A Bayesian Look at Nuisance parameters

A. P. DAWID
The City University, London

SUMMARY

The elimination of nuisance parameters has classically been tackled by various *ad hoc* devices, and has led to a number of attempts to define partial sufficiency and ancillarity. The Bayesian approach is clearly defined. This paper examines some classical procedures in order to see when they can be given a Bayesian justification.

Keywords ANCILLARY; BAYESIAN INFERENCE; D-ANCILLARY; D-SUFFICIENT; G-SUFFICIENT; NUISANCE PARAMETER; S-ANCILLARY; S-SUFFICIENT; SUFFICIENT.

1. INTRODUCTION

Problems with nuisance parameters, where we are interested in only a part of the parameter that governs the distribution of our data, are of prime practical importance, yet our theoretical understanding remains limited and confused. One attractive approach is to simplify the model, by reducing the data or by conditioning on some statistic. Attempts to justify such simplification may be based on generalization of the concepts of sufficiency and ancillarity, but this generalization may be made in many ways. Another approach is to be, or to act like, a Bayesian, and integrate out any unwanted parameters. This in itself leads to a particular form of generalized sufficiency and ancillarity which, while of little direct interest to the Bayesian, is useful as a standard for judging other definitions.

In this paper we use examples and theory to indicate both the similarities and the differences between the Bayesian and classical approaches. Section 2 introduces nuisance parameters. Section 3 and 4 describe the Bayesian approach to generalized sufficiency, and Section 5 some classical definitions. In Section 6 we illustrate various possible ways in which these properties can hold, alone or together. Section 7 introduces specialized Bayesian versions of generalized sufficiency and ancillarity which are of particular relevance for comparison with the classical approach, and Section 8 takes up this comparison. In particular, it is shown that, under certain conditions, the classical approach can be given a Bayesian justification only for very special prior distributions.

Notation.

A capital letter will normally be used to denote an uncertain quantity (random variable or parameter), and the corresponding small letter for a realized or hypothetical value. However, this convention is not rigid. We use the symbols f, $\bar{f}$ and π to denote probability densities, leaving the relevant variables to be understood from the context: thus $f(t|s,\theta)$ denotes the density at t for the distribution of T, conditional on $S = s$, when $\Theta = \theta$. Our manipulations with such densities will be informal and far from rigorous, though all can be made precise. Thus $f(x|\lambda) = f(t|\theta)f(x|t,\phi)$ means that the parameter Λ is equivalent to the pair (Θ,Φ); that the marginal distributions of T depend on Λ through Θ alone; and that the conditional distributions of X given T depend on Φ alone. Such concepts can be conveniently and accurately expressed using the notation of *conditional independence* (Dawid, 1979a): the above properties would read: $T \perp \Lambda \mid \Theta$, $X \perp \Lambda \mid (T,\Phi)$. However, as this notation is still relatively unfamiliar, it has been avoided in this paper.

2. NUISANCE PARAMETERS

Suppose we are interested in the value of some unknown quantity Θ, (which, like all other abstract quantities we shall consider, may have several components) and can conduct a statistical investigation to learn about Θ. The outcome of this investigation will be our data x, the realised value of a random variable X.

If we are fortunate, the distribution of X will be completely determined by the value of Θ; this is the state of affairs treated in greatest depth in the inference text-books. However, in most real problems such simplicity is an unattainable ideal, even after we have made simplifying assumptions, such as normality, in setting up the model. Instead, the distributions of X might be governed by a parameter Λ, which is in some way connected with Θ. The most common case, to which we shall restrict our attention in this paper, is that Θ gives only a partial description of the distributions of X, so that Θ is a function of Λ.

The usual approach to such a problem is to introduce a further parameter Φ which, combined with Θ, completes the specification of the distribution of X. Then the pair (Θ, Φ) may be taken to be Λ. For example, if our experiment consists of an unbiased measurement of Θ, where the measuring instrument is subject to a normally distributed error of unknown variance, it would be usual to take Φ to be this variance. In such a case, Φ would be designated as "the nuisance parameter", and inference about Θ becomes "elimination of Φ". This seems a natural and obvious stance, but it should be pointed out that there is an arbitrariness involved in the choice of nuisance parameter. For instance, in the above case, why not take, for Φ, the *coefficient of variation* of the distribution? There is, indeed, a whole host of possible choices of the nuisance parameter. For some purposes (in particular, Bayesian inference) this will make no difference to our inference about Θ but, as we shall see, it may frequently be important to recognise the arbitrary nature of the nuisance parameter.

3. THE BAYESIAN APPROACH

A coherent Bayesian B has no conceptual difficulty in making inference about Θ in the presence of nuisance parameters. The distributions of X depend only on Λ, so that the observation $X = x$ provides a likelihood function, $f(x|\lambda)$ say, for Λ. To use this coherently requires a prior distribution for Λ, which B can specify. He now derives, in the usual way, his posterior distribution for Λ and, being interested in Θ alone, simply summarises his posterior opinions about Θ by means of the implied marginal posterior distribution for Θ.

No specification of a nuisance parameter is neccesary for this calculation, and if such a choice is made —as it normally will be— it is for convenience alone. For example, knowledge of the real world problem at hand may often make it possible to choose a Φ for which it would be reasonable to take Θ and Φ as *a priori* independent. In the measurement problem of the previous section, this might pick out the variance, rather than the coefficient of variation; but is easy to think of similar problems, with the same normal family of distributions, where this preference might be reversed. In any event, such a choice of nuisance parameter serves merely to simplify the psychological problem of specifying one's prior distribution, and is in no way essential to the statistical analysis.

