ANCILLARIES AND CONDITIONAL INFERENCE

D.A.S. Fraser
Department of Statistics, University of Toronto,
Toronto, Canada M5S 3G3

ABSTRACT

Sufficiency has long been regarded as the primary reduction procedure to simplify a statistical model. And the assessment of this procedure typically 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 asymptotics shows however that special conditional procedures provide in wide generality the definitive reduction to a variable of the same dimension as the parameter. We begin with a discussion of two examples from the literature (Welch, 1939; Cox, 1958) which compare conditional and global inference methods and yet come to opposite assessments concerning the 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 techniques from likelihood asymptotics produce the definitive reduction for the initial stage of the analysis. We then examine the vector parameter case and find that for the elimination of nuisance parameters a marginalization step is needed, not the commonly recommended conditional calculation 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. 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 adherence to sufficiency and the inappropriate neglect of the conditioning procedures based directly
on available evidence.

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 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 for conditioning. It can be noted however that the two papers differ in their orientation towards statistics, the first being decision theoretic and the second being inferential. The conditional approach as discussed in the second paper does however violate 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 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 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. The implication is that sufficiency and the global repeated sampling principle have together been a major delaying factor to the recognition of the conditional approach. These two sections also include an overview of the methods provided by recent likelihood asymptotics;
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) considers two measurement instruments, both unbiased and normal but with different variances. The context also includes a random choice of which instrument to use to make a single measurement on a parameter $\theta$. The example is also discussed in Cox & Hinkley (1974, p.96), but despite its importance seems not to appear in most texts on statistics. A somewhat related example had been considered earlier by Welch (1939).

Cox initially considers the appropriate sample space for statistical inference, but he then develops this in terms of conditioning on an ancillary statistic (Fisher, 1925, 1934, 1935). A statistic is ancillary if its distribution is free of the full parameter in the problem. A related notion of reference set is introduced in Fisher (1961).

Cox notes 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; thus, $a$ is ancillary. The Fisher conditionality approach is to condition on the observed value of the ancillary $a$ and thus to use the normal model corresponding 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 & Hinkley (1974) consider the two measuring 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 $\sigma_0$, respectively. A 95% confidence interval based on the measurement instrument actually used has the form

$$
(y \pm 196\sigma_0) \quad \text{if} \quad a = 1,
$$

$$
(y \pm 1.96\sigma_0) \quad \text{if} \quad a = 2.
$$

(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 following 95% confidence interval:

$$
(y \pm 164\sigma_0) \quad \text{if} \quad a = 1
$$

$$
(y \pm 5\sigma_0) \quad \text{if} \quad a = 2.
$$

(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 does average to the desired overall 95% 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 $\alpha$ 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 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 that advertised that his 95% intervals are shorter on average.
His policy might be: give the clients with more accurate measuring instruments longer intervals and give the clients with less precise instruments shorter intervals. He 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 chooses to control the coverage probability associated with each instrument. This would presumably 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 using an optimality approach. The next example will display this strange trade off.

Let us consider the two measurement instruments example in Welch (1939). For this we have two measurements $y_1, y_2$ of $\theta$ 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 by 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 $\beta$ level confidence interval conditional on the ancillary $R$ is

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

the $\beta = 75\%$ acceptance region for testing a value $\theta$ corresponding to (2.3) is recorded in Figure 1a.
A likelihood ratio argument can be used to obtain the most powerful (often called, rather inappropriately, most accurate) unbiased or symmetric $\beta$-level interval:

$$\{\mathcal{I} \pm \frac{1 - R}{2}\} \quad \text{if} \quad R > \left(\frac{1 - \beta}{2}\right)^{1/2},$$

$$\left[\mathcal{I} \pm \left\{\frac{1 + R}{2} - \left(\frac{1 - \beta}{2}\right)^{1/2}\right\}\right] \quad \text{if} \quad R \leq \left(\frac{1 - \beta}{2}\right)^{1/2}. \tag{2.4}$$

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

$$\left(\mathcal{I} \pm \frac{1 - R}{2}\right) \quad \text{if} \quad R > (1 - \sqrt{\beta}),$$

$$\emptyset \quad \text{if} \quad R \leq (1 - \sqrt{\beta}). \tag{2.5}$$

This confidence interval is either the full range of possible $\theta$ values or the empty set; the acceptance region corresponding to this confidence interval (2.5) is recorded in Figure 1c. 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 $\theta$), the minimum average length 75% interval is the empty set 25% of the time (and then never covers $\theta$). These are certainly extraordinary and unpleasant properties that hopefully would not easily be explained away to a client.

