ANCILLARY STATISTICS: A REVIEW

M. Ghosh, N. Reid and D.A.S. Fraser
University of Florida and University of Toronto

ABSTRACT

Ancillary statistics, one of R. A. Fisher's most fundamental contributions to statistical inference, are statistics whose distributions do not depend on the model parameters. However, in conjunction with some other statistics, typically the maximum likelihood estimate, they provide valuable information about the parameters of interest.

The present article is a review of some of the uses and limitations of ancillary statistics. Due to the vastness of the subject, the present account is, by no means, comprehensive. The topics selected reflect our interest, and clearly many important contributions to the subject are left out.

We touch upon both exact and asymptotic inference based on ancillary statistics. The discussion includes Barndorff-Nielsen's p^* formula, the role of ancillary statistics in the elimination of nuisance parameters, and in finding optimal estimating functions. We also discuss some approximate ancillary statistics, Bayesian ancillarity and the ancillarity paradox.

Keywords: ancillarity paradox, approximate ancillary, asymptotic ancillarity, Bayesian ancillarity, estimating functions, hierarchical Bayes, location, location-scale, multiple ancillaries, nuisance parameters, p-values, P-ancillarity, S-ancillarity, saddlepoint approximation.

1 INTRODUCTION

Ancillary statistics are one of R.A. Fisher's many pioneering contributions to statistical inference. Over the decades, there has been a steady stream of research in this general area, albeit the topic does not enjoy as much popularity as some of Sir Ronald's other contributions such as p-values, randomization, sufficiency, maximum likelihood, just to name a few.

What are ancillary statistics? These are the statistics with distributions not depending on the model parameters. So, why are they useful? It was pointed out by Fisher (1925, 1934, 1935) that though an ancillary statistic by itself failed to provide any information about the parameter, yet in conjunction with another statistic, typically the maximum likelihood estimator (MLE), it could provide valuable information about the parameter.

To be specific suppose X has probability density function (pdf) $f_\theta(X)$, and the MLE $T \equiv T(X)$ of θ has pdf $g_\theta(T)$. We write $I(\theta) = E_\theta\big[-\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2}\big]$, the Fisher information contained in X and $I_T(\theta) = E_\theta\big[-\frac{\partial^2 \log g_\theta(T)}{\partial \theta^2}\big]$, the Fisher information contained in T. We assume any needed regularity conditions to justify these definitions. It is easy to show that $I(\theta) \geq I_T(\theta)$ with equality if and only if T is sufficient. In the above, and in what follows, we will not make a distinction between the random variables and the values that they assume, and use capital letters for both.

Thus, when the MLE T itself is not sufficient, there is loss of information in the Fisherian sense. How to recover this apparent loss of information? If U is an ancillary statistic, but (T, U) is sufficient, then U is then referred to as ancillary complement of T. Letting $h_\theta(T|U)$ be the conditional pdf of T given U, we have $I(\theta) = E_\theta[I_T(\theta|U)]$, where $I_T(\theta|U) = E_\theta\big[-\frac{\partial^2 \log h_\theta(T|U)}{\partial \theta^2}|U\big]$ is the Fisher information contained in the conditional distribution of T given U. According to Fisher, it is a mistake to calculate the information content of T with respect to the marginal distribution of T: the appropriate measure is $I_T(\theta|U)$ and not $I_T(\theta)$.

Example 1. Let (X_i, Y_i) $(i = 1, \cdots, n)$ be independent and identically distributed (iid) with common pdf

$$f_\theta(x, y) = \exp(-\theta x - y/\theta)\chi_{[x>0, y>0]}(x, y);\ \theta > 0,$$

where $\chi_A(t)$ is 1 if $t \in A$, and is 0 otherwise. This example is usually referred to as Fisher's gamma hyperbola (Fisher, 1956; Efron and Hinkley, 1978; Barndorff-Nielsen and Cox,

1994; Reid, 2003). Defining $T = (\sum_{i=1}^n Y_i / \sum_{i=1}^n X_i)^{\frac{1}{2}}$, $U = (\sum_{i=1}^n X_i)^{\frac{1}{2}}(\sum_{i=1}^n Y_i)^{\frac{1}{2}}$, it is easy to check that (i) T is the MLE of θ; (ii) U is ancillary; (iii) (T, U) is jointly minimal sufficient for θ. In this case, $I(\theta) = \frac{2n}{\theta^2}$, and $I_T(\theta) = \frac{2n}{\theta^2} \cdot \frac{2n}{2n+1}$, so that the loss of information is $I(\theta) - I_T(\theta) = \frac{2n}{(2n+1)\theta^2}$. However, according to Fisher, based on T, one should not report the information concerning θ as $I_T(\theta)$, but instead should report $I_T(\theta|U) = \frac{2n}{\theta^2} \cdot \frac{K_1(2U)}{K_0(2U)}$, where K_0 and K_1 are certain Bessel functions. For fixed n, $I_T(\theta|U)$ converges to zero as $U \to 0$ and to ∞ as $U \to \infty$.

Thus, following Fisher's recommendation, one needs to condition on an ancillary complement U to report the information content of T, the MLE, when T itself is not sufficient. Stretching Fisher's ideas further, one can even think of constructing inferential procedures in general, which are based on such conditioning. In Section 2 of this article, we will provide some examples to illustrate this.

However, situations may arise when an ancillary U may not exist. Indeed, Pena *et al.* (1992) have demonstrated this phenomenon for general discrete models. Basu pointed out that if $X_1, \cdots, X_n$ $(n \geq 2)$ are iid uniform (θ, θ^2), $\theta > 1$, then the MLE of θ is $T = [\max(X_1, \cdots, X_n)]^{1/2}$. But, in this case, there does not exist any ancillary complement U of T. On the other hand, it is shown in Basu (1964) that in many other situations there may exist multiple ancillary complements of the MLE T of θ, and it may not be clear which one to condition on.

One possible solution is to condition on the maximal ancillary statistic (Basu, 1959) if it exists. A maximal ancillary statistic is one such that every ancillary statistic is a function of the same. Indeed, such a statistic exists in some simple cases such as the location or location-scale models. However, in general, such a statistic may not exist.

We will introduce in Section 3 some of Basu's examples, and the response of Barnard and Sprott (1971), Cox (1971) and Cox and Hinkley (1974) on additional criteria needed to pick suitable ancillary complements of T.

Fraser (2004), in a recent article, has argued very strongly in favor of conditional inference, and has provided many compelling examples to illustrate his point. We consider in Section 4 two of his examples involving the general location or location-scale family of distributions. These examples are also considered elsewhere (see e.g. Reid, 1988). In both these examples, the dimension of the minimal sufficient statistic exceeds that of the parameter(s) of interest. However, Fraser has shown that when one conditions on a suitable ancillary statistic, it is possible to achieve the necessary dimension reduction, and the answer matches perfectly with the ones in the corresponding normal examples

where the dimension of the minimal sufficient statistic equals that of the parameter(s). Thus, in this problem, while dimension reduction through sufficiency is possible only in the normal case, the same is possible through ancillarity in more general situations where the former fails, and nothing is lost when it holds.

Section 5 introduces some approximate ancillary statistics, particularly those proposed by Efron and Hinkley (1978). In Section 6, we discuss briefly the role of ancillary statistics in deriving some higher order asymptotic results related to maximum likelihood estimation, generalized likelihood ratio tests and p-values. In particular, we introduce the p^*-formula of Barndorff-Nielsen (1983), and discuss some of its uses.

In Section 7, we bring out the issues of marginalization and conditioning with the objective of elimination of nuisance parameters, and discuss the role of some variations of ancillary statistics in this context. Section 8 introduces the notion of Bayesian ancillarity as given in Severini (1995). Section 9 contains some miscellaneous remarks which include the role of ancillary statistics in finding optimal estimating functions as discussed in Godambe (1976, 1980), as well as the ancillarity paradox of Brown (1990). Some concluding remarks are made in Section 10.

There are some earlier reviews of ancillary statistics. We would like to draw attention in particular to Fraser (1979) and Buehler (1982).

2 CONDITIONAL INFERENCE USING ANCILLARY STATISTICS

We begin with an example originally considered in Welch (1939), and subsequently revisited by many authors, such as Barndorff-Nielsen and Cox (1994), and most recently by Fraser (2004).

Example 2. Suppose X_1 and X_2 are iid uniform $(\theta-1, \theta+1)$, θ real. The minimal sufficient statistic is $(Y_1 = \min(X_1, X_2), Y_2 = \max(X_1, X_2))$. Let $T = \frac{1}{2}(Y_1 + Y_2)$ and $U = \frac{1}{2}(Y_2 - Y_1)$. Then (Y_1, Y_2) is one-to-one with (T, U). Also the dimension of the minimal sufficient statistic exceeds that of the parameter. Here the MLE of θ is any point in the interval $(Y_2 - 1, Y_1 + 1)$. In particular, T is an MLE of θ. Also, U is ancillary. The conditional pdf of T given $U = u$ is

$$\mathrm{f}_\theta(T|U) = [2(1-U)]^{-1} \chi_{[\theta-1+U<T<\theta+1-U]}(T), \tag{1}$$

where $\chi_A(T)$ is 1 or 0 according as $T \in A$ or $T \in A^c$. Based on this conditional pdf, a $100(1-\alpha)\%$ confidence interval for θ is given by $(T-(1-U)(1-\alpha), T+(1-U)(1-\alpha))$. It may be noted also that when U is close to 1, then θ is very precisely determined, while if U is close to zero, the conditional distribution of T given U is essentially uniform over the whole interval $(\theta-1, \theta+1)$. Thus while U carries no direct location information concerning θ, it carries extensive information on precision.

