THE ELUSIVE ANCILLARY

D. A. S. FRASER

SUMMARY

The concept of an ancillary statistic has been promoted by R.A. Fisher (e.g., 1925) as a basis for determining the accuracy of estimates and as a means for avoiding any loss of information. Basu (1959, 1964) has shown that a weaker definition of an ancillary leads to nonuniqueness. Barnard and Sprott (1971) have suggested a resolution of the nonuniqueness using invariance of the likelihood function. Cox (1971) suggests a more generally applicable resolution based on information. The two resolutions however may well be contradictory in larger contexts, and Cox's approach in fact provides a substantial argument against the validity of the ancillary concept for statistical inference. Ancillary statistic, Information, Accuracy, Conditional probability, Inference.

1. The Fisher Ancillary.

Fisher's continued promotion of the ancillary statistic dates from some of his earliest works on inference.

For example, Fisher (1925): "Since the original data cannot be replaced by a single statistic without loss of accuracy, it is of interest to see what can be done by calculating, in addition to our estimate, an ancillary statistic which shall be available in combination with our estimate in future calculations. If our two statistics specify the values of $\hat{\theta}/\theta_0$ and $\hat{\theta}_1^2/\theta_0^2$ for some central value of $\theta$, such as $\hat{\theta}$, then the variance of $\hat{\theta}/\theta_0$ over the sets of samples for which both statistics are constant, ... will ordinarily be of order $n^{-1}$ at least. With the aid of such an ancillary statistic the loss of accuracy tends to zero for large samples."

And Fisher (1934): "... the sets of observations which provide the same estimate differ in their likelihood functions, and therefore in the nature and quantity of the information they supply (and) yet when samples alike in the information they convey exist for all values of the estimate and occur with the same frequency for corresponding values of the parameter (then) a second case (occurs that) need involve no loss of information".

And also: "the loss (of information) may be recovered by using
as an ancillary statistic, in addition to the maximum likelihood estimate, the second and higher differential coefficients at the maximum. In general we can only hope to recover the total loss, by taking into account the entire course of the likelihood function."

These quotations give some indication of the complexity that surrounds Fisher's concept of an ancillary statistic. He depended more on the many examples throughout his writings and less on any formal delineation of the concept and its role. There are of course many people who urge earlier and sharper conceptualizations in statistics but the history of this and indeed certain other concepts shows the substantial disadvantages of what could be called premature conceptualization.

2. The Basu Ancillary.

Basu (1959) extracts one aspect of Fisher's characterization and defines an ancillary statistic as "one whose distribution is the same for all possible values of the unknown parameter." He investigates this form of the concept in some detail and finds that often there may be more than one maximal ancillary.

Basu (1964) then considers the role of an ancillary statistic in inference. He presents three examples that have more than one maximal ancillary, and notes that this produces difficulties for the recommended programmes of inference. He does suggest that the difficulties can be resolved by recognizing the difference between a real and a conceptual statistical experiment. But he does not elaborate on the difference and did "... not think it necessary to enter into a lengthy discussion on the reality or performability of a statistical experiment".

His examples do not clarify his intent -- in one case a coin determines a subexperiment, and in another case he hypothesizes a coin and then uses its nonexistence to discount the reality of the component experiments.

3. The Barnard-Sprott Resolution.

Barnard and Sprott (1971) argue "that if proper attention is paid to the likelihood function as the primary inference from an experiment and the role of an ancillary in describing its shape, (then) the difficulties raised by Basu disappear." They consider the family of possible likelihood functions and seek a location indicator $T$ (as estimate) and a shape or scale indicator $A$ (as ancillary).
They then obtain a unique ancillary for each of the three Basu examples.

Barnard and Sprott's procedure for choosing an ancillary is in accord with Fisher's recommendations (1925, 1934) concerning ancillaries. Fisher recommends a complex of statistics to determine "the entire course of the likelihood function". In the invariance case examined by Barnard and Sprott such a complex of statistics gives the orbit or shape ancillary.

In commenting on the Barnard and Sprott (1971) paper, Cox notes that it is "plausible that their particular choice of ancillaries partitions the overall distribution into components as distinct as possible in some sense. If this is indeed true, it provides a resolution of the non-uniqueness independent of group arguments."

4. The Cox Resolution.

Cox (1971) elaborates concerning his comments on the Barnard and Sprott paper, and recommends choosing "that ancillary statistic which separates the expected information into components that differ as widely as possible." Specifically the expected information is examined for each value of an ancillary and the variance of this expected information is used to measure the success of the ancillary in separating the expected information into the different components.

Let $\mathcal{I}(\theta)$ be the log-likelihood and

$$ I(\theta; a) = \mathbb{E} \left[ \frac{\partial^2 \mathcal{I}(\theta)}{\partial \theta^2} : a, \theta \right] $$

be the mean information given the value $a$ of an ancillary. Then Cox suggests using

$$ \text{Var}(I(\theta; a) : \theta) $$

as the measure of an ancillary and recommends choosing that ancillary for which the variance measure is largest.