Cox (1958) offers general support for the conditionality approach from Fisher (1961). Welch (1959) invokes optimality conditions and argues against conditionality using a similar example. Similar opposite viewpoints may be found in Fraser & 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 prob-
ability assessments lead generally to analyses that do not acknowledge clear and evident characteristics of particular investigations.

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 $(\theta, \sigma_0^2)$ with observed data $y^0$. The observed likelihood function is available immediately,

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

where $\phi$ is the standard normal density; it has maximum value at $y^0$, has normal shape, and is scaled by $\sigma_0$. The observed $p$-value function is

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

where $\Phi$ 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 $\theta$; it can be viewed as presenting the percentile position of the data $y^0$ relative to the distribution for $y$ that is indexed by $\theta$. In more general contexts we will 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.

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

7
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 $t$-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 have the measurement $y$ with model $f(y - \theta)$ where $f$ is specified, and the observed data value $y^0$.

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

$$L^0(\theta) = cf(y^0, \theta)$$  \hspace{1cm} (3.3)

and the observed $p$-value function is

$$p^0(\theta) = F(y^0; \theta)$$  \hspace{1cm} (3.4)

where $F$ is the cumulative distribution function corresponding 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

$$p^0(\hat{\theta}_L) = .975, \quad p^0(\hat{\theta}_U) = .025,$$

where we are assuming for convenience that the distribution shifts to the right with increasing $\theta$. Here 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 are available in wide generality and with little computational difficulty. A natural next step to that above 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 modelling.

3.3. Multiple measurements with location parameter. Consider a sample \((y_1, \ldots, y_n)\) from a distribution \(f(y - \theta)\). The residual vector \(a(y) = (y_1 - \overline{y}, \ldots, y_n - \overline{y})'\) is describing the distribution pattern within the sample and is easily seen to have a known distribution. To show this write \(y_i = \theta + e_i\) where \((e_1, \ldots, e_n)\) is a sample from the distribution \(f(e)\). Then \(a(y) = (y_1 - \overline{y}, \ldots, y_n - \overline{y})' = (e_1 - \overline{e}, \ldots, e_n - \overline{e})' = a(e)\), which explicitly shows that the distribution for \(a(y)\) is free of \(\theta\). The residual vector is sometimes called a configuration statistic; it is ancillary and in addition is directly presenting key observable characteristics of the underlying or latent errors. Also recall the discussion in Example 2.2.

Now consider observed data \((y_1^0, \ldots, y_n^0)\). The form of the model conditional on the observed configuration \(a(y^0) = a^0\) is easily obtained by change of variable,

\[ g(\overline{y} \mid a^0; \theta) = k f(\overline{y} + a^0_1 - \theta) \cdots f(\overline{y} + a^0_n - \theta), \]

where \(k\) is the norming constant. Then noting that the observed likelihood is

\[ L^0(\theta) = c f(y_1^0 - \theta) \cdots f(y_n^0 - \theta) \]

we see that

\[ g(\overline{y} \mid a^0; \theta) = L^0(\theta - \overline{y} + \overline{y}^0) \]

where the proportionality constant \(c\) in (3.5) is assumed to have been appropriately chosen. The observed \(p\)-value can then be written as an appropriate integral of observed likelihood:

\[ p^0(\theta) = \int_{-\infty}^{\overline{y}} g(\overline{y} \mid a^0; \theta)d\overline{y} = \int_{-\infty}^{\overline{y}} L^0(\theta - \overline{y} + \overline{y}^0)d\overline{y} = \int_{\theta}^{\infty} L^0(\theta)d\theta, \]

which happens also to be the Bayesian survival probability derived from a flat or uniform prior \(\pi(\theta) = k\). In the special case that the error density is normal, we have that (3.5) and (3.7) duplicate the results (3.1) and (3.2) for normal sampling. Thus we see that
sufficiency works essentially just for the simple normal, but that conditioning works also for the general error case and in doing so reproduces the special earlier result for the case of normal error.

When we examine a still more general case in the next subsection 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 quite generally available using just the observed likelihood $L_0(\theta)$ and the gradient of the log-likelihood $l(\theta; y)$ taken at the data point in a sensitivity direction $v$ for which the ancillary is constant in value. At this stage, it is of interest to see just what the vector $v$ is. If $a(y) = a^0$ then a point $y$ has projection $(y_1 - \bar{y}, \ldots, y_n - \bar{y}) = (a_1^0, \ldots, a_n^0)$ on the orthogonal complement $L^\perp(1)$ of the one-vector; the points with such fixed projection lie on the line $a^0 + L(1)$ 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.