On the other hand, the marginal pdf of T is

$$\begin{aligned} f_\theta(T) &= T-\theta+1 \text{ if } \theta-1<T<\theta \\ &= \theta+1-T \text{ if } \theta \leq T < \theta+1. \end{aligned}$$

If this marginal pdf is used to form, say, a confidence interval for θ, it can lead "confidence" in θ values that are absurd when U is close to 1. Then θ is essentially known exactly, but the unconditional confidence region for θ may lead to values which are impossible in the light of the data; see also Fraser (2004, pp 335-336).

Welch (1939), in a similar example, argued strongly against conditional inference by producing two different $100(1-\alpha)\%$ confidence intervals for θ based on two different optimality criteria, both with rather extreme properties. The first, a likelihood ratio based unbiased confidence interval for θ, gives full range of possible θ values for large U. The second, the shortest on average symmetric confidence interval, is either the full range of possible θ values or the empty set. The details are given in equations (2.4) and (2.5) of Fraser (2004, pp 335-336). However, the interval as mentioned earlier does not suffer from this problem, and the arguments put forward by Barndorff-Nielsen and Cox (1994, pp 34-35) seem quite convincing to favor a conditional inference.

Welch (1939) was arguing from a Neyman-Pearson viewpoint where the optimality property was based on maximizing the power subject to a constraint given by the test size. The phenomenon, however, is much more general; see Fraser and McDunnough (1980). If one optimizes a criterion subject to a constraint conditional on an ancillary statistic, and then decides to require the constraint only marginally, then one can always increase optimality, or at least not reduce it, by allowing more for some values of the ancillary statistic, and less for others; or in a confidence context, give bigger intervals in the high precision cases, and smaller intervals in the low precision cases, thus shortening the average interval length (Fraser, 2004, pp 335-336) at a given confidence level.

The next example provides an empirical Bayes (EB) scenario where conditioning with

respect to an ancillary statistic can produce quite a meaningful answer.

Example 3. This example appears in Hill (1990). Let $X_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, 1)$ and $\theta_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, A)$ $(i = 1, \cdots, k)$. Here $A(> 0)$ is known, but μ(real) is possibly unknown. Suppose one needs a confidence interval for one of the θ_i, say θ_1. Writing $B = (1+A)^{-1}$, the posterior distribution of θ_1 is $\mathrm{N}((1-B)X_1 + B\mu, 1-B)$. In an EB method, one estimates μ from the marginal distribution of $(X_1, \cdots, X_k)$. Since marginally $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, B^{-1})$, $\bar{X} = k^{-1}\sum_{i=1}^{k} X_i$ is complete sufficient for μ and the estimated posterior of θ_1 is $\mathrm{N}((1-B)X_1 + B\bar{X}, 1-B)$. Based on this, the shortest $100(1-\alpha)\%$ confidence interval for θ_1 is $(1-B)X_1 + B\bar{X} \pm z_{\alpha/2}\sqrt{1-B}$, where $z_{\alpha/2}$ is the upper $100\alpha\%$ point of the $\mathrm{N}(0,1)$ distribution.

It is clear that the above EB method does not account for the uncertainty due to estimation of μ. To see how an ancillary argument can overcome this, we may note that marginally $U = X_1 - \bar{X}$ is ancillary and $U \sim \mathrm{N}(0, \frac{k-1}{kB})$. It is easy to check also that $\theta_1 - \{(1-B)X_1 + B\bar{X}\}|U \sim \mathrm{N}(0, 1-B+Bk^{-1})$. Thus the shortest $100(1-\alpha)\%$ confidence interval for θ_1 based on this conditional distribution is $(1-B)X_1 + B\bar{X} \pm z_{\alpha/2}\sqrt{1-B+Bk^{-1}}$.

Alternatively, if one takes a hierarchical Bayesian (HB) approach where (i) $X_i|\theta_1, \cdots, \theta_k, \mu \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, 1)$, (ii) $\theta_1, \cdots, \theta_k|\mu \overset{\text{iid}}{\sim} \mathrm{N}(\mu, A)(A > 0)$, and (iii) $\mu \sim \text{uniform}(-\infty, \infty)$, it turns out that $\theta_1|X_1, \cdots, X_k, \mu \sim \mathrm{N}((1-B)X_1 + B\mu, 1-B)$ and $\mu|X_1, \cdots, X_k \sim \mathrm{N}(\bar{X}, (kB)^{-1})$. Together, they imply $\theta_1|X_1, \cdots, X_k \sim \mathrm{N}((1-B)X_1 + B\bar{X}, 1-B+Bk^{-1})$. Thus the $100(1-\alpha)\%$ confidence interval for θ_1 based on this hierarchical prior is the same as the one conditioned on the ancillary U. Noting that $Bk^{-1} = V(B\mu|X_1, \cdots, X_k)$, it may be noted that in this case ancillarity accounts for the uncertainty due to estimation μ as much as the HB procedure. While the above coincidence between the two procedures need not always be true, conditioning on an ancillary statistic, can often correct the problem faced by a naive EB procedure.

Datta *et al.* (2002) demonstrated this in a framework slightly more general than that of Hill. Once again, we consider the model where $X_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, 1)$ and $\theta_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, A)$. However, this time both μ and A are unknown. Let $S^2 = k^{-1}\sum_{i=1}^{k}(X_i - \bar{X})^2$. Since the X_i are marginally iid $\mathrm{N}(\mu, B^{-1})$, $(\bar{X}, S)$ is complete sufficient for μ, A, while $U = (X_1 - \bar{X})/S$ is ancillary. Consider the EB estimator $(1-\hat{B})X_1 + \hat{B}\bar{X}$ of θ_1, where $\hat{B} \equiv \hat{B}_d(S) = \min[\{(k-d)/(k-1)S^2\}, (k-d)/(k-1)]$ for some bounded $d > 1$ not depending on k. Then writing

$t \equiv t(\hat{B}, U.\alpha) = z_{\alpha/2}\left[1 + k^{-1}\left\{\frac{(1+z^2_{\alpha/2})\hat{B}^2}{4(1-\hat{B})^2} + \frac{(2U^2+4-d)\hat{B}}{2(1-\hat{B})}\right\}\right]$,
where $z_{\alpha/2}$ is the upper $100\alpha/2\%$ point of the N(0,1) distribution,

$$P(\theta_1 \in (1-\hat{B})X_1 + \hat{B}\bar{X} \pm t(1-\hat{B})^{1/2}|U] = 1 - \alpha + O_p(k^{-3/2}).$$

This shows that with a suitable bias correction, asymptotically an EB confidence interval, conditioned on an ancillary statistic, meets the target coverage probability quite accurately.

3 NONUNIQUENESS OF ANCILLARY STATISTICS

Basu (1964,1992) pointed out many difficulties with the actual use of ancillary statistics. First two statistics U_1 and U_2 may be individually ancillary, but (U_1, U_2) may not jointly be so. Thus, in the case of a controversy as to which one of U_1 and U_2 should determine the reference set, the dilemma cannot be solved by conditioning on (U_1, U_2) jointly. The following simple example illustrates this.

Example 4. Let

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \overset{\text{iid}}{\sim} \text{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$$

$i = 1, \cdots, n$, where $\rho \in (-1, 1)$ is unknown. We let $U_1 = \sum_{i=1}^n X_i^2$, $U_2 = \sum_{i=1}^n Y_i^2$ and $W = \sum_{i=1}^n X_i Y_i$. It is easy to recognize both U_1 and U_2 as ancillary, each having the χ^2_n distribution. But jointly (U_1, U_2) is not ancillary as it is readily checked by calculating $\text{corr}(U_1, U_2) = \rho^2$ which depends on ρ. Thus, while $\frac{W}{U_1}$ and $\frac{W}{U_2}$ are both unbiased estimators of ρ (unconditionally or conditionally), $V(\frac{W}{U_1}|U_1) = \frac{1-\rho^2}{U_1}$ and $V(\frac{W}{U_2}|U_2) = \frac{1-\rho^2}{U_2}$. It is tempting to opt for the larger one of U_1 and U_2 as the ancillary statistic in this example, but then the choice of the ancillary statistic becomes data-dependent, which is counter to the usual frequentist paradigm. However, the problem disappears when one conditions on an approximate ancillary statistic (see Section 6).

Cox (1971) suggested a way to deal with multiple ancillaries. By the identity $I(\theta) = E[I_T(\theta|U)]$, Cox argued that the basic role of conditioning on an ancillary U is to discriminate between samples with varying degrees of information. In the presence of multiple ancillaries, choose that U for which $I_T(\theta|U)$ is most variable, i.e., $V_\theta[I_T(\theta|U)]$ is maxi-

mum. Unfortunately, in most instances $V_\theta[I_T(\theta|U)]$ is a function of the unknown θ, and there may not be a unique U which maximizes $V_\theta[I_T(\theta|U)]$ for all θ. Moreover, in Example 4, since $V_\theta[I_T(\theta|U_1)] = V_\theta[I_T(\theta|U_2)]$, the Cox method will fail to distinguish between U_1 and U_2.

