Basu on Ancillarity

Philip Dawid

1 The origins of ancillarity

The term "ancillary statistic" was introduced by R. A. Fisher (Fisher 1925) in the context of maximum likelihood estimation. Fisher regarded the likelihood function as embodying all the information that the data had to supply about the unknown parameter. At a purely abstract level, this might be regarded as simply an application of the sufficiency principle (SP), since as a function of the data the whole likelihood function (*modulo* a positive constant factor — a gloss we shall henceforth take as read) is minimal sufficient; but that principle says nothing about what we should do with the likelihood function when we have it. Fisher went beyond this stark interpretation, regarding the actual form of the likelihood function as itself somehow embodying the appropriate inference. In some cases, such as full exponential families, the maximum likelihood estimator (MLE) is itself sufficient, fully determining the whole course of the likelihood function; but more generally it is only in many-one correspondence with the likehood function, so that two different sets of data can have associated likelihood functions whose maxima are in same place, but nevertheless differ in shape. Initially, for Fisher, an *ancillary* statistic (from the Latin "ancilla", meaning handmaiden) denoted a quantity calculated from the data which "lent support" to the MLE, by providing additional information about the shape of the likelihood function, over and above the position of its maximum — for example, higher derivatives of the log-likelihood at the MLE. If we regard the spikiness of the likelihood function as telling us something about the (data-dependent) precision of the MLE, we might select a suitable ancillary statistic to quantify this precision: this appears to have been Fisher's original motivation. According to this understanding of an ancillary statistic as describing the shape of the likelihood function, it is necessarily a function of the minimal sufficient statistic. Ideally, the MLE together with it handmaiden would fully determine the likelihood function, the pair then constituting a minimal sufficient statistic.

Fisher (1934) then considered the working out of these general concepts in the special case of a location model, where the MLE fully determines the location of the likelihood function, but is entirely uninformative as to its its shape; while the configuration statistic, *i.e.*, the set of pairwise differences between the observations, constitutes an ancillary statistic, fully determining the shape of the likelihood function, but uninformative about its location. For this model (though, for Fisher's original definition, not necessarily more generally) it is also true that the distribution of this ancillary statistic is entirely independent of the value of the unknown location parameter; furthermore, the conditional distribution of the maximum likelihood estimator, given the configuration, has a density

P. Dawid (✉)
Statistical Laboratory, Centre for Mathematical Sciences, CB3 0WB, UK
e-mail: apd@statslab.cam.ac.uk

that has the same shape as the likelihood function. At a certain point, Fisher decided that it was such properties, rather than his original handmaiden conception, that were of crucial general importance, and from that point on the word "ancillary" was used to mean "having a distribution independent of the parameter". Associated with this was the somewhat vague idea of a "conditionality principle" (CP), whereby it is the conditional distribution of the data, given the ancillary statistic, that is regarded as supplying the appropriate "frame of reference" for determining the precision of our estimate. As a simple example lending support to this principle, suppose we first toss a fair coin, and then take 10 observations if it lands heads up, or 100 if it lands tails up. The coin toss does not depend upon the parameter (it is ancillary in the revised sense, although not necessarily in the original sense), and so cannot, of itself, be directly informative about it; but it does determine the precision of the experiment subsequently performed, and it does seem eminently sensible to condition on the number of observations actually taken to obtain a realistic measure of realised precision.

At an abstract level, CP can be phrased as requiring that any inference should be (or behave as if it were) conducted in the frame of reference that conditions on the realised value of an ancillary statistic. One can attempt to draw analogies between this CP and the sufficiency principle, SP, which tells us that our inference should always be (or behave as if it were) based on a sufficient statistic. But is important to note that in either case there may be a choice of statistics of the relevant kind, and we would like to be able to apply the principle simultaneously for all such. Considering first the case of sufficiency, suppose T_1 is sufficient and, in accordance with SP, we are basing our inference on T_1 alone. If now T_1 is a function of T_2, then T_2 is also sufficient: but the property that our inference should be based on T_2 alone is automatically inherited from this property holding for the "smaller" statistic T_1, so we do not need to take any explicit steps to ensure this. In particular, if we can find a *smallest* sufficient statistic T_0, a function of any other sufficient statistic, then basing our inference on T_0 will automatically satisfy SP with respect to *any* choice of sufficient statistic. It is well known that, subject only to mild regularity conditions, such a smallest ("minimal") sufficient statistic can generally be found. Hence it is pretty straightforward to satisfy SP simultaneously with respect to every sufficient statistic: simply base inference on the minimal sufficient statistic.