In his original 1925 paper on information Fisher suggests the use of the second derivative $\frac{\partial^2 \mathcal{I}}{\partial \theta^2}$ as an ancillary statistic measuring information, and in part considers how the mean conditional variance

$$ \mathbb{E} \left[ \text{Var} \left( \frac{\partial^2 \mathcal{I}(\theta)}{\partial \theta^2} : a, \theta \right) : \theta \right] $$

measures the inadequacy of a grouping of sample points. The common formula for variance about regression gives
\[ \text{Var}\left( \frac{2^2 \lambda}{\lambda \sigma^2} \right) = E \text{Var}\left( \frac{2^2 \lambda}{\lambda \sigma^2} : a \right) + \text{Var} \ E\left( \frac{2^2 \lambda}{\lambda \sigma^2} : a \right). \]

Thus maximizing the variance of the conditional information

\[ \text{Var}(I(\theta:a):\theta) \]

is equivalent to minimizing the mean conditional variance of the observed information \(-\frac{2^2 \lambda}{\lambda \sigma^2}\). The Cox criteria then becomes one of assembling together points where observed information is as nearly the same as possible, which is the intent in Fisher's original paper.

The observed information \(-\frac{2^2 \lambda}{\lambda \sigma^2}\) gives a measure of the curvature of the likelihood function. Thus in a more technical sense the Cox procedure tends to agree with the Barnard and Sprott procedure of grouping on the basis of similarly shaped likelihood functions.

5. **Other Segregation Methods.**

Cox notes that his procedure for segregating points on the basis of the variance of the mean information is "rather arbitrary", but nevertheless consistent with the general objective for ancillary statistics -- complete segregation of samples on the basis of the information they provide. Consider some alternative measures that can be used for selecting an ancillary.

For any value \(a\) of an ancillary the mean log-likelihood slope

\[ E\left( \frac{2\lambda}{\lambda \sigma} : a, \theta \right) = 0 \]

is equal to zero. Then in accord with the Barnard and Sprott recommendation to group on the basis of the likelihood function it seems reasonable to group together points that have approximately the same likelihood slope \( |\frac{2\lambda}{\lambda \sigma}| \). The formula for variance about regression gives

\[ \text{Var}\left| \frac{2\lambda}{\lambda \sigma} \right| = I(\theta) - \left[ E\left| \frac{2\lambda}{\lambda \sigma} \right| \right]^2 = E \text{Var}\left( \frac{2\lambda}{\lambda \sigma} : a \right) + \text{Var} \ E\left( \frac{2\lambda}{\lambda \sigma} : a \right). \]

Thus the grouping together of points having approximately the same likelihood slope can be accomplished by maximizing the variance of the conditional mean slope \( |\frac{2\lambda}{\lambda \sigma}| \); this is a simple variation on maximizing the variance of the conditional mean curvature \(-\frac{2^2 \lambda}{\lambda \sigma^2}\).

Alternatively we could consider grouping more directly on the basis of information. In his original paper on information Fisher
presents a first-order formula for the information lost in using only the maximum likelihood from a conditional distribution:

\[
\frac{\text{Var}[\theta^2 I/\partial \theta^2 : a, \theta]}{E[-\theta^2 I/\partial \theta^2 : a, \theta]}
\]

The mean value

\[
E \frac{\text{Var}[\theta^2 I/\partial \theta^2 : a, \theta]}{E[-\theta^2 I/\partial \theta^2 : a, \theta]}
\]

then provides a measure of how much an ancillary fails to be a full complement to the maximum likelihood estimate. This formula from Fisher then suggests a variation on the Cox procedure, a variation that seems more specific to the purposes of an ancillary. The choice of an ancillary is then based on minimizing the mean value of the conditional variance of the observed information as a proportion of the conditional mean of the observed information.

The invariance grouping of sample points, the maximum separation of the information, and the minimum loss of information in using the ancillary as sole ancillary may well be contradictory criteria in larger contexts than the simple multinomial examined by Basu, Barnard and Sprott, and Cox.


The Basu ancillary is concerned with grouping sample points together so that the marginal distribution of the groups is independent of the parameter. The Barnard and Sprott and the Cox procedures are concerned in addition with grouping points together on the basis of common likelihood and information properties; in doing this these procedures are attempting to restore some of the ancillary properties envisaged by Fisher.

The argument used with these procedures is one of mathematically assembling points in various ways and then assessing the groups on the basis of some general criteria. It should be noted that this is the type of argument that is used in the Birnbaum analysis to deduce the likelihood principle from the sufficiency and conditionality principles. Thus the argument used with the present procedures for choosing an ancillary is an argument that reduces the problem to the likelihood function alone and thereby voids the problem of finding an ancillary. In effect the methods solve the problem by causing it to vanish.
7. A Case Against Ancillaries.

