Statistica Sinica **20** (2010), 1309-1332

ANCILLARY STATISTICS: A REVIEW

M. Ghosh[1], N. Reid[2] and D. A. S. Fraser[2]

[1] *University of Florida and* [2] *University of Toronto*

Abstract: In a parametric statistical model, a function of the data is said to be ancillary if its distribution does not depend on the parameters in the model. The concept of ancillary statistics is one of R. A. Fisher's fundamental contributions to statistical inference. Fisher motivated the principle of conditioning on ancillary statistics by an argument based on relevant subsets, and by a closely related argument on recovery of information. Conditioning can also be used to reduce the dimension of the data to that of the parameter of interest, and conditioning on ancillary statistics ensures that no information about the parameter is lost in this reduction.

The present review article is an attempt to illustrate various aspects of the use of ancillarity in statistical inference. Both exact and asymptotic theory are considered. Without any claim of completeness, we have made a modest attempt to crystalize many of the basic ideas in the literature.

Key words and phrases: Ancillarity paradox, approximate ancillary, estimating functions, hierarchical Bayes, local ancillarity, location, location-scale, multiple ancillaries, nuisance parameters, P-ancillarity, p-values, S-ancillarity, saddlepoint approximation.

1. Introduction

Ancillary statistics were one of R.A. Fisher's many pioneering contributions to statistical inference, introduced in Fisher (1925) and further discussed in Fisher (1934, 1935). Fisher did not provide a formal definition of ancillarity, but following later authors such as Basu (1964), the usual definition is that a statistic is ancillary if its distribution does not depend on the parameters in the assumed model. Some authors (e.g., Lehmann and Scholz (1992)) demand in addition that ancillary statistics should be functions of minimal sufficient statistics, as a way of narrowing the class of ancillary statistics.

The development of Fisher's ideas on ancillarity between 1925 and 1935 is reviewed in Stigler (2001), and the various, somewhat vague, aspects of ancillarity used by Fisher are still quite useful. First, although an ancillary statistic by itself provides no information about the parameter, it may provide a means of recovering information lost by reducing the data to the maximum likelihood estimator. To be specific suppose X has pdf $f_\theta(X)$, and the MLE $T \equiv T(X)$ of

θ has pdf $g_\theta(T)$. We write $I(\theta) = E_\theta\{-\partial^2 \log f_\theta(X)/\partial\theta^2\}$, the Fisher information contained in X and $J(\theta) = E_\theta\{-\partial^2 \log g_\theta(T)/\partial\theta^2\}$, the Fisher information contained in T, implicitly assuming any needed regularity conditions to justify these definitions. It is easy to show that $I(\theta) \geq J(\theta)$ with equality if and only if T is sufficient.

Thus, when the MLE T itself is not sufficient, there is loss of Fisher information. This information can be recovered by conditioning on an ancillary statistic U, in the sense that $I(\theta) = E_\theta\{J(\theta|U)\}$, where $J(\theta|U)$ is the Fisher information contained in $h_\theta(T|U)$, the conditional distribution of T given U:

$$J(\theta|U) = E_\theta\Big[- \Big\{ \frac{\partial^2 \log h_\theta(T|U)}{\partial\theta^2} \Big\} \mid U \Big].$$

It is assumed in this definition that the pair (T, U) is sufficient, and then U is referred to as an ancillary complement to T. According to Fisher, the appropriate measure of information in T is $J(\theta|U)$ and not $J(\theta)$.

Example 1. Let (X_i, Y_i) $(i = 1, \ldots, n)$ be i.i.d. with common pdf

$$f_\theta(x, y) = \exp(-\theta x - \frac{y}{\theta}) 1_{[x>0, y>0]};\ \theta > 0.$$

This example is usually referred to as Fisher's gamma hyperbola (Efron and Hinkley (1978); Barndorff-Nielsen and Cox (1994); Reid (2003)). Defining $T = (\sum_{i=1}^n Y_i / \sum_{i=1}^n X_i)^{1/2}$, $U = (\sum_{i=1}^n X_i)^{1/2} (\sum_{i=1}^n Y_i)^{1/2}$, 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, $I(\theta) = 2n/\theta^2$ and $J(\theta) = (2n/\theta^2)\{2n/(2n+1)\}$, so that the loss of information is $I(\theta) - J(\theta) = (2n)/\{(2n+1)\theta^2\}$. However, according to Fisher, one should not report the information in T as $J(\theta)$, but instead should report $J(\theta|U) = \{(2n)/\theta^2\}\{K_1(2U)/K_0(2U)\}$, where K_0 and K_1 are Bessel functions, and their ratio recovers on average the loss of information.

In later work on the location model, Fisher (1934) showed that the configuration statistic, $(X_1 - T, \ldots, X_n - T)$, where T is an estimator of the location parameter, is ancillary, and that conditional inference for the location parameter is simply obtained from the likelihood function (Cox and Hinkley (1974, Chap. 4)). The configuration statistic defines in this case a 'relevant subset' of the sample space, and this relevant subset argument was developed in further detail in Fraser (1968, 1979). Cox's (1958) paper set out the details of the relevant subset argument most concretely.

Another role for ancillary statistics is to reduce the dimension of the sample space to that of the parameter space, thus providing a distribution that can provide direct inference statements for the parameter. While this is closely related

to the relevant subsets and information recovery aspects, it is subtly different. The dimension reduction argument focusses more directly on the conditional distribution that is left, rather than the marginal distribution that is ignored. This dimension reduction aspect has proved to be extremely useful for asymptotic theory based on the likelihood function, which interestingly was also anticipated by Fisher (1925), who argued that the higher order derivatives of the log-likelihood function could often serve as what we would now call approximately ancillary statistics. Fraser's (2004) survey of conditional inference emphasizes the dimension reduction aspect of ancillarity.

Fisher (1935) also invoked ancillarity for the elimination of nuisance parameters in the context of the 2×2 table, although making this notion of ancillarity precise is even more difficult than the ordinary notion of ancillarity for the full parameter.

