis. Indeed optimality criteria and global probability assessments lead generally to analyses that do not acknowledge clear and evident characteristics in particular circumstances.

3. SCALAR PARAMETER MEASUREMENT EXAMPLES

3.1 Measurement with Known Normal Error

Consider a very simple example with known normal measurement error: Let y be normal (θ, σ_0^2) with observed data y^0. The observed likelihood function is available immediately,

$$L^0(\theta) = c\exp\left\{-\frac{1}{2\sigma_0^2}(y^0-\theta)^2\right\} = c\phi\left(\frac{y^0-\theta}{\sigma_0}\right), \tag{3.1}$$

where ϕ is the standard normal density. It has maximum value at y^0, has normal shape and is scaled by σ_0, and it displays how much probability sits at the data point under various possible θ values. The observed p-value function is

$$p^0(\theta) = \Phi\left(\frac{y^0-\theta}{\sigma_0}\right), \tag{3.2}$$

where Φ is the standard normal cumulative distribution function. This records the left tail probability at the data point y^0 when the parameter has the value θ; it can be viewed as presenting the percentile position of the data y^0 relative to the distribution for y that is indexed by θ. In more general contexts we can typically interpret "left" in the sense of smaller maximum likelihood value.

An end user might be interested in a right tail or a two-tailed p value, but we take the left p value as in (3.2) as the elemental or primitive inference summary from which the others can be derived; this is in accord with the conventional definition for a distribution function. The p value records the percentile position of the data point relative to the distribution indexed by θ.

Suppose now that we are in the sampling context with data $(y_1^0, \ldots, y_n^0)$. The familiar sufficiency argument gives a reduction to the sample average $\overline{y}$. The observed likelihood and observed p value $p^0(\theta)$ are then available as $L^0(\theta)$ in (3.1) and $p^0(\theta)$ in (3.2), but with y^0 replaced by $\overline{y}^0$ and σ_0 replaced by $\sigma_0/\sqrt{n}$. The likelihood function and the p-value function give two complementing assessments of the unknown θ.

3.2 Measurement with Known Nonnormal Error

Suppose now that we know the shape and scaling of the error distribution, say the logistic or even the Student distribution with 7 degrees of freedom often cited as having an appropriate thickness in the tails. Let $f(e)$ be the error density and suppose for convenience that $f(e)$ has been centered at $e = 0$. For an asymmetric distribution there would be arbitrariness in the centering choice, but this has no effect of substance on the considerations here. We thus consider the measurement y with model $f(y-\theta)$ together with observed data value y^0.

For some of the discussion we can be still more general and consider y with model $f(y;\theta)$ together with observed data y^0. Then, as in Section 3.1, we have that the observed likelihood function is

$$L^0(\theta) = cf(y^0, \theta) \tag{3.3}$$

and the observed p-value function is

$$p^0(\theta) = F(y^0; \theta), \tag{3.4}$$

where F is the cumulative distribution function that corresponds to f. Confidence intervals are available immediately by the standard inversion of (3.4); for example, the central 95% interval $(\hat{\theta}_L, \hat{\theta}_U)$ is obtained by solving

$$p^0(\hat{\theta}_L) = 0.975, \quad p^0(\hat{\theta}_U) = 0.025,$$

where we assume for convenience that the distribution shifts to the right with increasing θ. For the moment we are examining just the case with a single measurement y.

A primary theme in this paper is that observed likelihood and observed p-value functions are primary inference elements, and are available in wide generality and with little computational difficulty. Toward this, a natural next step is to consider a sampling situation or, more generally, a multiple response situation. With nonnormal $f(y;\theta)$, or with varying $f_i(y_i,\theta)$, the simple reduction by sufficiency is almost never available. We will see, however, that definitive conditioning is readily available, and for this we first examine the case with direct location modeling.

3.3 Multiple Measurements with Location Parameter

Consider a sample $(y_1,\ldots,y_n)$ from a distribution $f(y-\theta)$. The residual vector $a(y)=(y_1-\overline{y},\ldots,y_n-\overline{y})'$ describes the pattern within the sample and is easily seen to have a fixed parameter-free distribution. To make this transparent we write $y_i=\theta+e_i$, where $(e_1,\ldots,e_n)$ is a sample from the error distribution $f(e)$. Then $a(y)=(y_1-\overline{y},\ldots,y_n-\overline{y})'=(e_1-\overline{e},\ldots,e_n-\overline{e})'=a(e)$; this clearly shows that the distribution for $a(y)$ depends only on the error sample $(e_1,\ldots,e_n)$ and is thus free of the parameter θ. The residual vector is sometimes called a *configuration statistic*: It is ancillary and, in addition, directly presents key observable characteristics of the underlying or latent errors; recall the discussion in Example 2.2.

Now consider observed data $(y_1^0,\ldots,y_n^0)$. From this we know that the ancillary $a(y)$ has observed value $a^0=a(y^0)$ and then, in accord with the conditionality approach, we work with the conditional model given the observed configuration $a(y^0)=a^0$. This conditional model can be derived in various ways and can be expressed as a density for, say, $\overline{y}$ given a^0; that is,

$$g(\overline{y}|a^0;\theta)=kf(\overline{y}+a_1^0-\theta)\cdots f(\overline{y}+a_n^0-\theta),$$

where k is the norming constant and in most applications would be obtained by numerical integration at the same time as a probability of interest was calculated by the appropriate numerical integration.

The usual derivation of a conditional model requires the calculation of a Jacobian to the new variables, here $\overline{y}$ and $a(y)$. This can be presented quite simply here by noting that the new variables are both linear and in fact are orthogonal: $\overline{y}$ records position in the direction of the one-vector, and $a(y)$ records position in the directions of the orthogonal complement $\mathcal{L}^{\perp}(1)$ of the one-vector. In effect we are finding the distribution of one coordinate given the remaining coordinates, all after an orthogonal transformation. The Jacobian is thus constant and the conditional density up to a norming constant is available as just the full density reexpressed in terms of the new variables.

For an alternative expression for the conditional distribution we note that the observed likelihood is

$$L^0(\theta)=cf(y_1^0-\theta)\cdots f(y_n^0-\theta) \tag{3.5}$$

and we can thus write

$$g(\overline{y}|a^0;\theta)=L^0(\theta-\overline{y}+\overline{y}^0), \tag{3.6}$$

where the proportionality constant c in (3.5) is taken to be the appropriate norming constant k just described.

The observed p value is then obtained as the appropriate integral of the conditional model:

$$\begin{aligned} p^0(\theta) &= \int_{-\infty}^{\overline{y}^0} g(\overline{y}|a^0;\theta)\,d\overline{y} \\ &= \int_{-\infty}^{\overline{y}^0} L^0(\theta-\overline{y}+\overline{y}^0)\,d\overline{y} \\ &= \int_{\theta}^{\infty} L^0(\overline{\theta})\,d\overline{\theta}. \end{aligned} \tag{3.7}$$

Note that this has been expressed as an integral of observed likelihood and in fact happens to be the Bayesian survival probability derived from the flat or uniform prior $\pi(\theta)=k$. Also note that for the special case with normal error density we have that (3.5) and (3.7) duplicate the results (3.1) and (3.2) for the normal case. We thus see that sufficiency works essentially just for the simple normal model, but that conditioning works in the general case and in doing so reproduces the special earlier result for the normal case.

When we examine a still more general case in the next section, we will see that for implementation we do not need to know the full ancillary or full configuration statistic. It suffices to know just the nature of the conditioning at the observed data point. In fact we will see that highly accurate p values are available quite generally using just the observed likelihood $L^0(\theta)$ and the gradient of the log-likelihood $l(\theta;y)$ calculated at the data point in what we call a *sensitivity* direction, a direction, say v, in which the ancillary is constant in value. At this stage, it is easy and of interest to see what such a vector would be like. If $a(y)=a^0$, then a point y has projection $(y_1-\overline{y},\ldots,y_n-\overline{y})=(a_1^0,\ldots,a_n^0)$ on the orthogonal complement $\mathcal{L}^{\perp}(1)$ of the one-vector. The points with such fixed projection lie on the line $a^0+\mathcal{L}(1)$ and a tangent to the line is of course in the direction of the one-vector; thus $v=1$ or some

multiple of it. This vector tells us what the ancillary looks like near the observed data point; it just happens here in this location model case that the tangent vector is the same vector at all the possible data points. More generally, for accurate inference we do not need the appropriate ancillary explicitly; it suffices to have just its tangent vector at the data point, and we will see that this is easily obtained.

3.4 Multiple Measurements with Scalar Parameter

With a location model $f(y-\theta)$ we see that a change in the parameter θ causes a shift of the distribution by a corresponding amount. We can refer to this as y change caused by θ change, and write $dy/d\theta$, which, for this simple location model, has the value 1; for this we should understand clearly that the derivative is taken for a fixed value of the error quantity $c = y - \theta$. For a more general case with distribution function $F(y;\theta)$, we note that a small increment δ to the parameter from a value θ causes a shift of the distribution by an amount $v\delta$ at a point y, where

$$v = -\frac{\partial F(y;\theta)/\partial\theta}{\partial F(y;\theta)/\partial y}.$$

For this we take the probability position of the point y to be given by its p value $F(y;\theta)$, and we hold this mathematically fixed as we examine how θ change causes y change, using the total differential of F. Correspondingly we call $v = v(\theta)$ the sensitivity of y relative to θ. Indeed this agrees with the sensitivity mentioned for the location case in the preceding section.

When we speak of the probability position or the p value of a point y we are presenting the same information as the traffic monitor when he or she asserts that you are driving at the 99.5 percentile; the statistical position relative to other cars would be clearly understood.

Now consider independent measurements $y_1, \ldots, y_n$, where y_i has model $f_i(y_i;\theta)$ with distribution function $F_i(y_i;\theta)$. A change δ in θ causes in the manner just described a change $v_i\delta$ in the coordinate y_i; this gives the sensitivity

$$v_i(\theta) = -\frac{\partial F_i(y_i;\theta)/\partial\theta}{\partial F_i(y_i;\theta)/\partial y_i} \tag{3.8}$$

for the ith coordinate. With a data point $(y_1^0, \ldots, y_n^0)$ we could then reasonably be interested in the sensitivity vector

$$v = \{v_1(\hat\theta^0), \ldots, v_n(\hat\theta^0)\}' \tag{3.9}$$

at the observed data y^0 which describes change corresponding to change in θ at the maximum likelihood value $\theta = \hat\theta^0$.

As a simple example consider the regression model with independent coordinates and $y_i = \beta x_i + e_i$, where the errors have a known distribution and the covariate values x_i are known. The effect of change in β on the response vector is then given as $v = x$, which is the very simple design matrix. A second example is given at the end of this subsection.

In any case, likelihood theory establishes v as the tangent vector to an approximate ancillary suitable for highly accurate likelihood inference. Whether the physical suggestion of sensitivity under parameter change has persuasive value, it does provide the basis for the arguments that lead to the ancillary property (Fraser and Reid, 1995, 2001).

Recent likelihood inference theory focuses on the likelihood function and in wide generality produces results that have high accuracy as opposed to the first-order accuracy when standard normality is ascribed to the score or maximum likelihood departure measures. By high accuracy we mean that the approximation errors are of order $O(n^{-3/2})$, where n is the sample size or some equivalent indicator of data dimension, and being based on likelihood, the approximations can have extraordinary accuracy even with very small samples.

For these recent likelihood approximations we need two special first-order departure measures. Let $L(\theta)$ be the observed likelihood and let $\ell(\theta)$ be the observed log-likelihood. If we then write

$$\frac{L(\theta)}{L(\hat\theta)} = \exp\{\ell(\theta) - \ell(\hat\theta)\} = \exp\left(\frac{-r^2}{2}\right) \tag{3.10}$$

and solve for r with an appropriate sign we obtain

$$r = \operatorname{sgn}(\hat\theta - \theta)[2\{\ell(\hat\theta) - \ell(\theta)\}]^{1/2}, \tag{3.11}$$

which is called the signed likelihood root. The second departure measure is a standardized maximum likelihood departure

$$q = \operatorname{sgn}(\hat\theta - \theta)|\varphi(\hat\theta) - \varphi(\theta)|\hat\jmath_{\varphi\varphi}^{1/2}, \tag{3.12}$$

where $\hat\jmath_{\varphi\varphi} = -(\partial^2/\partial\varphi^2)\ell(\theta; y^0)|_{\theta=\hat\theta^0}$ is the corresponding observed information. This has certain rather special features that turn out to be very important: The standardization is with respect to observed and not the

usual expected information, and the departure is calculated in terms of a special reparameterization $\varphi(\theta)$. The use of the special parameterization $\varphi(\theta)$ is essential; it needs to be obtained as the gradient

$$\varphi(\theta) = \frac{d}{dv}\ell(\theta; y)\Big|_{y=y^0} \tag{3.13}$$

of likelihood at the data point and calculated in the sensitivity direction v discussed above. For (3.13) a directional derivative d/dv is defined by

$$\frac{d}{dv}h(y) = \frac{d}{dx}h(y + xv)\Big|_{x=0}.$$

Certainly we would expect likelihood at and near a data point to be important, and the use of the sensitivity direction as being a plausible way to examine likelihood near the data point, but for some background motivation and details, see Fraser and Reid (1993, 1995, 2001). We do note that $\varphi(\theta)$ can be replaced by any increasing affine equivalent $a\varphi(\theta)+b$ without altering q, but any further modification of the reparameterization can destroy the high accuracy. The special reparameterization will be called the exponential reparameterization, because it takes the role of a canonical parameter of a closely approximating exponential model (Fraser and Reid, 1993).

The observed p value $p^0(\theta)$ for testing θ with observed data y^0 is then given by

$$p^0(\theta) = \Phi(r^0) + \left(\frac{1}{r^0} - \frac{1}{q^0}\right)\varphi(r^0) \tag{3.14}$$

or

$$p^0(\theta) = \Phi\left\{r^0 - \left(\frac{1}{r^0}\right)\log\left(\frac{r^0}{q^0}\right)\right\}, \tag{3.15}$$

where r^0 and q^0 refer to the observed values obtained from (3.11) and (3.12). These formulas (3.14) and (3.15) for combining the likelihood ratio and maximum likelihood departure measures are from Lugannani and Rice (1980) and Barndorff-Nielsen (1986) as derived in particular contexts; the p value has third-order accuracy and conforms to appropriate ancillary conditioning (Fraser and Reid, 2001).

In the special normal case described in Section 3.1, the quantities r and q are both equal to $(\overline{y} - \theta)/(\sigma_0/\sqrt{n})$. Both formulas (3.14) and (3.15) have numerical difficulties near $\theta = \hat{\theta}^0$, where both r and q are equal to zero. Of course, we are usually not interested in p values near the maximum likelihood value, but simple bridging formulas are available (Fraser, Reid, Li and Wong, 2003).

In the location model context in Section 3.2, the reparameterization $\varphi(\theta)$ becomes the familiar score parameter

$$\varphi(\theta) = -\frac{\partial}{\partial\theta}\ell(\theta; y^0) = -\ell_\theta(\theta; y^0),$$

where the subscript θ denotes differentiation with respect to θ. Formulas (3.14) and (3.15) then give third-order approximations to (3.7).

Now to illustrate the accuracy of the approximations (3.14) and (3.15), consider a sample from the density function $\theta\exp\{-\theta y\}$ on the positive axis. For a coordinate y_i we obtain the log-likelihood $\ell_i(\theta) = \log\theta - \theta y_i$ and the log-likelihood gradient is $\varphi_i(\theta) = -\theta$. From this we obtain the overall log-likelihood

$$\ell(\theta) = n\log\theta - \theta\sum y_i.$$

A natural pivotal for the ith coordinate is $z_i = \theta y_i$. This has a fixed distribution, of course, with distribution function $F(z_i) = 1 - \exp(-z_i)$. For the vector case this gives the n-dimensional pivotal $(y_1\theta, \ldots, y_n\theta)$. From this we obtain the sensitivity vector

$$v(y, \theta) = \left(-\frac{y_1}{\theta}, \ldots, -\frac{y_n}{\theta}\right)'.$$

If we examine this at $(y^0, \hat{\theta}^0)$ we obtain the related sensitivity vector

$$v(y) = v(y; \hat{\theta}^0) = \left(-\frac{y_1^0}{\hat{\theta}^0}, \ldots, -\frac{y_n^0}{\hat{\theta}^0}\right)'$$

and the related reparameterization

$$\varphi(\theta) = \sum_1^n\left(-\frac{y_i^0}{\hat{\theta}^0}\right)(-\theta) = c\theta.$$

Because the model is exponential, this $\varphi(\theta)$ is, of course, just the exponential parameter of the initial model, and the sensitivity vector in this case, where a full sufficiency reduction is available, has no effect on the calculation as all the possible directions yield the same reparameterization. For a numerical illustration, consider the extreme case of a sample of $n = 1$ from this very nonnormal distribution and examine the data point $y = 1$ relative to the parameter value $\theta = 10$. The familiar signed likelihood ratio r has value -3.6599. With the common normal approximation, this gives the p value 0.000126. Alternatively the maximum likelihood departure q, which has value -9, with a normal approximation clearly gives an unrealistic approximation. If, however, we use r and q in (3.14) we obtain the p value 0.000046 which agrees very

closely with the exact p value 0.000045. As the model here is a location model in mild disguise, the calculations also provide an approximation to (3.7). The present type of calculation using (3.14) or (3.15) can be surprisingly accurate even for extremely small samples and extremely nonnormal distributions; for a range of numerical examples, see Fraser, Wong and Wu (1999).

3.5 Condition to Separate Main Effects

Our examples in this section were concerned with a scalar parameter θ, and we began with the case of normal error with known scaling. Sufficiency provided the reduction to the sample average and we obtained likelihood and p values directly. We then considered nonnormal location models, followed by general models that describe independent coordinates of a vector response. We found that conditional methods produced the accurate p values, while sufficiency methods typically are not available. We also saw that when sufficiency was available the conditional methods reproduced the same result as sufficiency. In Appendix A we show this holds more generally: That is, if sufficiency is available to simplify a problem, then in wide generality conditioning produces the same result. Thus we hardly need sufficiency; it can be replaced by conditioning. Indeed historically the extreme focus on sufficiency has distracted appropriate attention from serious consideration of conditional methods.