### 3.4. Multiple measurements with scalar parameter.

With a location model $f(y - \theta)$ we see that a change in the parameter $\theta$ causes a shift of the distribution by a corresponding amount. We could even write that the corresponding sensitivity velocity $v$ is $dy/d\theta = 1$ where we understand clearly that the derivative is taken for fixed value of the error $e = y - \theta$. For a more general context with distribution function $F(y; \theta)$ we can note that a small increment $\delta$ from a value $\theta$ causes a shift of the distribution by an amount $v\delta$ at a point $y$ where

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

the probability position of the point $y$ is given by its p-value $F(y; \theta)$; we thus call $v = v(\theta)$ the sensitivity of $y$ relative to change at $\theta$; indeed this gives the sensitivity as described
for the location case in the previous subsection.

When we speak of the probability position or the $p$–value of a point $y$ we are presenting the same type of information as the police officer when he says you are driving at the 99.5 percentile and then gives you a major ticket; the position relative to other cars is 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 $\delta$ in $\theta$ causes in the manner just described a change $v_i \delta$ in the coordinate $y_i$ giving the sensitivity

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

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

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

corresponding to change in $\theta$ at the maximum likelihood value $\theta = \hat{\theta}_0$. In any case, asymptotic theory establishes $v$ as the tangent vector to an approximate ancillary suitable for recent likelihood inference. Whether the physical suggestion of sensitivity under parameter change has any persuasive value, it does provide the basis for the arguments that derive the ancillary property (Fraser & Reid, 2001).

Recent likelihood inference theory focuses on the likelihood function itself and in wide generality produces results that have third-order accuracy as opposed to the first-order accuracy when normality is assumed for the score or maximum likelihood departure measures. By third-order we mean that the approximation errors are of order $O(n^{-3/2})$ where $n$ is the sample size or an equivalent indicator of data dimension.

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

$$\frac{L(\theta)}{L(\hat{\theta})} = \exp \{\ell(\theta) - \ell(\hat{\theta})\} = e^{-r^2/2}$$

(3.10)
and solve for \( r \) with an appropriate sign we obtain

\[
    r = \text{sgn}(\hat{\theta} - \theta) \left[ 2 \{ \ell(\hat{\theta}) - \ell(\theta) \} \right]^{1/2}
\]

(3.11)
called the signed likelihood root. The second departure measure is a standardized maximum likelihood departure based on a special exponential-type reparameterization \( \varphi(\theta) \),

\[
    q = \text{sgn}(\hat{\theta} - \theta) \left\| \varphi(\hat{\theta}) - \varphi(\theta) \right\|^{1/2} \hat{j}_{\varphi \varphi}
\]

(3.12)
where \( \hat{j}_{\varphi \varphi} = -\left( \partial^2 / \partial \varphi^2 \right) \ell(\theta; y^0) \mid \theta = \hat{\theta} \) is the corresponding observed information calculated in the \( \varphi \) parameterization. The special choice of parameterization \( \varphi(\theta) \) is essential: it needs to be an exponential reparameterization as obtained from the gradient

\[
    \varphi(\theta) = \frac{d}{dv} \ell(\theta; y) \mid_{y = y^0}
\]

(3.13)
of likelihood at the data point and calculated in an appropriate direction \( v \); the directional derivative \( d/dv \) is defined by

\[
    (d/dv)h(y) = (d/dx)h(y + xv) \mid_{x=0}.
\]

The reparameterization does depend in minor ways on the data point \( y^0 \) but as part of this does provide the input needed to calculate the highly accurate distribution function values at that data point; for background and details see Fraser & Reid (1993, 1995, 2001).

We do note that \( \varphi(\theta) \) can be replaced by an affine equivalent \( a\varphi(\theta) + b \) without altering \( q \). The observed \( p \)-value for testing \( \theta \) is then given as

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

(3.14)
or

\[
    p^0(\theta) = \Phi\left\{ r^0 - (1/r^0) \log(r^0/q^0) \right\}
\]

(3.15)
using saddlepoint type combining formulas due to Lugannani & Rice (1980) and Barndorff-Nielsen (1986); the \( p \)-value has third order accuracy and conforms to appropriate ancillary conditioning (Fraser & Reid, 2001).
In the special normal case described in Subsection 3.1, the quantities \( r \) and \( q \) are both equal to \((\bar{y} - \theta)/(\sigma_0/\sqrt{n})\). 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, Wong, 2002).