Kalbfleisch (1975, 1982) classified ancillarity as being "experimental" or "mathematical". According to his criterion, "the former are ancillary by virtue of the experimental design", for example the random sample size regardless of the chosen parametric model. He contrasts those with "mathematical ancillaries" which often depend on the structure of the assumed parametric model as in Example 1. Lehmann (1981) and Lehmann and Scholz (1992) showed the connection of ancillarity with other statistical concepts including sufficiency, group families, completeness, and mixture experiments, in addition to information and conditionality as mentioned earlier.

In this paper we review these various aspects of ancillarity, largely in the context of key examples from the literature. The goal is to try to clarify the various aspects of ancillarity, and to highlight the importance of conditioning in both exact or approximate inference. In many standard treatments of statistical inference, the concept of ancillarity is presented as problematic, usually by means of some examples. However, it seems essential in non-Bayesian inference to condition on some features of the data, and we hope the examples also clarify why this is the case.

In Section 2 we discuss the role of conditionality in two classic examples due to Welch (1939) and Cox (1958) and a relatively new example due to Hill (1990). These examples illustrate respectively the importance of conditioning. In Section 3 we give three puzzling examples due to Basu (1964), and discuss suggestions of Barnard and Sprott (1971) and Cox (1971) toward their resolution. In Section 4, we discuss the role of ancillary statistics in deriving some higher order asymptotic results related to maximum likelihood estimation and p-values. In particular, we introduce the p^*-formula of Barndorff-Nielsen (1983), and indicate the role of approximately ancillary statistics in this formula.

In Section 5, we consider the issue of conditioning with the objective of elimination of nuisance parameters, and discuss the role of ancillary statistics in

this context. Some extended definitions of ancillarity are given: in particular, the notion of S-ancillarity (Sandved (1965); Sverdrup (1966)), P-ancillarity (Bhapkar (1989, 1991)), and the connection to the theory of estimating functions (Godambe (1976)). The related concept of Bayesian ancillarity (Severini (1995)) is discussed, and we give a brief description of Brown's (1990) ancillarity paradox.

Buehler (1982) proposed properties of ancillary statistics as a means of more formally assessing the information recovery, and relevant subsets aspects of ancillarity, and discussed this through a large collection of examples. The emphasis in our paper, on the other hand, is an examination of the role of ancillary statistics statistics in exact and approximate likelihood inference. In some examples the argument is clearer if the data is first reduced to the minimal sufficient statistic, and the ancillary taken to be a component of this statistic. This approach is emphasized in Cox and Hinkley (1974), for example. However in some examples, such as the location-scale model, it is easier to work with the full data vector, and then reduce the conditional model by sufficiency if this is available. This is the approach emphasized in Fraser (2004).

2. The Case for Ancillarity

If there is an emerging consensus about the role of ancillarity in statistical inference, it stems from the notion that conditioning on ancillary statistics makes the resulting inference more relevant to the observed data. In this sense ancillary conditioning is a sort of 'halfway-house' between Bayesian and frequentist inference. The most compelling arguments for ancillary conditioning in the literature come from consideration of simple but highly illustrative examples, and we review some of them in this section. These examples illustrate the three different, although related, arguments in favor of ancillary conditioning discussed in the Introduction: (i) ancillary statistics provide relevant subsets; (ii) ancillary statistics give the right measure of variation; (iii) ancillary statistics provide a means of dimension reduction.

Example 2. One of the most compelling, if somewhat artificial, example is Cox's (1958) example of two measuring instruments, which has been discussed by many authors. A particularly thorough discussion is given in Berger and Wolpert (1984, Chap. 2); Cox and Hinkley (1974, Chap. 4) give detailed calculations in the context of hypothesis testing,; Fraser (2004) gives a discussion in terms of confidence intervals that makes the case for conditioning even more starkly. The model is that of observing a random pair (X, U), where X follows a normal distribution with mean μ and variance σ_U^2, and U follows a Bernoulli distribution with $P(U = 1) = 0.5 = P(U = 0)$. The importance of conditioning on the observed value of U is emphasized by assuming that σ_0^2 is much smaller than σ_1^2,

i.e., the measurement was either taken with a very precise or a very imprecise measuring instrument, but in either case we know which measuring instrument was used. Although it is possible to construct a more powerful test of $\mu = 0$ and a confidence interval for μ with shorter expected length by not conditioning on the observed value of U, it is clear that the resulting unconditional inference about μ is irrelevant for any particular measurement or set of measurements from the more accurate instrument. The power of the test and the expected length of the confidence interval need to be calculated under the joint distribution of (X, U) for this argument to be correct, and at least in this example it seems clear that the unconditional evaluation of power and expected length is inappropriate. Unconditional confidence intervals are typically longer in more precise contexts and shorter in less precise contexts than the conditional intervals (Fraser and McDunnogh (1980)).

This is an example of a 'relevant subsets' argument; the relevant subset of the sample space is that determined by the possible values of X and the observed value of U. It can easily be generalized to a model where the probability that $U = 1$ is unknown, but unrelated to μ; to a model with a random sample size; to a regression setting, where covariates are selected by a random mechanism; and so on as long as the parameters determining the ancillary statistic (sample size, or covariates) are completely uninformative about the parameters of interest.

Next we consider an example originally given in Welch (1939), and subsequently revisited by many authors, such as Barndorff-Nielsen and Cox (1994), and most recently Fraser (2004).

Example 3. Suppose X_1 and X_2 are i.i.d. uniform $(\theta - 1, \theta + 1)$, θ real. Let $T = (Y_1 + Y_2)/2$ and $U = (Y_2 - Y_1)/2$, where $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$. The dimension of the minimal sufficient statistic (T, U) exceeds that of the parameter. The MLE of θ is any random variable in the interval $(Y_2 - 1, Y_1 + 1)$; in particular, T is a MLE of θ. U is ancillary, and the conditional pdf of T given U is

$$f_\theta(T|U) = \{2(1 - U)\}^{-1} 1_{[\theta - 1 + U < T < \theta + 1 - U]}(T). \tag{2.1}$$

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, θ is very precisely determined.