Consider two distinct populations $A_1, A_2$ of experimental units. And suppose that a treatment produces a reaction with probability $(1+0)/2$ for a unit from the first population and produces a reaction with probability $(2+0)/4$ for a unit from the second population. Let $x_1$ be the number of positive reactions for a sample of $n_1$ from the first population and let $x_2$ be the number of positive reactions for a sample of $n_2$ from the second population. The statistical model then is a product of two binomial models with probabilities $(1+0)/2$ and $(2+0)/4$ determined by the single parameter $\theta$.

Now suppose that the two populations are in fact mixed together and that their sizes are in the ratio one to two. Consider a random sample of $n$ from the combined populations and suppose that $n_1$ belongs to the first population and $n_2$ belongs to the second population.

The principles of conditional probability prescribe that we condition on what we know concerning the population sampling -- specifically that we examine the model given $n_1$ and given $n_2$. We thus obtain the product of two binomial models as described for the direct sampling of the individual populations.

Alternatively consider two distinct populations $B_1, B_2$ of experimental units, and suppose that a treatment produces a reaction with probability $(1+0)/3$ for a unit from the first population and produces a reaction with probability $(2+0)/3$ for a unit from the second population. Let $x_1$ be the number of positive reactions for a sample of $m_1$ from the first population and let $x_2$ be the number of positive reactions for a sample of $m_2$ from the second population. The statistical model then is a product of two binomial models with probabilities $(1+0)/3$ and $(2+0)/3$ determined by the single parameter $\theta$.

Now suppose that the two populations are in fact mixed together and that their sizes are in the ratio one-to-one. Consider a random sample of $n$ from the combined population and suppose that $m_1$ belongs to the first population and $m_2$ belongs to the second population.

The principles of conditional probability again prescribe that we condition on what we know concerning the population sample -- specifically that we examine the model given $m_1$ and given $m_2$. We thus obtain the product of two binomial models as described for the direct sampling of the individual populations.

Note however that the multinomial model for the sampling from
the mixed populations $A_1$ and $A_2$ is the same as from the mixed populations $B_1$ and $B_2$:

\[
\begin{array}{ccc}
B_1 & B_2 \\
A_1 & x_1 & n-x_1 & n_1 \\
A_2 & n_2-x_2 & x_2 & n_2 \\
m_1 & m_2
\end{array}
\]

from

\[
\begin{array}{ccc}
(1+0)/6 & (1-0)/6 & 1/3 \\
(2-0)/6 & (2+0)/6 & 1/3 \\
1/2 & 1/2
\end{array}
\]

If we know the investigation is within populations $A_1$, $A_2$ then we use the first product binomial. If we know the investigation is within populations $B_1$, $B_2$ then we use the second product binomial.

If however we do not know about a population separation $A_1$ to $A_2$ or $B_1$ to $B_2$ then the ancillary discussions of Barnard and Sprott, and Cox prescribe that we choose one or other of the two separations depending on similar likelihoods or on information separation -- and in fact prescribe that we choose the $A_1$ to $A_2$ separation because of its preferred characteristics.

Clearly we could be in a $A_1$ to $A_2$ situation or in a $B_1$ to $B_2$ situation and accordingly have the conditional models determined for us. But if we are in the position of not knowing which applies -- if either, then no valid principle of inference can argue that we act as if we are in that situation for which the inference procedures would be more attractive. Inference cannot be preference in an objective analysis.

Conditioning on the sample size does of course provide the common paradigm for the presentation of the ancillary concept. And the ancillary concept accordingly acquires much of the force that goes with the paradigm. Part of the difficulty lies in the mathematical approach that treats all probabilities in the same way. In fact, however, different probabilities can be associated with quite distinct physical operations and any valid analysis must acknowledge the physical distinction.

If we have a probability space then any realization leads to conditional probabilities. If however we have a class of probability
measures on a measurable space, then a realization even if it has constant probability is not grounds for the conditional model. The indeterminacy of the ancillary is the cautioning counter example.

8. Addendum.
Consider a probability space \((S,A,P)\) and a class \(G\) of continuous transformations \(\Theta\) that map \(S\) into \(S\) and form a group. Suppose that an observed response \(X\) has been obtained from some unknown transformation applied to a realization \(E\) on the probability space. The observed \(X\) identifies the set \(GX\) as the set of possible values for the realization \(E\). Probability principles accordingly prescribe the conditional model \(\{S,A,P_{GX}\}\) given the event \(GX\). This reduction has been viewed as an ancillary reduction by some commentators. In fact it is the basic reduction of probability theory. An ancillary would need a class of probability measures. The model however presents a class of random variables on a probability space.

References