4. VECTOR PARAMETER MEASUREMENT EXAMPLES

4.1 Measurements with Normal Error

Consider the case of a sample $(y_1, \ldots, y_n)$ from the normal (μ, σ^2) distribution and let $(y_1^0, \ldots, y_n^0)$ be the observed data. The observed likelihood function is

$$L^0(\mu, \sigma) = c\sigma^{-n} \exp\left\{-\frac{(s^0)^2}{2\sigma^2} - \frac{n(\overline{y}^0 - \mu)^2}{2\sigma^2}\right\}, \tag{4.1}$$

where $s^2 = \sum(y_i - \overline{y})^2$. We could be interested in various parameter components, but we choose just the simple location parameter μ. From a general viewpoint we might want a likelihood for μ; there are recent developments for this (e.g., Fraser, 2003), but to address them here would take us from the main theme of this paper. A p value, however, is directly available and widely accepted; that is,

$$p^0(\mu) = H\left(\frac{\overline{y}^0 - \mu}{s^0/(n^2 - n)^{1/2}}\right), \tag{4.2}$$

where H is the Student $(n - 1)$ distribution function. This can be argued in various ways. The statistic $(\overline{y}, s)$ is minimal sufficient and is the sole data ingredient needed for the likelihood $L(\mu, \sigma; y_1, \ldots, y_n)$; for fixed μ, $t = n^{1/2}(\overline{y} - \mu)/s_y$ has uniqueness properties as a continuous function of $(\overline{y}, s)$ with distribution free of the nuisance parameter σ. Whatever the basis, we here take the t quantity as the appropriate input for the p value.

4.2 Measurements with Known Error Shape

Consider $y_1, \ldots, y_n$, where $y_i = \mu + \sigma e_i$ and the e_i form a sample from some known error distribution $f(e)$. To have a sensible definition of μ and σ we require that $f(e)$ be appropriately centered and scaled.

The standardized residuals $d_i = (y_i - \overline{y})/s$ describe simple characteristics of a sample $(y_1, \ldots, y_n)$, free of location and scale. It is straightforward to see that $d = (d_1, \ldots, d_n)'$ has a fixed distribution, free of μ and σ. Accordingly it is ancillary in the conventional sense. We can also note that $d(y^0) = d(e^0)$, where e^0 records the realized underlying errors; thus the underlying standardized errors are directly observable. Accordingly $d(y)$ can be viewed as the appropriate configuration statistic.

The observed likelihood function is

$$L^0(\mu, \sigma) = c\sigma^{-n} \prod_{i=1}^{n} f\{\sigma^{-1}(y_i^0 - \mu)\}. \tag{4.3}$$

The conditional distribution of the response vector given the standardized residuals is obtained by change of variable; it has probability element

$$c\sigma^{-n} \prod_{i=1}^{n} f\{\sigma^{-1}(\overline{y} + sd_i^0 - \mu)\} s^n \frac{d\overline{y}\, ds}{s^2},$$

which can be rewritten as

$$L^0\left(\overline{y}^0 + s^0\frac{\mu - \overline{y}}{s}, \frac{s^0\sigma}{s}\right)\frac{d\overline{y}\, ds}{s^2}, \tag{4.4}$$

where the constant in the likelihood L^0 is taken to be the appropriate norming constant. We thus see that any probability for $(\overline{y}, s)$ can be presented as an appropriate integral of observed likelihood.

Also in the particular case that $f(e)$ is the standard normal $\phi(e)$ as in Section 4.1, we have that (4.4)

reproduces the normal distribution for $\overline{y}$ and the scaled chi square distribution for s^2.

For testing a value of μ free of the nuisance parameter σ, the statistic $t = n^{1/2}(\overline{y} - \mu)/s_y$ has uniqueness properties as a continuous function with distribution free of the nuisance parameter σ. The corresponding p value is

$$p^0(\mu) = \int_{t \le t^0} L^0\left\{\overline{y}^0 + s^0\frac{\mu - \overline{y}}{s}, \frac{s^0\sigma}{s}\right\}\frac{d\overline{y}\,ds}{s^2}, \tag{4.5}$$

which is readily evaluated by numerical integration. We also see that (4.5) can be rewritten as

$$p^0(\mu) = \int_{\tilde{\mu}=\mu}^{\infty}\int_{\sigma=0}^{\infty} L^0(\tilde{\mu}, \sigma)\,\frac{d\tilde{\mu}\,d\sigma}{\sigma}, \tag{4.6}$$

which gives a simple expression for the p value as a direct integral of likelihood, indeed in the form of a survival posterior probability using the prior σ^{-1}. Highly accurate approximations for (4.5) or (4.6) are also easily available; see Section 4.4.

4.3 Exponential Model and Canonical Parameters

Consider an exponential model with natural or canonical parameters (ψ, λ):

$$\begin{aligned} &f(s_1, s_2; \psi, \lambda) \\ &\quad = \exp\{\psi s_1 + \lambda s_2 - \kappa(\psi, \lambda)\}h(s_1, s_2). \end{aligned} \tag{4.7}$$

This type of model is frequently mentioned when inference for a parameter ψ in the presence of a nuisance parameter λ is under discussion. If sampling is part of the background, then the coefficients of ψ and λ in the exponent of (4.7) form the minimal sufficient statistic or likelihood statistic. We anticipated this in (4.7) by writing (s_1, s_2) to suggest the sufficient statistic under sampling. In this sampling case, however, the support density $h(s_1, s_2)$ typically is available only by integration from some original composite density for the sample; by contrast, the likelihood ingredient $\kappa(\psi, \lambda)$ is quite typically available explicitly.

For testing a value ψ free of the nuisance parameter λ, the conditional distribution of s_1 given the nuisance score s_2 is often advocated. It is of course free of λ, but its density for direct calculation needs the typically unavailable density factor $h(s_1, s_2)$. However, for discussion here let $f(s_1|s_2; \psi)$ designate this conditional density. The p value for ψ is then given as

$$p^0(\psi) = \int^{s_1^0} f(s_1|s_2^0; \psi)\,ds_1, \tag{4.8}$$

where the lower limit is the lower end of the range of the variable. Some details for such calculations for the gamma mean problem can be found in Fraser, Reid and Wong (1997). The p value in (4.8) is presented as a conditional p value, conditional on the nuisance parameter score. It is also, however, a marginal p value, just a matter of whether it is being considered from the conditional or the overall marginal viewpoint: If it has a uniform distribution given any value for the condition, then it has that same uniform distribution marginally.

In wide generality, as will be seen in the next section, p values free of nuisance parameters are not available by such conditional calculations, but are obtained free of the nuisance parameter by a marginalization that eliminates the effect of the nuisance parameter. They are available by the conditional argument as just indicated only for very special model types such as the exponential described here; in such cases, the conditional p value is also a marginal p value, so there is no conflict with the marginal approach now being recommended. Conditioning above is then an alternative route to the same end by a different argument, but suitable just for certain special cases.

4.4 Location Model and Canonical Parameters

Consider a location model on the plane and let (y_1, y_2) be the variable with location (ψ, λ) and error density $f(e_1, e_2)$. We could examine the rather special case with independent normal errors, but for interest assume something more general, where say $f(e_1, e_2)$ is rotationally symmetic as for example with the Student density $\pi^{-1}(1 + e_1^2 + e_2^2)^{-2}$. A still more general case would proceed in the same manner. Also suppose that we are interested in the component parameter ψ. For a general context, see Fraser (2003).

For a sample of n we can reasonably consider the residual vectors for each coordinate, $d_1 = (y_{11} - \overline{y}_1, \ldots, y_{1n} - \overline{y}_1)'$ and $d_2 = (y_{21} - \overline{y}_2, \ldots, y_{2n} - \overline{y}_2)'$, as providing the data pattern free of location characteristics. It follows that $d_1(y_1, y_2) = d_1(e_1, e_2)$ and $d_2(y_1, y_2) = d_2(e_1, e_2)$, thus showing that the distribution for (d_1, d_2) is free of (ψ, λ), and also showing that the residual characteristics of the underlying errors are directly calculable from the observed data vectors.

In the presence of observed data $\{(y_{1i}^0, y_{2i}^0)\}$ we have that the conditional distribution of $(\overline{y}_1, \overline{y}_2)$ given the observed residuals is available by change of variable:

$$f(\overline{y}_1, \overline{y}_2 | d_1^0, d_2^0; \psi, \lambda) = k \prod_{i=1}^{n} f(\overline{y}_1 + d_{1i}^0 - \psi, \overline{y}_2 + d_{2i}^0 - \lambda).$$

As the observed likelihood is

$$L^0(\psi, \lambda) = c \prod_{i=1}^{n} f(y_{1i}^0 - \theta, y_{2i}^0 - \theta), \tag{4.9}$$

we find that we can then rewrite the conditional distribution as

$$f(\overline{y}_1, \overline{y}_2 | d_1^0, d_2^0, \psi, \lambda) = L^0(\psi - \overline{y}_1 + \overline{y}_1^0, \lambda - \overline{y}_2 + \overline{y}_2^0). \tag{4.10}$$

Again the arbitrary constant in the likelihood would be taken equal to the norming constant. This reduced model is a two-dimensional location model with parameter (ψ, λ).

Under a requirement of moderate continuity for the variables under study it is straightforward to see that $\overline{y}_1$ is the essentially unique variable free of λ. The corresponding marginal distribution is

$$f(\overline{y}_1 - \psi | d_1^0, d_2^0) = \int_{-\infty}^{\infty} L^0(\psi - \overline{y}_1 + \overline{y}_1^0, t)\, dt$$

and the essentially unique p value for assessing ψ is

$$p^0(\psi) = \int_{\psi}^{\infty} \int_{-\infty}^{\infty} L^0(\tilde{\psi}, \lambda)\, d\tilde{\psi}\, d\lambda, \tag{4.11}$$

which, in this pure location case, is equal to the Bayesian survival probability based on the flat prior in the location parameterization. The p values for various ψ values can then be obtained by numerical integration of likelihood. Highly accurate approximations to (4.11) are available and discussed in the next section.

For a more general approach to location parameterization, see Fraser and Yi (2002), and for the interplay of frequentist and Bayesian methods, see Fraser and Reid (2003).

4.5 Multiple Measurements: Interest and Nuisance Parameters

With the location model in the preceding section we see that a change in the parameter (ψ, λ) causes a corresponding translation of the distribution $f(y_1 - \psi, y_2 - \lambda)$ on the plane. For a sample of n, the effect is particularly simple: A change in ψ causes a shift in the first coordinate n-vector by the corresponding multiple of the one-vector for that coordinate. A change in λ similarly causes a shift in the second coordinate vector by the corresponding multiple of the one-vector for that second coordinate. This sensitivity connection between the parameter and the distribution for the response seems obvious and natural here in the location context, but for its more general version some discussion is needed.

Suppose that ψ and λ are scalars, and that independent y_i have a common distribution with distribution function $F(y; \psi, \lambda)$ and density function $f(y; \psi, \lambda)$. Then, as in Section 3.4, we examine how a change in (ψ, λ) shifts the distribution. We do this by examining the p value $F(y_i; \psi, \lambda)$ for the ith coordinate and seeing how, for fixed value of this pivotal, the distribution shifts at a point y_i. From the total differential of the p value we obtain

$$(v_{i1}, v_{i2}) = \frac{\partial y_i}{\partial(\psi, \lambda)} = \left(-\frac{\partial F(y_i; \theta)/\partial \psi}{\partial F(y_i; \theta)/\partial y_i}, -\frac{\partial F(y_i; \theta)/\partial \lambda}{\partial F(y_i; \theta)/\partial y_i} \right).$$

If we then consider all n coordinates, we obtain an array of two sensitivity vectors

$$V = \begin{pmatrix} v_{11} & v_{12} \\ \vdots & \vdots \\ v_{n1} & v_{n2} \end{pmatrix} = (v_1, v_2), \tag{4.12}$$

which describes how (ψ, λ) affects the distribution. Quite reasonably we are concerned with this effect for an observed data point y^0 at the corresponding maximum likelihood parameter value $\hat{\theta}^0$. Let V in (4.12) be evaluated for $(y, \theta) = (y^0, \hat{\theta}^0)$. As a simple example consider $y = X\beta + \sigma e$, where the error is a sample from a known distribution and the design matrix X is given. The sensitivity vector array V then has a vector for each parameter coordinate and simple calculation gives $V = (X, \hat{e}^0)$, where $\hat{e}^0$ is the fitted standardized error vector. This leads to accurate inference even with nonnormal error and extends easily to nonlinear regression; for examples, see Fraser, Wong and Wu (1999).

For the two parameter case as indicated by (4.12), general theory (Fraser and Reid, 2001) then shows that

there is an approximate ancillary $a(y)$ of dimension $n-2$ for which the tangent vectors V at the data point y^0 are given by (4.12). This then leads to highly accurate third-order p values for scalar components of the parameter θ. The calculations for the p values for assessing say ψ need just the observed log-likelihood $\ell^0(\psi,\lambda)$ and the observed log-likelihood gradient

$$\begin{aligned}\varphi'(\theta) &= \{\varphi_1(\theta), \varphi_2(\theta)\} = \frac{d}{dV}\ell(\theta; y)\Big|_{y=y^0} \\ &= \left\{\frac{d}{dv_1}\ell(\theta; y)\Big| y^0, \frac{d}{dv_2}\ell(\theta; y)\right\}\Big|_{y=y^0},\end{aligned} \tag{4.13}$$

using directional derivatives as defined after (3.13). We refer to this as the exponential parameterization, being the canonical parameter of some best fitting exponential model near the data point. For inference concerning ψ we can then calculate a first departure measure given by the signed likelihood ratio

$$r^0(\psi) = \operatorname{sgn}(\hat\psi^0 - \psi)\{2[\ell^0(\hat\theta) - \ell^0(\hat\theta_\psi)]\}^{1/2}, \tag{4.14}$$

where $\hat\theta_\psi$ is the maximum likelihood value under the constraint $\psi(\theta) = \psi$, and we can calculate a second departure measure given as a special standardized maximum likelihood departure

$$\begin{aligned}q^0(\psi) &= \operatorname{sgn}(\hat\psi^0 - \psi) \\ &\quad \cdot |\chi(\hat\theta) - \chi(\hat\theta_\psi)|\left\{\frac{|\hat j_{\varphi\varphi}|}{j_{(\lambda\lambda)}(\hat\theta_\psi)}\right\}^{1/2}.\end{aligned} \tag{4.15}$$

In this $\chi(\theta)$ is a rotated coordinate of $\varphi(\theta)$ that agrees with $\psi(\theta)$ at $\hat\theta_\psi$ and acts as a surrogate for $\psi(\theta)$ at $\hat\theta_\psi$, and the full and nuisance informations are recalibrated in the φ parameterization, as indicated by the use of parentheses around $\lambda\lambda$. Further details are recorded in Appendix C; also see the regression examples in Fraser, Wong and Wu (1999) and Fraser, Monette, Ng and Wong (1994). The p value $p^0(\psi)$ is then given by (3.15) in Section 3.4.

The p value just discussed corresponds to the use of the special conditional model given the approximate ancillary with tangent vectors V, followed by a marginalization to eliminate the nuisance parameter. This two-step simplification corresponds closely to that found for the location model in Section 4.4, and the present p value provides an approximation to that given by (4.11). The present p value also can provide an approximation to the Student p value at (4.2), or to the location scale p value at (4.5) or to the exponential model p value at (4.8). We can thus note that the present approach using sensitivity vectors V covers the simple cases where sufficiency can be used and covers the general cases as developed in Sections 4.3 and 4.4, where sufficiency is not available.

5. SOME CONDITIONING AND MODELING CRITERIA

5.1 The Two Measurement Instruments Example

In Section 2 we discussed two examples that involved measurement instruments, as presented by Cox (1958) and, earlier, by Welch (1939). Our theme, in contrast with that in Welch, was that conditioning is appropriate and proper for both examples.

For the earlier example (Welch, 1939), the two instruments were identical and both were used in a single investigation. The conditioning under discussion used Fisher's configuration statistic and provided the background for the succession of examples in Sections 3 and 4. We develop further aspects of conditioning on configuration statistics in the next section. For the other example (Cox, 1958), only one of the instruments was actually used. This raises a serious issue. Should the modeling include probability structure for measurements that were never taken? Cox comes out quite firmly in support of the use of the appropriate conditional model, the model for the measurement that was actually made. Surprisingly there seems to have been little subsequent support for such an approach. We develop some further aspects of this modeling in Section 5.3.

5.2 Conditioning Directions V

The examples in Sections 3 and 4 all involved a primary role for continuity: how a change in the parameter shifts the response distribution, in particular, how it shifts the distribution in the neighborhood of the observed data. At the present time this theory is now available for the case of discrete distributions. The concern with the model in the neighborhood of the data does seem data dependent, but at the observed data is where the model form is of particular importance. In substance this is not dissimilar to standardization of a maximum likelihood departure $\hat\theta - \theta$ by an observed information, information at the data point of interest rather than expected information, thus giving $q = (\hat\theta - \theta)\hat j^{1/2}$. Theoretically this type of standardization has strong support.

The examples in Sections 3 and 4 all consider how a change in the parameter shifts the response distribution. In the context of independent scalar coordinates, the coordinate p values $F_i(y_i;\theta)$ provide the direct continuity link that describes how a parameter change affects a coordinate y_i; see (3.8) and (4.12) for details.

Now, more generally, suppose that the coordinates are vector-valued with dimension say equal to the dimension p of the parameter. A change in the parameter will lead to an altered distribution, but this in itself does not prescribe a point-by-point movement of the distribution; something more is needed. For the ith coordinate let $z_i(y_i;\theta)$ be some appropriate pivotal quantity. With $p>1$ there may not be an obvious unique choice for this pivotal; we would then seek one that best describes how the ith variable measures or relates to the parameter being measured. A basis for this choice is discussed elsewhere. Here we assume that it is given or has been chosen on a natural or what-if basis.

The pivotal allows us to examine how a θ change affects or moves the data point y. For this we let y be the np-dimensional vector obtained by stacking the y_i and similarly let z be the np-dimensional vector obtained by stacking the z_i. Then taking the total differential of the pivotal we obtain

$$V=-z_y^{-1}(y^0;\hat\theta^0)z_{;\theta}(y^0;\hat\theta^0), \tag{5.1}$$

where the Jacobian matrices are, respectively, $np\times np$ and $np\times p$, and are evaluated at the data point y^0 and the corresponding maximum likelihood value $\hat\theta^0$; the subscripts indicate differentiation with repect to the argument before or after the semicolon.