On the other hand, the marginal pdf of T is

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

This may lead to absurd inference for θ when U is close to 1. Then θ is essentially known exactly, but an unconditional confidence region for θ may lead to values which may be quite different from this exact value.

Welch (1939) argued against conditional inference by producing two different $100(1-\alpha)\%$ confidence intervals for θ, one which is based on a more powerful test, and one which has shorter expected length. Explicit expressions for Welch's intervals are given in (2.4) and (2.5) of Fraser (2004). Fraser shows, for extreme values of U, that Welch's intervals may be either the full parameter space or the empty set, but the interval based on (2.1) will not have this extreme behavior. In general the requirements of power or average length are at odds with the requirement of conditioning, although as a reviewer has pointed out, conditional tests may not be less powerful than unconditional tests, in settings where no uniformly most powerful unconditional test exists (Barnard (1982); Severini (1995)).

The next example provides an empirical Bayes (EB) scenario where conditioning with respect to an ancillary statistic can produce quite a meaningful answer.

Example 4 (Hill (1990))**.** Let $X_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, 1)$ and $\theta_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, A)\,(i = 1, \ldots, k)$. 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 $(X_1, \ldots, X_k)$. Since marginally $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, B^{-1})$, $\bar{X} = k^{-1}\sum_{i=1}^{k} X_i$ is a complete sufficient statistic 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 ancillarity argument can overcome this, we may note that marginally $U = X_1 - \bar{X}$ is ancillary and $U \sim \mathrm{N}(0, (k-1)/(kB))$. 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

1. $X_i|\theta_1, \ldots, \theta_k, \mu \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, 1)$,
2. $\theta_1, \ldots, \theta_k|\mu \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, A)(A > 0)$, and
3. $\mu \sim \text{uniform}(-\infty, \infty)$,

it turns out that $\theta_1|X_1, \ldots, X_n, \mu \sim \mathrm{N}((1-B)X_1 + B\mu, 1-B)$ and $\mu|X_1, \ldots, X_n \sim \mathrm{N}(\bar{X}, (kB)^{-1})$. Together, they imply $\theta_1|X_1, \ldots, X_n \sim \mathrm{N}\{(1-B)X_1 + B\bar{X},$

$1 - B + Bk^{-1}\}$. 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 $Bk^{-1} = V(B\mu|X_1, \ldots, X_n)$, it may be noted that in this case ancillarity accounts for the uncertainty due to estimation of μ as much as does 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.

Examples 3 and 4 illustrate the argument that the conditional variance given the ancillary statistic of the estimator of the parameter of interest is a more appropriate measure of variability than is the unconditional variance. This is similar to the relevant subsets argument, and in simple cases is nearly as compelling, but does not seem to be as readily accepted.

The third role of ancillary conditioning, reduction of dimension, is most clearly useful in the higher order approximations discussed in Section 4, but it is already apparent in Example 2. There the minimal sufficient statistic is of dimension 2, and the parameter of interest is of dimension 1: conditioning on the ancillary statistic provides a 1-dimensional distribution for inference about θ. Example 2 is a location model, and this reduction in dimension is available in a general location model by conditioning on the residuals $U = (X_1 - \hat{\theta}, \cdots, X_n - \hat{\theta})$ where $\hat{\theta}$ is the maximum likelihood estimator, although any location-equivariant estimator will do. Fisher called this ancillary statistic a configuration statistic, and argued that it also defines a relevant subset of the sample space for inference; this line of argument was extended and generalized in Fraser's (1968) structural inference. Efron and Hinkley (1978) argued for conditioning on the configuration statistic to get a more appropriate assessment of the variance of the maximum likelihood estimator, and showed how this could be extended to approximate ancillarity. These asymptotic arguments are summarized in Section 4, but first we turn to several classical examples that seem to raise red flags around ancillary conditioning.

3. Ancillary Puzzles

Often there are problems associated with ancillary statistics. First, situations may arise when an ancillary U may not exist. Indeed, Pena, Rohatgi and Szekely (1992) have demonstrated this phenomenon for general discrete models. In Ghosh (1988, p.2), Basu considered the example where $X_1, \ldots, X_n$ ($n \geq 2$) are i.i.d. uniform (θ, θ^2), $\theta > 1$. The MLE of θ is $T = \{\max(X_1, \ldots, X_n)\}^{1/2}$, while the minimal sufficient statistic is $\{\min(X_1, \ldots, X_n),\ \max(X_1, \ldots, X_n)\}$. Basu pointed out that, in this example, there does not exist any ancillary complement U of T. It may be noted that in this example, the dimension of the minimal

sufficient statistic exceeds that of the parameter. 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 is not at all clear which one to condition on. Moreover, 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 resolved by conditioning on (U_1, U_2) jointly. Basu illustrated this with an example. Stigler (2001) pointed out that earlier Edgeworth (1893) and Pearson (1896) considered this example also, though from a somewhat different perspective.

Example 5. Let

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

$i = 1, \ldots, n$, 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_n^2 distribution, but jointly (U_1, U_2) is not ancillary as $\mathrm{corr}(U_1, U_2) = \rho^2$ depends on ρ. Thus, while W/U_1 and W/U_2 are both unbiased estimators of ρ (unconditionally or conditionally), $V(W/U_1|U_1) = (1 - \rho^2)/U_1$ and $V(W/U_2|U_2) = (1 - \rho^2)/U_2$. 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 *entirely* data-dependent, which is counter to the usual frequentist paradigm.

Cox (1971) suggested a way to deal with multiple ancillaries in this problem. By the identity $I(\theta) = E\{J(\theta|U)\}$, 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 $J(\theta|U)$ is most variable, i.e., $V_\theta\{J(\theta|U)\}$ is maximum. Unfortunately, in most instances $V_\theta\{J(\theta|U)\}$ is a function of the unknown θ, and there may not be a unique U which maximizes $V_\theta\{J(\theta|U)\}$ for all θ. Moreover, in Example 4, since $V_\theta\{J(\theta|U_1)\} = V_\theta\{J(\theta|U_2)\}$, the Cox method will fail to distinguish between U_1 and U_2.