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

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

and (3.14) and (3.15) 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 reparameterization \( \varphi_i(\theta) = -\theta \). From this we obtain the overall log likelihood

\[
\ell(\theta) = n \log \theta - \theta \sum_i y_i.
\]

A natural pivotal for the \( i \)th coordinate is \( \theta y_i \); this has a fixed distribution of course and the distribution function \( F_i(y_i; \theta) = 1 - \exp(-\theta y_i) \). For the vector case this gives the \( n \) dimensional pivotal \((y_1 \theta, \ldots, y_n \theta)\). From this we obtain the velocity 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 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 reparameterization

\[
\varphi(\theta) = \sum_{i=1}^n \left( -\frac{y_i^0}{\hat{\theta}^0} \right)(-\theta) = c\theta.
\]
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 \(-3.6599\), which with the standard normal distribution function gives the approximate \( p \)-value \( 0.00126 \). Alternatively, we calculate \( q = -9 \) and use it with \( r \) in (3.14) to obtain the \( p \)-value \( 0.00046 \). These approximate \( p \)-values can be compared to the exact \( p \)-value \( 0.00045 \). 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 \& Wu (1999).

3.6. **Condition to separate main effects.** Our examples in this section were concerned with a scalar parameter \( \theta \), 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 describing independent coordinates of a vector response. We found that conditional methods produced the accurate \( p \)-values while sufficiency methods were typically 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 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 attention from the serious consideration of conditional methods.

4. **VECTOR PARAMETER MEASUREMENT EXAMPLES**

14
4.1. Measurements with normal error. Consider the case of a sample \((y_1, \ldots, y_n)\) from the normal \((\mu, \sigma^2)\) distribution and let \((y_1^0, \ldots, y_n^0)\) be the observed data. The observed likelihood function is

\[
L^0(\mu, \sigma) = c\sigma^{-n} \exp \left\{ -\frac{(s^0)^2}{2\sigma^2} - \frac{n(\bar{y}^0 - \mu)^2}{2\sigma^2} \right\}
\]

(4.1)

where \(s^2 = \Sigma(y_i - \bar{y})^2\). We could be interested in various parameter components, but we choose just the simple location parameter \(\mu\). From a general viewpoint we might want a likelihood for \(\mu\); there are recent developments for this (for example, Fraser 2002), 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:

\[
p^0(\mu) = H \left( \frac{\bar{y}^0 - \mu}{s^0/\sqrt{n^2 - n}} \right),
\]

(4.2)

where \(H\) is the Student \((n - 1)\) distribution function. This can be argued in various ways. The statistic \((\bar{y}, s)\) is minimal sufficient and is the sole data ingredient needed for the likelihood \(L(\mu, \sigma; y_1, \ldots, y_n)\); for fixed \(\mu\), \(t = n^{1/2}(\bar{y} - \mu)/s_y\) has uniqueness properties as a continuous function of \((\bar{y}, s)\) with distribution free of the nuisance parameter \(\sigma\). Whatever the basis, we take the \(t\) quantity as the appropriate quantity for providing 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 \(f(e)\). In order to have a sensible definition of \(\mu\) and \(\sigma\) we require that \(f(e)\) be appropriately centered and scaled.

The standardized residuals \(d_i = (y_i - \bar{y})/s\) 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 \(\mu\) and \(\sigma\). Accordingly it is ancillary in the conventional sense. But 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 \(d(y)\) is the appropriate configuration statistic.
The observed likelihood function is

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

(4.3)

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

\[ c\sigma^{-n} \prod_{i=1}^{n} f\{\sigma^{-1}(\bar{y} + s\delta_i^0 - \mu)\} s^n \cdot \frac{d\bar{y}ds}{s^2} \]

which can be rewritten as

\[ L^0(\bar{y}^0 + s^0(\mu - \bar{y})/s, s^0\sigma/s) \cdot \frac{d\bar{y}ds}{s^2}. \]  

(4.4)

We thus see that any probability for \((\bar{y}, s)\) 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 Subsection 4.1, we have that (4.4) reproduces the normal distribution for \(\bar{y}\) and the scaled chi square distribution for \(s^2\).