Example 5. The next example of Basu (1964) involves a random variable X assuming values $1, 2, \cdots, 6$ such that

$$P_\theta(X = j) = \begin{cases} (j-\theta)/12, & j = 1, 2, 3; \\ (j-3+\theta)/12, & j = 4, 5, 6. \end{cases}$$

where $\theta \in [-1, 1]$. Here the MLE of θ is given by $T(X)$, where $T(1) = T(2) = T(3) = -1$ and $T(4) = T(5) = T(6) = 1$. There are six possible ancillary complements of T given by

X	1	2	3	4	5	6
$U_1(X)$	0	1	2	0	1	2
$U_2(X)$	0	1	2	0	2	1
$U_3(X)$	0	1	2	1	0	2
$U_4(X)$	0	1	2	2	0	1
$U_5(X)$	0	1	2	1	2	0
$U_6(X)$	0	1	2	2	1	0

A natural question concerns which ancillary complement to choose under the given circumstance. Basu left this example with a question mark. Also, there is no clearcut choice among $U_1, \cdots, U_6$ if one computes the information content of T based on its conditional distributions given these six ancillary statistics since no strict inequality exists between the $I_T(\theta|U_j)$ $(j = 1, \cdots, 6)$ for all θ and all values of U_j. However, after some tedious algebra, one can check that $V_\theta[I_T(\theta|U_1)] > V_\theta[I_T(\theta|U_j)]$ $(j = 2, \cdots, 6)$ for all θ, so that following Cox's guidelines, one chooses U_1 as the ancillary statistic to condition on. From another point of view (Barnard and Sprott, 1971), under the transformation $gX = X + 3 \pmod 6$ which gives the induced transformation $g^*\theta = -\theta$, it turns out that the only ancillary statistic unaffected by this transformation is U_1. Thus, with this constraint as well, U_1 seems to be the most appropriate one.

Example 6. Basu's third example deals with $X \sim$ uniform$[\ \theta, \theta + 1)$, $0 \le \theta < \infty$. The

sample space is $\mathcal{X} = [\,0, \infty)$, and the likelihood function

$$L(\theta) = \begin{cases} 1, & \text{if } X - 1 < \theta \leq X; \\ 0, & \text{otherwise.} \end{cases}$$

Thus, every point in the interval $(X-1, X]$ is an MLE of θ. One such choice is $T = [X]$, the integer part of X. Let $\phi(X) = X - [X]$. Then $\phi(X) \sim$ uniform$[0, 1)$, and is ancillary. Since $X = [X] + \phi(X)$, $([X], \phi(X))$ is one-to-one with the minimal sufficient X. So $\phi(X)$ is the ancillary complement of [X]. Note that

$$[X] \begin{cases} = [\theta], & \text{if } \phi(X) \geq \phi(\theta) \Leftrightarrow \theta \leq X < [\theta] + 1; \\ = [\theta + 1] = [\theta] + 1, & \text{if } \phi(X) < \phi(\theta) \Leftrightarrow [\theta] + 1 \leq X < \theta + 1. \end{cases}$$

Also, it is easy to check that

$$\begin{aligned} \mathrm{P}_\theta[[X] = [\theta] | \phi(X)] = 1, & \quad \text{if } \phi(\theta) \leq \phi(X); \\ \mathrm{P}_\theta[[X] = [\theta + 1] | \phi(X)] = 1, & \quad \text{if } \phi(\theta) > \phi(X). \end{aligned}$$

Thus, the conditional distribution of the MLE $[X]$ given $\phi(X)$ is degenerate at $[\theta]$ or $[\theta + 1]$ depending on whether $\phi(X) \geq \phi(\theta)$ or $\phi(X) < \phi(\theta)$. This changes the status of $[X]$ from a random variable to an unknown constant. However, Barnard and Sprott (1971) did not find any anomaly in this. In their view, the likelihood is defined in $[X]$ in the ratio $1 - \phi(X) : \phi(X)$. Thus $[X]$ measures position of the likelihood, and $\phi(X)$ measures its shape in the sense of the proportion into which $[X]$ divides the likelihood. Thus, holding $\phi(X)$ fixed will also result in holding $[X]$ fixed as well.

4 LOCATION AND LOCATION-SCALE FAMILIES OF DISTRIBUTIONS

The models discussed in this section are treated in more detail in Fraser (2004). It is well-known that for most examples of the location or location-scale family of distributions (other than the normal), the dimension of the minimal sufficient statistic exceeds that of the parameter of the distribution. However, for the location family of distributions, inference about the location parameters can be based on the conditional distribution of the sample average given some suitable ancillary statistic. For the location-scale family

of distributions, similar calculations can be carried out from the conditional distribution of a t-statistic. The end result is identical to the result in the normal distribution case where the dimension of the minimal sufficient statistic is equal to the dimension of the parameter. Thus, conditioning on an ancillary statistic reduces the dimensionality in general location or location-scale problems, while nothing is lost when the distribution is normal. We make this more specific in the following two examples.

Example 7. Let $X_1, \cdots, X_n$ be iid $\mathrm{N}(\theta, 1)$, θ real. Then any inference for θ will be based on $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ which is the minimal sufficient statistic.

Suppose now we dispense with the normality assumption and assume that $X_1, \cdots, X_n$ ($n \geq 2$) are iid with common pdf $f(X - \theta)$, θ real, and f is some specified pdf other than normal. The the maximal ancillary statistic is $(U_1, \cdots, U_n)$, where $U_j = X_j - \bar{X}$, $j = 1, \cdots, n$. (Note that only $n - 1$ of $U_1, \cdots, U_n$ are linearly independent.) Then the conditional pdf of $\bar{X}$ given $U_1, \cdots, U_n$ is

$$f_\theta(\bar{X}|U_1, \cdots, U_n) = k \prod_{j=1}^{n} f(\bar{X} + U_j - \theta), \tag{2}$$

where k is the normalizing constant, which typically depends on the U_j. This conditionality argument reduces the dimension of the problem from n to 1. In the normal case, the conditional density given in (2) becomes the unconditional density of $\bar{X}$ since by Basu's theorem (Basu, 1955), the complete sufficient statistic $\bar{X}$ is independent of the ancillary $U_1, \cdots, U_n$. Fraser (2004) illustrated this dimension reduction through p-values. One can work instead with the conditional pdf of $\hat{\theta}$, the MLE of θ, given the ancillary statistics $(U_1, \cdots, U_n)$, since $\hat{\theta}$ is also location equivariant and in the normal case, $\hat{\theta}$ equals $\bar{X}$.

Example 8. We now extend the previous problem to location-scale family of distributions. Once again we begin with $X_1, \cdots, X_n$ ($n \geq 2$) which are iid $\mathrm{N}(\theta, \sigma^2)$. Then the usual inferential procedure for θ is based on the t-pivot $T = \sqrt{n}(\overline{X} - \mu)/S$, where $S^2 = (n-1)^{-1} \Sigma_{i=1}^{n} (X_i - \overline{X})^2$.

Instead, now we begin with the location-scale family of densities, where $X_1, \cdots, X_n$ are iid with a common pdf $f(X|\theta, \sigma) = \sigma^{-1} f((X - \theta)/\sigma)$, where f is some specified pdf other than normal. Then $U_j = (X_j - \bar{X})/S$ ($j = 1, \cdots, n$) are ancillary (only $n - 2$ of the U_j are linearly independent for $n \geq 3$) and the joint pdf of $\bar{X}$ and S conditional on

the U_j's is given by

$$f(\bar{X}, S|U_1, \cdots, U_n; \theta, \sigma) = k\sigma^{-n} \prod_{j=1}^{n} f\left(\frac{\bar{X} + SU_j - \theta}{\sigma}\right) S^{n-2}, \tag{3}$$

where k is the normalizing constant. The dimensionality of the problem is now reduced to 2. In the normal case, the complete sufficient $(\bar{X}, S)$ is independent of the ancillary $U_1, \cdots, U_n$, and the conditional pdf given in (3) becomes the unconditional pdf.

5 APPROXIMATE ANCILLARIES

Efron and Hinkley (1978) introduced the notion of approximate ancillarity in the context of conditioning with respect to the observed Fisher information. They motivated their approach with the following example from Cox (1958).

Example 9. Suppose there are two instruments each designed to measure some unknown θ. Suppose using instrument k, the observed measurements $X_1, \cdots, X_n$ of θ are iid $\mathrm{N}(\theta, \sigma_k^2)$, $k = 0, 1$, where σ_0^2 and σ_1^2 are known and unequal. When a measurement is obtained, one notes also the instrument used, so that the observed data are $(X_1, U_1), \cdots, (X_n, U_n)$, where $U_j = k$ if U_j is obtained using instrument k. The choice between the two instruments is decided by a flip of an unbiased coin so that $\mathrm{P}(U_j = 1) = \mathrm{P}(U_j = 0) = \frac{1}{2}$. Clearly U_j's are ancillary.