For conditional inference with an approximate ancillary, the measurement vectors V represent the directions of change along which the appropriate conditional model is defined. They give tangent vectors to an approximate second-order ancillary (Fraser and Reid, 2001). General theory (Fraser and Reid, 1993, 1995) shows that a second-order ancillary suffices for third-order likelihood inference.

The directional vectors V lead to an exponential-type recalibration of the parameter. The exponential-type parameterization for the ith coordinate model is available as the gradient of log-likelihood

$$\varphi_i'(\theta)=\frac{\partial}{\partial y_i'}\ell(\theta;y_i^0), \tag{5.2}$$

which is recorded here as a p-dimensional row vector. For the full model the appropriate reparameterization is obtained by combining these components using the sensitivity vectors V in (5.1),

$$\varphi'(\theta)=\sum_{i=1}^n\varphi_i'(\theta)V_i=\ell_{;V}(\theta;y^0), \tag{5.3}$$

where V_i is the $p\times p$ block of the matrix V that corresponds to the ith observation y_i and the right-hand term of (5.3) is an array of p directional derivatives.

For inference concerning a scalar parameter $\psi(\theta)$, it then suffices for third-order inference to act as if the model is exponential with observed likelihood $\ell(\theta;y^0)=\ell^0(\theta)$ and with canonical parameter $\varphi(\theta)$ from (5.3). In particular, the observed p-value function $p^0(\psi)$ is given by (3.14) or (3.15) using $r(\psi)$ and $q(\psi)$ given by (4.14) and (4.15). For a variety of examples in a regression context, see Fraser, Wong and Wu (1999) and Fraser, Monette, Ng and Wong (1994).

5.3 Modeling the Actual Data Production

As mentioned in Section 5.1, the Cox (1958) example recommended that only the measurements that were actually made should be modelled or, put another way, that the full model should not be describing measurements that were not made. We now develop this in more detail.

Consider a succession of measurements on a parameter θ and suppose that for each there is a direct measurement relationship to the parameter, as discussed in Sections 3, 4 and 5.2. For illustrative purposes a succession of three models, say M_1, M_2 and M_3, will suffice. Let y_1, y_2 and y_3 be the corresponding data. Many issues can be involved in the modeling of such a context. Here we focus on the goal of statistical inference for the parameter in question and propose three modeling criteria:

I. Provide a model for each measurement that has been made.
II. Do not provide a model for measurements that were not made.
III. Do not provide a model otherwise for the process or procedure that led to the choice of a particular measurement process.

These seem reasonably natural and persuasive, but have some rather striking implications.

EXAMPLE 5.1. Consider Example 2.1 concerning the two measurement instruments and suppose we have data $y=y^0$ and $a=a^0=2$ (the second instrument is chosen). By criterion III, we do not model the coin

toss used to choose the instrument. By criterion II, we do not model the measurement process for the first instrument. By criterion I, we do model the measurement process for the second instrument. We then have data y^0 and a normal model with mean θ and standard deviation σ_0. A 95% confidence interval is given as $(y^0 \pm 1.96\sigma_0)$.

EXAMPLE 5.2 (Meta-analysis). Consider the meta-analysis of three investigations concerning a parameter θ. In practice the precise definition of θ may vary from investigation to investigation, and various factors such as reliability of measurements may arise. For our illustration here we assume that these are not at issue. By criterion III, we do not model the process by which the particular investigations were selected. For example, the data with investigation M_1 may have suggested some interesting range of values for θ, but were inconclusive for this, thus leading to the choice of a more comprehensive or demanding investigation M_2. Or, the data with M_1 might have been very strongly conclusive for the interesting range, leading to no further investigation. Also M_3 might only have been performed in the case of conflicting results from M_1 and M_2. By criteria I and II, we model exclusively the investigations that have actually been made and in doing so make reference to repeated sampling just for the corresponding measurement models. Accordingly, our composite model is the product formed from the individual models. In particular, this would say that the randomness in model M_2 is not influenced by the results from the investigation M_1. That is, M_1 and M_2 are taken as statistically independent. We note of course that if M_1 had produced a different outcome, we might have had a different investigation in place of M_2 or indeed have had no second or subsequent investigations. This is in accord with criterion I: We are concerned with the randomness in the measurement processes that have been performed, and not with randomness in other possible investigations that in fact did not take place. The repeated sampling reference is for measurements that have been made and does not embrace repeated sampling in a global sense that might embrace many possible other models, none of which has corresponding data values.

In conclusion, we note that the use of the product model for the analysis of M_1, M_2 and M_3 as just described is the common procedure for meta-analysis. We return to this consideration of meta-analysis in Section 7.

6. SOME FAMILIAR ANCILLARY EXAMPLES

We are concerned with conditional inference theory and how it relates to the ancillarity principle that specifies the use of the conditional model given the observed value of an appropriate ancillary statistic. In Sections 3 and 4 we noted that conditional methods could be used quite generally to replace sufficiency and, in addition, to provide definitive inference methodology in a much broader context. As part of this we used continuity and a notion of a measurement sensitivity to motivate the related results from recent likelihood asymptotics. In Section 2 we examined the Cox two measuring instruments example and noted that there was something stronger than ancillarity involved, that only measurements that were actual made should be modelled. This led in Section 5 to criteria for models for inference, in particular criteria for isolating certain components, that is, the components that correspond to measurements that were actually made. This went significantly beyond just conditioning on an observed ancillary.

In this section we examine some of the commonly cited ancillary examples. A survey of such ancillary examples can be found in Fraser (1979, pages 54–68 and 76–86) and in Buehler (1982); see also Reid (1995) for a general discussion of conditional inference. Here we examine these examples from the viewpoint of what the proper model for inference should be in the presence of data and for this we use the criteria from Section 5. We also compare these models for inference with the result of invoking ancillarity within models that are global (encompassing all possible data that might have been observed) and thus violate criteria II and III.

EXAMPLE 6.1 (Random choice of sample size). Consider the repeated measurement unit assessment of a parameter θ and suppose that the number of repetitions n is random with known density $p(n)$. In accord with criteria I and II, we would model the specific measurement units that were performed, and in accord with criterion III, we would not model the process that leads to the sample size n. This gives the inference model $\prod_1^n f(y_i;\theta)$ plus the corresponding data. From the global repeated sampling viewpoint, however, we would examine the composite model $p(n)\prod_1^n f(y_i;\theta)$ with data $(n; y_1^0, \ldots, y_n^0)$. For this full model, n is an ancillary statistic and the corresponding ancillary reduction gives the just described inference model. The two viewpoints lead to the same reduced model. More

generally we can consider a distribution $p(n;\lambda)$ for n with dependence on a parameter λ free of θ. The criteria again give the model $\prod_1^n f(y_i;\theta)$ with data $(y_1^0,\ldots,y_n^0)$.

EXAMPLE 6.2 (Sampling from a mixed population). Consider two populations A_1 and A_2 of relative sizes q_1 and q_2 that are intermixed and the elements of which are not easily distinguishable. A parameter θ may have the same value in each population and yet distributionally express itself differently: $f_1(y;\theta)$ and $f_2(y;\theta)$ in A_1 and A_2, respectively. We consider a random sample of n from the mixed population, yielding observed numbers n_1 and n_2 from the populations A_1 and A_2. The inference model would describe the data $(y_1^1,\ldots,y_{n_1}^1)$ and $(y_1^2,\ldots,y_{n_2}^2)$ from the random sampling of n_1 elements from A_1 and n_2 elements from A_2 (with n_1 and n_2 fixed at their observed values). By criterion III we would omit the hypergeometric model that yields (n_1,n_2). However, if we consider the full global model, we can note that the allocation (n_1,n_2) has a fixed distribution and is ancillary. The corresponding conditional model is that just described: n_1 observations randomly sampled from A_1 and n_2 observations randomly sampled from A_2. Accordingly the reduced model conditional on the ancillary coincides with the inference model. Note that in the full global model the indicator variables that describe which n_1 elements of A_1 are chosen, and which n_2 elements of A_2 are chosen, with given n_1, n_2, have a fixed distribution with probabilities $1/(Nq_1)^{(n_1)}(Nq_2)^{(n_2)}$ and are thus also ancillary. Conditioning on this ancillary just gives the assessment of specified units in each population and thus can be viewed as 100% sampling of particular subsets of A_1 and A_2. Thus, this use of ancillarity seems to go too far and eliminates the inference assessment available from finite population sampling (Fraser, 1979). Some consideration of this issue in terms of labels for sample elements was given by Godambe (1982, 1985).

EXAMPLE 6.3 (Random regression input). Consider a regression model $y = X\beta + \sigma e$, where the rows X_i of the $n \times r$ design matrix have been generated randomly from some distribution $g(x_1,\ldots,x_r)$ for input variables. The inference model again would be for fixed X even in the context where g depends on a parameter λ with range free of θ. More specifically, the inference model concerning θ would be the model for the actual measurements made. From the ancillarity viewpoint we note that for the first case the variable X has a fixed distribution and is thus ancillary. The corresponding conditional model then agrees with the inference model just described.

EXAMPLE 6.4 (A 2×2 table; Fisher, 1956, page 47). The offspring in a breeding experiment can be classified by phenotype based on two genetic characteristics (A, a) and (B, b) that show complete dominance. The relative proportions for AB, Ab, aB and ab are 9, 3, 3 and 1 if there is no linkage and are $2+\theta$, $1-\theta$, $1-\theta$ and θ in the presence of a linkage parameter θ, where $\theta = 1/4$ corresponds to the no linkage case. The proportions for A : a or for B : b are the standard 3 : 1 of dominant to recessive phenotypes. Let n_{11}, n_{12}, n_{21} and n_{22} be the data for n offspring in a particular mating with say $(n_{1\cdot}, n_{2\cdot}) = (n_{11}+n_{12}, n_{21}+n_{22})$ designating row totals and $(n_{\cdot 1}, n_{\cdot 2})$ designating column totals.

If the data are assembled in terms of the A phenotype, we then have that n_{11} is binomial $\{n_{1\cdot}, (2+\theta)/3\}$ and n_{21} is binomial $\{n_{2\cdot}, (1-\theta)\}$. Alternatively, if the data are assembled in terms of the B phenotype, we then have that n_{11} is binomial $\{n_{\cdot 1}, (2+\theta)/3\}$ and n_{12} is binomial $\{n_{\cdot 2}, (1-\theta)\}$. We thus obtain two different inference modelings based on two different classifications of the data, by A phenotype or by B phenotype, each classification corresponding to a particular viewpoint concerning the context in which the parameter θ is being investigated.

From the ancillary viewpoint we can note that the row totals $n_{1\cdot}, n_{2\cdot}$ have a binomial allocation with probabilities in the ratio 3 : 1, and thus are ancillary; this gives a reduced model that coincides with the inference model based on assembly by A phenotype. Also we can note that the column totals $n_{\cdot 1}, n_{\cdot 2}$ have a 3 : 1 binomial allocation and are thus ancillary; the corresponding reduced model coincides with the inference model based on assembly by B phenotype. We do note, however, that the combination of the row totals and the column totals is not ancillary. Thus the ancillarity approach gives two different modelings and provides no preference for one over the other.

EXAMPLE 6.5 (Bivariate correlation). A continuous example closely analogous to the preceding example is provided by data from a bivariate normal distribution for (x, y) with means 0, variances 1 and correlation ρ. If we examine the data labelled by the x values, we have that the y values are normal with mean ρx and variance $1-\rho^2$. Alternatively, if we examine the data labelled by the y values, we have that the x values are normal with mean ρy and variance

$1-\rho^2$. Accordingly, we obtain two different inference modelings that correspond to two different assemblies or classifications of the data, by x or by y. By contrast we can note that from the full model ancillary viewpoint we have that the $x_1, \ldots, x_n$ are ancillary, and the corresponding conditional model examines y's for fixed x's and agrees with the first inference model above. In a parallel way we note that the $y_1, \ldots, y_n$ are ancillary in the full model with a conditional model that agrees with the second inference model above. Again we have conflicting ancillaries and ancillarity alone does not provide a resolution. Indeed ancillarity itself creates the conflict between the two conditional resolutions. We could also rotate our coordinates through an angle of $\pi/2$ and in effect use $w_i = x_i + y_i$, $z_i = x_i - y_i$; the independent coordinates w_i and z_i could then be examined more transparently using the approximate ancillary approach in Section 3.4.

For the first three examples, our model for the inference approach and the ancillarity approach are in agreement. For the final two examples, the model for the inference approach required a particular assembly of the data, by choice of phenotype or by choice of input variable. Without this choice of how to assemble the data, the ancillarity approach produces conflicting recommendations. It thus seems that invoking ancillarity also requires some specification of how the data are to be assembled for analysis.

We do note that the two approaches lead to the same observed likelihood function, even in the context of conflicting ancillaries. If, however, we wish to go beyond just observed likelihood, we find that different ancillaries can produce different distributions for possible likelihood functions and can produce different confidence assessments and different p values. Accordingly, some additional specification is needed and indeed should not have been omitted at the initial modeling stage. This leads to the use of measurement directions as introduced in Sections 3.4 and 4.5, which use continuity and express how parameter change can produce an effect at a data point.

7. ARE GLOBAL REPEATED SAMPLING PROPERTIES WANTED?

We have been considering ancillary statistics and how they lead naturally to conditional inference given an observed value of the ancillary. However, our initial examples from Cox (1958) and Welch (1939) included some discussion of overall or global sampling properties, where repetitions of some complete process were being considered. Cox argued that the conditional approach should take precedence over global properties, and Welch argued that the global properties invalidated the conditional approach. This leads to the focal issue: What probabilities are the appropriate probabilities for presenting inference conclusions from context and data information?

With the modeling criteria in Section 5.3, we viewed the individual measurement probabilities as the primary ingredients, with frequency interpretations based on repetitions of the individual measurement processes. This supports the Cox viewpoint for the two measurement instrument example. Our earlier discussion in Section 2 viewed the global probabilities as artificial in that they used probabilities for measurement units that might have been used, but in fact were not.

At the heart of the global approach is the calculation of probabilities for repetitions of the full process under a fixed value for the parameter. This allows the calculation of global operating characteristics for the full investigation under consideration. On the surface this seems hard to argue against or, at least to argue against it is counter to present culture. Of course it is telling a story, but perhaps not the relevant story for the purposes of statistical inference.

From the global viewpoint there seems little alternative to that of repetitions under a fixed parameter value, without say putting weights on the possible parameter values and using a Bayesian-type argument. Of course this Bayesian approach has given a wealth of possible answers to wide ranging problems, in contrast to the range of answers from the traditional optimality approach, but this same wealth is of course available more directly, and without pretense, by weighted likelihood and integration. For some recent discussion, see Fraser (1972), Fraser and Reid (2003) and Fraser and Yi (2002).

Here we examine some aspects of global and conditional probabilities without resort to probabilities or weights on the various values for the parameter.

EXAMPLE 7.1 (Meta-analysis). As part of the discussion of inference modeling in Section 5.3 we considered conditional inference and metanalysis for three investigations of a scalar parameter θ. For some comparisons with global probabilities we now examine an even simpler case that involves two measurements of the parameter θ: a first measurement y_1 is unbiased and normal with standard deviation σ_0 say equal to 1; a second measurement is unbiased and normal with standard deviation $\sigma_0/100 = 0.01$. We also suppose

that some threshold value $\theta = \theta_0$ is of interest and for simplicity and convenience take this value here to be zero.

If there had just been the first measurement, say $y_1 = y_1^0$, the p value or significance function for θ would be

$$p_1(\theta) = \Phi(y_1^0 - \theta) \tag{7.1}$$

and the p value for the threshold would be $p_1 = \Phi(y_1^0)$. However, with the two measurements the weighted average $y = (y_1 + 10000y_2)/10001$ would be the appropriate combined estimate and the p value or significance function for θ would be

$$p_2(\theta) = \Phi\{100(y_2^0 - \theta)\}, \tag{7.2}$$

where, as a reasonable approximation and simplification, we ignore y_1 because of the very large weight on y_2 in the weighted average y; the p value for the threshold would be $p_2 = \Phi(100y_2^0)$. In summary, with just the first measurement the significance function is a reverse standard normal distribution function centered on the data y_1^0, while with two measurements it is a reverse normal distribution function centered at the value y_2^0 but scaled much more tightly around that value, indeed by a factor of 100 to 1. Also the p value for the threshold $\theta = 0$ changes from $\Phi(y_1^0)$ to $\Phi(100y_2^0)$ in going from the one- to the two-measurement situation.

Now consider an experimental context for these two investigations. The investigator is particularly interested in the threshold value $\theta = 0$. He or she makes a first measurement of θ and obtains a value $y_1^0 = 1.1$, suggesting in a very informal way that perhaps the true value for θ is above the threshold. As a result he or she decides to take a second high precision measurement and obtains $y_2^0 = -0.1$; this new significance function is very tight and substantially left of the origin. We suggest that both the preliminary and the subsequent p values represent appropriate expressions of the information at the respective times. We also note that these seem in agreement with the meta-analysis approach.

Now suppose that if the first measurement had been negative with a p value less than $1/2$ then no follow-up measurement would have been deemed appropriate. Consider the global probability assessment of this for the null situation $\theta = 0$. With the first measurement the initial p values are uniform $(0, 1)$; with probability $1/2$ the pivotal p value is greater than $1/2$ leading to the follow-up combined p value, which is approximately uniform $(0, 1)$. The global probability distribution for the reported p value is then piecewise uniform with density $3/2$ on $(0, 1/2)$ and density $1/2$ on $(1/2, 1)$.

We believe that the individual p values $\Phi(y_1^0)$ and $\Phi(100y_2^0)$ provide the appropriate inference presentation for the particular cases as they arose in time, and that the nonuniform global p value is a consequence of the seemingly inappropriate use of an overall marginal assessment of the p values for this two measurement situation. Also recall the earlier Example 2.1.

From a raw global approach we thus note that it is possible to obtain p values biased to the left by deliberately taking follow-up measurements when an initial p value is high. The inappropriateness of the use of global probabilities is again to be emphasized.