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

\[ p^0(\mu) = \int_{t \leq t_0} L^0\{\bar{y}^0 + s^0(\mu - \bar{y})/s, s^0\sigma/s\} \frac{d\bar{y}ds}{s^2}, \]  

(4.5)

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

\[ p^0(\mu) = \int_{\mu=\mu}^{\infty} \int_{\sigma=0}^{\infty} L^0(\bar{\mu}, \sigma) \frac{d\bar{\mu}d\sigma}{\sigma}, \]  

(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 \(\sigma^{-1}\). Highly accurate approximations for (4.5) or (4.6) are also easily available; see Subsection 4.4.
4.3. **Exponential model and canonical parameters.** Consider an exponential model with natural or canonical parameters $(\psi, \lambda)$:

$$f(s_1, s_2; \psi, \lambda) = \exp \{\psi s_1 + \lambda s_2 - \kappa(\psi, \lambda)\} h(s_1, s_2).$$

(4.7)

This type of model is frequently mentioned when inference for a parameter $\psi$ in the presence of a nuisance parameter $\lambda$ is under discussion. If sampling is part of the background, then the coefficients of $\psi$ and $\lambda$ in the exponent of (4.7) form the minimal sufficient statistic or likelihood statistic. We have 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 $h(s_1, s_2)$ would typically be available only by integration from some original composite density for the sample; by contrast, the likelihood ingredient $\kappa(\psi, \lambda)$ is generally available explicitly.

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

$$p^0(\psi) = \int_{s_1^0} f(s_1|s_2^0; \psi)ds_1.$$  

(4.8)

where the lower limit is the lower end of the appropriate range. Some details for such calculations for the gamma mean problem may be found in Fraser, Reid & Wong (1997). The $p$-value in (4.8) is presented as a conditional $p$-value, conditional on the appropriate ancillary. 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 of the ancillary then it has that same distribution regardless of the ancillary, that is, marginally.

In wide generality, as will be seen in the next section, $p$-values free of nuisance parameters are not available by a full conditional calculation but are obtained free of the
nuisance parameter when calculated by a marginalization that eliminates the effect of the
nuisance parameter. They are available by a conditional argument as just described only
for very special model types such as the exponential; in such cases, the conditional p-value
is also a marginal p-value, so there is no conflict with developments in the next Subsec-
tion. Conditioning here is then an alternate route to the same end by a different argument
suitable 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 \((\psi, \lambda)\) 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 symmctric as for example
the Student density \(\pi^{-1}(1 + e_1^2 + e_2^2)^{-2}\); a still more general case still would proceed in the
same manner. Also suppose that we are interested in the component parameter \(\psi\). For a
general context see Fraser (2002).

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 \(d_1\) and \(d_2\) is free of \((\psi, \lambda)\), and also
showing that the residual characteristics of the underlying errors are directly available
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:

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

The observed likelihood is

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

and we can then write

\[
f(\overline{y}_1, \overline{y}_2 \mid d_1^0, d_2^0, \psi, \lambda) = L^0(\psi - \overline{y}_1 + \overline{y}_1^0, \lambda - \overline{y}_2 + \overline{y}_2^0).
\]
This reduced model is a two-dimensional location model with parameter \((\psi, \lambda)\).

Under a requirement of moderate continuity for the variables under study it is straightforward to see that \(\bar{y}_1\) is the essentially unique variable free of \(\lambda\); the corresponding marginal distribution is

\[
f(\bar{y}_1 - \psi \mid d_1^0, d_2^0) = \int_{-\infty}^{\infty} L^0(\psi - \bar{y}_1 + \bar{y}_1^0, t)dt,
\]

and the essentially unique \(p\)-value for assessing \(\psi\) is

\[
p^0(\psi) = \int_{\psi}^{\infty} \int_{-\infty}^{\infty} L^0(\tilde{\psi}, \lambda)d\tilde{\psi}d\lambda,
\]

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 \(\psi\) values can then be obtained by numerical integration of likelihood. Highly accurate approximations to (4.11) are available and discussed in the next subsection.