Also, there does not exist any ancillary function of the minimal sufficient statistic $(U_1 + U_2, W)$; in particular, $U_1 + U_2$ is not ancillary. However, as disussed in Cox and Hinkley (1974), $U_1 + U_2$ is approximately ancillary for small ρ. This can be partially justified by noting that $E(U_1 + U_2) = 2n$ and $V(U_1 + U_2) = 4n(1 + \rho^2)$. Approximate ancillarity in this example is discussed in Section 4.

The next example due to Basu (1964) is one of multiple ancillaries where there is no clearcut choice of which one to condition on without invoking further conditions.

Example 6. Consider a random variable X assuming values $1, \ldots, 6$ such that

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

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 is which ancillary complement one chooses under the given circumstance. Basu left this example with a question mark. However, if one computes the information content of T based on its conditional distribution given these six ancillary statistics, then it turns out that for $X = 1$ or 4, the maximum information content lies in the conditional distributions given U_1 or U_4. For $X = 2$ or 5, this is for U_1 or U_6; while for $X = 3, 6$, this is for U_1 or U_3. Thus, considering all three situations, U_1 seems to be the most suitable ancillary statistic. From another point of view (Barnard and Sprott (1971); Cox and Hinkley (1974)), under the transformation $gX = X + 3 \pmod 6$, so that the induced transformation on the parameter space is $g^*\theta = -\theta$, it turns out that the only ancillary statistic unaffected by this transformation is U_1. Finally, if one uses Cox's (1971) criterion, it turns out that $V_\theta\{J(\theta|U_1)\} > V_\theta\{J(\theta|U_i)\}$ for all $i = 2, \ldots, 6$. From all these considerations, U_1 seems to be the most appropriate ancillary statistic in this example.

Basu's next example brings out an anomaly which one may encounter in the use of ancillary statistics.

Example 7. Basu's third example deals with $X \sim \text{uniform}[\,\theta, \theta+1)$, $0 \le \theta \le \infty$. The sample space is $\mathcal{X} = [\,0, \infty)$, and the likelihood function is

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

Thus, every point in the interval $(X - 1, X]$ is a MLE of θ. One such choice is $T = [X]$, the integer part of X. Let $\phi(X) = X - [X]$. Then $\phi(X) \sim \text{uniform}[0, 1)$,

and is ancillary. Since $X = [X] + \phi(X)$, $\{[X], \phi(X)\}$ is a one-to-one function of the minimal sufficient statistic 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

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

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 anamoly 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.

Traditionally the definition of an ancillary statistic has been a statistic with a distribution free of the model parameter, although as we saw in the measuring instrument example it is natural to say free of the model parameters of interest. Some writers insist that the ancillary statistic should be a component of the minimal sufficient statistic, and others not, but that distinction usually has no effect on the resulting inference if the ancillary statistic is required to have maximal dimension: whether one conditions first and then makes the sufficiency reduction, or gets the sufficient statistic first and then conditions, one is usually led to the same conditional distribution with a maximal ancillary statistic. However the definition is perhaps too narrow to capture the roles of the ancillary statistic outlined in Section 2, and we think this strict emphasis on distribution has led many readers to conclude that a theory of inference that insists on conditioning on ancillary statistics is more problematic than it really is. The approximate theory outlined in Section 4 shows that by emphasizing the conditional distribution after conditioning on an ancillary statistic, rather than the marginal distribution, leads to a fruitful theory of likelihood based inference.

4. Approximations and Ancillarity

In Section 2 we saw that one role of an ancillary statistic is to give a more relevant estimate of the information in the observed sample. This is extended

in the notion of approximate ancillarity, first discussed in Efron and Hinkley (1978). Assume we have an independent sample $X = (X_1, \ldots, X_n)$ from a scalar parameter model $f(x;\theta)$ with log-likelihood function $\ell(\theta)$. Efron and Hinkley (1978) showed that there is an approximately ancillary statistic U such that

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

where $j(\hat{\theta}) = -\ell''(\hat{\theta})$ is the observed Fisher information. This is the basis for the often-repeated claim that the observed information is a better estimate of the variance of the maximum likelihood estimator than the expected information. They also showed that

$$\sqrt{n}\left(\frac{j(\hat{\theta})}{I(\hat{\theta})} - 1\right) \xrightarrow{d} N(0, \gamma_\theta^2), \tag{4.1}$$

where

$$\gamma_\theta = \frac{(\nu_{20}\nu_{02} - \nu_{11})^{3/2}}{\nu_{20}^{3/2}}$$

was called the *statistical curvature* of the model, and

$$\nu_{jk} = E\left\{\left(\frac{\partial \ell}{\partial \theta}\right)^j \left\{\frac{\partial^2 \ell}{\partial \theta^2} + E\left(\frac{\partial \ell}{\partial \theta}\right)^2\right\}^k\right\}.$$

It follows from (4.1) that the statistic

$$U = \frac{1 - j(\hat{\theta})/I(\hat{\theta})}{\gamma_{\hat{\theta}}}$$

is approximately ancillary in the sense that $\sqrt{n}U$ has a limiting standard normal distribution; U has come to be known as the Efron-Hinkley ancillary. It is first-order ancillary, i.e., the normal approximation to the distribution of U has relative error $O(n^{-1/2})$. Skovgaard (1986) 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).

Example 8. In the measuring instruments example of Cox (1958), introduced in Section 2, assume that using instrument k, $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma_k^2)$ where σ_0^2 and σ_1^2 are known and unequal. The data are $(X_1, U_1), \ldots, (X_n, U_n)$, where X_i is the ith measurement and U_i is an indicator that takes the value 1 if the first instrument is used. The observed and expected Fisher information are, respectively,

$$I(\theta) = (n/2)(\sigma_0^{-2} + \sigma_1^{-2}), \quad j(\hat{\theta}) = (n - U_\cdot)\sigma_0^{-2} + U_\cdot \sigma_1^{-2},$$

where $U_\cdot = \sum U_i$ records the number of times the first measuring instrument is used. As the maximum likelihood estimate of θ is $\hat{\theta} = \sum_{j=1}^{n} X_j \sigma_{U_j}^{-2} / \sum_{j=1}^{n} \sigma_{U_j}^{-2}$, we have that

$$V(\hat{\theta} \mid U_1, \ldots, U_n) = j^{-1}(\hat{\theta})$$

exactly in this case, and that this is indeed the appropriate estimator of the variance.