EXAMPLE 7.2 (AR1 models). The typical autoregressive model is used for data that arrives sequentially in time and as such seems appropriate for consideration here from our present conditional viewpoint. For this we examine now a very simple case with just two measurements that illustrates some of the key issues. Consider normal $(0, \sigma_0^2)$ errors with an autoregressive parameter θ and two observations. Thus $y_1 = e_1$ and $y_2 = \theta y_1 + e_2$, where e_1 and e_2 are normal $(0, \sigma_0^2)$. The log-likelihood function is

$$\ell(\theta) = -\frac{1}{2\sigma_0^2}(y_2 - \theta y_1)^2. \tag{7.3}$$

This has the maximum likelihood value $\hat{\theta} = y_2/y_1$, which has a standard Cauchy distribution centered at the point θ.

Now consider the inference modeling viewpoint from Section 5. The first y_1 does not measure θ, but it does determine the precision for the second measurement y_2. By criterion III, we do not model y_1. Then by criterion I, we do model y_2 and by criterion II, we model y_2 only for its particular measurement situation. This gives the model y_2 is normal $(\theta y_1; \sigma_0^2)$, and this produces the same likelihood function (7.3) as does the global model, and the maximum likelihood value is just the same $\hat{\theta} = y_2/y_1$. We observe, however, that the maximum likelihood value is now normal $(\theta; \sigma_0^2/y_1^2)$, where the y_1 value is taken at its observed value. The issue we have mentioned before becomes more transparent here. Do we use the actual measurement process model with its normal distribution or do we use some average of possible measurement situations that typically did not occur, leading to the Cauchy analysis? We know that the normal distribution describes the actual measurement that was made and leads to a normal analysis. But the persuasive global approach would

want to include modelings for other measurements that were never made and thus argue for the Cauchy analysis.

From the present viewpoint we prefer the measurement model approach, conditioning on preceding measurements. Of course there may be cases where the global probabilities are wanted, but for direct statistical inference with observed data the conditional approach seems appropriate. Also it avoids the usual and well-known singularities that arise with the marginal approach in the neighborhood of $\theta = 1$. It now seems clear that these singularities arise precisely from the inclusion of a wealth of possible models that apply to measurements that were in fact never made.

The preceding is arguing in support of conditioning in the time series context; this is of course not a common recommendation, but has been suggested on several occasions by Professor Jim Durbin. Perhaps the only way to argue against it is to make some preliminary assumption that only the global repeated sampling principle will be entertained.

Now consider briefly the global repeated sampling approach and how it interacts with various common optimality criteria. The examples in Section 2 show how a search for optimality leads to a trade-off between different measurement situations. In particular we saw how a precise measurement instance could be given a longer confidence interval so that a much shorter interval could be given in a less precise instance. Optimality in the global framework can lead to results in particular instances that are contrary to the available evidence. Alternatively, by overstating and by understating in particular instances it is possible to increment toward some optimality goal on the global scale. This clearly argues against the appropriateness of the optimality applied on the global scale; this has been asserted very gently by Cox (1958).

APPENDIX A: CONDITIONING REPLACES SUFFICIENCY TO SEPARATE MAIN EFFECTS

Consider the case of continuous variables and suppose there is a sufficient statistic $s(y)$ that has the same dimension p as the parameter. Also suppose for ease of argument that the conditioned variable, say $t(y)$ given $s(y)$, has constant dimension which would then be $n - p$. It follows from sufficiency that the distribution of $t(y)$ given $s(y)$ is parameter-free: Let $u(y)$ be a coordinate-by-coordinate sequential probability integral transformation of $t(y)$ as obtained from the conditional distribution given $s(y)$; for example, the probability integral transformation for the first coordinate, the probability integral transformation of the second coordinate conditional on the first and so on. Of course there are many such transformations obtained even by varying the order of the coordinates. It follows that the conditional distribution of $u(y)$ given $s(y)$ is uniform on a unit cube and thus does not depend on $s(y)$. It follows that u and s are independent and thus that $f(s;\theta) = f(s|u;\theta)$, showing that a conditional model equivalent to the given model is available. This result does not depend on the choice of the probability integral transformation. This says that an analysis using sufficiency can be duplicated by a conditional analysis. For a simple example consider (y_1, y_2) from the normal (θ, σ_0^2). The model for $\overline{y}$ is normal$(\theta, \sigma_0^2/2)$; the conditional model for $\overline{y}$ given the configuration $y_2 - y_1$ is also normal$(\theta, \sigma_0^2/2)$. If, however, we are without normality, then sufficiency is typically not available, but the conditional analysis remains available and is routine. Accordingly we support the conditional approach and suggest that there is little need for sufficiency methods for inference in the continuous case. Of course they can be convenient in special cases, but they do not provide the methodological sanction needed for general contexts; they should be viewed as an expediency for the special cases. For the typical discrete case, sufficiency can be convenient, but some simple invariance notions typically suffice.

APPENDIX B: MARGINALIZATION TO ELIMINATE PARAMETERS

Conditioning is often suggested as a means to eliminate nuisance parameters, but in general contexts marginalization is the effective method and conditioning can be viewed as an expediency when special model structure is available. Consider two examples. For a continuous exponential model,

$$\exp\{y_1\psi + y_2\lambda - c(\psi, \lambda)\}h(y_1, y_2), \tag{B.1}$$

the conditional distribution of $y_1|y_2$ depends on ψ only and is thus free of λ. For a continuous location model,

$$f(y_1 - \psi, y_2 - \lambda), \tag{B.2}$$

the marginal distribution of y_1 depends on ψ only and is thus free of λ. In each case we have a special model type with specialized variables and parameters, and these are often referred to as canonical variables and parameters.

Now consider the first example, where conditioning provides freedom from the nuisance parameter, and suppose we are testing ψ. Let $u(y_1, y_2)$ be a probability integral transformation of $y_1|y_2$ obtained from the λ-free conditional distribution for testing ψ. Then for the tested ψ, the distribution of $u|y_2$ is free of y_2; thus u is independent of y_2. It follows that the marginal distribution of u is λ-free and gives p values that agree with those from the initial conditional variable.

Recent likelihood asymptotics (e.g., Fraser and Reid, 1993; Fraser, Reid and Wu, 1999) shows that for a general asymptotic model with continuous variables, the testing of a parameter value $\psi(\theta) = \psi$ is available from a marginal distribution obtained by integrating over a nuisance parameter based conditional distribution as in the second example, which follows the pattern for the location model as discussed in Section 4.4.

APPENDIX C: THE PARAMETER REEXPRESSION

The third-order p values obtained from (3.14) or (3.15) using the signed likelihood ratio $r(\psi)$ in (4.14) and the maximum likelihood departure $q(\psi)$ in (4.15) are based on an exponential type reparameterization $\varphi(\theta)$ in (3.13), (4.13) or (5.3). The full information determinant calculated in the new parameterization is available as

$$|_{J(\theta\theta)}| = |_{J\theta\theta}(\hat\theta)||\varphi_\theta(\hat\theta)|^{-2},$$

using the Jacobian $\varphi_\theta(\theta) = \partial\varphi(\theta)/\partial\theta'$. The nuisance information determinant somewhat similarly takes the form

$$|_{J(\lambda\lambda)}(\hat\theta_\psi)| = |j_{\lambda\lambda}(\hat\theta_\psi)||\varphi_{\lambda'}(\hat\theta_\psi)|^{-2} = |j_{\lambda\lambda}(\hat\theta_\psi)||X|^{-2},$$

where the right-hand determinant uses $X = \varphi_{\lambda'}(\hat\theta_\psi)$ with $|X| = |X'X|^{-1/2}$, which in the regression context records the volume on the regression surface as a proportion of the corresponding volume for regression coefficients; in the preceding formula this changes the scaling for the nuisance parameter to that derived from the φ parameterization. The expressions above are for the case where θ' is given as (ψ, λ') with an explicit nuisance parameterization; the more general version is recorded in Fraser, Reid and Wu (1999). The rotated coordinate $\chi(\theta)$ in the φ parameterization is obtained from the gradient vector of $\psi(\theta)$ at $\hat\theta_\psi$ and has the form

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat\theta_\psi)}{|\psi_{\varphi'}(\hat\theta_\psi)|}\varphi(\theta),$$

where the row vector multiplying $\varphi(\theta)$ is the unit vector obtained from the gradient $\psi'_\varphi(\hat\theta_\psi)$ and is obtained from

$$\begin{aligned}\psi_{\varphi'}(\theta) &= \frac{\partial\psi(\theta)}{\partial\varphi'}\\ &= \frac{\partial\psi(\theta)}{\partial\theta'}\left(\frac{\partial\varphi(\theta)}{\partial\theta'}\right)^{-1}\\ &= \psi_{\theta'}(\theta)\varphi_{\theta'}^{-1}(\theta);\end{aligned}$$

in this we take $\psi_{\varphi'}$ to be the Jacobian of the column vector ψ with respect to the row vector φ' and, for example, would have $(\psi_{\varphi'})' = \psi'_\varphi$ for the transpose of the first Jacobian.

ACKNOWLEDGMENTS

The author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada. Very special thanks go to Leon Gleser for many discussions that led through many revisions to the present paper, and to referees for many helpful suggestions and comments.

REFERENCES

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized, signed log likelihood ratio. *Biometrika* **73** 307–322.

BROWN, L. D. (1990). An ancillarity paradox which appears in multiple linear regression (with discussion). *Ann. Statist.* **18** 471–538.

BUEHLER, R. J. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Assoc.* **77** 581–594.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.

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

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

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. London Ser. A* **144** 285–307.

FISHER, R. A. (1935). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.* **98** 39–82.

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

FISHER, R. A. (1961). Sampling the reference set. *Sankhyā Ser. A* **23** 3–8.

FRASER, D. A. S. (1972). Bayes, likelihood or structural. *Ann. Math. Statist.* **43** 777–790.

FRASER, D. A. S. (1979). *Inference and Linear Models*. McGraw–Hill, New York.

FRASER, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90** 327–339.

FRASER, D. A. S. and MCDUNNOUGH, P. (1980). Some remarks on conditional and unconditional inference for location–scale models. *Statist. Hefte* **21** 224–231.

FRASER, D. A. S., MONETTE, G., NG, K. W. and WONG, A. (1994). Higher order approximations with generalized linear models. In *Multivariate Analysis and Its Applications* (T. W. Anderson, K. T. Fang and I. Olkin, eds.) 253–262. IMS, Hayward, CA.

FRASER, D. A. S. and 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. and REID, N. (1995). Ancillaries and third order significance. *Utilitas Math.* **47** 33–53.

FRASER, D. A. S. and REID, N. (2001). Ancillary information for statistical inference. *Empirical Bayes and Likelihood Inference. Lecture Notes in Statist.* **148** 185–209. Springer, New York.

FRASER, D. A. S. and REID, N. (2003). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Inference* **103** 263–285.

FRASER, D. A. S., REID, N., LI, R. and WONG, A. (2003). p-value formulas from likelihood asymptotics: Bridging the singularities. *J. Statist. Res.* **37** 1–15.

FRASER, D. A. S., REID, N. and WONG, A. (1997). Simple and accurate inference for the mean of the gamma model. *Canad. J. Statist.* **25** 91–99.

FRASER, D. A. S., REID, N. and WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264.

FRASER, D. A. S., WONG, A. and WU, J. (1999). Regression analysis, nonlinear or nonnormal: Simple and accurate p values from likelihood analysis. *J. Amer. Statist. Assoc.* **94** 1286–1295.

FRASER, D. A. S. and YI, G. Y. (2002). Location reparameterization and default priors for statistical analysis. *J. Iranian Statist. Soc.* **1** 55–78.

GODAMBE, V. P. (1982). Ancillarity principle and a statistical paradox. *J. Amer. Statist. Assoc.* **77** 931–933.

GODAMBE, V. P. (1985). Discussion of "Resolution of Godambe's paradox," by C. Genest and M. J. Schervish. *Canad. J. Statist.* **13** 300.

LUGANNANI, R. and RICE, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490.

REID, N. (1995). The roles of conditioning in inference (with discussion). *Statist. Sci.* **10** 138–157, 173–196.

WELCH, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Statist.* **10** 58–69.

Comment

Ronald W. Butler

1. INTRODUCTION

Can we now put to rest the unthinking and unqualified use of "global repeated sampling properties" as a means for probability computation and inference? Professor Fraser has forcefully and eloquently stated the case against the use of this principle when the model structure would suggest otherwise. In Section 7 paragraph 3 he concedes that "On the surface this (principle) seems hard to argue against…." However, after a careful reading of this paper, one must conclude from the multitude of examples and discussion that the unconditional and blanket use of this principle is seriously flawed. There are many modeling situations which would qualify for its use, particularly nonparametric modeling settings; however, the models presented here clearly do not.

My comments are divided into two parts. First some consideration of what these ideas about ancillarity and conditional inference might mean for predictive inference. This is followed by the bulk of the discussion, which presents a numerical example for a curved exponential family. No exact ancillaries are known for this example, but it will be shown that (i) the likelihood ancillary is particularly appropriate, (ii) the approximate p value suggested in (3.14) agrees with that in Barndorff-Nielsen (1990), expression (1.2), and (iii) the "sensitive direction" points tangent to the manifold created by holding the likelihood ancillary fixed at the data. The findings of the example pose further questions.

2. PREDICTIVE INFERENCE

The dual problem to parametric inference is predictive inference for unobserved z. Criterion II in Section 5 needs slight modification if z is to be inferred from observed y^0 using a parametric model. Criterion III seems particularly relevant to this setting: ignore the reason why z has not been observed, whether it be in the future or the past, or perhaps because it is

Ronald W. Butler is Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, USA (e-mail: butler@stat.colostate.edu).

a random effect in a model and therefore can never be observed.

The problem is dual because, rather than conditioning on ancillary statistics to make the inference more relevant to the model and data at hand, the reference set might now be a fixed value for the sufficient statistic. In the simplest setting considered in Butler (1986), suppose there is a sufficient statistic $s(y)$ of fixed dimension that agrees with that of the model parameter θ. If a generic value z is adjoined with y in (y, z) space, then the evidence for this z value should be with respect to the reference set $s(y^0, z)$. In entertaining the value of z as the unobserved value, then $s(y^0, z)$ conveys all relevant information about θ and hence about the state of the model used to make the predictive inference. In the same way that conditioning on ancillary a is used to convey "key observable characteristics of the underlying error" for a location model, conditioning on $s(y^0, z)$ provides complete information about the current state of understanding for θ within the parametric model setting. Fixing this understanding allows the model structure to make the prediction and also have it relevant to the current level of understanding about the model. Professor Fraser's arguments in Appendix A were also particularly interesting and relevant because they relate to determining the ancillary values in (y, z) space for prediction. The approach in Butler (1986) worked instead with orthonormal coordinates that are locally orthogonal to $s(y^0, z)$.

Barnard (1986) also suggested a pivotal approach to prediction, which is the dual procedure to the parametric inference in Section 4.2. Working with the location–scale model, his approach also used the marginalization step to remove dependence on all parameters to determine the marginal distribution of an ancillary $a(y, z)$. This ancillary is now transformed into predictive pivot $p(y, z)$ and predictive ancillary $q(y)$, with the latter quantity offering evidence for model criticism derived from data y. The conditional distribution of $p(y, z)$ given $q(y)$ evaluated at the data $y = y^0$ now provides the predictive extrapolation.

Based on the discussion above, it seems likely that the inferential structure proposed by Professor Fraser can neatly accommodate the dual problem of prediction. Other predictive approaches that attempted to extend higher-order asymptotic methods beyond the restriction of sufficiency include Butler (1989), Vidoni (1995) and Barndorff-Nielsen and Cox (1996). The first paper suggested that conditioning on the proper reference set [e.g., the maximum likelihood estimator (MLE) $\hat{\theta}(y^0, z)$] provides a more generally applicable principle than the restriction due to predictive sufficiency. For an overview of this issue and others, see Bjørnstad (1996).

3. GAMMA EXPONENTIAL EXAMPLE

An example is given that is similar to that considered by Pedersen (1981). The example is used to consider (and partially answer) the following questions and speculations:

1. What sort of ancillary, affine or likelihood, should be used for inference about θ and in what format?
2. Which ancillary is "more ancillary"?
3. What are the relationships between these ancillaries and the "sensitive" or ancillary directions suggested in the paper? Are there any deeper connections between the results of this paper and the suggestions of Barndorff-Nielsen (1990)?

3.1 Model and Ancillaries

A $(2, 1)$ curved exponential family may be defined by supposing that $y_1 \sim \text{Exponential}(\theta)$ independently of $y_2 \sim \text{Exponential}(e^\theta)$. To keep numerical computation simple, suppose the data are $y_1^0 = 1$ and $y_2^0 = 2$. The MLE is

$$\hat{\theta}^0 = \text{LambertW}(1/2) \simeq 0.3517$$

and solves an equation which, when rearranged, allows y_2 to be expressed in terms of $\hat{\theta}$ and y_1 as

$$y_2 = e^{-\hat{\theta}}(1/\hat{\theta} + 1 - y_1). \tag{1}$$

Two ancillaries are considered. The first is an affine ancillary a as discussed in Efron and Hinkley (1978) and Barndorff-Nielsen (1990), and sometimes named after the former authors. If vector $y = (y_1, y_2)'$ has mean μ_θ and covariance Σ_θ, then the affine ancillary is computed as the MLE of the Studentized vector or

$$\begin{aligned} a^2 &= (y - \mu_{\hat{\theta}})' \Sigma_{\hat{\theta}}^{-1} (y - \mu_{\hat{\theta}}) \\ &= (\hat{\theta} y_1 - 1)^2 + (e^{\hat{\theta}} y_2 - 1)^2 \end{aligned} \tag{2}$$

and $a^0 \simeq 1.954$. To compute the p^* density for conditionality resolution $\hat{\theta}|a$, the transformation $(y_1, y_2) \to (\hat{\theta}, a)$ needs to be inverted from (2), which leads to

$$y_1 = 1/\hat{\theta} - |a|/\sqrt{1 + \hat{\theta}^2} \tag{3}$$

with y_2 given in (1).