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

4.5. Multiple measurements: interest and nuisance parameters. With the location model in the preceding section we see that a change in the parameter \((\psi, \lambda)\) 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 \(\psi\) causes a shift in the first coordinate \(n\)-vector by the corresponding multiple of the one-vector for that coordinate. A change in \(\lambda\) 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 \(\psi\) and \(\lambda\) 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 Subsection 3.4, we examine how a change in \((\psi, \lambda)\) shifts the distribution. We do this by
examining the p-value $F(y_i; \psi, \lambda)$ for the $i$-th 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

\[
(v_1, v_2) = \frac{\partial y_i}{\partial (\psi, \lambda)} = \left( - \frac{\partial F(y_i; \theta)/\partial \psi}{\partial F(y_i; \theta)/\partial y_i}, - \frac{\partial F(y_i; \theta)/\partial \lambda}{\partial F(y_i; \theta)/\partial y_i} \right).
\]

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

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

which describe how $(\psi, \lambda)$ affects the distribution. We are certainly concerned with this effect at an observed data point $y^0$ and corresponding maximum likelihood parameter value $\hat{\theta}^0$. Let $V$ in (4.12) be evaluated for $(y, \theta) = (y^0, \hat{\theta}^0)$. General theory (Fraser & Reid, 2001) then shows that there is an approximate ancillary $a(y)$ of dimension $n - 2$ for which the tangent vectors at the data point $y^0$ are given by $V$. This suffices to give third order $p$-values for the components of $\theta = (\psi, \lambda)$. The calculations for the $p$-values need just the observed log likelihood $\ell^0(\psi, \lambda)$, and the observed log likelihood gradient

\[
\varphi'(\theta) = \{\varphi_1(\theta), \varphi_2(\theta)\} = \frac{d}{dV} \ell(\theta; y) \big|_{y=y^0}
\]

\[
= \left\{ \frac{d}{dv_1} \ell(\theta; y) \big|_{y^0}, \frac{d}{dv_2} \ell(\theta; y) \big|_{y^0} \right\} \bigg|_{y=y^0}
\]

using directional derivatives as defined after (3.13). For inference concerning $\psi$ we then calculate a first departure measure given by the signed likelihood ratio

\[
r^0(\psi) = \text{sgn}(\hat{\psi}^0 - \psi) \cdot \left\{ 2 \left[ \ell^0(\hat{\theta}) - \ell^0(\hat{\theta}_\psi) \right] \right\}^{1/2},
\]

and a second departure measure given by a special standardized maximum likelihood departure

\[
q^0(\psi) = \text{sgn}(\hat{\psi}^0 - \psi) \left[ \chi(\hat{\theta}) - \chi(\hat{\theta}_\psi) \right] \left\{ \frac{|\hat{J}_{\varphi \psi}|}{\hat{J}(\varphi \psi)(\hat{\theta}_\psi)} \right\}^{1/2};
\]

in this $\chi(\theta)$ is a rotated coordinate of $\varphi(\theta)$ that agrees with $\psi(\theta)$ to first derivative 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 $\varphi$ 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 & Wu (1999) and Fraser, Monette, Ng & 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 Subsection 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 more general cases as developed in Subsections 4.3 and 4.4, where sufficiency is not available.

5. **SOME CONDITIONING AND MODELLING CRITERIA**

5.1. **The two measurement instruments example.** In Section 2 we discussed two examples involving measurement instruments, as presented in Cox (1958) and in the earlier 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 provides the background for the succession of examples in Sections 3 and 4. We develop further aspects of conditioning on configuration statistics in the next subsection. For the other example (Cox, 1958), only one of the instruments was actually used. This raised a serious issue. Should the modelling 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 modelling in Subsection 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 neighbourhood of the observed data. At the present time this theory is not available for the case of discrete distributions. The concern with the model in the neighbourhood of the data does seem data dependent. But at the observe data is where the model form is of particular importance, and in substance is not dissimilar to the 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 \( q = (\hat{\theta} - \theta)^{1/2} \). 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 dsrtibution. In the context of independent scalar coordinates, the coordinate \( p \)-values \( F_i(y_i; \theta) \) provide the direct continuity link showing 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 \( i \)th 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 \( i \)th variable measures or relates to the parameter being measured. A basis for this choice will be 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 \( \theta \) 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

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

(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 respect 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 & Reid, 2001).
General theory (Fraser & 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
for the $i$th coordinate model. The exponential type parameterization is available as the
gradient of log likelihood

$$\varphi_i'(\theta) = \frac{\partial}{\partial y_i'} \ell(\theta; y^0_i)$$

(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):

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

(5.3)

where the $V_i$ is the $p \times p$ block of the matrix $V$ that corresponds to the $i$th 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 infer-
ence 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 the $r(\psi)$ and $q(\psi)$ given by (4.14) and (4.15). For a variety
of examples in a regression context see Fraser, Wong & Wu (1999) and Fraser, Monette, Ng & Wong (1994).