The log-likelihood function in this case is given by

$$\log L(\theta) = l(\theta) = \text{constant} - \sum_{j=1}^{n} \log \sigma_{U_j} - \frac{1}{2} \sum_{j=1}^{n} \frac{(X_j - \theta)^2}{\sigma_{U_j}^2}, \tag{4}$$

Then the MLE of θ is $\hat{\theta} = \sum_{j=1}^{n} X_j \sigma_{U_j}^{-2} / \sum_{j=1}^{n} \sigma_{U_j}^{-2}$. Also, the (expected) Fisher information is given by

$$E\Big[-\frac{\partial^2 l}{\partial \theta^2}\Big] = \sum_{j=1}^{n} E(\sigma_{U_j}^{-2}) = \frac{n}{2}(\sigma_0^{-2} + \sigma_1^{-2}), \tag{5}$$

while the observed Fisher information

$$j(\hat{\theta}) = -\frac{\partial^2 l}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} = \sum_{j=1}^{n} \sigma_{U_j}^{-2} = (n - U)\sigma_0^{-2} + U\sigma_1^{-2} \tag{6}$$

where $U = \sum_{j=1}^{n} U_j$, the number of times instrument 1 was used.
Note also in this case

$$V(\hat{\theta}|U_1, \cdots, U_n) = [U\sigma_1^{-2} + (n-U)\sigma_0^{-2}]^{-1} = j^{-1}(\hat{\theta}) \tag{7}$$

which equals the reciprocal of the observed Fisher information given in (6), but is different from the reciprocal of the expected Fisher information given in (5).

In this example it seems clear that $V(\hat{\theta} \mid U_1, \ldots, U_n)$ is the correct variance estimate, since it is known which instrument was used. In general we do not have

$$V(\hat{\theta} \mid U) = j^{-1}(\hat{\theta}),$$

but Efron and Hinkley (1978) show that in i.i.d. sampling from a suitably regular one-parameter model this holds approximately, in the sense that

$$V(\hat{\theta} \mid U) = j^{-1}(\hat{\theta})\{1 + O_p(n^{-1})\},$$

and further that the discrepancy between $j(\hat{\theta})$ and $I(\theta)$ can be expressed in terms of the statistical curvature of the model:

$$\sqrt{n}\left(\frac{j(\hat{\theta})}{I(\hat{\theta})} - 1\right) \sim N(0, \gamma_\theta^2), \tag{8}$$

where the curvature $\gamma_\theta = (\nu_{20}\nu_{02} - \nu_{11})^{\frac{3}{2}}/(\nu_{20})^{\frac{3}{2}}$, and

$$\nu_{jk} = \nu_{jk}(\theta) = E\Big[\Big(\frac{\partial \log f}{\partial \theta}\Big)^j \Big\{\frac{\partial^2 \log f}{\partial \theta^2} + E\Big(\frac{\partial \log f}{\partial \theta}\Big)^2\Big\}\Big]^k,$$

and f is the given density of the observations.

Result (8) also suggests that an approximately ancillary statistic is given by

$$Q = \frac{1 - j(\hat{\theta})/I(\hat{\theta})}{\gamma_{\hat{\theta}}}, \tag{9}$$

in the sense that $\sqrt{n}Q$ has a limiting standard normal distribution, and (9) has come to be known as the Efron-Hinkley ancillary statistic. It is first-order ancillary; i.e. the normal approximation to the distribution of Q has relative error $O(n^{-1/2})$. Skovgaard (1985) showed that the relative error is actually $O(n^{-1})$, in a moderate deviation neighborhood

of an arbitrary fixed point θ_0 in the interior of the parameter space; this is called second order local ancillarity. Local ancillarity was introduced in Cox (1980).

Approximate ancillary statistics are closely tied to higher order approximations, but are not usually of particular interest in themselves. Fraser and Reid (1993, 1995, 2001) use the theory of higher order approximations to avoid the explicit calculation of a second order ancillary statistic, but instead find a direct expression for an approximation for the conditional distribution given a second order ancillary. This construction is described as well in Reid (2003) and Severini (2000, Ch. 7.5.3).

Example 4 (revisited). Cox and Hinkley (1974, p.34) suggest $U' = U_1 + U_2 = \Sigma(X_i^2 + Y_i^2)$ as an approximate ancillary statistic for this example, as it has mean $2n$ and variance $4n(1+\rho^2)$, so its first moment is free of ρ and its second moment approximately so, at least for small ρ. Wang (1993) suggested a standardized version $(U'-2n)/2\sqrt{\{W^2+n^2\}}$, which has both mean and variance independent of ρ. Defining ancillary statistics through constancy of moments is not the same as local or approximate ancillarity, although to first order it is the same for asymptotically normally distributed statistics.

The Efron-Hinkley ancillary statistic for this example can be calculated from (9), but the explicit expression is not very informative. Since its claim to ancillarity is that it has mean 0 and variance 1, and is asymptotically normally distributed, it is likely to be equivalent to Wang's modification of U'. We can also embed the model in a two-parameter exponential family and compute the directed likelihood ancillary. Either of these ancillary statistics can be used for higher order appoximations to the distribution of the maximum likelihood estimator, although the detailed calculations are somewhat cumbersome. Reid (2003) illustrates the construction of Fraser and Reid (1993, 1995) on this example and derives higher order appoximations to the distribution of the maximum likelihood estimator. The detailed calculations, however, are also cumbersome.

6 ASYMPTOTICS AND ANCILLARITY

Barndorff Nielson's famous p^*-formula is an approximation to the density of the MLE given an approximate ancillary. This result has been the cornerstone of much subsequent research including the derivation of generalized likelihood ratio tests, p-values and even various new likelihoods. To use it for numerical integration to get p-values, however, requires information on the ancillary, at least locally. It turns out that when the MLE is itself a sufficient statistic, this approximation is for the unconditional density of the

MLE. On the other hand, if the MLE is not sufficient, but together with an ancillary statistic constitutes the minimal sufficient statistic, then this approximation holds for the density of the MLE conditional on its ancillary complement.

We will illustrate Barndorff-Nielsen's general result with examples. First, we begin with the regular exponential family of densities where the MLE is minimal sufficient. For simplicity of exposition, we consider only the one-parameter exponential family, although similar results are available for the multiparameter exponential family of densities as well. The following example is discussed in detail in Ghosh (1994, pp 75-76) and Severini (2000, pp 184-185).

Example 10: Let $X_1, \ldots, X_n$ be iid with the common pdf (with respect to some σ-finite measure μ)

$$f_\theta(X) = \exp[\theta X - \psi(\theta) + c(X)]$$

Then the cumulant generating function is $K_\theta(t) = \psi(t+\theta) - \psi(\theta)$.

The saddlepoint approximation to the pdf of $\bar{X}$ (see for example Daniels, 1954; Reid, 1988)

$$f_\theta(\bar{X}) \doteq \frac{\sqrt{n}\exp[n\{K_\theta(\hat{\lambda}) - \hat{\lambda}\bar{X}\}]}{[2\pi K_\theta''(\hat{\lambda})]^{1/2}},$$

where $K_\theta'(\hat{\lambda}) = \bar{X}$. Since $K_\theta'(t) = \psi'(t+\theta)$, $K_\theta''(t) = \psi''(t+\theta)$, $\bar{X} = \psi'(\hat{\lambda}+\theta)$, i.e. $\hat{\lambda}+\theta = (\psi')^{-1}(\bar{X}) = \hat{\theta}$, where $\hat{\theta}$ is the MLE of θ. Then $K_\theta(\hat{\lambda}) = \psi(\hat{\lambda}+\theta) - \psi(\theta) = \psi(\hat{\theta}) - \psi(\theta)$, and $K_\theta''(\hat{\lambda}) = \psi''(\hat{\lambda}+\theta) = \psi''(\hat{\theta})$. Hence, the saddlepoint approximation to the pdf of $\bar{X}$ is

$$\begin{aligned} f_\theta(\bar{X}) &= \frac{\sqrt{n}\exp[n\{\psi(\hat{\theta}) - \psi(\theta)\} - n(\hat{\theta}-\theta)\bar{X}]}{[2\pi\psi''(\hat{\theta})]^{1/2}} \\ &= \frac{\sqrt{n}\exp[n\{\theta\bar{X} - \psi(\theta)\} - n\{\hat{\theta}\bar{X} - \psi(\hat{\theta})\}]}{[2\pi\psi''(\hat{\theta})]^{1/2}} \\ &= n\frac{L(\theta)/L(\hat{\theta})}{\sqrt{2\pi}(-l_{\theta\theta})^{1/2}}, \end{aligned}$$

where $L(\theta) = \exp[n\{\theta\bar{X} - \psi(\theta)\}$ is the likelihood function, and $l_{\theta\theta} = \frac{d^2}{d\theta^2}l(\theta)$, where $l(\theta) = \log L(\theta)$. In this example $l_{\theta\theta} = -n\psi''(\theta)$. Since $\psi'(\hat{\theta}) = \bar{X}$, $d\bar{X}/d\hat{\theta} = \psi''(\hat{\theta}) = n^{-1}j(\hat{\theta})$, where $j(\hat{\theta})$ is the observed Fisher information, namely, $-l_{\theta\theta}$ evaluated at $\theta = \hat{\theta}$. Hence, the pdf $p_\theta(\hat{\theta})$ of the MLE $\hat{\theta}$ of θ is approximated by

$$f_\theta(\hat{\theta}) \doteq \frac{1}{\sqrt{2\pi}} \frac{L(\theta)}{L(\hat{\theta})} j(\hat{\theta})^{1/2}.$$

The error of approximation is $O(n^{-1})$. It turns out that the renormalized density $p^*_\theta(\hat{\theta}) = cL(\theta)/L(\hat{\theta})j(\hat{\theta})^{1/2}$ reduces the error of approximation to $O(n^{-3/2})$. The multiparameter generalization of this result is given by

$$f_\theta(\hat{\theta}) = c|j(\hat{\theta})|^{1/2} L(\theta)/L(\hat{\theta}),$$

where $|A|$ denotes the determinant of the matrix A.