The second ancillary is a likelihood ancillary. It is defined through the process of completing the $(2, 1)$ curved exponential family so it is $(2, 2)$ with the

addition of another parameter $\chi > 0$. This is most simply done by assuming that $y_2 \sim \text{Exponential}(\chi e^{\theta})$ with the value $\chi = 1$ creating the curved exponential family. The likelihood ancillary is now based on the likelihood ratio test that $\chi = 1$. If $l_a(\theta, \chi)$ denotes the log-likelihood under the alternative, then the ancillary a_χ assumes the value

$$\begin{aligned} \tfrac{1}{2}a_\chi^2 &= l_a(\tilde{\theta}, \tilde{\chi}) - l_a(\hat{\theta}, 1) \\ &= -\ln\hat{\theta} + 1/\hat{\theta} - 1 - (1-\hat{\theta})y_1 \\ &\quad - \ln y_1 - \ln(1/\hat{\theta} + 1 - y_1), \end{aligned} \tag{4}$$

where $(\tilde{\theta}, \tilde{\chi})$ denotes the MLE under the alternative. In (4), any dependence on y_2 has already been replaced with y_1 using (1). Let the sign of a_χ^0 be $\text{sgn}(\tilde{\chi} - 1) = \text{sgn}(0.3517 - 1) = -1$ so that $a_\chi^0 \simeq -1.546$.

3.2 Which Ancillary and in What Format?

The format to be used for inference is the p^* density. It uses the likelihood shape to approximate the conditional density of $\hat{\theta}|a;\theta$ as the normalized $(d\hat{\theta})$ version of

$$p^\dagger(\hat{\theta}|a;\theta) = \sqrt{j_{\hat{\theta}}/(2\pi)}\exp\{l(\theta;\hat{\theta},a) - l(\hat{\theta};\hat{\theta},a)\}.$$

In the case of the conditioning on the observed affine ancillary a^0, plots of p^* (dashed), $p^\dagger$ (dotted) and the true density $f(\hat{\theta}|a^0;\theta)$ (solid) are shown in Figure 1 and are obtained through the inverse transformation $\hat{\theta}|a^0 \to (y_1, y_2)$ given in (3). As θ moves from $\theta = 4$ (top left), $2, 1,$ to $1/2$ (bottom right), the accuracy of p^* and $p^\dagger$ diminish markedly.

Compare this with the use of p^* and $p^\dagger$ when conditioning instead on the observed likelihood ancillary a_χ^0. Figure 2 shows the same quantities as its counterparts in Figure 1 as concerns the assessment of accuracy of p^* and $p^\dagger$ for their respective true densities. However, the true conditional densities are different in the two sets of plots since Figure 1 fixes $a = a^0$, while Figure 2 fixes $a_\chi = a_\chi^0$. Fixing a_χ^0 rather than affine a is a considerably more difficult computation since the inverse transformation $\hat{\theta}|a_\chi^0 \to (y_1, y_2)$ requires selecting the correct y_1 roots in (4) over a fine grid of $\hat{\theta}$ values. The true joint density of $(\hat{\theta}, a_\chi)$ has also been computed the same way but with the additional complication of a Jacobian determination based on implicit differentiation.

3.3 Which Ancillary Is "More Ancillary"?

The normalization constants $(d\hat{\theta})$ of the joint densities $f(\hat{\theta}, a_\chi^0;\theta)$ and $f(\hat{\theta}, a^0;\theta)$ provide the marginal densities $f(a_\chi;\theta)$ and $f(a;\theta)$, which should not

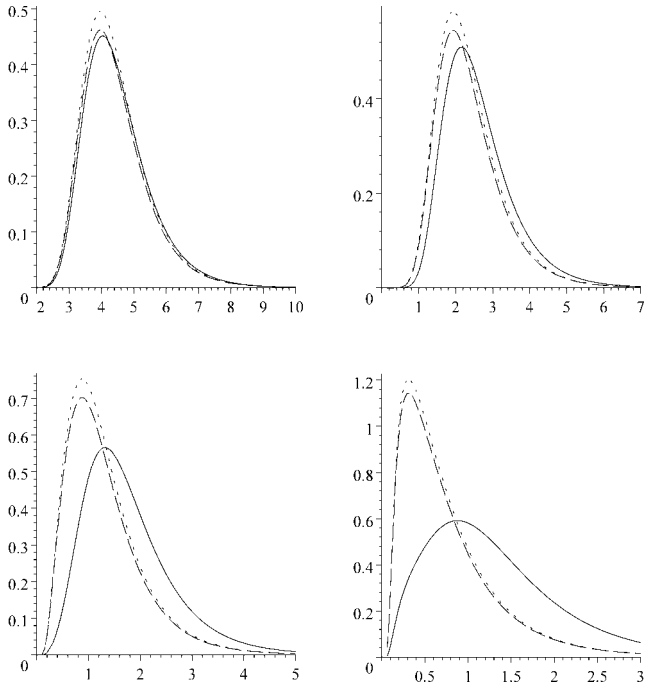

FIG. 1. *Densities for $\hat{\theta}$ in the gamma exponential example when conditioning on the affine ancillary $a^0 = 1.954$. The plots show a range of accuracy from good to poor and depict the exact density $f(\hat{\theta}|a^0;\theta)$ (solid), $p^\dagger(\hat{\theta}|a;\theta)$ (dotted) and $p^*(\hat{\theta}|a;\theta)$ (dashed) for $\theta = 4, 2, 1$ and $1/2$, respectively.*

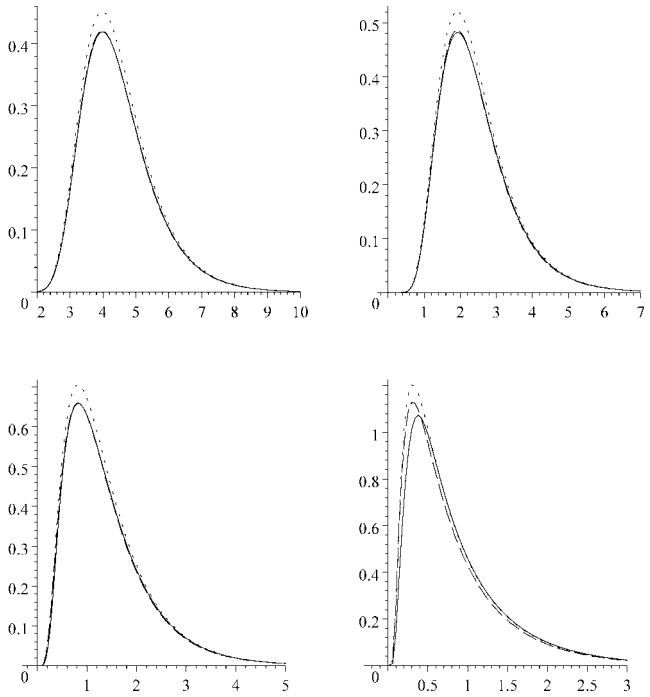

FIG. 2. *Densities for $\hat{\theta}$ when conditioning on the likelihood ancillary $a_\chi^0 = -1.546$. In each plot, $f(\hat{\theta}|a_\chi^0;\theta)$ (solid), $p^\dagger(\hat{\theta}|a_\chi^0;\theta)$ (dotted) and $p^*(\hat{\theta}|a_\chi^0;\theta)$ (dashed) are shown.*

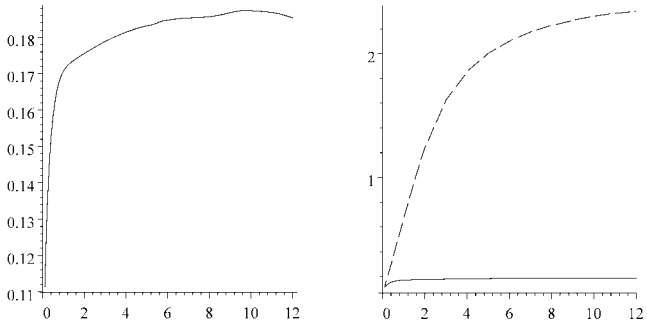

FIG. 3. *Marginal likelihood plots for* $f(a_\chi^0;\theta)$ (*solid*) *and* $f(a;\theta)$ (*dashed*) *versus* θ, *where* a_χ^0 *and* a *are the likelihood and affine ancillaries, respectively.*

show extraordinary dependence on θ if a and a_χ are "good ancillaries." Figure 3 plots $f(a_\chi^0;\theta)$ (solid) and $f(a^0;\theta)$ (dashed) versus θ. These plots show the marginal evidence about θ contained in each of the observed ancillaries. The observed likelihood ancillary is clearly more ancillary as revealed by the comparison in the right plot. All numerical computations for the likelihood ancillary here and in the previous subsection used the grid $\hat\theta \in \{0.02(0.04)9.98, 10\frac{1}{16}(\frac{1}{16})12, 12\frac{1}{8}(\frac{1}{8})16\}$. The superior performance of the likelihood ancillary was previously suggested in the asymptotics of Barndorff-Nielsen and Wood (1996). This superior performance can now be confirmed using a sample size of $n = 1$ for this dataset and model.

3.4 "Sensitive" Directions, *p* Value Computations and *r** Connections

For this example, the ancillary direction is computed as

$$v' = -(y_1/\hat\theta, y_2),$$

which leads to the data dependent parameterization

$$\varphi(\theta) = \theta y_1/\hat\theta + e^\theta y_2.$$

Computation of the standardized maximum likelihood departure value leads to

$$q(\theta) = \operatorname{sgn}(\hat\theta - \theta)|y_1(1-\theta/\hat\theta) + y_2(e^{\hat\theta} - e^\theta)| \cdot \sqrt{j_{\hat\theta}}|y_1/\hat\theta + e^{\hat\theta} y_2|^{-1}, \tag{5}$$

where

$$j_{\hat\theta} = 1/\hat\theta^2 + 1/\hat\theta + 1 - y_1.$$

At this junction, quite remarkably, it can be shown for any data (y_1^0, y_2^0), that $q(\theta)$ is analytically the same as the value for the standardized maximum likelihood departure u suggested in Barndorff-Nielsen (1990) as (1.4) and computed as in (5.5). We return to the implications of this equivalence below, but first pause to tabulate some p values in Table 1.

TABLE 1
p values $p^0(\theta)$ *for the various methods listed in the rows*

	θ				
Method[a]	**1/2**	**3/4**	**1**	**3/2**	**2**
Exact (trapezoidal)	0.189	0.0689	0.0194	0.0^3489	0.0^5120
(3.14) with q	0.238	0.0864	0.0239	0.0^3583	0.0^5140
Skovgaard	0.259	0.0990	0.0289	0.0^3796	0.0^5219
Normal	0.325	0.130	0.0392	0.0^2112	0.0^5315

[a]"Exact" refers to trapezoidal summation for $\Pr(\hat\theta < \hat\theta^0 | a_\chi^0;\theta)$, (3.14) accounts for the sensitive direction as well as Barndorff-Nielsen's (1990) value of u, Skovgaard (1996) computes u using the author's approximate sample space derivatives and Normal uses the normal approximation to r in (3.11).

Even for this $n = 1$ setting, the sensitive direction approach and that using Skovgaard's (1996) approximate sample space derivatives show remarkable accuracy, particularly for large θ. Taking the inference for θ further, the exact confidence interval by inverting $\Pr(\hat\theta < \hat\theta^0 | a_\chi^0;\theta)$ gives $(0.0276, 0.664)$ while (3.14) gives $(0.0446, 0.717)$ and Skovgaard's method gives $(0.0478, 0.748)$.

The analytical equivalence of Fraser's $q(\theta)$ with u from Barndorff-Nielsen's (1990) approach, which explicitly conditions on a_χ^0, suggests that the sensitive direction in which the directional derivative is taken in (3.13) to define $\varphi(\theta)$ is tangent to the manifold $\{(y_1, y_2) : a_\chi(y_1, y_2) = a_\chi^0\}$. This is indeed the case. Implicit differentiation of (1) to determine $\partial y_2/\partial y_1$, holding a_χ^0 fixed, requires the determination of $\partial\hat\theta/\partial y_1$ through (4). After long computations,

$$\partial y_2/\partial y_1 = \hat\theta y_2/y_1 = v_2/v_1, \tag{6}$$

the direction of v. At the data this slope is 0.7035.

Is this example merely a coincidence or are there any greater generalities to these agreements? To be tangent to the likelihood ancillary curve, the curve must be a solution to the differential equation in (6), which is complicated by the dependence of $\hat\theta$ on (y_1, y_2). General differential equation theory (see Ross, 1974, Theorem 1.1) only guarantees a local solution to (6) at the data, but this is all that is required for a local ancillary. This seems to say that the sensitive direction

approach has greater mathematical generality when a likelihood ancillary does not exist. If it does exist, when is the sensitive direction equal to or "close" to the direction of the likelihood ancillary curve? To what extent can the equivalence between Fraser's $q(\theta)$ and Barndorff-Nielsen's u be asserted with or without nuisance parameters in curved exponential families or in other classes of models?

Comment

Ib M. Skovgaard

INTRODUCTION

Fraser is to be thanked for his persistence in emphasizing the importance of conditional inference and ancillarity. A quick glance at the reference list reveals not only his immense stamina, but also the moderate amount of support from others. A few have taken the theory further, first of all Barndorff-Nielsen, but mainly in terms of asymptotic solutions while still leaving the question whether conditional inference given ancillary statistics has any logical justification. Despite scattered attempts to resolve this problem of frequentist inference, it seems that the majority of the statistical community has given up on the idea after some severe knock-outs around 1960. I am referring first to Basu (1959), who pointed out the lack of unique maximal ancillaries, thus raising not only the question which one to condition on, but more importantly why the argument for conditioning on one does not apply equally well to the other. Second I refer to Birnbaum (1962), who showed that conditioning on ancillaries as a principle together with basing inference on sufficient statistics implies the likelihood principle, which essentially is only met by orthodox Bayesian inference.

Despite these difficulties and the lack of general approval of any kind of ancillarity principle, conditioning on (some) ancillaries is used frequently in practice, almost unconsciously. In a clinical trial running over a certain period and allocating the incoming patients randomly to one of two treatments, say, the sample size is not given in advance and is an ancillary statistic. Few people would hesitate to consider sample size fixed when analyzing the data, and indeed it does seem very artificial to take into account that the trial might have comprised 100 patients if only 50 participated. This problem is conceptionally almost identical to Cox's artificially looking measuring instrument example. The only difference is that the distribution of the number of patients is not known, but this is hardly important for the argument and the distribution could probably be modeled reasonably if it were considered of importance.

My point is that problems of conditioning are not artificial philosophic problems of limited practical relevance, but should be considered more seriously in statistical practice. Fraser keeps reminding us of this, and on the main issues I agree entirely. I also agree that he has some good and very accurate asymptotic solutions through the methods he describes. There are still open problems and questions, however, conceptionally as well as asymptotically, and I do not find his solutions and arguments entirely convincing in all respects as I will try to substantiate below.

Initially let me point out, though, that my comments deal entirely with problems of frequentist inference. Bayesian inference (in its orthodox setting with a proper prior) avoids these problems and contradictions. In my view, Bayesian inference of this kind is obviously correct, but the problem is whether you can come up with a prior on which you want to base your conclusions. My experience is that this is rarely the case, and it would be a pity to give up the idea that reasonable inference can be made without a prior. The ancillarity problem is a central theme in the pursuit of the logic of such inference.

CONDITIONAL OR OPTIMAL INFERENCE?

If you believe that this is a relevant question, then you are already defeated if you support conditional inference. The point is that if you behave sensibly, nobody should be able to convince you that it is not optimal. In other words, if conditioning on ancillaries is the proper way, then this ought to drop out as an optimal method. Presently this is unfortunately not so. The question arises then whether the criteria used

Ib M. Skovgaard is Professor, Department of Natural Sciences, The Royal Veterinary and Agricultural University, DK-1871 Frederiksberg C, Denmark (e-mail: ims@kvl.dk).

for optimality are reasonable. Fraser suggests in Section 5.3 the modeling criteria I–III to help resolve the problem. In my taste these are too vague and imprecise to be of much help. To make a model for measurements that were made and not for measurements that were not made sounds reasonable, but to what extent does it restrict the modeling? Does it mean that censoring mechanisms, sample selection and biological variation (beside measuring errors) should not be modeled, for example? Some of Fraser's examples in Section 6 illustrate the problem with the limitation of his principles.

The best attempt I have seen so far toward a resolution of the problem is a recent article by Sundberg (2003) that quantified the intuitive feeling that, in the measuring instrument experiment, for example, the *relevant* variance is the one attached to the instrument we have actually used. Before reviewing this idea, let me digress a little to some simple basic considerations that must be kept in mind.

Several different probability distributions that describe the same events can all be correct, but some are more useful or more informative than others. Consider, for example, the probability that a man who belongs to a particular age group dies of a heart attack within a year. The frequency in the male population at that age is one answer, and indeed, if you check the distribution, it will turn out to be correct. Another distribution has probabilities either 1 or 0 for each individual: 1 if that person dies of a heart attack; 0 if not. Again this is a correct probability distribution and it is perfectly accurate, but it is not useful since we cannot use it for prediction. Suppose, however, that we could measure some variable, say cholesterol in the blood, that could distinguish to some extent between those who die and those who do not. Then we could ascribe probabilities closer to 1 or 0 than the population average and in this way bring us part of the way toward the accurate distribution. No doubt that this distribution is both more accurate and more useful than the population average. A quantification of the increased precision is the variance of the prediction probabilities, P_i, where i labels the individual. Let Z_i denote the indicator of death by heart attack, let p denote the population average and assume that $\mathrm{E}(Z|P) = P$, according to the requirement of "correctness." Then the squared prediction error is

$$\mathrm{SPE} = \mathrm{E}(Z - P)^2 = p(1-p) - \mathrm{var}\, P,$$

which decreases with increasing variance of the predictor.

Now we can leap to the confidence intervals which predict the event that the parameter is inside the interval with a certain probability, namely the confidence level, say 0.95. While this may be correct by a certain method, other methods might give more useful predictors. Here is where the conditioning on ancillaries comes in: It gives more useful prediction of the degree of confidence. This is intuitively obvious in Cox's example with the two measuring instruments. The "optimal" confidence interval, by whatever current optimality criterion, does not agree with the conditional confidence interval which uses the standard deviation of the measuring instrument actually used. Then, just as we can point out that some person with high cholesterol has a higher risk of heart attack, we can point out that some of the optimal confidence intervals have higher or lower chance of capturing the parameter. Such improved predictions can be made uniformly in the parameter and based on the same information that was used to construct the confidence intervals, so it ought to be clear that conditioning in this case is more useful, if not more correct. Now the hope is that the optimal method, in terms of usefulness (or relevance in the setting of Sundberg's paper), "automatically" is the conditional one, so there is no need for principles of ancillarity, only for optimizing with respect to the appropriate criteria.