5.3. Modelling the actual data production. As mentioned in Subsection 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 $\theta$ and suppose that for each there is a direct measurement relationship to the parameter, as discussed in Sections 3, 4 and Subsection 5.2. For illustrative purposes a succession of three models, say $M_1$, $M_2$, $M_3$, will suffice; let $y_1$, $y_2$, $y_3$ be the corresponding data. Many issues can be involved in the modelling of such a context. Here we focus on the goal of statistical inference for the parameter in question, and propose three modelling 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 $\theta$ and standard deviation $\sigma_0$. 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 $\theta$. In practice the precise definition of $\theta$ 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 \( \theta \), but was 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 have corresponding data values.

In conclusion, we note that the use of the product model for the analysis of \( M_1 \), \( M_2 \), \( M_3 \) 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, the components that corresponded to the 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 may be found in Fraser (1979; pp. 54-68, pp. 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 as to what the proper model for inference should be in the presence of data and for this 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 are violating criteria II and III.

**Example 6.1.** Random choice of sample size. Consider the repeated measurement unit assessment of a parameter \( \theta \), and suppose that the number of repetitions \( n \) is random with known density \( p(n) \). In accord with criteria I, II we would model the specific measurement units that were performed, and in accord with criterion III we would not model the process leading to the sample size \( n \). This gives the inference model \( \prod_{i=1}^{n} f(y_i; \theta) \) plus the corresponding data. From the global repeated sampling viewpoint, however, we would examine the composite model \( p(n) \prod_{i=1}^{n} f(y_i; \theta) \) 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 \( \lambda \).
free of $\theta$. The criteria again give the model $\prod_{i=1}^{n} f(y_i; \theta)$ 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 are not easily distinguishable. A parameter $\theta$ 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 yielding $(n_1, n_2)$. 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 describing 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 has been considered 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 $\lambda$ with range free of $\theta$. More specifically, the inference model concerning $\theta$ 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 $2 \times 2$ table (Fisher, 1957, p.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 $AB, Ab, aB, ab$ are 9, 3, 3, 1 if there is no linkage and are $2 + \theta, 1 - \theta, 1 - \theta, \theta$ in the presence of a linkage parameter $\theta$, 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 $n_{11}, n_{12}, n_{21}, n_{22}$ be the data for $n$ offspring in a particular mating with say $(n_1, n_2) = (n_{11} + n_{12}, n_{21} + n_{22})$ designating row totals and $(n_1, n_2)$ designating column totals.

If the data are assembled in terms of the $A$ phenotype we then have that $n_{11}$ is binomial \{n_{11}, (2 + \theta)/3\} and $n_{21}$ is binomial \{n_{21}, (1 - \theta)\}. Alternatively, if the data are assembled in terms of the $B$ phenotype we then have that $n_{11}$ is binomial \{n_{11}, (2 + \theta)/3\} and $n_{12}$ is binomial \{n_{21}, (1 - \theta)\}. We thus obtain two different inference modellings 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 $\theta$ is being investigated.

From the ancillary viewpoint we can note that the row totals $n_1, n_2$ have a binomial allocation with probabilities in the ratio 3 to 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 $n_1, n_2$ 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 modellings and provides no preference of 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 \(\rho\). If we examine the data labelled by the \(x\) values, we have that the \(y\) values are normal with mean \(\rho x\) 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 \(\rho y\) and variance \(1 - \rho^2\). Accordingly, we obtain two different inference modelings corresponding to two different assemblies or classifications of the data, by \(x\) or by \(y\).

By contrast we can note that with 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 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 Subsection 3.4.

For the first three examples, our model for inference approach and the ancillarity approach are in agreement. For the final two examples, the model for 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 modelling stage. This leads to the use of measurement
directions as introduced in Subsections 3.4 and 4.5 which use continuity and express how
a parameter change has 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 condi-
tional inference given an observed value of the ancillary. Our initial examples however,
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 modelling criteria in Subsection 5.3, we viewed the individual measurement
probabilities as the primary ingredients, with frequency interpretations based on repeti-
tions 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 pretence, by weighted likelihood and integration; for some recent discussion see Fraser (1972), Fraser & Reid (2002), and Fraser & 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 modelling in Subsection 5.3 we considered conditional inference and meta analysis for three investigations of a scalar parameter \( \theta \). For some comparisons with global probabilities we now examine an even simpler case involving two measurements of the parameter \( \theta \): a first measurement \( y_1 \) is unbiased and normal with standard deviation \( \sigma_0 \) say equal to 1; a second measurement is unbiased and normal with standard deviation \( \sigma_0/100 = .01 \). 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 \( \theta \) would be

\[
p_1(\theta) = \Phi(y_1^0 - \theta)
\]