We now revisit the location-scale family of densities to understand the role of ancillary statistics in this formulation. This is considered in Reid (1988).

Example 8(continued). The likelihood $L(\theta, \sigma)$ is given by $L(\theta, \sigma) = \sigma^{-n} \prod_1^n f((X_i - \theta)/\sigma)$. We now make the transformation $U_i = (X_i - \hat{\theta})/\hat{\sigma}$, where $(\hat{\theta}, \hat{\sigma})$ is the MLE of (θ, σ). In this example,

$$\begin{aligned} \frac{L(\theta, \sigma)}{L(\hat{\theta}, \hat{\sigma})} &= \frac{\sigma^{-n} \prod_1^n f((X_i - \theta)/\sigma)}{\hat{\sigma}^{-n} \prod_1^n f((X_i - \hat{\theta})/\hat{\sigma})} \\ &= (\hat{\sigma}/\sigma)^n \prod_1^n f\left(\frac{\hat{\sigma} U_i + \hat{\theta} - \theta}{\sigma}\right) / \prod_1^n f(U_i) \end{aligned}$$

Note that $f(\hat{\theta}, \hat{\sigma}, U_1, \ldots, U_n | \theta) = \sigma^{-n} \prod_1^n f\left(\frac{\hat{\theta} + \hat{\sigma} U_i - \theta}{\sigma}\right) \hat{\sigma}^n$. This leads to

$$\begin{aligned} f(\hat{\theta}, \hat{\sigma} | U_1, \ldots, U_n, \theta, \sigma) &= c(U_1, \cdots, U_n) \sigma^{-n} \hat{\sigma}^{n-2} \prod_1^n f\left(\frac{\hat{\theta} + \hat{\sigma} U_i - \theta}{\sigma}\right) \\ &= c(U_1, \ldots, U_n) |j(\hat{\theta})|^{1/2} \frac{L(\theta, \sigma)}{L(\hat{\theta}, \hat{\sigma})}. \end{aligned}$$

Barndorff-Nielsen (1983) and later Skovgaard (1990) gave a somewhat heuristic proof that the above provides an expression for the conditional density of the MLE in a more general framework. In particular, suppose that the minimal sufficient statistic is one-to-one with $(\hat{\theta}, U)$, where $\hat{\theta}$ is the MLE of the parameter of interest, and U is ancillary. Then Barndorff-Nielsen's approximation to the conditional density of $\hat{\theta}$ given U is given by

$$p^*_\theta(\hat{\theta} | U, \theta) = c(U, \theta) \frac{L(\theta; \hat{\theta}, U)}{L(\hat{\theta}; \hat{\theta}, U)} |j(\hat{\theta})|^{1/2}. \tag{10}$$

Davison (1988) gave an interpretation of the right hand side of (10) as the Laplace approximation of the normalized likelihood $L(\theta)/\int L(\theta)d\theta$. Begin with the approximation $L(\theta) = \exp[l(\theta)] \doteq \exp[n\{l(\hat{\theta}) + (\theta - \hat{\theta})^T l_\theta(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T j(\hat{\theta})(\theta - \hat{\theta})]$, where $l_\theta(\hat{\theta})$ denotes the gradient vector of $l(\theta)$. Since $l_\theta(\hat{\theta}) = 0$, we get

$$\begin{aligned} L(\theta) &\doteq \exp[l(\hat{\theta})]\exp[-\frac{1}{2}(\theta - \hat{\theta})^T j(\hat{\theta})(\theta - \hat{\theta})] \\ &= L(\hat{\theta})\exp[-\frac{1}{2}(\theta - \hat{\theta})^T j(\hat{\theta})^{1/2}(\theta - \hat{\theta})]. \end{aligned}$$

Hence, $\int_{-\infty}^{\infty} L(\theta)d\theta = L(\hat{\theta})(2\pi)^{p/2}|j(\hat{\theta})|^{-1/2}$. Thus,

$$\frac{L(\theta)}{\int_{-\infty}^{\infty} L(\theta)d\theta} \doteq \frac{L(\theta)}{L(\hat{\theta})(2\pi)^{p/2}}|j(\hat{\theta})|^{1/2}.$$

We can view the role of the ancillary U as providing a complementing statistic to $\hat{\theta}$, in order that the p^* approximation given in equation (10) is defined on a sample space that is of the same dimension as the parameter space. This approximation, however, will only be valid for inference about θ if U is either exactly ancillary or approximately ancillary. If U is second order ancillary, then the renormalized p^* approximation will have relative error $O(n^{-3/2})$, while if U is just first order ancillary, it will have relative error $O(n^{-1})$.

When θ is a scalar parameter, the p^* approximation given in (10) can be re-expressed in terms of the density of the signed likelihood root

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)[2\{l(\hat{\theta}) - l(\theta)\}]^{1/2},$$

assuming the transformation from $\hat{\theta}$ to r is one-to-one, although the dependence of r on $\hat{\theta}$ is suppressed in the notation. Inference about θ is then readily obtained from the distribution function $F(r \mid U; \theta)$, for example the p-value for testing that $\theta = \theta_0$ is $F(r^0(\theta_0) \mid U; \theta_0)$. This distribution function can be approximated to $O(n^{-3/2})$, using a technique due to Lugannani and Rice (1980) and Barndorff-Nielsen (1986). The resulting approximation is

$$F(r \mid U; \theta) = \Phi(r^*)\{1 + O(n^{-3/2})\} \tag{11}$$

where $r^* = r + r^{-1}\log(q/r)$, $q = \{l_{;\hat{\theta}}(\hat{\theta}) - l_{;\hat{\theta}}(\theta)\}j^{-1/2}(\hat{\theta})$, and $l_{;\hat{\theta}} = \partial\ell(\theta; \hat{\theta}, U)/\partial\hat{\theta}$ is a sample space derivative with the ancillary statistic U held fixed. A statistic Q that does not require the determination of an explicit expression for U is developed in Fraser and

Reid (1993, 1995, 2001).

7 ELIMINATION OF NUISANCE PARAMETERS

7.1 EXTENDED DEFINITIONS OF ANCILLARITY

We begin with a model parameterized by $\theta = (\psi, \lambda)$, where ψ is the parameter of the interest, and λ is the nuisance parameter. The simplest way to eliminate the nuisance parameter λ is to substitute $\hat{\lambda}_\psi$, the constrained maximum likelihood estimate of λ for fixed ψ. While the resulting profile likelihood $L_p(\psi) = L(\psi, \hat{\lambda}_\psi)$ has many of the first order asymptotic properties of a likelihood (see especially Barndorff-Nielsen and Cox, 1994, Ch. 3.4), this solution is unsatisfactory more generally as it does not allow for errors of estimation of λ. An alternative, relatively simple, solution is Bayesian: assign a prior to (ψ, λ), find the posterior density and hence the posterior marginal density for ψ by integration. However, this requires a choice of priors, which may be difficult for high-dimensional nuisance parameters.

Other methods do exist. For example, if (T, U) is minimal sufficient and the marginal distribution of T involves ψ only, and does not involve the nuisance parameter λ, then inference about ψ can be based on the marginal distribution of T. The following example illustrates this.

Example 11. This is the famous Neyman-Scott (1948) situation. Let $(X_{i1}, X_{i2})^T$ be independently distributed $N_2\left[\begin{pmatrix}\mu_i\\ \mu_i\end{pmatrix}, \sigma^2 I_2\right]$, $i = 1, \ldots, n$, where $\mu_1, \ldots, \mu_n, \sigma^2$ are all unknown. Here $\psi = \sigma^2$ and $\lambda = (\mu_1, \cdots, \mu_n)$. If inference is based on the marginal likelihood of σ^2 based on the paired differences $X_{i1} - X_{i2}$, then $\hat{\sigma}^2_{MMLE} = \frac{1}{2n}\sum_1^n (X_{i1} - X_{i2})^2$, the marginal MLE of σ^2, converges in probability to σ^2 as $n \to \infty$. In contrast, the MLE $\hat{\sigma}^2_{MLE} = \frac{1}{4n}\sum_1^n (X_{i1} - X_{i2})^2$ of σ^2 converges in probability to $\frac{1}{2}\sigma^2$ and not σ^2 as $n \to \infty$. However, in general, it may not be possible to find a statistic T whose marginal distribution does not depend on the nuisance parameters.