There are several beneficial side effects of optimizing rather than conditioning on ancillaries as a matter of principle. First of all, it avoids the contradictions pointed out by Birnbaum and Basu, as mentioned in the Introduction. Second, a reasonable optimality criterion will be continuous with respect to the model, so that slight model changes will not alter the inference dramatically and so that ideally approximate (asymptotic) ancillarity may drop out as an (nearly) optimal result even if no exact ancillary has been found. This might be the case even for discrete data for which the current higher-order asymptotic approximations do not hold.

The above scenario is my understanding of the idea of Sundberg's paper, which has other arguments and a lot more detail; in particular, the superiority of conditional variance as a predictor of the actual squared error of the estimate. There are still open questions such as whether mean squared error of squared prediction error is the proper quantity to optimize and how to optimize tests and confidence intervals, but I think Sundberg's paper presents a breakthrough with regard to convincing arguments for conditioning.

ASYMPTOTIC SOLUTIONS

Since the paper by Barndorff-Nielsen (1986) the main problem in deriving highly asymptotically accurate p values has been to construct an (asymptotic) ancillary statistic that, jointly with the maximum likelihood estimate, is sufficient and for which we can calculate the local changes of the likelihood function and its first derivative with an infinitesimal change of the maximum likelihood estimate, not only at the observed value, but at any (hypothetical) value of the estimate. Quite remarkably these so-called sample–space derivatives are all that is needed to calculate the highly accurate approximations to the tail probabilities that constitute the p values. Let me contrast Fraser's way of achieving this with my own (Skovgaard, 1996), starting with the latter. Since we need only the local change of the likelihood function, it suffices to know the derivatives of a sufficient statistic. A representation of this is the (infinite) number of derivatives of the log-likelihood function at that point, the first one being the score function, which vanishes at that point, of course. The derivative of the score function with respect to the estimate at this point is necessarily the observed Fisher information. Now the changes of all derivatives of higher order may be approximated by regressing them on the score statistic. This gives all information that is needed and provides an explicit solution. Fraser deliberately discards the fact that local changes are required only for a sufficient statistic and describes instead the local changes of the entire set of observations. While this, at first, may seem unnecessarily complicated, it does make natural constructions more readily available, essentially by keeping the quantiles in the estimated distribution fixed, as he shows in the present paper and as discussed below. Furthermore, since a solution is obtained this way, the construction of ancillary directions on the (bigger) entire observation space can hardly be said to be a drawback. While my suggestion is fairly easy and general, Fraser's is undoubtedly better suited for location models and some other group models. It should be noted, though, that for group models the local changes of the log-likelihood are also fairly easily written down exactly given the maximal invariant ancillary statistic, thus providing the same exact sample space derivatives as obtained by Fraser.

There is little doubt that Fraser's suggestion of sample-space derivatives, [see, e.g., (4.12)], in combination with the Lugannani–Rice approximation (3.14) or the Barndorff-Nielsen approximation (3.15), provides highly accurate p values for a large number of situations. Also, I would be surprised if they turned out to be much different in practice from alternatives along the same direction, such as the r^*-type statistics given by Jensen (1997) and Skovgaard (1996) or the original and principal, but less operational, version by Barndorff-Nielsen (1986). So the following queries really concern details regarding the principal differences between the various approaches. First a little more background on Fraser's method.

Fraser argues that the pivots $F(y_i, \theta)$ should be kept constant (at least locally) along any level surface of the conditioning statistic as a function of the maximum likelihood estimate, $\hat{\theta}$, when this is plugged into the pivot; see the equation defining the tangent vectors v_1 and v_2 just above (4.12) and recall the analogy with the location models. This has some good sides: This pivot is a natural choice which is close to the measurement process in a heuristic sense and its use to define the tangent directions is an excellent way to make use of a pivot, after many years of less successful attempts going back to Fisher's arguments in favor of fiducial inference. The method also raises some questions, however.

The first question is a bit technical and has to do with the existence of the ancillary statistic and whether it agrees with the statistic $(F(y_1; \hat{\theta}), \ldots, F(y_n; \hat{\theta}))$, considering only the case of independent replications. This statistic formally gives the tangent directions above (4.12) and Fraser's subsequent reference to Fraser and Reid (2001) suggests that the same tangent vectors arise from their approximate ancillary. Hence I deduce that their approximate ancillary statistic is the vector of quantiles in the estimated distribution, or am I wrong about that? If so, my problem is then whether conditioning on this statistic may not exclude certain parameter values or even, in some cases, exclude all but one, such that the conditional distribution is degenerate. In other words, the set of observation vectors that gives rise to a certain value of the maximum likelihood estimate, say θ_1, corresponds to a certain set of quantile vectors, but this will not in general be the same set of quantiles that correspond to another estimate θ_2, I suspect. So what is approximate here, the ancillarity of the statistic, the tangent vectors of the approximate ancillary statistic, or the existence of the statistic so that it is merely a technical device for obtaining unconditional p values of high quality?

The second question regards Fraser's claim that we hardly need sufficiency (Section 3.6 and Appendix A), because whatever we may achieve by sufficiency reductions may also be achieved by conditioning. In the

same vein one might argue that conditional p values are superfluous because they are also valid unconditionally and may therefore be obtained without conditioning. I do not agree with either of these arguments: sufficiency *restricts* the choice of method in a way that conditioning does not and conditioning restricts the permissible results compared to unconditional methods, and such restrictions may be useful because they guide our method of inference.

This leads to the third question: Does the p value as obtained here by Fraser depend on the data in other ways than through the sufficient statistic? I suspect that it may and that we do not quite agree whether this is an advantage. I highly respect the viewpoint that aspects other than the model should be taken into account (e.g., robustness considerations), but whether such aspects should enter the model-based part of the inference directly is another matter.

Let me conclude by emphasizing that the foregoing comments are of little concern compared with the excellent results obtained. I congratulate Fraser for achieving these results along the lines of conditioning and ancillarity that he has stubbornly pursued since the early days of his scientific career.

Comment

Rudolf Beran

The concepts of sufficiency, ancillarity and conditional inference are parts of a classical statistical theory that treats data as a random sample from a probability model with relatively few parameters. In discussing Don Fraser's paper, I will consider the place of these and related concepts in the evolution of statistics.

1. THE EVOLUTION OF STATISTICAL THEORY

Reliance on probability theory in statistical writing spans the spectrum from none to fixed effects models to random effects models to Bayesian reasoning. One factor is the extent to which an author regards probability as a feature of the natural world. For a Bayesian, probability measures the strength of opinions, which are modelled by a sigma algebra. At the other end of the spectrum, illustrated by Tukey's (1977) *Exploratory Data Analysis*, data-analytic algorithms are basic reality and probability models are hypothetical constructs.

A second factor is the technological environment in which an author is writing. Until the late 1950s, the tools available to a statistician consisted of mathematics, logic, mechanical calculators and simple computers. Because calculation was laborious, writers on statistical theory thought in terms of virtual data governed by probability models that involved relatively few parameters. Indeed, the great intellectual advances made in probability theory during the twentieth century made this approach the technology of choice. Thus, the hotly debated statistical theories formulated in Wald's (1950) *Statistical Decision Functions*, Fisher's (1956) *Statistical Methods and Scientific Inference* and Savage's (1954) *The Foundations of Statistics* shared a common reliance on relatively simple probability models.

After 1960, results on weak convergence of probability measures provided the technology for major development of asymptotic theory in statistics. Notable achievements by 1970 included (a) the clarification of what is meant by asymptotic optimality, (b) the understanding, through Le Cam's work, that risks in simple parametric models can approximate risks in certain more general models, (c) the discovery of superefficient estimators whose asymptotic risk undercuts the information bound on sets of Lebesgue measure zero and (d) the remarkable discovery, through the James–Stein estimator, that superefficient estimators for parameters of sufficiently high dimension can dominate classical estimators globally. These findings set the stage for the vigorous subsequent development of robustness, of nonparametrics and of biased estimators in models with many or an infinite number of parameters. Theoretical study of Efron's (1979) bootstrap benefited from the evolution in asymptotic theory. In turn, the bootstrap and iterated bootstrap provided intuitive algorithms for realizing in statistical practice the benefits of erudite asymptotic improvements.

Mathematical logicians investigating the notion of proof had greatly refined the concept of algorithm by

Rudolf Beran is Professor, Department of Statistics, University of California, Davis, California 95616, USA (e-mail: beran@wald.ucdavis.edu).

mid-century (cf. Berlinksi, 2000). Through the technological development of digital computers, programming languages, video displays, printers and numerical linear algebra, stable computational algorithms enriched the statistician's toolbox. In consequence, a wider range of statistical procedures, numerical and graphical, became feasible. Case studies and experiments with artificial data offered nonprobabilistic ways to understand the performance of statistical procedures. The fundamental distinctions among data, probability model, pseudorandom numbers and algorithm returned to prominence. The extent to which deterministic pseudorandom sequences can imitate properties of random variables received more attention (cf. Knuth, 1969). It became clear once again that data are not certifiably random. Computing technology provided a new environment in which to extend and reconsider statistical ideas developed with probability technology. The bootstrap is a case in point.

From our present technological standpoint, statistics is the development, study and implementation of algorithms for data analysis.

How is a data-analytic algorithm to be understood? One answer, offered by Brillinger and Tukey (1984), addressed the gap between statistical theory and data-analytic techniques:

> If our techniques have no hypotheses, what then do they have? How is our understanding of their behavior to be described?
> As a generalization of an umbra within a penumbra. Here there are at least three successively larger regions, namely:
>
> 1. An inner core of proven quality (usually quite unrealistically narrow) . . . .
> 2. A middle-sized region of understanding, where we have a reasonable grasp of our technique's performance . . . .
> 3. A third region, often much larger than the other two, in which the techniques will be used . . . .

For example, the inner core of understanding could be an analysis under a simple probability model; the middle core could be asymptotic analyses and simulations under substantially more general probability models together with salient case studies; and the outer core would contain data analyses that use the techniques. In reality, data consist of scientific and other contexts as much as numerical observations.

For some statistical problems, such as classification of handwritten digits, probability models may not generate effective procedures. Breiman (2001) observed:

> If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

His paper emphasized algorithmic modeling techniques that treat the data mechanism as essentially unknown.

How are data-analytic algorithms to be implemented? One answer, offered by the Omega-hat project (www.omegahat.org), focuses on open-source development of the next generation of statistical computing paradigms, environments and software. The project provides an optionally typed language that extends both S and JAVA, and a customizable, multithreaded interpreter; it encourages participation by those wanting to extend computing capabilities in one of the existing languages for statistical computing, by those interested in distributed or web-based statistical software and by those interested in the design of new statistical languages.

This answer recognizes that software provides a powerful new medium for expressing statistical ideas. The Introduction to McLuhan's (1964) book *Understanding Media*: *The Extensions of Man* began:

> In a culture like ours, long accustomed to splitting and dividing all things as a matter of control, it is sometimes a bit of a shock to be reminded, in operational and practical fact, that the medium is the message.

In other words, the nature of a medium has at least as much effect on human activity as does its content, which itself is just an older medium that is being expressed through the newer medium. In this manner, leading edge statistical computing environments stand to influence core ideas about statistics.

Fraser's paper examines pros and cons of conditional versus unconditional inference in classical probability models for data where the parameter of interest is one-dimensional. His examples indicate that these approaches can yield procedures with differing probabilistic properties. A diversity of answers is to be expected once we recognize the difference between data and probability model. In my discussion, I will consider (a) the construction of simultaneous confidence sets, a problem that intrinsically has multiple answers with different properties, and (b) estimation of the means in a two-way layout with one observation per combination of factor levels, a typical multiparametric problem where neither sufficiency nor

ancillarity is of immediate help. I will argue that, within the statistical environment created through technological advances in asymptotic theory and computing, the role of ancillarity and sufficiency is narrow.

Narrow is not the same as none. For example, Hájek's convolution theorem in locally asymptotically normal families has a form in which asymptotic sufficiency, asymptotic ancillarity and Basu's theorem suggest a heuristic interpretation of necessary and sufficient conditions for consistency of parametric bootstrap distributions. The interested reader is referred to Beran (1997) and, for a tangentially related nonparametric discussion, to van Zwet and van Zwet (1999).

2. SIMULTANEOUS CONFIDENCE SETS

Coverage probability under a probability model does not, by itself, determine a confidence set. A further design goal, whether minimum expected geometrical size or equal conditional coverage probabilities, is needed to construct the confidence set. Geometrical size may be of interest if the confidence set is to serve as a set-valued estimator of the parameter. Equal conditional coverage probabilities may be of interest if the conditioning variable reflects a real feature in the data. An experienced statistician selecting a data-analytic algorithm will consider the context, and aims of the analysis as well as probability models.

The construction of simultaneous confidence sets has raised issues analogous to those in Fraser's second section. Consider a statistical model in which a sample X_n of size n has joint probability distribution $P_{\theta,n}$, where $\theta \in \Theta$ is unknown. The parameter space Θ is an open subset of a metric space, whether of finite or infinite dimension. Of interest is the parametric function $\tau = T(\theta)$, where T is a specified function on Θ. Suppose that τ has *components* $\{\tau_u = T_u(\theta) : u \in U\}$, U being a metric space, which jointly determine τ. For each u, let C_u denote a confidence set for the component τ_u. By simultaneously asserting the confidence sets $\{C_u : u \in U\}$, we obtain a simultaneous confidence set C for the components $\{\tau_u\}$.

If the components $\{\tau_u\}$ are deemed logically similar, the statistician may wish to construct the confidence sets $\{C_u\}$ in such a way that

$$(1) \qquad P_{\theta,n}[C_u \ni \tau_u] \text{ is the same } \forall\, u \in U$$

and

$$(2) \quad P_{\theta,n}[C_u \ni \tau_u,\ \forall\, u \in U] = P_{\theta,n}[C \ni \tau] = \beta.$$

Property (1) is called *balance*. It reflects the wish that the confidence set C treat the logically similar components τ_u in an even-handed way while controlling the simultaneous coverage probability (2). The balance constraint is a cousin to the equal conditional coverage probability condition treated in Fraser's second section.

One general approach starts with a *root* $R_{n,u} = R_{n,u}(X_n, \tau_u)$ for each component τ_u. The root may or may not be an exact pivot. Let $\mathcal{T}_u$ and $\mathcal{T}$ denote, respectively, the ranges of $\tau_u = T_u(\theta)$ and $\tau = T(\theta)$. Every point in $\mathcal{T}$ can be written in the component form $t = \{t_u : u \in U\}$. The simultaneous confidence sets to be considered are

$$(3) \quad C = \{t \in \mathcal{T} : R_{n,u}(X_n, t_u) \le c_u(\beta),\ \forall\, u \in U\}.$$

The technical problem is to devise critical values $\{c_u(\beta)\}$ so that, to a satisfactory approximation, C is balanced and has simultaneous coverage probability β for the $\{\tau_u\}$.

Let $H_{n,u}(\cdot, \theta)$ and $H_n(\cdot, \theta)$ denote the left-continuous cumulative distribution functions (c.d.f.'s) of $R_{n,u}$ and of $\sup_{u\in U} H_{n,u}(R_{n,u}, \theta)$, respectively. If θ were known and the two c.d.f.'s just defined were continuous in their first arguments, an oracle choice of critical values for the component confidence sets would be $c_u(\beta) = H_{n,u}^{-1}[H_n^{-1}(\beta, \theta), \theta]$. The oracle component confidence set

$$(4) \qquad \begin{aligned} C_u &= \{t_u \in \mathcal{T}_u : R_{n,u}(X_n, t_u) \le c_u(\beta)\} \\ &= \{t_u \in \mathcal{T}_u : H_{n,u}(R_{n,u}, \theta) \le H_n^{-1}(\beta, \theta)\} \end{aligned}$$

has coverage probability $H_n^{-1}(\beta, \theta)$ for τ_u. The oracle simultaneous confidence set C, defined through (3), has coverage probability β for τ by definition of $H_n(\cdot, \theta)$. In historically influential special cases, this oracle construction can be carried out because neither $H_{n,u}$ nor H_n depends on the unknown θ.

EXAMPLE 1. Suppose that X_n has a $N(A\gamma, \sigma^2 I_n)$ distribution, where the vector γ is $p \times 1$ and the matrix A has rank p. The unknown parameter $\theta = (\gamma, \sigma^2)$ is estimated by $\hat{\theta}_n = (\hat{\gamma}_n, \hat{\sigma}_n^2)$ from least squares theory. Suppose that the root

$$(5) \qquad R_{n,u} = |u'(\hat{\gamma}_n - \gamma)|/\hat{\sigma}_{n,u},$$

where u is a p-dimensional vector and $\hat{\sigma}_{n,u}^2 = u'(A'A)^{-1}u\hat{\sigma}_n^2$. The roots $\{R_{n,u}\}$ are identically distributed, each having a t distribution, folded over at the origin, with $n - p$ degrees of freedom.

Suppose that U is a subspace of R^p of dimension q. Then $\sup_{u\in U} R_{n,u}$ is a continuous pivot (cf. Miller,

1966, Chapter 2, Section 2). In this instance, the oracle balanced simultaneous confidence set defined by (3) and (4) coincides with Scheffé's simultaneous confidence intervals for the linear combinations $\{u'\gamma : u \in U\}$.

EXAMPLE 2. Specializing to a balanced one-way layout, suppose that U consists of all pairwise contrasts. The parameter γ is just the vector of means in this case of the linear model. Then $\sup_{u\in U} R_{n,u}$ is a continuous pivot (cf. Miller, 1966, Chapter 2, Section 1). In this instance, the oracle balanced simultaneous confidence set defined by (3) and (4) coincides with Tukey's simultaneous confidence intervals for all pairwise differences in means.