The case of ancillarity appears very similar, though with the functional ordering reversed. Suppose S_1 is ancillary, and, in accordance with CP, we are basing inference on the conditional distribution of the data, given S_1. If now S_2 is a function of S_1, then S_2 is also ancillary; and the property that inference is conditioned on S_2 is automatically inherited from this property holding for the "larger" statistic S_1. This analysis suggests that — in close analogy with the case of the minimal sufficient statistic — we should aim to identify a *largest* ancillary statistic S_0, of which every ancillary statistic would be a function. Then conditioning on S_0 would automatically satisfy CP, simultaneously with respect to every choice of ancillary statistic.

2 Enter Basu

The above analysis appears unproblematic, and might be thought to make a compelling case for always conditioning on the largest ancillary statistic — an apparently straightforward enterprise. But then along comes Basu, and suddenly things are not so clear!

Basu presented theory and counter-examples to show that in general there is *no* unique largest ancillary statistic, conditioning on which would allow us to apply the conditionality principle unambiguously. Typically there will be a multiplicity of ancillary statistics that are maximal, *i.e.*, cannot be expressed as a non-trivial function of any other ancillary; and in this case no single largest ancillary can exist. Even in what would seem to be the simplest special case, of two independent observations from the normal location model, having $X_i \sim \mathcal{N}(\theta, 1)$ $(i = 1, 2)$, there is no largest ancillary: for any $c \in [-\infty, +\infty]$, the statistic S_c defined as $X_1 - X_2$ if $X_1 + X_2 > c$, $X_2 - X_1$ otherwise, is

ancillary (Basu 1959). But knowing S_c for all c we can recover the full data (X_1, X_2) — which is clearly *not* ancillary. This possibility arises because two statistics can each be marginally ancillary, while not being jointly ancillary. Another example of this phenomenon occurs for the bivariate normal distribution with standard normal marginals and unknown correlation coefficient: the data on either variable singly are ancillary, but this clearly fails when both variables are combined.

Some have argued that such examples merely show that we should not have abandoned Fisher's original conception that an ancillary statistic should itself be a function of the minimal sufficient statistic — which does not hold in the above examples. But Basu has other examples that are not subject to this criticism. One simple example (Basu 1964) involves the outcome X of a single throw of a die, where, for some value of $\theta \in [0, 1]$, the probabilities of obtaining the scores 1–6 are respectively, $(1/12) \times (1 - \theta, 2 - \theta, 3 - \theta, 1 + \theta, 2 + \theta, 3 + \theta)$. Then X itself is minimal sufficient, but there are 6 non-equivalent maximal ancillary functions of X: for example, one such is $Y_1 = 0$ if $X = 1$ or 4, $Y_1 = 2$ if $X = 2$ or 5, $Y_1 = 3$ if $X = 3$ or 6; another is $Y_6 = 0$ if $X = 1$ or 6, $Y_6 = 2$ if $X = 2$ or 5, $Y_6 = 3$ if $X = 3$ or 4. We thus have a choice of ancillaries to condition on. For any such choice, the conditional distribution of X is confined to two possible values, so looking like a biased coin-toss; but the bias is a different function of θ in each case, and there is no clear reason to prefer one of these choices rather than another. Since there is no largest ancillary here, the simple interpretation of the conditionality principle, as enjoining us to make inference in the reference set obtained by conditioning on any ancillary, appears non-operational. Attempts — *e.g.*, Cox (1971), Kalbfleisch (1975) — have been made to restrict CP to apply only to certain ancillaries, but these are either unconvincingly *ad hoc* or fail fully to resolve the difficulty.