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 \( \theta \) would be

\[
p_2(\theta) = \Phi\{100(y_2^0 - \theta)\},
\]

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 makes a first measurement of \( \theta \) and obtains a value \( y_1^0 = 1.1 \) suggesting to him in a very informal way that perhaps the true value for \( \theta \) is above the threshold. As a result he decides to take a second high precision measurement and obtains \( y_2^0 = -.1 \); his 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 one half, 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 one half the pivotal \( p \)-value is greater than one half 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 non-uniform 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 \(\theta\) and two observations; thus \(y_1 = e_1, y_2 = \theta y_1 + e_2\) 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 \(\theta\).

Now consider the inference modelling viewpoint from Section 5. The first \(y_1\) does not measure \(\theta\), 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. And the persuasive global approach would want to include modellings 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 neighbourhood 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 proposed on several occasion by Professor Jim Durbin. Perhaps the only way to argue against it is to preface the argument with an assumption of some type of global repeated sampling principle.

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. Or by overstating and by understating in particular instances it is possible to increment towards 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 clearly in Cox (1958).

8. ACKNOWLEDGEMENTS

The author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada. Very special thanks go to the Editor and to referees for many helpful suggestions and comments.

APPENDIX A. Conditioning replaces sufficiency to separate main effects.

Consider the case of continuous variables and suppose there is a sufficient statistic $s(y)$ having 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 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 \( f(s; \theta) = f(s|u) = f(s|t; \theta) \); that is, the marginal and conditional models are equal. 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 \((\theta, \sigma_\theta^2)\). The model for \( \overline{y} \) is normal\((\theta, \sigma_\theta^2/2)\); the conditional model for \( \overline{y} \) given the configuration \( y_2 - y_1 \) is also normal\((\theta, \sigma_\theta^2/2)\). If, however, we depart from normality then sufficiency is generally 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

\[
\exp\{y_1\psi + y_2\lambda - c(\psi, \lambda)\} h(y_1, y_2),
\]  

\[ (B.1) \]
the conditional distribution of \( y_1|y_2 \) depends on \( \psi \) only and is thus free of \( \lambda \). For a continuous location model

\[
f(y_1 - \psi, y_2 - \lambda)
\]

(B.2)

the marginal distribution of \( y_1 \) depends on \( \psi \) only and is thus free of \( \lambda \). In each case we have a special model type with specialized variables and parameters; these are often referred to as canonical variables and parameters.

Now consider the first example where conditioning provides the freedom from the nuisance parameter, and suppose we are testing \( \psi \). Let \( u(y_1, y_2) \) be a probability integral transformation of \( y_1|y_2 \) obtained from the \( \lambda \)-free conditional distribution for testing \( \psi \). Then for the tested \( \psi \), 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 \( \lambda \) free and gives \( p \)-values that agree with those from the initial conditional variable.

Recent likelihood asymptotics (for example, Fraser & Reid 1993, Fraser, Reid & 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 Subsection 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 in the new parameterization is available immediately

\[
|J(\lambda\lambda)| = |J_{\theta\theta}(\hat{\theta})|\varphi(\hat{\theta})|^{-2}
\]

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

\[
|J(\lambda\lambda)(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)| \cdot |\varphi(\hat{\theta}_\psi)|^{-2} = |j_{\lambda\lambda}(\hat{\theta}_\psi)| \cdot |X'X|
\]

36
where the right hand determinant uses $X = \phi_X(\hat{\theta}_\psi)$ which in the regression context records the volume on the regression surface as a proportion of the volume for the regression coefficients; in the preceding formula this changes the scaling for the nuisance parameter to that derived from the $\varphi$ parameterization. The expressions above are for the case where $\theta'$ is given as $(\psi, \lambda')$ with an explicit nuisance parameterization; the more general version is recorded in Fraser, Reid & Wu (1999). The rotated coordinate $\chi(\theta)$ is obtained from the gradient vector of $\psi(\theta)$ at $\hat{\theta}_\psi$ and in the $\varphi$ parameterization becomes

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

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

$$\psi_{\varphi'}(\theta) = \partial\psi(\theta)/\partial\varphi' = (\partial\psi(\theta)/\partial\theta') \cdot (\partial\varphi(\theta)/\partial\theta')^{-1} = \psi_{\varphi'}(\theta)\varphi_{\varphi'}^{-1}(\theta);$$

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

REFERENCES