The exact pivots used by Tukey and Scheffé in constructing their respective balanced simultaneous confidence intervals do not exist in most probability models. However, bootstrap techniques enable more general construction of simultaneous confidence sets that behave asymptotically like oracle simultaneous confidence sets. Suppose that $\hat{\theta}_n$ is a consistent estimator of θ. Replacing θ by $\hat{\theta}_n$ in the oracle critical values that appear in (4) yields bootstrap simultaneous confidence sets for the $\{\tau_u\}$. A Monte Carlo approximation to the bootstrap critical values requires only one round of bootstrap sampling. Computation of the supremum over U may require further approximations when the cardinality of U is not finite. In practice, the case of a finite number of components $\{\tau_u\}$ is both approachable and important. Theorem 4.1 in Beran (1988) provides sufficient conditions under which the bootstrap simultaneous confidence set is asymptotically balanced and has asymptotic overall coverage probability β.

The balance condition (1) on the simultaneous confidence sets is a design element that can be modified at will. Technically speaking, we could seek specified proportions among the componentwise coverage probabilities. (I am not aware of a problem where this would be useful.) The Tukey and Scheffé exact constructions and, more generally, the bootstrap construction are readily modified to handle this design goal. On the other hand, balance has not been found compelling in situations where the components $\{\tau_u\}$ are not logically comparable.

EXAMPLE 3. Given an independent identically distributed sample from the $N(\mu, \sigma^2)$ distribution, it is easy to construct a balanced simultaneous confidence set of coverage probability β for the pair (μ, σ^2). However, this is not a popular procedure, no doubt because the parameters μ and σ^2 are logically dissimilar.

The discussion in this section illustrates how advances in asymptotic and computer technology have given statisticians the ability to explore beyond the statistical principles of earlier eras, principles whose formulation captures, as in amber, the technological environment of their times.

3. MULTIPARAMETRIC TWO-WAY LAYOUTS

Consider a high-dimensional two-way layout with one observation per combination of factor levels. Factor k has p_k levels $\{t_{kj} : 1 \le j \le p_k\}$, which may be nominal or ordinal. Such a two-way layout is associated with experimental designs, gray-scale images and gene chips. Subscripting is arranged so that, for an ordinal factor, the factor levels are a strictly increasing function of subscript. A simple probability model asserts that

$$(6) \quad y_{ij} = m_{ij} + \varepsilon_{ij}, \quad 1 \le i \le p_1,\ 1 \le j \le p_2,$$

where the $\{y_{ij}\}$ are the observations, $m_{ij} = \mu(t_{1i}, t_{2j})$ and the errors $\{\varepsilon_{ij}\}$ are independent, identically distributed $N(0, \sigma^2)$ random variables. The function μ and the variance σ^2 are unknown. A basic problem is to estimate the means $\{m_{ij}\}$ and σ^2.

For the means in model (6), the minimum variance unbiased (MVU) estimator and the minimum quadratic risk location equivariant estimator both coincide with the raw data. This estimator is unacceptable in contexts such as image processing or estimation of response surfaces. Indeed, Stein (1956) showed that the MVU is inadmissible under quadratic loss whenever the number of factor-level combinations $p = p_1 p_2$ exceeds 2. Neither reduction by sufficiency nor by ancillarity suggests a satisfactory estimator of the means in model (6). A partial exception to this claim holds for the one-way layout with nominal factor levels, but does not handle ordinal factor levels (cf. Beran, 1996).

What does work is regularization, the use of a constrained fit to the means that trades bias for variance so as to achieve lower risk in estimating the means of the two-way layout. Regularization is an estimation strategy for models that have many or an infinite number of unknown parameters—models that play a prominent role in modern statistics. A regularized fit is typically constructed in three stages. First, we devise a candidate class of constrained mean estimators that individually expresses competing prior notions about the unknown means. Second, we estimate the risk of each candidate estimator under a general model that does *not* assume any of the prior notions in step one.

Third, we define the regularized fit to be a candidate fit that minimizes estimated risk or a related criterion. This regularized fit may be interpreted as the trend discernible in the noisy observations.

In the two-way layout, let y denote the $p \times 1$ vector obtained by ordering the observations $\{y_{ij}\}$ in mirror dictionary order: The first subscript runs faster than the second subscript. Let m denote the similarly vectorized means $\{m_{ij}\}$. Model (6) asserts that the distribution of y is $N(m, \sigma^2 I_p)$. For $k = 1, 2$, define the $p_k \times 1$ vector u_k and the $p_k \times p_k$ matrices J_k, H_k by

$$\begin{aligned} u_k &= p_k^{-1/2}(1, 1, \ldots, 1)', \\ J_k &= u_k u_k', \\ H_k &= I_{p_k} - J_k. \end{aligned} \tag{7}$$

For each k, the symmetric idempotent matrices J_k and H_k have rank (or trace) 1 and $p_k - 1$, respectively. They are thus orthogonal projections that decompose R^{p_k} into two mutually orthogonal subspaces of dimensions 1 and $p_k - 1$. The identity $I_{p_k} = J_k + H_k$ implies that

$$\begin{aligned} m &= (I_{p_2} \otimes I_{p_1})m \\ &= P_0 m + P_1 m + P_2 m + P_{12} m, \end{aligned} \tag{8}$$

where $P_0 = J_2 \otimes J_1$, $P_1 = J_2 \otimes H_1$, $P_2 = H_2 \otimes J_1$ and $P_{12} = H_2 \otimes H_1$. Equation (8) gives, in projection form, the analysis of variance (ANOVA) decomposition of a complete two-way layout of means into overall mean, main effects and interactions.

Certain penalized least squares criteria generate a class of candidate estimators by restricting, in varying degree, the ANOVA decomposition. Let A_k be any matrix with p_k columns such that $A_k u_k = 0$. Examples of such *annihilator* matrices are $A_k = H_k$, suitable when factor k is nominal, and A_k equal to the dth difference matrix, suitable when factor k is ordinal with equally spaced factor levels $\{t_{kj}\}$. Let $B_k = A_k' A_k$ and define $Q_1 = J_2 \otimes B_1$, $Q_2 = B_2 \otimes J_1$ and $Q_{12} = B_2 \otimes B_1$. Let $A = \{A_1, A_2\}$ and let $\nu = (\nu_1, \nu_2, \nu_{12})$ be any vector in $[0, \infty]^3$. The *candidate penalized least squares* (PLS) estimator of m is $\hat{m}_{\mathrm{PLS}}(\nu, A) = \arg\min_m S(m, \nu, A)$, where

$$\begin{aligned} &S(m, \nu, A) \\ &\quad = |y - m|^2 + m'(\nu_1 Q_1 + \nu_2 Q_2 + \nu_{12} Q_{12})m. \end{aligned} \tag{9}$$

The symmetric matrix B_k has spectral decomposition $U_k \Lambda_k U_k'$, where $\Lambda = \mathrm{diag}\{\lambda_{ki}\}$ is diagonal with $0 = \lambda_{k1} \le \lambda_{k2} \le \cdots \le \lambda_{kp_k}$ and the eigenvector matrix U_k is orthonormal with first column equal to u_k. Let $f_{ij}(\nu) = [1 + \nu_1 \lambda_{1i} e_{2j} + \nu_2 e_{1i} \lambda_{2j} + \nu_{12} \lambda_{1i} \lambda_{2j}]^{-1}$, where $e_{k1} = 1$ and all other $\{e_{kj}\}$ vanish. Vectorize the $\{f_{ij}(\nu)\}$ in mirror dictionary order to obtain the vector $f(\nu)$ and let $z = (U_2 \otimes U_1)' y$. It follows readily that the candidate PLS estimator is the shrinkage estimator

$$\hat{m}_{\mathrm{PLS}}(\nu, A) = (U_2 \otimes U_1)\,\mathrm{diag}\{f(\nu)\} z. \tag{10}$$

Let $\xi = (U_2 \otimes U_1)' m$ and let $\mathrm{ave}(h)$ denote the average of the components of vector h. The normalized quadratic risk of the (usually biased) candidate estimator (10) is

$$\begin{aligned} &p^{-1} \mathrm{E} |\hat{m}_{\mathrm{PLS}}(\nu, A) - m|^2 \\ &\quad = \mathrm{ave}\big[f^2(\nu)\sigma^2 + \big(1 - f(\nu)\big)^2 \xi^2 \big]. \end{aligned} \tag{11}$$

The operations inside the average are performed componentwise, as in the S language.

Having devised a variance estimator $\hat{\sigma}^2$ by some form of pooling, say, we may estimate the risk (11) by

$$\begin{aligned} &\hat{r}(A, \nu) \\ &\quad = \mathrm{ave}\big[f^2(\nu)\hat{\sigma}^2 + \big(1 - f(\nu)\big)^2 (z^2 - \hat{\sigma}^2) \big]. \end{aligned} \tag{12}$$

This is just Stein's unbiased risk estimator with σ^2 replaced by $\hat{\sigma}^2$. For a specified class $\mathcal{A}$ of annihilator pairs A, we define the *adaptive PLS* estimator of m to be the candidate PLS estimator with smallest estimated risk:

$$\begin{aligned} &\hat{m}_{\mathrm{PLS}} = \hat{m}_{\mathrm{PLS}}(\hat{\nu}, \hat{A}) \\ &\qquad \text{where } (\hat{\nu}, \hat{A}) = \operatorname*{arg\,min}_{(A, \nu) \in \mathcal{A} \times [0, \infty]^3} \hat{r}(A, \nu). \end{aligned} \tag{13}$$

This adaptive estimator is an empirical approximation to the oracle candidate PLS estimator that minimizes the unknown risk (11) over $(A, \nu) \in \mathcal{A} \times [0, \infty]^3$.

Computational algorithms, case studies, and multiparametric asymptotics for $\hat{m}_{\mathrm{PLS}}$ were developed by Beran (2002). Under model (6), subject to restrictions on the richness of the annihilator class $\mathcal{A}$ and to assumptions that ensure consistency of $\hat{\sigma}^2$, the risk of the adaptive PLS estimator $\hat{m}_{\mathrm{PLS}}$ converges to that of the oracle candidate estimator as the number of factor-level combinations tends to infinity. By construction, this limiting risk cannot exceed that of the MVU estimator. In case studies, it is not unusual for the adaptive PLS estimator to reduce risk by a factor of 3 or more over that of the MVU estimator. For two-way layouts with nominal factors, the adaptive PLS estimator generated by $A_k = H_k$ essentially coincides with the multiple shrinkage estimator studied by Stein (1966). For two-way layouts with ordinal factors, the adaptive

PLS estimator based on local polynomial annihilators can be strikingly more efficient than the MVU estimator and is akin to spline fits in two-way functional data analysis.

The foregoing discussion of the two-way layout illustrates a technology developed over the past five decades for better estimation in multiparametric and nonparametric models. The role of sufficiency and ancillarity has been inconsequential in this substantial portion of modern statistics.

ACKNOWLEDGMENT

This research was supported in part by NSF Grant DMS-03-00806.

Rejoinder

D. A. S. Fraser

1. INTRODUCTION

My appreciation and thanks go to the reviewers for their thoughtful and careful comments on ancillaries and conditional inference. It is a special delight to be able to participate further in ongoing discussion of the topics.

As part of this, I express my sincere thanks to the Editor, George Casella, for advice on the final versions of the paper and for arranging discussants who have a wide spectrum of views. I wish also to acknowledge the very large contributions of the previous Editor, Leon Gleser, for his encouragement over many decades to bring to written form an examination of ancillaries and conditional inference; indeed his support stems from his days as a graduate student at Stanford, where I had the good fortune and opportunity to argue closely with him.

The three discussants express very different views with very little overlap of the points they raise. Rudy Beran expresses the view that "statistics is the development, study and implementation of algorithms for data analysis." While I fully share Rudy's enthusiasm for such data algorithms and their importance, I hesitate on such a catholic view that the whole of statistics is algorithms. Indeed the claim is not dissimilar to that of decision theory in the mid-twentieth century, that statistics is just deductive behavior or the application of decision rules: just change "data algorithm" to "decision algorithm." The escape from that decision theory philosophy is only partial at best and perhaps an overemphasis on data algorithms would assist the escape. Surely there is a large place for determining what is known in any context of interest and for not being pressed into such extreme discipline directions. Indeed it is now possible in some generality to report the total inference information from a model-with-data investigation; see Sections 2 and 3.

Ron Butler enquires concerning predictive inference and the extent to which the conditioning methods can be applied. Clearly they can be applied, and one direction involves treating a probability for a future observation as just a parameter of interest for the original model and then proceeding with the conditioning approach.

Ron also examines in detail how the sensitivity directions approach works in comparison with some alternative ancillary methods. He does find that they uncover familiar ancillaries, thus allowing a more predictable and mechanical access to the methods based on such ancillaries. Of course, in addition, the sensitivity approach provides an easy and direct access to the new high accuracy approximation methods, which are thus available quite generally for wide areas of application.

He then reports on a simulation to compare various methods for a simple exponential example and finds that they compare favorably with an exact calculation. We find that the exact distribution for the conditional case can be examined directly and present simulation results that show the new methods are even closer to the truth than Ron's calculations suggest; see Sections 4 and 5.

Ib Skovgaard notes the widespread lack of professional statistical approval for conditioning and that, despite this, conditioning is in fact frequently used in practice. He provides a persuasive example. While he does not directly address the stigmata connected with conditioning or the social origins of the stigmata, he does speak positively of many aspects of conditioning, and indeed asks whether or how a conditioning imperative might be derived from some optimality approach. He then comments on aspects of conditioning,

and raises insightful and serious questions about many aspects of the recommended methodology; see Sections 6–9. Section 10 then provides a brief overview.

2. STATISTICS IS . . . ALGORITHMS FOR DATA ANALYSIS

Data-analysis algorithms are undoubtedly making major contributions to statistical methods and will continue to do so, often in areas neglected or underexamined by traditional mainstream statistics. However, the suggestion that they form all of statistics seems more a measure of the enthusiasm and optimism of those involved. In particular, the implied suggestion that statistical models and data have largely been superceded by the algorithms neglects immense important areas of statistics. Consider the use of many generalized linear models or the analysis of categorical data arrays which just now are amenable to the recent higher-order likelihood methods. Would you want the possible benefits of a new drug therapy to be evaluated by an algorithm?

Rudy provides four examples that illustrate an "experienced statistician selecting a data-analytic algorithm. . . ." All four examples involve normality with independent errors and common variance, a very specialized textbook-type formulation. All yield quite easily what can be called a presentation of the total inference information. What Rudy presents are innovative ways to repackage this total inference information for specific interests or purposes, something to which he has made substantial contributions.

The examples could at least have involved nonnormality. We all acknowlege that data rarely come to us as if from the skinny-tailed normal. Then if the analysis is based on the nonnormal case, we would find that ancillary inference procedures are needed. Indeed these alternatives are the focus of the paper and are available in the literature; for a range of examples in the general regression context, see Fraser, Wong and Wu (1999).

This is not to say that repackaging the total inference information for specific purposes in not important; it certainly is, but it is secondary to the considerations here.

3. SUFFICIENCY AND ANCILLARITY

Rudy notes, "The role of sufficiency and ancillarity has been inconsequential in this substantial portion of modern statistics" and then discusses the four examples mentioned above. All the examples have independent normal errors with common variance, and because of the rotational symmetry of the composite normal error distribution, everything factors into independent normal pieces, each addressing a different orthogonal parameter or addressing pure error with mean zero: the simplicity of the familiar analysis of variance context!

The examples are to illustrate how to repackage the full inference information to focus on particular interest parameters. This is of course an important area and one to which Rudy has made strong contributions. It is also one that the paper addresses, for scalar parameters; see Fraser (2003) for more general cases. However, neither sufficiency nor ancillarity is concerned with this information repackaging. If the examples had departed from the over simple normal, then ancillarity would play a major role, as described in the paper. To ignore the then available structure and default to marginal methods is to blatently throw away the well defined information. Of course the repackaging is still needed, but you address it from what you know.

In the paper a prominent theme concerning sufficiency is that in broad generality it is not needed for statistical analysis and, indeed, that it has lulled the theoretical side of statistics into complacency so that effective alternatives are not discussed or investigated. So with regard to sufficiency, I fully support Rudy's view that it is not needed, and indeed go further and suggest that its widespread acceptance has been seriously damaging to statistics.

What then ancillarity? For such normal examples ancillarity works, but can be ignored because with simple normal error everything factors into independent pieces as described above. Such cases do not well represent real cases, but we stick with them for no obvious good reasons beyond the methodological simplicity.

Rudy makes frequent references to optimality. Optimality has of course a strong appeal. Express the desired properties in the form of an optimality criterion together with the related modeling, and we have a mathematical problem that is often very challenging and an obtained solution has all the stature of optimality. However, as Cox (1958) mentioned rather gently, "With (respect) to certain . . . long-run properties, the unconditional (procedure) may be in order, although it may be doubted whether the specification of desired properties is . . . very sensible." Put more bluntly, unconditional analysis allows a trade-off between the known case you have in hand and other cases that might have occurred but did not, thus allowing one to optimize the chosen optimality criterion over a broader context at the expense of the present context. For many, the message seems lost in the medium.

4. PREDICTION

Ron discusses predictive inference and the extent to which the conditional methods can "accommodate the dual problem of prediction." He mentions the use of sufficiency (Butler, 1986, 1989) to develop a pivotal quantity to obtain inference for a future observation and also the use of conditioning (Barnard, 1986) to develop a pivotal quantity for such purposes; some steps for the latter may be found in Fraser and Haq (1969, 1970). These and other methods suggested by Ron seem very promising. An alternative is to treat a probability for a future observation as yet another parameter of the original model and follow the likelihood routes described in the paper.

5. TRUE p VALUES

Ron discusses several kinds of ancillary that have been used for conditional inference, and then examines them for an exponential model example. One ancillary is the first-order ancillary affine ancillary (Efron and Hinkley, 1978; Barndorff-Nielsen, 1986). Another ancillary is the second-order ancillary given by the likelihood ratio statistic for testing the given model in a larger embedding model, often available in simple examples with exponential form that allows embedding in a saturated model; this leads to third-order inference. An extension allows second-order inference and uses a locally defined score variable coupled with directions obtained from the mean value of those score variables (Fraser and Reid, 2001). Another second- and higher-order ancillary is that obtained from the sensitivity directions. It yields third-order inference and for familiar exponential-type examples, as Ron notes, agrees with the preceding ancillary. For many other examples, however, the exponential structure is not available and yet the sensitivity ancillary is easily accessible (see, e.g., Fraser, Reid and Wu, 1999; Fraser, Wong and Wu, 1999).