In the presence of a choice of maximal ancillaries to condition on, CP could nevertheless be rescued if the conditioning in fact had no effect (or had the same effect in all cases). In another strand of his work, Basu found conditions for this to hold. Thus let T be a *complete* (and hence also minimal) sufficient statistic. Basu (1955) showed that T must be independent of any ancillary statistic, for any value of the parameter. It follows that any inference based only on the marginal distribution of T, which of course respects SP, will also respect CP, with respect to any possible choice of ancillary statistic. This applies, for example, in the example above of two observations from the normal location model, in which the minimal sufficient statistic $\overline{T} = \frac{1}{2}(X_1 + X_2)$ is complete, hence independent of any ancillary (including S_c), so that any inference based on T alone will automatically satisfy CP. However, the general usefulness of this construction is limited, since in many problems (such as the biased die example above) the minimal sufficient statistic is not complete, and its conditional distribution does depend on which ancillary is conditioned on — so that this particular escape route is blocked off.

A related result (though with less direct relevance for CP) is that, under additional conditions, any statistic which is independent of a sufficient statistic is ancillary. In Basu (1955) this was asserted as true without further conditions; a counter-example and corrected version were given in Basu (1958).

3 Ancillarity and likelihood

In the light of these depressing results, one might reach the depressing conclusion that, however appealing CP may seem, there is no general way of satisfying it. And that conclusion is essentially correct if we take a frequentist approach to inference, since we end up with entirely different sample spaces, with entirely different properties, by conditioning on different ancillaries. However, this does not mean that there is no way of making inferences that respect CP. For example, suppose we agree, along with Fisher, that the message in the observation $X = x_0$ is entirely encapsulated in the likelihood function for the parameter θ that this observation generates: $L_0(\theta) \propto \text{Prob}(X = x_0 \mid \theta)$ (we here suppose, purely for simplicity of exposition, that the sample space is discrete). For any ancillary

statistic $S = s(X)$, the conditional probability of $X = x$, given $S = s := s(x)$, is $\text{Prob}(X = x \mid \theta)/\text{Prob}(S = s)$, where the denominator does not involve θ by ancillarity of S; so, for any data x_0 the likelihood function computed in this conditional frame of reference, $L_0^*(\theta) \propto \text{Prob}(X = x_0 \mid S = s_0; \theta)$ (with $s_0 = s(x_0)$), will be identical with the full-data likelihood function, $L_0(\theta) \propto \text{Prob}(X = x_0 \mid \theta)$. It follows that any inference that is based purely on the properties of the observed likelihood function — for instance, the maximum likelihood estimate, the curvature of the log-likelihood at its maximum, a Bayesian posterior based on a fixed prior distribution,... — will be entirely unaffected if we condition on an ancillary statistic, and hence will automatically satisfy CP.

One of the most significant results in this area was Birnbaum's theorem (Birnbaum 1962). This showed that any general method of inference about some parameter θ, applicable across a range of experimental setups, will satisfy both the sufficiency and the conditionality principles if and only if it depends only on the observed likelihood function — *i.e.*, it satisfies the likelihood principle, LP. Basu's investigations led him down this same path, and he did fully accept LP. However, even this was not enough for him, and in many of his works — *e.g.*, Basu (1975); Basu (1977) — he argued that the only sensible way of satisfying LP is to be, or at least act like, a Bayesian with a fixed proper prior distribution. But that is another story.

References

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377–80.

Basu, D. (1958). On statistics independent of sufficient statistics. *Sankhyā*, **20**, 223–6.

Basu, D. (1959). The family of ancillary statistics. *Sankhyā*, **21**, 247–56.

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

Basu, D. (1975). Statistical information and likelihood. *Sankhyā, Series A*, **37**, 1–71.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, **72**, 355–66.

Birnbaum, A. (1962). On the foundations of statistical inference (with Discussion). *Journal of the American Statistical Association*, **57**, 269–326.

Cox, D. R. (1971). The choice between alternative ancillary statistics. *Journal of the Royal Statistical Society, Series B*, **33**, 252–252.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–25.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A*, **144**, 285–307.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, **62**, 251–9.

Springer