There are several other approximate ancillary statistics that have been suggested in the literature. Skovgaard (1986) showed that a second order local ancillary statistic suffices to construct density and distribution function approximations accurate to third order, i.e., with relative error $O(n^{-3/2})$ in a moderate deviation region. Barndorff-Nielsen and Cox (1994, Chap. 7.2) discuss an approximate ancillary statistic based on a likelihood ratio statistic; they call this a directed likelihood ancillary. Suppose the statistical model forms a *curved exponential family*

$$f(y;\theta) = \exp\{a_1(\theta)t_1(y) + a_2(\theta)t_2(y) - c(\theta) - d(y)\},$$

where for simplicity of notation we assume that a_1 and a_2 are scalar functions of a scalar parameter θ: these define the curve in the full parameter space where the pair (a_1, a_2) is unrestricted. An example is a normal distribution with mean θ and variance θ^2. If we wanted to test the fit of the curved model, relative to the full model, we could use a likelihood ratio statistic $W = 2\{\ell(\hat{a}_1, \hat{a}_2) - \ell\{a_1(\hat{\theta}), a_2(\hat{\theta})\}\}$, where $(\hat{a}_1, \hat{a}_2)$ maximizes the log-likelihood over the unconstrained parameter space. The statistic W is asymptotically distributed as χ_1^2 under the "null" hypothesis that the curved model is correct, and its signed square root is asymptotically normal, hence ancillary to first order. It is also locally ancillary to second order. Further adjustments to this directed likelihood ancillary can be made to improve the order of accuracy of this approximation, although this first step is adequate for use in the p^* and r^* approximations described below. Note that this "hypothesis test" is preliminary to the desired inference for the parameter θ. The use of ancillary statistics in goodness-of-fit testing of an assumed model is discussed in Cox and Hinkley (1974, Chap. 2), and this is an asymptotic extension of that idea.

Example 5 (continued)**.** 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 is approximately so. 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 (4.1), 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 (1993) 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.

The role of ancillarity in the theory of asymptotic inference is most explicit in Barndorff-Nielsen's p^* approximation to the density of the maximum likelihood estimator. This approximation is

$$p^*(\hat{\theta} \mid u; \theta) = c|j(\hat{\theta})|^{1/2}\exp\{l(\theta; \hat{\theta}, u) - l(\hat{\theta}; \hat{\theta}, u)\}, \tag{4.2}$$

where we have assumed that there is a one-to-one transformation from the sample vector y to the pair $(\hat{\theta}, u)$, and this is explicitly indicated in the argument of the log-likelihood function. The renormalizing constant $c = c(\theta, u)$ can be shown to be equal to $(2\pi)^{d/2}$ where d is the dimension of θ. If the underlying model is a full exponential family, then (4.2) is a version of the saddlepoint approximation to the distribution of the minimal sufficient statistic, and no ancillary statistic is needed. The saddlepoint approximation is given in Daniels (1954), and a simple derivation of (4.2) is given in Durbin (1980).

Another special case is of particular interest in connection with ancillarity: if the underlying model is a transformation family, then (4.2) gives the exact conditional distribution of $\hat{\theta}$, and U is the maximal invariant on the group of transformations. This transformation family version of p^* was derived in Barndorff-Nielsen (1980), generalizing Fisher's (1934) result for location families. In general for transformation families, the maximal invariant for the group provides a natural ancillary statistic; in Example 5 above, this argument was used to choose among ancillary statistics.

We can view the role of U as providing a complementing statistic to $\hat{\theta}$, in order that the p^* approximation is defined on a sample space that is of the same dimension as the parameter space. Using this approximation will lose information about θ however, unless U has a distribution free of θ, i.e., is ancillary. Since p^* is an approximation to the density of $\hat{\theta}$, it suffices that U be approximately ancillary. If U is second order ancillary then the p^* approximation has relative error $O(n^{-3/2})$, while if U is just first order ancillary the p^* approximation

has relative error $O(n^{-1})$. Verifying these results requires specification of the ancillary statistic U; a good technical reference is Skovgaard (1990).

When θ is a scalar parameter, the p^* approximation can be re-expressed as the approximation to 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 $F(r|U;\theta)$, 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 also be approximated to $O(n^{-3/2})$, using a technique due to Lugannani and Rice (1980). The resulting approximation is

$$F(r|U;\theta) = \Phi(r^*)\{1 + O(n^{-3/2})\}, \tag{4.3}$$

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 simpler statistic Q that does not require the determination of an explicit expression for U, but leads to an r^* approximation with relative error $O(n^{-3/2})$, is developed in Fraser and Reid (1993, 1995). Skovgaard (1996) suggests a statistic $\tilde{Q}$ that also avoids specification of an ancillary statistic, and leads to an r^* approximation with relative error $O(n^{-1})$; among a number of suggested versions equivalent to this order Skovgaard's seems to be the most accurate in examples. The connection between the three versions of r^* are further developed in Reid and Fraser (2008).

Although the notion of approximately ancillary statistics appears to introduce even more possibilities for ancillary statistics, the p^* and r^* approximations provide very accurate approximations to the density and distribution of the maximum likelihood estimator; from that point of view the choice of particular ancillary statistic is not crucial. McCullagh (1984) shows that for scalar parameters, all choices of approximate ancillary lead to the same p^* approximation to $O(n^{-1})$. For implementation of the r^* approximation (4.3), the approach of Fraser and Reid (1995) requires only specification of ancillary directions, which can be much simpler than finding the explicit form of the ancillary statistic: see for example Brazzale, Davison and Reid (2007, Chap. 8).

5. Elimination of Nuisance Parameters

5.1. Extended definitions of ancillarity

It may be noted that the function of ancillary statistics in the presence of nuisance parameters is quite different from what was discussed earlier. The main objective of standard ancillarity is recovery of the loss of information or more

generally, probability calculations conditional on a relevant subset. However, in the presence of nuisance parameters, their elimination without any loss of information is the primary goal.