An alternative method is the so called conditional likelihood approach. Suppose the joint density of the minimal sufficient statistic (T, U) is given by

$$f(T, U; \psi, \lambda) = f(T|U; \psi) f(U; \psi, \lambda) \tag{12}$$

Then the inference can be based on the conditional density $f(T|U; \psi)$ which does not

involve λ. In the Neyman-Scott example, noting that (X_{i1}, X_{i2}) is one-to-one with $(X_{i1}+X_{i2}, X_{i1}-X_{i2})$, by the independence of $X_{i1}+X_{i2}$ with $X_{i1}-X_{i2}$, it follows that conditional on the $X_{i1}+X_{i2}$'s $(i=1,\ldots,n)$, the conditional MLE of σ^2 based on the distributions of the $X_{i1}-X_{i2}$'s is the same as its marginal MLE, and is consistent.

One possible drawback of a conditional likelihood approach is that the conditioning variable U may contain information about ψ which is lost when it is held fixed. Hence, it may be appropriate to require that the distribution of U, the conditioning statistic does not contain any information about ψ in the presence of λ. In such cases, U is said to be ancillary for ψ in the presence of λ.

If for example, the distribution of U depends only on λ, then U does not contain any information about ψ. This is a very stringent requirement, and does not hold in general. In the Neyman-Scott problem, the $X_{i1}+X_{i2}$ are independent $\mathrm{N}(2\mu_i, 2\psi)$.

Example 12. Let $Y_1, Y_2, \cdots, Y_n$ be iid with common pdf

$$p(Y|\psi,\lambda) = \frac{\Gamma(\psi+Y)}{\Gamma(Y+1)\Gamma(\psi)}\lambda^Y(1-\lambda)^\psi,$$

where $\psi > 0$ and $0 < \lambda < 1$. For fixed ψ, $U = \sum_{j=1}^n Y_j$ is sufficient for λ so that the conditional distribution of $Y_1, \cdots, Y_n$ given U depends only on ψ. However, U has pdf

$$p(U|\psi,\lambda) = \frac{\Gamma(n\psi+U)}{\Gamma(U+1)\Gamma(n\psi)}\lambda^U(1-\lambda)^{n\psi},$$

which is not ancillary for ψ in the usual sense. Indeed, the Fisher information contained in the distribution of U depends on both ψ and λ.

The fact that U is not ancillary in the usual sense, has led to the notion of S-ancillarity (Barndorff-Nielsen, 1973, 1976). A statistic U is said to be S-ancillary for ψ in the presence of λ if the family of pdf's $\{f(U;\psi,\lambda); \lambda \in \Lambda\}$ remains the same for each ψ. More specifically, if U is S-ancillary, then for every ψ_0, ψ_1 and λ_0, there exists $\lambda_1 = h(\psi_0,\psi_1,\lambda_0) \in \Lambda$ such that $f(U;\psi_0,\lambda_0) = f(U;\psi_1,\lambda_1)$.

We consider a simple example.

Example 13 (Severini, 2000, p 282). $X_i \overset{ind}{\sim}$ Poisson$(\exp(\lambda+\psi Z_i))$, $i=1,\ldots,n$. Then writing $\phi = \sum_{i=1}^n \exp(\lambda+\psi Z_i)$, $U = \sum_1^n X_i$ is S-ancillary. Also, then the joint conditional distribution of the X_i given U is multinomial $(U; \frac{\exp(\psi Z_1)}{\sum_1^n \exp(\psi Z_i)}, \ldots, \frac{\exp(\psi Z_n)}{\sum_1^n \exp(\psi Z_i)})$.

Another approach to defining ancillarity in the presence of a nuisance parameter is based on the notion of partial information for ψ. This led to the development of

partial ancillary (P-ancillary) statistics (Bhapkar, 1989;1991). Denoting the information matrix for the minimal sufficient statistic (T, U) as $I^{T,U}(\psi, \lambda)$, one partitions the same as $\begin{pmatrix} I^{T,U}_{\psi\psi} & I^{T,U}_{\psi\lambda} \\ I^{T,U}_{\lambda\psi} & I^{T,U}_{\lambda\lambda} \end{pmatrix}$, where $I^{T,U}_{\psi\psi} = E[-\frac{\partial^2 f(T,U;\psi,\lambda)}{\partial\psi^2}]$, and the other elements of $I^{T,U}(\psi, \lambda)$ are similarly defined. The information matrix $I^U((\psi, \lambda)$ for U is similarly partitioned. Then the partial information for ψ based on (T, U) is

$$I^{T,U}_{\psi\psi.\lambda} = I^{T,U}_{\psi\psi} - I^{T,U}_{\psi\lambda}(I^{T,U}_{\lambda\lambda})^{-1} I^{T,U}_{\lambda\psi}, \tag{13}$$

while the partial information for ψ based on U alone is

$$I^{U}_{\psi\psi.\lambda} = I^{U}_{\psi\psi} - I^{U}_{\psi\lambda}(I^{U}_{\lambda\lambda})^{-1} I^{U}_{\lambda\psi}. \tag{14}$$

Once again we begin with the assumption that the joint density of the minimal sufficient statistic (T, U) can be factorized as in (12). This leads to the identities

$$I^{T,U}_{\psi\psi} = E[-\frac{\partial^2 f(T|U,\psi)}{\partial\psi^2}|U] + I^{U}_{\psi\psi}; \tag{15}$$

$$I^{T,U}_{\psi\lambda} = I^{U}_{\psi\lambda};\ I^{T,U}_{\lambda\lambda} = I^{U}_{\lambda\lambda}. \tag{16}$$

It follows from (13)-(16) that

$$I^{T,U}_{\psi\psi.\lambda} = E[-\frac{\partial^2 f(T|U,\psi)}{\partial\psi^2}|U] + I^{U}_{\psi\psi.\lambda}. \tag{17}$$

We say that U is partial ancillary (P-ancillary) for ψ if the partial information for ψ based on the joint distribution of T and U is the same as that in the conditional distribution of T given U, or equivalently $I^{U}_{\psi\psi.\lambda} = 0$.

Example 13 (Continued). In this example $U = \sum_{j=1}^{n} X_j \sim \text{Poisson}(\sum_{j=1}^{n} \exp(\lambda + \psi Z_j))$. Hence,

$$I^U(\psi, \lambda) = \begin{pmatrix} \frac{\{\sum Z_j \exp(\lambda+\psi Z_j)\}^2}{\sum \exp(\lambda+\psi Z_j)} & \sum Z_j \exp(\lambda + \psi Z_j) \\ \sum Z_j \exp(\lambda + \psi Z_j) & \sum \exp(\lambda + \psi Z_j) \end{pmatrix}$$

This leads immediately to $I^{U}_{\psi\psi.\lambda} = 0$, i.e. the S-ancillary U is also P-ancillary.

The following example shows that P-ancillarity does not necessarily imply S-ancillarity.

Example 14 (Severini, 2000, p 284). Let $X_i \overset{ind}{\sim} \text{N}(\lambda + \psi Z_i, 1)$, $i = 1, \ldots, n$, where

$\lambda > 0$. Then the minimal sufficient statistic $(\sum_{i=1}^n Z_i X_i, \sum_{i=1}^n X_i)$ is one-to-one with $(\sum_{i=1}^n Z_i(X_i - \bar{X}), \bar{X})$, and the conditional distribution of $\sum_{i=1}^n Z_i(X_i - \bar{X})$ given $\bar{X}$, which is the same as its marginal distribution due to independence, does not involve λ. Since $\bar{X} \sim N(\lambda + \psi\bar{Z}, n^{-1})$, $I^{\bar{X}}(\psi, \lambda) = n \begin{pmatrix} \bar{Z}^2 & \bar{Z} \\ \bar{Z} & 1 \end{pmatrix}$ so that $I^{\bar{X}}_{\psi\psi.\lambda} = 0$.

Hence, $\bar{X}$ is P-ancillary. However, $\bar{X}$ is not S-ancillary. To see this, we may note that if $\psi = -1$, $\bar{X} \sim \mathrm{N}(\lambda - \bar{Z}, 1/n)$ so that the mean of $\bar{X}$ is any number greater than $-\bar{Z}$ which may be negative. Thus $\bar{X}$ cannot be S-ancillary.

The next example shows that even though (12) holds, U may not be either P- or S-ancillary. This example is a simplified version of one given in Severini (2000, p 280).

Example 15. Let $X_1, \ldots, X_n$ be iid with common gamma pdf

$$f(X : \psi, \lambda) = \exp(-\lambda X) X^{\psi-1} \lambda^{\psi} / \Gamma(\psi),$$

where $X > 0$, $\psi > 0$, and $\lambda > 0$. Then the minimal sufficient statistic for (ψ, λ) is $(T = \prod_{i=1}^n X_i, U = \sum_{i=1}^n X_i)$. Also, for fixed ψ, U is sufficient for λ. Hence, the conditional distribution of T given U does not depend on λ. However, U is not either P- or S-ancillary.

To see this, first note that U has pdf $f(U; \psi, \lambda) = \exp(-\lambda U) U^{n\psi-1} \lambda^{n\psi} / \Gamma(n\psi)$. Hence, the information matrix based on U is given by $I^U(\psi, \lambda) = \begin{pmatrix} \frac{d^2 \log \Gamma(n\psi)}{d\psi^2} & -n/\lambda \\ -n/\lambda & n\psi/\lambda^2 \end{pmatrix}$. Clearly, $I^U_{\psi\psi.\lambda} \neq 0$ so that U is not P-ancillary.