For a simple exponential model example, Ron considers three methods for calculating an approximate p value for a particular data point: the signed likelihood ratio (SLR) method using approximate normality, the Skovgaard (1996) method using implicit conditioning, and the sensitivity directions method using approximate conditioning. He also obtains an exact p value based on the numerical integration of the distribution for $\hat{\theta}$.

The example Ron uses is a $(2, 1)$ exponential model formed by y_1 with an exponential distribution that has rate parameter θ and by y_2 with an exponential distribution that has rate parameter e^{θ}. The observed data point is taken to be $(1, 2)$. For testing say $\theta = 1/2$, the observed p value calculated using the sensitivity directions aproach is 0.238, or 0.23771 to extra places. For various θ values $(1/2, 3/4, 1, 3/2, 2)$ Ron records p values obtained by the three methods; some of these are reproduced here in Table 1. He finds that the sensitivity direction approach is closer to the exact

TABLE 1
Observed significance probability from the data point $(1, 2)$ *for testing* θ*, where there are two exponential variables with rate parameters* $(\theta, \exp\theta)$

	θ				
Method	**0.5**	**0.75**	**1**	**1.5**	**2**
MLE					
Integration[a]	0.189	0.0689	0.0194	0.0^3489	0.0^5120
True	0.18747	0.09063	0.04336	0.0^2759	0.0^376
(2σ)	(0.00242)	(0.00178)	(0.00126)	(0.0^354)	(0.0^317)
SLR					
Normal	0.325	0.130	0.0392	0.0^2112	0.0^5315
True	0.1610	0.0518	0.0136	0.0^337	NA
(2σ)	(0.0023)	(0.0014)	(0.0^371)	(0.0^312)	NA
Skovgaard					
Second order	0.259	0.0990	0.0289	0.0^3796	0.0^5219
Sensitivity					
Third order	0.23771	0.08641	0.02391	0.0^35829	0.0^51404
True	0.24201	0.08691	0.02461	0.0^382	NA
(2σ)	(0.0027)	(0.0017)	(0.0^396)	(0.0^318)	NA

[a]From Butler.

than the Skovgaard approach or the signed likelihood approach.

To obtain true values here, we resort to simulations (Wong, 2003) and obtain such values to any accuracy by sampling. Accordingly with $N = 100{,}000$ repetitions we find for the special data point $(1, 2)$ that the true p value is 0.24201 with a 2σ simulation limit 0.0027. The simulated p values are obtained by conditionally measuring departure of data from true and then simulating to get the true probability position of the particular data point in question. For the data point $(1, 2)$, Table 1 records the true p value with 2σ limit corresponding to various p values obtained by the various methods mentioned above.

The sensitivity directions approach yields a value very close to the true, although clearly not within the tight 2σ simulation limits. This departure can be attributed to the approximation involved in the conditioning and the very small sample size $n = 1$. The Skovgaard approach also yields a reasonable p value, but substantially farther from the true; it too could be assessed against its own implicit way of measuring departure of data from true, but consistent with the conditional theme of the paper, we use the measure of departure given by the sensitivity directions. It seems clear in the example that the sensitivity directions approach gives remarkable accuracy.

6. CONDITIONING IMPLIES SUFFICIENCY

Ib Skovgaard mentions that "conditioning on ... ancillaries is used frequently in practice, almost unconsciously" and cites a persuasive example. He also mentions the darker side that a "majority of the statistical community (have) given up on the idea (of using ancillaries)."

As part of this discussion he cites the nonuniqueness of maximal ancillaries and the well known Birnbaum result that the principles of sufficiency and conditioning together imply the likelihood principle. What is less widely known is that the conditioning principle alone implies the likelihood principle (Evans, Fraser and Monette, 1985, 1986). The details of the derivation provide key insights to the role of sufficiency in the earlier argument to the likelihood principle: that it lumps together sample space points, ignoring the integrity of the underlying variables, treating the model as just frequencies attached to unassociated points and ignoring structure other than provided by the minimalist statistical model; for some recent views on this issue, see McCullagh (2002), and for an earlier and less structured view, see Fraser (1968). The details also provide yet another strong statement concerning the role and appropriateness of sufficiency. The proof from conditioning to likelihood also indicates how, in the minimalist statistical model, the nonuniqueness of ancillaries follows from the arbitrary lumping together of sample points. Indeed this can be used to create a nominal proof for quite arbitrary results (Evans, Fraser and Monette, 1985, 1986).

All of which points back to the minimalist statistical model as being at the root of most of the apparent difficulties in statistical inference. Of course, in real examples continuity and integrity of variables are implicitly included in much the way, as Ib notes, that a lot of conditioning is done in applications without really noticing it.

Thus ignore sufficiency, add continuity and integrity of variables as explicit parts of the statistical model, and be prepared to condition widely and sensibly. As Ib notes, issues "of conditioning ... should be considered more seriously in practice." Ib also has some concerns about detail which I address below.

7. CONDITIONING OR OPTIMALITY?

This section title is a small variation on Ib's: just replace "optimal inference" with "optimality." Somehow the term optimal inference seems to be two words in contradiction: either you get the total inference concerning the unknown or you do not. Perhaps you should target getting it all, even though subsequently you might package and target it on specific parameter characteristics; recall the various examples mentioned by Rudy that deal with focussed final inference. I feel skeptical generally about seeking a measure and then optimizing with respect to that measure, but here seeking a measure of total inference information without first having some understanding of total information does seem like putting things in the wrong order. Surely you would want the inference material assembled before you try to measure it numerically. How successful has the measuring of information been? Not that some measures of information have not been abundantly useful and, as emphasized in the paper, the related use of optimization allows you to trade higher value in one context against a lower value in other contexts, so you are not presenting things as they are. Can we expect an optimization approach then to tell us what the total inference is in a particular context? It seems unlikely.

Ib then describes an appealing approach that involves squared error of prediction and a notion of

relevance. This is persuasive and its development is promising, the preceding discussion notwithstanding.

8. ASYMPTOTIC SOLUTIONS?

Many inference methods in statistics have evolved in the asymptotic context following patterns found with exponential models, and location and transformation models. Exponential models provide the pattern for obtaining accurate p values from observed likelihood functions (Lugannani and Rice, 1980) using (3.14) and using a Fourier inversion for distribution functions to advance an earlier saddlepoint inversion for density functions (Daniels, 1954). The high third-order accuracy for such results was then extended for scalar parameter and variable models from the exponential case to the general asymptotic case: this used a technical modification of the Wald-type ingredient q, and was implicit in Barndorff-Nielsen (1986), explicit in Fraser (1990) and, in alternate form (3.15), in Barndorff-Nielsen (1991).

With nuisance parameters in addition to a scalar interest parameter, an integration over a nuisance parameter distribution allows the preceding to be applied to a scalar pivot for a scalar interest parameter (Fraser and Reid, 1993; implicit in Barndorff-Nielsen, 1986). Collectively this covers the case of an asymptotic model with variable and parameter of the same dimension. From the details, particularly the construction of the nominal or operational parameter $\varphi(\theta)$, it is seen that the third-order accuracy needs only the likelihood and the gradient of likelihood at the observed data point. This is a remarkable and powerful fact with far-reaching implications for statistical methodology.

Location and transformation models by contrast provide the pattern for extending these p-value methods to cases with the dimension of the variable larger than that of the parameter. The mechanism involves quite generally the use of ancillaries, exact or approximate or implicit. Exact ancillaries are widely and directly available with location and transformation models; indeed most ancillaries have their origins in this context. Approximate ancillaries are then developed by restricting attention to parameter values close to the observed maximum likelihood value; for this there are different approaches. Ib works with the distribution of parameter derivatives of the likelihood function. By contrast the paper examine how local changes in the parameter affect individual coordinates; this in fact reproduces the ancillaries in the location and transformation case. Both approaches generate approximate ancillaries and thus enable the use of the approximation methods described in the preceding paragraph. How can they be compared?

Refering to his approach, Ib notes, “This gives all [the] information needed and provides an explicit solution.” He then adds, “Fraser deliberately discards the fact that local changes are only required for a sufficient statistic and describes local changes [for] the entire set of observation.” However, “deliberately discards ... [a] fact”: What fact? That the weak likelihood or sufficiency principle says one needs only to look at the sufficient statistic? Perhaps “fact” only in the context of total belief in the likelihood-sufficiency principle. A major claim of the paper is that sufficiency is widely an inappropriate principle: in effect it works exclusively with frequencies at data points with the minimalist statistical model, and ignores continuity and coordinate integrity and the direct effect of parameter change on individual measurements or coordinates. Would this make sense for a surveyor or an astronomer? So rather than discarding a “fact,” there is the assertion that there is not such fact and that other substantial facts are being ignored.

Whatever the merits or demerits of the construction procedures, one can of course see how the end results perform. For this, consider an example where multiple ancillaries are present: a covariance matrix in normal sampling. A covariance matrix can have a positive lower triangular square root and this generates a standard ancillary from the obvious transformation model. However, take a rotation of the coordinates and then apply the preceding method; a different ancillary is obtained, and these are different from the ancillary obtained from the likelihod analysis proposed by Ib, which does not favor an order for the coordinates. The context could determine a preference, based on a choice of how you view the coordinates as measuring the parameters, and then the other ancillaries would not be appropriate. Perhaps the seemingly hidden integrity of coordinate variables is more fact than sufficiency.

9. SOME TECHNICAL QUESTIONS

9.1 Constant Pivots?

The sensitivity directions describe what the ancillary looks like at the data point. These directions are obtained by seeing how a change in θ causes a change in y for a fixed pivot, examining this coordinate by coordinate of course. Ib then suggests that this “argues that the pivots $F(y_i, \theta)$ should be kept constant (at least locally) along (a contour) of the conditioning statistic

as a function of the maximum likelihood estimate $\hat{\theta}$ when this is plugged into the pivot." He then questions whether the implied conditioning "statistic agrees with $(F(y_1;\hat{\theta}),\ldots,F(y_n;\hat{\theta}))$" and expresses concern that the conditional distribution from the latter statistic might be degenerate. I share Ib's concern for this statistic, but do note that in general it does not generate the sensitivity directions ancillary.

Consider the simple example of the standard symmetric normal with center on a circle of known radius ρ. A natural pivot is $\{y_1-\rho\cos(\alpha), y_2-\rho\sin(\alpha)\}$, where α is the polar angle. At a data point $(y_1, y_2) = (r\cos a, r\sin a)$ it generates the sensitivity direction $(-\sin a, \cos a)$, which is tangent to the circle of radius r through the data point and is thus tangent to the familiar ancillary r for this problem. On the other hand, the mentioned statistic is equivalent to $(y_1-\rho\cos a, y_2-\rho\sin a)$, which can be rewritten as $(1-\rho/r)(y_1, y_2)$ and is seen to be one–one equivalent to the data point itself. As a conditioning statistic, it is as Ib suspected degenerate.

In general, the estimated residuals or, more generally, the estimated pivots do not generate the ancillary conditioning; the complication for that route lies in the gradient of $\hat{\theta}$ with respect to the data point.

9.2 We Hardly Need Sufficiency

Ib questions the "claim that we hardly need sufficiency..." and raises several related issues. The paper shows that a method for obtaining a p value from a sufficient statistic can be duplicated by a conditional approach, so there is no need to work from a sufficient statistic because the same can be duplicated otherwise. Whereas a conditional p value is also a marginal p value, he then "in the same vein" suggests that conditional p values would be superfluous because they would be available without conditioning. Agreed. However, they would not be based on the conditional structure that makes the departure measure sensible for the particular data point of interest.

Thus, reaffirmation for the initial claim and rejection for the in-the-same-vein claim. Ib rejects both and suggests two views, both of which I agree with:

1. Sufficiency restricts the choice of method (with bad effects).
2. Conditioning restricts the permissible results (for good reasons).

My views are given in parentheses. So the crunch is the adherence to sufficiency: one view against and one view for. Of course sufficiency has heavy traditions and lots of believers, but is there any real basis for the belief? It is not visible.

9.3 Does the Conditional p Value Depend on Just the Sufficient Statistic?

If we do not care about sufficiency, then the question is academic. If the sufficient statistic has the same dimension as the parameter, then it is a nonissue as discussed in the paper. If the dimension of the sufficient statistic is larger, then available procedures in the context of sufficiency include conditioning to bring the dimension down to that of the parameter and then the discussion in the paper is applicable. Thus, without loss of generality, it can depend on the sufficient statistic, but relevant information for making a sensible choice of conditional measure of departure may have been lost.

10. DISCUSSION

The theme in the paper is that ancillarity and conditioning lead to a wealth of highly accurate inference procedures. For the familiar special cases that have available ancillaries, the procedures give accurate approximations to the corresponding p value, and for the wide range of more general cases, the procedures use natural approximate ancillaries and give again highly accurate p values.

It has always been my feeling that there must be logic and structure to natural processes, viewed here as including statistical reasoning. However, much in statistics has worked from the minimalist model, often using optimization to trade off a present instance against other cases that might have arisen but have not. The related recommendation that you stand by and act by rules suggests you have given up on finding substance to statistical thinking and are relying on an external decision or algorithmic approach. So certainly there was persistance in the search for structure in the statistical context, against of course strongly held views opposing such structure. This never particularly bothered me and may even have supported the search. When Ib mentions "stubbornly" it suggests overt forces to be resisted. I have not seen overt forces, so hardly acted stubbornly. But substantial structure? Clearly evident.

ACKNOWLEDGMENTS

Very special thanks go to the Editors, past and present, to many referees, to the discussants and to colleagues, particularly Nancy Reid, Tom DiCiccio and Augustine Wong, for many very helpful discussions, suggestions and comments that led to numerous improved revisions. The support of the Natural Sciences

and Engineering Research Council of Canada is acknowledged.

ADDITIONAL REFERENCES

BARNARD, G. (1986). Discussion of "Predictive likelihood inference with applications," by R. W. Butler. *J. Roy. Statist. Soc. Ser. B* **48** 27–28.

BARNDORFF-NIELSEN, O. (1990). Approximate interval probabilities. *J. Roy. Statist. Soc. Ser. B* **52** 485–496.

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

BARNDORFF-NIELSEN, O. and COX, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2** 319–340.

BARNDORFF-NIELSEN, O. and WOOD, A. T. A. (1996). On large deviations and choice of ancillary for p^* and r^*. *Bernoulli* **4** 35–63.

BASU, D. (1959). The family of ancillary statistics. *Sankhyā Ser. A* **21** 247–256.

BERAN, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83** 679–686.

BERAN, R. (1996). Stein estimation in high dimensions: A retrospective. In *Research Developments in Probability and Statistics* (E. Brunner and M. Denker, eds.) 91–110. VSP, Utrecht.

BERAN, R. (1997). Diagnosing bootstrap success. *Ann. Inst. Statist. Math.* **49** 1–24.

BERAN, R. (2002). Adaptively denoising two-way layouts. Preprint. Available at www.stat.ucdavis.edu/~beran/two.pdf.

BERLINSKI, D. (2000). *The Advent of the Algorithm.* Harcourt, New York.

BIRNBAUM, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326.

BJØRNSTAD, J. F. (1996). On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791–806.

BREIMAN, L. (2001). Statistical modeling: The two cultures (with discussion). *Statist. Sci.* **16** 199–231.

BRILLINGER, D. R. and TUKEY, J. W. (1984). Spectrum analysis in the presence of noise: Some issues and examples. In *The Collected Works of John W. Tukey II. Time Series*: *1965–1984* (D. R. Brillinger, ed.) 1001–1141. Wadsworth, Monterey, CA.

BUTLER, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 1–38.

BUTLER, R. W. (1989). Approximate predictive pivots and densities. *Biometrika* **76** 489–501.

DANIELS, H. E. (1954). Saddlepoint approximation in statistics. *Ann. Math. Statist.* **25** 631–650.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65** 457–482.

EVANS, M., FRASER, D. A. S. and MONETTE, G. (1985). Mixtures, embedding and ancillarity. *Canad. J. Statist.* **13** 1–6.

EVANS, M., FRASER, D. A. S. and MONETTE, G. (1986). On the sufficiency–conditionality to likelihood argument. *Canad. J. Statist.* **14** 1–31.

FRASER, D. A. S. (1968). A black box or a comprehensive model. *Technometrics* **10** 219–229.

FRASER, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77** 65–76.

FRASER, D. A. S. and HAQ, M. S. (1969). Structural probability and prediction for the multivariate model. *J. Roy. Statist. Soc. Ser. B* **31** 317–331.

FRASER, D. A. S. and HAQ, M. S. (1970). Inference and prediction for the multilinear model. *J. Statist. Res.* **4** 93–109.

JENSEN, J. L. (1997). A simple derivation of r^* for curved exponential families. *Scand. J. Statist.* **24** 33–46.

KNUTH, D. E. (1969). *The Art of Computer Programming* **2**. Addison–Wesley, Reading, MA.

MCCULLAGH, P. (2002). What is a statistical model? (with discussion). *Ann. Statist.* **30** 1225–1310.

MCLUHAN, M. (1964). *Understanding Media*: *The Extensions of Man.* McGraw–Hill, New York.

MILLER, R. (1966). *Simultaneous Statistical Inference.* McGraw–Hill, New York.

PEDERSEN, B. V. (1981). A comparison of the Efron–Hinkley ancillary and the likelihood ratio ancillary in a particular example. *Ann. Statist.* **9** 1328–1333.

ROSS, S. L. (1974). *Differential Equations*, 2nd ed. Xerox College Publishing, Lexington, MA.

SAVAGE, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

SKOVGAARD, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2** 145–165.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press, Berkeley.

STEIN, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Research Papers in Statistics* (F. N. David, ed.) 351–364. Wiley, New York.

SUNDBERG, R. (2003). Conditional statistical inference and quantification of relevance. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 299–315.

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison–Wesley, Reading, MA.

VAN ZWET, E. W. and VAN ZWET, W. R. (1999). A remark on consistent estimation. *Math. Methods Statist.* **8** 277–284.

VIDONI, P. (1995). A simple predictive density based on the p^* formula. *Biometrika* **82** 855–863.

WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.

WONG, A. (2003). Calibration of conditional p values. Technical report, York Univ.