To illustrate, we begin with a model parameterized by $\theta = (\psi, \lambda)$, where ψ is the parameter of the interest, and λ is the nuisance parameter. In such cases, in order to draw inferences regarding the parameter of interest ψ, one approach is to eliminate the nuisance parameter λ.

A standard approach is the so-called conditional likelihood approach. Suppose the joint density of the minimal sufficient statistic (T, U) is given by

$$f(T, U; \psi, \lambda) = f(T|U, \psi) f(U; \psi, \lambda). \tag{5.1}$$

Then the inference is based on the conditional density $f(T|U, \psi)$ which does not involve λ.

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.

The above requirement is met if the marginal density of U does not depend on λ. This, however, does not happen, in general, as the following example shows.

Example 9. Let $X_1, \ldots, X_n$ be i.i.d. with common pdf

$$f(X; \psi, \lambda) = \frac{\Gamma(\psi + X)}{\Gamma(X + 1)\Gamma(\psi)} \lambda^X (1 - \lambda)^\psi,$$

where $\psi > 0$ and $0 < \lambda < 1$. For fixed ψ, $U = \sum_{i=1}^n X_i$ is sufficient for λ so that the conditional distribution of $X_1, \ldots, X_n$ given U depends only on ψ. However, U has pdf

$$f(U; \psi, \lambda) = \frac{\Gamma(n\psi + U)}{\Gamma(U + 1)\Gamma(n\psi)} \lambda^X (1 - \lambda)^\psi,$$

which depends on both ψ and λ, and is not ancillary for ψ in the usual sense. Indeed, the Fisher information contained in 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 (Sandved (1965); Sverdrup (1966)). 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, U is S-ancillary if and only if there exists a reparameterization of (ψ, λ) into (ψ, ϕ) such that the marginal distribution of U depends only on ϕ. The following example given in Severini (2000) illustrates this.

Example 10 (Severini (2000, Example 8.3)). $X_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}\{\exp(\lambda + \psi Z_i)\}$, $i = 1, \ldots, n$. 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 $(U; p_1, \ldots, p_n)$, where $p_i = \{\exp(\psi Z_i)\}/\sum_{i=1}^n \exp(\psi Z_i)$.

However, S-ancillary statistics need not always exist. The following simple example illustrates this.

Example 11 (Severini (2000, Example 8.7)). Let $X_i \overset{\text{i.i.d.}}{\sim} \text{N}(\lambda + \psi Z_i, 1)$, $i = 1, \ldots, n$, where we restrict the parameter space to $\lambda > 0$. The log-likelihood is given by $l(\psi, \lambda) = -(n/2)(\bar{X} - \lambda - \psi\bar{Z})^2$, and $\bar{X}$ is a P-ancillary statistic. To see that an S-ancillary statistic does not exist, note that for $\psi = 0$, $\bar{X} \sim \text{N}(\lambda, 1/n)$ so that the mean is positive, while if $\psi = -1$, $\bar{X} \sim \text{N}(n\lambda - \bar{Z}, 1/n)$ so that the mean of $\bar{X}$ is any number greater than $-\bar{Z}$. Thus $\bar{X}$ cannot be S-ancillary for λ.

5.2. 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. Ferreira and Minder (1981) provided examples to show how statistics satisfying Godambe's definition of ancillarity could still be useful for inference about the parameter of interest. According to Godambe's formulation, let $Y_1, \ldots, Y_n$ be independent with pdf's $f(Y_i|\psi, \lambda_i)$, where ψ is the parameter of interest, while the λ_i are the nuisance parameters. Let $g(Y_i, \psi)$ be a function of Y_i and ψ, the parameter of interest, which satisfies $E\{g(Y_i, \psi, \lambda_i)\} = 0$. Then $g(\boldsymbol{Y}, \psi) = \sum_{i=1}^n g(Y_i, \psi)$ is called an unbiased estimating function, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

Godambe (1976) defined an optimal unbiased estimating function as the minimizer of $E\{g^2(\boldsymbol{Y}|\psi)/E\{\partial g(\boldsymbol{Y}, \psi)/\partial\psi)\}^2\}$. Earlier (Godambe (1960)), he showed that that without any nuisance parameters, the score function was the optimal unbiased estimating function. In the presence of nuisance parameters, he showed that if the joint density $f(\boldsymbol{Y}|\psi, \lambda_1, \ldots, \lambda_n)$ factors as

$$f(\boldsymbol{Y}|\psi, \lambda_1, \ldots, \lambda_n) = f(\boldsymbol{Y}|U, \psi) f(U|\psi, \lambda_1, \ldots, \lambda_n),$$

where U (possibly vector-valued) is a complete sufficient statistic for the nuisance parameter vector $(\lambda_1, \ldots, \lambda_n)$, then the conditional score function $\partial \log f(\boldsymbol{Y}|U, \psi)/\partial\psi$ is the optimal unbiased estimating function. He also showed that the information contained in the conditional distribution of $\boldsymbol{Y}$ given U is the same as that contained in its unconditional distribution.

In Example 9, $U = \sum_{i=1}^{n} Y_i$ is a complete sufficient statistic for the nuisance parameter λ, and so the conditional score function based on the conditional pdf

$$f(\boldsymbol{Y}|\psi) = \prod_{i=1}^{n} \binom{\psi + X_i - 1}{X_i} \Big/ \binom{n\psi + U - 1}{U}$$

is the optimal unbiased estimating function.

The above optimality of the conditional score function led to the more general notion of P-ancillarity (partial ancillarity) due to Bhapkar (1989, 1991). Here ancillarity in the presence of a nuisance parameter is based on the notion of partial information for ψ. In order to define partial information, we partition the information matrix for (ψ, λ) into submatrices according to the partition of the parameter. Then the partial information for ψ is given by $I_{\psi\psi.\lambda} = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}$. This is due to the fact that $I_{\psi\psi.\lambda}$ is the information content in the conditional distribution of T given U. We say that U is partial ancillary (P-ancillary) for ψ is $I_{\psi\psi.\lambda} = 0$.