In order to show that U is not S-ancillary, one may note that if for some $(\psi_0, \psi_1 (\neq \psi_0), \lambda_0)$, there exists λ_1 such that the pdf of U under (ψ_0, λ_0) is the same as that under (ψ_1, λ_1), then

$$n\psi_0/\lambda_0 = E_{\psi_0,\lambda_0}(U) = E_{\psi_1,\lambda_1}(U) = n\psi_1/\lambda_1;$$

$$n\psi_0/\lambda_0^2 = V_{\psi_0,\lambda_0}(U) = V_{\psi_1,\lambda_1}(U) = n\psi_1/\lambda_1^2.$$

The above imply that $\psi_0 = \psi_1$ which is a contradiction. Hence, U is not S-ancillary.

7.2 APPROXIMATE ANCILLARITY

The definition of ordinary ancillarity in the presence of nuisance parameters is not at all straightforward, as we have seen in the previous subsection. It is equally difficult to formalize the notion of approximate ancillarity in the nuisance parameter setting.

However, it is possible to extend the asymptotic approximations outlined in Section 6 to the nuisance parameter setting, using an approximate version of (11).

Barndorff-Nielsen (1986) showed that starting from the p^* approximation (10) for the maximum likelihood estimate of the full parameter θ, we can write, to $O(n^{-3/2})$

$$p(\hat{\theta} \mid U; \theta) = p(r^*_\psi \mid U) p(\hat{\lambda}_\psi \mid r^*_\psi, U; \lambda)$$

where r^*_ψ is a modification of the likelihood root $r_\psi = \text{sign}(\hat{\psi} - \psi)[2\{l_p(\hat{\psi}) - l_p(\psi)\}]^{1/2}$, based on the profile log likelihood. This leads directly to approximate inference for ψ based on $\Phi(r^*_\psi)$, which can be shown to be standard normal with relative error $O(n^{-3/2})$. The construction of r^*_ψ is difficult, as it requires various sample space derivatives for fixed U: Fraser and Reid (1995) provide an alternative version that does not use explicit expressions for the approximate ancillary statistic; see also Severini (2000, Ch. 7.5.3).

Approximate ancillary statistics can also be constructed using covariances, as suggested in Skovgaard (1996) and Severini (1999); see Severini (2000, Ch. 7.5.4,5).

8 BAYESIAN ANCILLARITY

As noted in the previous section, neither S-ancillarity nor P-ancillarity of a statistic U implies that the distribution of U does not depend on ψ, and depends only on λ. Also, in Example 15, U is neither S-ancillary nor P-ancillary. Indeed, it is not even appropriate to say that U does not contain any information about ψ in the presence of λ without specifically defining what "information" really means. To overcome this problem, Severini (1995) brought in the notion of Bayesian ancillarity. We shall observe that as a consequence of his definition, by introducing a suitable prior, the marginal distribution of U will indeed not depend on ψ. The details are described below.

Severini defines a statistic U to be Bayes ancillary if with respect to *some* prior distribution, the posterior distribution of ψ based on the conditional distribution T given U is the same as the posterior distribution of ψ based on the joint distribution of (T, U). We first consider the case when there is no nuisance parameter λ. Using $p(\cdot|\cdot)$ as a generic symbol for a conditional pdf, and $p(\cdot)$ as a generic symbol for a marginal pdf, U is Bayes ancillary if and only if

$$\frac{p(T, U|\psi)p(\psi)}{\int p(T, U|\psi)p(\psi)d\psi} = \frac{p(T|U, \psi)p(\psi)}{\int p(T|U, \psi)p(\psi)d\psi}.$$

Writing $p(T,U|\psi) = p(T|U,\psi)p(U|\psi)$, the above simplifies to

$$p(U|\psi) = \frac{\int p(T,U|\psi)p(\psi)d\psi}{\int p(T|U,\psi)p(\psi)d\psi} = p(U),$$

that is the marginal of U does not depend on ψ. An alternative way of saying this is that here U does not contain any information about ψ.

In the presence of a nuisance parameter λ, we begin with the minimal sufficient (T,U) for (ψ,λ), and assume as before that $p(T,U|\psi,\lambda) = p(T|U,\psi)p(U|\psi,\lambda)$. Once again, invoking the definition of Bayesian ancillarity, U is Bayesian ancillary if and only if

$$\frac{\int p(T,U|\psi,\lambda)p(\lambda|\psi)p(\psi)d\lambda}{\int\int p(T,U|\psi,\lambda)p(\lambda|\psi)p(\psi)d\lambda d\psi} = \frac{p(T|U,\psi)p(\psi)}{\int p(T|U,\psi)p(\psi)d\psi}.$$

Since $p(T,U|\psi,\lambda) = p(T|U,\psi)p(U|\psi,\lambda)$, the above simplifies to

$$\int p(U|\psi,\lambda)p(\lambda|\psi)d\lambda = \frac{\int\int p(T|U,\psi)p(U|\psi,\lambda)p(\lambda|\psi)p(\psi)d\lambda d\psi}{\int p(T|U,\psi)p(\psi)d\psi},$$

i.e. $p(U|\psi) = \int p(T|U,\psi)p(U|\psi)p(\psi)d\psi / \int p(T|U,\psi)p(\psi)d\psi$. Once again, the marginal pdf of U does not involve ψ, and U is ancillary in the usual sense.

It may be noted in Example 12 that with the prior $p(\lambda|\psi) = \lambda^{-1}(1-\lambda)^{-1}$ which is improper and does not depend on ψ, it follows that $p(U|\psi) = \Gamma(U)/\Gamma(U+1) = U^{-1}$ which does not depend on ψ.

9 FURTHER REMARKS

9.1 Ancillarity and Optimal Estimating Equations

Godambe (1976, 1980) also considered the concepts of sufficiency and ancillarity in the presence of nuisance parameters, and tied these ideas to the theory of optimal estimating functions.

Godambe's formulation is as follows: Let $X_1, \cdots, X_n$ be independent observations with densities $f(X_i|\psi,\lambda)$, where once again ψ is the parameter of interest, and λ is the nuisance parameter. Let $g(X_i,\psi)$, a function of X_i and the parameter of interest satisfy $E[g(X_i,\psi)|\psi,\lambda] = 0$. Then $g(X_i,\psi)$ is called an unbiased estimating function. Let $g(\boldsymbol{X},\psi) = \sum_{i=1}^n g(X_i,\psi)$, where $\boldsymbol{X} = (X_1,\cdots,X_n)^T$.

Godambe (1960) defined the optimal unbiased estimating function as the minimizer of $E\left[g^2(\boldsymbol{X},\psi)/\{E(\partial g(\boldsymbol{X},\psi)/\partial\psi)\}^2\right]$. He showed that without any nuisance parameter, the usual score function is the optimal unbiased estimating function. In the presence of a nuisance parameter, Godambe (1976) showed that if the joint density $f(\boldsymbol{X}|\psi,\lambda)$ factorizes as

$$f(\boldsymbol{X}|\psi,\lambda) = f(\boldsymbol{X}|U,\psi)f(U|\psi,\lambda),$$

where U is complete sufficient for the nuisance parameter λ for fixed ψ, then the conditional score function $\partial \log f(\boldsymbol{X}|U,\psi)/\partial\psi$ is the optimal unbiased estimating function. He also showed that the information about ψ contained in the conditional distribution of X given U is the same as that contained in its unconditional distribution.

The following two examples illustrate this.

Example 16. Let $X_1,\ldots,X_n$ be iid gamma with common pdf

$$f(X|\psi,\lambda) = \left[\lambda^{\psi}\Gamma(\psi)\right]^{-n}\prod_{i=1}^{n}\{X_i^{\psi-1}\exp(-X_i/\lambda)\}$$

Let $U=\sum_1^n X_i$. Note that the conditional pdf of $X_1,\ldots,X_{n-1}$ given U is

$$f(X_1,\ldots,X_{n-1}|U,\psi) = \frac{\Gamma(n\psi)}{\Gamma^n(\psi)}\frac{(\prod_1^n X_i)^{\psi-1}}{U^{n\psi-1}}$$

which does not depend on λ. Also marginally U has the pdf

$$f(U|\psi,\lambda) = \exp(-U/\lambda)\frac{U^{n\psi-1}}{\lambda^{n\psi}\Gamma(n\psi)}$$

so that for fixed ψ, the family of distributions of U is complete.

Example 17. Let $X_1,\ldots,X_n$ be iid negative binomial with common pf

$$f(X|\psi,\lambda) = \binom{\psi+X-1}{X}\lambda^X(1-\lambda)^{\psi}$$

Again take $U = \sum_1^n X_i$. Then the conditional pdf of $X_1, \ldots, X_n$ given U is

$$f(X_1, \ldots, X_{n-1}|U, \psi) = \frac{\prod_1^n \binom{\psi + X_i - 1}{X_i}}{\binom{n\psi + U - 1}{T}}$$

which does not depend on λ. Further, the marginal pf of U is

$$f(U|\psi, \lambda) = \binom{n\psi + U - 1}{T} \lambda_0^U (1 - \lambda_0)^{n\psi}$$

So, for fixed ψ, the family of distributions of U is complete.