Example 10 (Continued)**.** In this example

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

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

In general, S-ancillarity need not not be the same as P-ancillarity. For instance, in Example 7, U is P-ancillary but not S-ancillary. Also, (Severini (2000, pp.282-285)) has produced a Gamma distribution example where the conditioning variable U is neither S-ancillary nor P-ancillary.

5.3. Bayesian ancillarity

As noted in Sections 5.1 and 5.2, S-ancillarity or P-ancillarity of a statistic U does not imply that the distribution of U does not depend on ψ, and depends only on λ. A natural question is whether one can find an alternative definition of ancillarity that ensures that the marginal distribution of U does not depend on ψ and depends only on λ.

To this end, Severini (1995) proposed the notion of Bayesian ancillarity. We observe as a consequence of his definition that, 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). In what follows, we use $p(\cdot|\cdot)$ as a generic symbol for a conditional pdf, and $p(\cdot)$ as a generic symbol for a marginal pdf.

First with no nuisance parameter, U is Bayes ancillary 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

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

that is, the marginal of U does not depend on ψ. So, U is ancillary in the usual sense.

In the presence of a nuisance parameter λ, suppose (T, U) is minimal sufficient 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

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

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

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

Equivalently, $p(U|\psi) = \int p(T|U,\psi)\,p(U|\psi)\,p(\psi)\,d\psi / \int p(T|U,\psi)\,p(\psi)\,d\psi$. Once again, the marginal pdf of U given ψ does not involve ψ, and U is ancillary in the usual sense. In Example 5, if $\pi(\lambda|\psi) \propto \lambda^{-1}(1-\lambda)^{-1}$, then $\int p(U|\psi,\lambda)p(\lambda|\psi)d\lambda = \Gamma(U)/\Gamma(U+1) = U^{-1}$ which shows that U is Bayes ancillary with respect to this prior.

5.4. Approximate ancillarity in the presence of nuisance parameters

The definition of ordinary ancillarity in the presence of nuisance parameters is not at all straightforward, as we have seen in the previous subsections. While it is possible to formalize the notion of approximate ancillarity in the nuisance parameter setting, as is done for S-ancillarity in Severini (1993), the development quickly gets very technical. However, it is possible to extend the asymptotic approximations outlined in Section 4 to the nuisance parameter setting, using an approximate version of (4.3).

We start with the p^* approximation (4.2) for the distribution of the full maximum likelihood estimator $\hat{\theta}$, conditional on an approximate ancillary statistic

U. The goal is to find an approximation that can be used for inference about the parameter of interest ψ, without specifying a value for the nuisance parameter λ. One way to approach this is to consider a p^* approximation for inference about λ, in a model where ψ is held fixed. An approximate ancillary statistic is needed for constructing this, and the resulting approximation is the conditional density of $\hat{\lambda}_\psi$, given the original ancillary statistic U and a further ancillary statistic U_ψ, say. Thus we have the partition

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

where U_ψ is the approximate ancillary statistic needed for the p^* approximation to the conditional density of $\hat{\lambda}_\psi$, and $p(U_\psi \mid U; \theta)$ is the ratio of the two p^* approximations. Barndorff-Nielsen (1986) showed that U_ψ can be transformed to a quantity r^*_ψ that has, to $O(n^{-3/2})$ a standard normal distribution. Further a constructive expression for r^*_ψ is available that combines $r_\psi = \text{sign}(\hat{\psi} - \psi)\{2\{l_p(\hat{\psi}) - l_p(\psi)\}\}^{1/2}$, from the profile log likelihood, with a related quantity q_ψ as

$$r^*_\psi = r_\psi + \frac{1}{r_\psi} \log\left(\frac{q_\psi}{r_\psi}\right).$$

This leads directly to approximate inference for ψ based on $\Phi(r^*_\psi)$, which has relative error $O(n^{-3/2})$ conditionally on U and unconditionally. The construction of r^*_ψ requires differentiation of the log-likelihood function on this sample space, with U fixed. Fraser and Reid (1995) show how to compute these derivatives without first obtaining an explicit expression for the approximate ancillary statistic U; see also (Severini (2000, Chap. 7.5))and Brazzale, Davison and Reid (2007, Chap. 2)). It is possible to avoid the ancillary statistic entirely by a method suggested in Skovgaard (1996), although the resulting approximation has relative error $O(n^{-1})$ instead of $O(n^{-3/2})$.

Example 12. Suppose $X_1, \ldots, X_n$ are i.i.d. from the $\text{N}(\mu, \sigma^2)$ distribution, with $\mu = \psi$ the parameter of interest. The expressions for r and q are given by

$$r = \text{sign}\ (q) \left[n \log\left\{ 1 + \frac{n(\hat{\mu} - \mu^2)}{\hat{\sigma}^2} \right\} \right]^{1/2},$$
$$q = \frac{n(\hat{\mu} - \mu)/\hat{\sigma}}{1 + \{n(\hat{\mu} - \mu)^2\}/\hat{\sigma}^2},$$

which are simple functions of the t-statistic $t = \sqrt{n}(\hat{\mu} - \mu)/\{\hat{\sigma}/(n-1)\}$. Expressions for r and q for general location-scale models are given in Barndorff-Nielsen and Cox (1994, Ex. 6.20). The detailed construction of U_ψ mentioned above is not needed for this example (Barndorff-Nielsen and Cox (1994, Ex. 6.11)). The

following very simple series expansion for r^* was derived by Sartori (2003) and Iglesias-Gonzalez (2007):

$$r^* = t - (t + t^3)/(4n) + O(n^{-2}). \tag{5.2}$$

Expansion (5.2) is still valid when μ is replaced by $X\beta$, with β a vector of p unknown parameters.