One of the limitations of Godambe's (1976) result is that he had to assume the existence of a complete sufficient statistic for λ which did not depend on ψ. While this is available for the regular exponential family of distributions, this need not be true in general. Lindsay (1982) showed that if $U = U(\psi)$, then the conditional score function for ψ depends on λ as well, and hence, is only locally optimal at the true λ.

9.2 Brown's Ancillarity Paradox

Brown (1990) introduced an ancillarity paradox (essentially an admissibility paradox) in the context of multiple linear regression. His main theme was to show via (in)admissibility results that procedures which are admissible conditional on some ancillarity statistics may unconditionally fail to be so.

We begin with the following simple example of Brown.

Example 18. Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ known positive definite. Let $\boldsymbol{U} \in R^p$ with $||\boldsymbol{U}|| > 0$. Let $\theta = \boldsymbol{U}^T\boldsymbol{\mu}$. The usual estimator of θ is $\boldsymbol{U}^T\boldsymbol{X}$. Under squared error loss, Cohen (1966) has shown that $\boldsymbol{U}^T\boldsymbol{X}$ is an admissible estimator of $\boldsymbol{U}^T\boldsymbol{\mu}$ for fixed $\boldsymbol{U}$. However, if $\boldsymbol{U}$ is random. writing $\boldsymbol{\Omega} = E(\boldsymbol{U}\boldsymbol{U}^T)$, and assuming it to be positive definite, Brown showed that $\boldsymbol{U}^T\boldsymbol{X}$ is dominated by $\boldsymbol{U}^T\boldsymbol{\delta}(\boldsymbol{X})$, under squared error loss, where

$$\boldsymbol{\delta}(\boldsymbol{X}) = \boldsymbol{X} - \frac{\rho}{\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$$

$0 < \rho < 2(p-2)$, $p \geq 3$.

Brown established a similar phenomenon in a multiple regression problem.

Example 19. Let $\boldsymbol{X} \sim \mathrm{N}_p(\alpha 1_p + \boldsymbol{Z}\boldsymbol{\beta}, \sigma^2 I_p)$, where $\boldsymbol{Z}$ $(p \times p)$ is the design matrix and $\boldsymbol{\beta}$ $(k \times 1)$ regression vector, 1_p is the p-component vector of 1's, and I_p is the identity matrix of order p. We assume that $p > k+1$, and $\boldsymbol{Z}$ is a full rank matrix. The objective is to estimate α under the squared error loss.

Let $\bar{X} = p^{-1}1_p^T\boldsymbol{X}$, $\bar{\boldsymbol{Z}} = p^{-1}1_p^T\boldsymbol{Z}$ and $\boldsymbol{S} = (\boldsymbol{Z} - 1_p\bar{\boldsymbol{Z}}^T)^T(\boldsymbol{Z} - 1_p\bar{\boldsymbol{Z}}^T)$. Here $\bar{X}$ is a scalar, $\bar{\boldsymbol{Z}}^T$ is a row vector of dimension k and $\boldsymbol{S}$ is a $k \times k$ matrix, positive definite with probability 1. The usual estimator $\hat{\alpha} = \bar{X} - \bar{\boldsymbol{Z}}^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimator $\boldsymbol{\beta}$ is admissible under squared error loss. However, if it is assumed that k-dimensional components of $\boldsymbol{Z}$ are iid $\mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_k)$, then $\hat{\alpha}$ ceases to be an admissible estimator of α under squared error loss.

What Brown's examples demonstrate is that conditional inference could potentially be in conflict with unconditional inference. However, it appears that there are no fundamental or conceptual difficulties associated with this conclusion. This was brought out by several discussants of his paper; see also the optimality discussion preceding Example 3 in Section 2. Another interesting example of ancillarity paradox in the context of finite population sampling appears in Godambe (1982).

10 SUMMARY AND CONCLUSION

Ancillary statistics is a fascinating vast topic. Over several decades, this area has witnessed phenomenal amount of research, both exact and asymptotic. It is impossible to cover each and every facet of the subject in a single review. Fraser (2004), in his recent article, has discussed many aspects of conditioning and ancillarity, and has argued strongly in favor of conditional inference. We share this view. But, in this article, we have tried to bring out both the uses and difficulties of ancillary statistics, taking more or less a neutral standpoint. Also, as the title says, this is only a selective review, and we are aware that many important contributions are omitted. We offer my sincerest apologies to these authors. Finally, we strongly feel that research on ancillary statistics has not reached a saturation level, and anticipate new and surprising results in this general area.

11 ACKNOWLEDGEMENTS

Ghosh's research was supported in part by NSF Grant SES-0317589. Reid and Fraser's research was supported in part by Discovery Grants from NSERC. The work began when

Ghosh delivered the Basu Memorial Lecture at 2004 JSM, where Reid was the discussant.

REFERENCES

Barnard, G.A. And Sprott, D.A. (1971). A note on Basu's examples of anamalous ancillary statistics. In *Foundations of Statistical Inference*. Eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto, pp 163-170.

Barndorff-Nielsen, O.E. (1973). On M-ancillarity. *Biometrika*, **60**, 447-455.

Barndorff-Nielsen, O.E. (1976). Noninformation. *Biometrika*, **63**, 567-571.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

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

Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika*, **78**, 557-563.

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

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377-380.

Basu, D. (1959). The family of ancillary statistics. *Sankhya*, **21**, 247-256.

Basu, D. (1964). Recovery of ancillary information. *Sankhya*, **26**, 3-16.

Basu, D. (1992). Learning statistics from counter examples: ancillary statistics. In *Bayesian Analysis in Statistics and Econometrics*. Eds. P.K.Goel and N.S. Iyengar. Springer-Verlag, New York, pp 217-223.

Bhapkar, V.P. (1989). Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Statist. Plann. Inf*, **21**, 139-160.

Bhapkar, V.P. (1991). Loss of information in the presence of nuisance parameters and partial sufficiency. *J. Statist. Plann. Inf*, **28**, 185-203.

Brown, L.D. (1990). An ancillarity paradox which appears in multiple regression models. *Annals of Statistics*, **18**, 471-538.

Buehler, R.J. (1982). Some ancillary statistics and their properties. *J. Amer. Statist.Assoc.*, **77**, 581-594.

Cohen, A. (1966). All admissible linear estimators of the mean vector. *Ann. Math. Statist.*, **37**, 458-463.

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

Cox, D.R. (1971). The choice between alternative ancillary statistics. *J. Roy. Statist. Soc., B*, **33**, 251-252.

Cox, D.R. (1980). Local ancillarity. *Biometrika*, **67**, 273-278.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman & Hall, New York.

Datta, G.S., Ghosh, M., Smith, D.D. and Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scand. J. Stat.*, **29**, 139-152.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.*, **25**, 631-650.

Davison, A.C. (1988). Approximate conditional inference in generalized linear models. J. Roy. Statist. Soc., B, **50**, 445-461.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**, 457-487.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700-725.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc., A*, **144**, 285-307.

Fisher, R.A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc., B*, **98**, 39-54.

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

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

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

Fraser, D.A.S. (2004). Ancillaries and conditional inference. (with discussion). *Stat. Science.*, **19**, 332-369.

Fraser, D.A.S. and McDunnough, P. (1980). Some remarks on conditional and unconditional inference in location-scale models. *Statistische Hefte*, **21**, 224-231.

Fraser, D.A.S. and Reid, N. (1993). Third order assymptotic models: likelihood functions leading to accurate approximation of distribution functions. *Statistica Sinica*,

3, 67-82.

Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitias Mathematica*, **47**, 33-53.

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

Ghosh, J.K. (1994). *Higher Order Asymptotics*. Institute of Mathematical Statistics, Hayward, California.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208-1212.

Godambe, V.P. (1976). Conditional likelihood and unconditional optimal estimating equations. *Biometrika*, **63**, 277-284.

Godambe, V.P. (1980). On sufficiency and ancillarity in the presence of nuisance parameters. *Biometrika*, **67**, 155-162.

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

Hill, J.R. (1990) A general framework for model based statistics. *Biometrika*, **77**, 115-126.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc., B*, **34**, 1-41.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika*, **69**, 503-512.

Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Prob.*, **12**, 475-490.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, **16**, 1-32.

Pena, E.A., Rohatgi, V.K., and Szekely, G.J. (1992). On the non-existence of ancillary statistics. *Statistics and Probability Letters*, **15**, 357-360.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statist. Sci.*, **3**, 213-238.

Reid, N. (1995). The roles of conditioning in inference. *Stat. Sci.*, **10**, 138-157.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, **31**, 1695-1731.

Severini, T.A. (1995). Information and Conditional Inference. *J. Amer. Statist. Assoc.*, **90**, 1341-1346.

Severini, T.A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika*, **86**, 235-247.

Severini, T.A. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

Skovgaard, I.M. (1985). A second order investigation of asymptotic ancillarity. *Ann. Statist.*, **13**, 534-551.

Skovgaard, I.M. (1990). On the density of minimum contrast estimators. *Ann. Statist.*, **18**, 779-789.

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

Wang, S. (1993). Saddlepoint approximations in conditional inference. *J. Applied Probability*, **30**, 397-404.

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