This model can be generalized in a number of directions: expressions for r and q in general regression-scale models can be obtained explicitly from summary formulae for q given in, for example, Brazzale, Davison and Reid (2007, Chap. 6), and formulae for nonlinear regression are given in Fraser, Wong and Wu (1999). The essential point is that the expressions derived, using notions of approximate ancillarity, provide a means of calculating a pivotal quantity, r^*, which like the t-statistic in normal theory models, provides inference for the parameter of interest with explicit specification of the nuisance parameter. The normal approximation holds both conditionally on the approximate ancillary statistic and unconditionally. From the point of view of the asymptotic theory, the conditional distribution given an ancillary statistic is more useful than the precise construction and definition of ancillary statistics.

5.5. Brown's ancillarity paradox

Brown (1990) introduced a very interesting 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 become so.

We begin with the following simple example of Brown.

Example 13. 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 $\boldsymbol{\Sigma} = E(\boldsymbol{U}\boldsymbol{U}^T)$, 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 14. 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 $L(\alpha, a) = (a - \alpha)^2$, $a \in R^1$.

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 square error loss. However, if it is assumed that k-dimensional components of $\boldsymbol{Z}$ are i.i.d. $\mathrm{N}(\boldsymbol{0}, \sigma^2 I_k)$, 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. Another interesting example of ancillarity paradox in the context of finite population sampling appears in Godambe (1982).

6. Conclusion

The topic of ancillarity continues to intrigue, at least in part because any satisfactory frequentist theory of inference must incorporate conditioning, but a wholly Bayesian approach that automatically conditions on the data raises other problems, including the meaning of, and choice of, prior probabilities. In this paper we have surveyed, through examples, various aspects of ancillarity and their relation to the theory of inference.

Acknowledgement

This paper is based on the Basu Memorial Lecture, presented by Ghosh and discussed by Reid at JSM 2004. We are grateful to the referees and an associate editor for helpful comments on an earlier version. The research was partially supported by NSF Grant SES-0631426 and NSA Grant MSPF-076-097, and by the Natural Sciences and Engineering Research Council of Canada.

References

Barnard, G. A. (1982). Conditionality versus similarity in the analysis of 2×2 tables. In *Statistics and Probability: Essays in Honor of C.R. Rao*. (Edited by G. Kallianpur, P. R. Krishnaiah and J. K. Ghosh). North Holland, Amsterdam.

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

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. *Biometrika*, **67**, 293-310.

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. and Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, New York.

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

Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. IMS Lecture Notes-Monograph Series, Volume 6. Institute of Mathematical Statistics, Hayward.

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

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

Brazzale, A. R., Davison, A. C. and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small Sample Statistics*. Cambridge University Press, Cambridge.

Brown, L. D. (1990). An ancillarity paradox which appears in multiple regression models. *Ann. Statist.* **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.

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

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.

Durbin, J. (1980). Approximations for densities of sufficient estimators *Biometrika* **67**, 311-333.

Edgeworth, F. Y. (1893). Exercises in the calculation of errors. *Philosophical Magazine* (*Fifth Series*) **36**, 98-111.

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

Ferreira, P. E. and Minder, C. E. (1981). On a concept of ancillarity. *Biometrika* **68**, 344.

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. Ser. A* **144**, 285-307.

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

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley and Sons, New York.

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

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.* **19**, 333-369.

Fraser, D. A. S. and McDunnogh, P. (1980). Some remarks on conditional and unconditional inference for location-scale models. *Statist. 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. *Statist. 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., Wong, A. C. M. and Wu, J. (1999). Regression analysis, nonlinear or nonnormal simple and accurate p-values from likelihood analysis. *J. Amer. Statist. Assoc.* **94**, 1286-1295.

Ghosh, J. K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu.* Springer-Verlag, New York.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood equation. *Ann. Math. Statist.* **31**, 1208-1211.

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.

Iglesias-Gonzalez, S. (2007). Highly accurate tests for the mixed linear model. Ph.D. Dissertation, University of Toronto.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62**, 251-268.

Kalbfleisch, J. D. (1982). Ancillary statistics. In *Encyclopedia of Statistical Sciences*, **V1**. (Edied by S. Kotz, N. L. Johnson and C.B. Read), 77-81. Wiley, New York.

Lehmann, E. L. (1981). An interpretation of completeness and Basu's theorem. *J. Amer. Statist. Assoc.*, **76** 335-340.

Lehmann, E. L. and Scholz, F. W. (1992). Ancillarity. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu.* (Edited by M. Ghosh and P. K. Pathak). Ims Lecture Notes and Monograph Series **17**, 32-51.

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

McCullagh, P. (1984). Local sufficiency. *Biometrika* **71**, 233-244.

Pearson, K. P. (1896). Mathematical contributions to the theory of evolution, III: regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lond. Ser. A* **187**, 253-318.

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

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

Reid, N. and Fraser, D. A. S. (2008). Mean likelihood and higher order approximations. Submitted for publication.

Sandved, E. (1965). A principle for conditioning on an ancillary statistic. *Skandinavisk Aktuarietidskrift*, **49**, 39-47.

Sartori, N. (2003). A note on likelihood asymptotics in normal linear regression. *Ann. Inst. Math. Statist.* **55**, 187-195.

Severini, T. A. (1993). Local ancillarity in the presence of a nuisance parameter. *Biometrika* **81**, 649-661.

Severini, T. A. (1995). Comment on "The roles of conditioning in inference" by N. Reid. *Stat. Science* 187-189.

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

Skovgaard, I. M. (1986). Successive improvement of the order of ancillarity. *Biometrika* **73**, 516-519.

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.

Stigler, S. M. (2001). Ancillary history. *State of the Art in Probability and Statistics* **36**, 555-567.

Sverdrup, E. (1966). The present state of decision theory and Neyman-Pearson theory. *Rev. Inst. Internat. Statist.* **34**, 309-333.

Wang, S. (1993). Saddlepoint approximations in conditional inference. *J. Appl. Probab.* **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.

Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.

E-mail: ghoshm@stat.ufl.edu

Department of Statistics, University of Toronto, Toronto, ON M5S 1A1, Canada.

E-mail: reid@utstat.utoronto.ca

Department of Statistics, University of Toronto, Toronto, ON M5S 1A1, Canada.

E-mail: dfraser@utstat.utoronto.ca

(Received August 2005; accepted June 2009)