In general, for any choice of Φ, B's distribution of Λ can be re-expressed as a joint distribution for (Θ,Φ), which can then be decomposed into the marginal distribution for Θ (which may be easy to specify, and will not depend on which Φ is used) and a conditional distribution for Φ given Θ (which may not, and of course will). Representing parameter-densities by the symbol π, Bayes'theorem gives

$$\begin{aligned} \pi(\theta,\phi|x) &\propto \pi(\theta,\phi)f(x|\theta,\phi) \\ \text{and so}\quad \pi(\theta|x) &\propto \int f(x|\theta,\phi)\pi(\theta,\phi)d\phi \\ &= \int f(x|\theta,\phi)\pi(\theta)\,\pi(\phi|\theta)d\phi \\ &= \bar{f}(x|\theta)\,\pi(\theta) \end{aligned}$$

where $\bar{f}(x|\theta) = \int f(x|\theta,\phi)\;\pi(\phi|\theta)d\phi$ gives the density of *B's* coherent distributions for X given only that $\Theta = \theta$, which we shall call the *marginal model* (for B). The marginal model does not depend on the choice of nuisance parameter Φ. As a function of θ, $\bar{f}(x|\theta)$ is known as the *marginal likelihood* of Θ, based on data $X = x$.

From the point of view of the single Bayesian B, the marginal likelihood is as good as any ordinary likelihood, but there are differences so far as the whole scientific community is concerned. There will normally be a good measure of agreement about the full model $f(x|\theta,\phi)$ (is not this what we really mean by a model?); but the marginal model $\bar{f}(x|\theta)$ is constructed by an operation involving *B's* subjective opinions, through $\pi(\phi|\theta)$, and so does not appear to share the objectivity of $f(x|\theta,\phi)$.

Armed with the marginal model, we can consider such concepts as sufficiency and ancillarity in it. We shall call T *marginally sufficient* for Θ(for B), if it is sufficient in the marginal model, and similarly for marginal ancillarity. Note that these concepts depend on the prior distribution, but only through the *conditional* distributions for Λ given Θ, the marginal prior distribution for Θ being arbitrary. Thus a collection $\boldsymbol{B}$ of Bayesians, with various prior distributions $\{\Pi_B : B \in \boldsymbol{B}\}$ for Λ, will all agree on the marginal model, and so agree whether or not a statistic T is marginally sufficient or ancillary, as long as they agree on the model and on the distributions of Λ given Θ (in which case we shall call $\boldsymbol{B}$ a *bevy* of Bayesians). An alternative statement of this last condition is that, for the family $\{\Pi_B\}$ of distributions, regarded as a model with "data" Λ and "parameter" B, Θ is a sufficient "statistic".

4. MARGINAL SUFFICIENCY

Suppose T is marginally sufficient for Θ (for B). Then $\bar{f}(x|\theta)$ has the form $a(x)\bar{f}(t|\theta)$ where $\bar{f}(t|\theta)$ is the marginal density of T given $\Theta = \theta$. Thus $\pi(\theta|x) \propto \pi(\theta)\ \bar{f}(x|\theta) \propto \pi(\theta)\ \bar{f}(t|\theta)$, whence $\pi(\theta|x) = \pi(\theta|t)$, and *B's* posterior marginal distribution for Θ depends on T alone, just as in the case of ordinary sufficiency with no nuisance parameters. Under some regularity conditions, the converse will hold. Our definition is therefore in accord with those of Raiffa and Schlaifer (1961) and Lindley (1965).

In a sense, marginal sufficiency is unimportant: B will get the same posterior distribution for Θ whether he bases it on the complete data X or on T alone, and for this very reason there is little point in his reducing his data to T before processing. However, it is often necessary to discard some data in the interests of manageability, and if B knows that he can do this in such a way that he loses no information about Θ, so much the better.

This raises the question: How can B know that T is marginally sufficient? It seems that he must first either evaluate his posterior for Θ, and discover its dependence on T alone, in which case it is too late to use the knowledge, or else calculate the marginal model, which seems to be as laborious as a full analysis. However, as we shall see, B may be able to profit from certain special structure in his model and prior to deduce that a statistic is marginally sufficient.

5. GENERALIZED SUFFICIENCY

It is not only, nor indeed primarily, the Bayesian who is motivated to simplify his problem of inference about Θ by discarding data. One obvious motivation for reducing the data to some statistic T is the possibility of eliminating nuisance parameters by satisfying the following definition:

Definition 5.1 (Basu, 1977). A statistic T is *Θ-oriented* if its sampling distribution is entirely determined by the value of Θ.

However, this property does not in itself justify one in discarding all the data but T, since one might be throwing away information relevant to inference about Θ. The Bayesian has, in marginal sufficiency, a coherent theory to tell him when he can reduce his data without essential loss. From the classical point of view, a variety of *ad hoc*, more or less intuitively reasonable ideas has been put forward, intended to identify properties of sampling distributions which serve to justify such reduction of the data.

A good account of these ideas is given by Barndorff-Nielsen (1978, Chapter 4). (See also Basu, 1977, 1978; Dawid, 1975) We shall concentrate on just two approaches, specializing Barndorff-Nielsen's definitions slightly.

5.1 *G-sufficiency*

This concept was introduced by Barnard (1963). The essence is as follows. Let the model be given by the family $P = \{P_\lambda\}$ of distributions for data X, and suppose these distributions are equivariant under the action of exact homomorphic transformation groups, G, acting on X, and $\bar{G}$ acting on Λ. That is to say, if $X \sim P_\lambda$, and $g \in G$, then $g \circ X \sim P_{\bar{g} \circ \lambda}$ (for further background see, for example, Dawid, Stone and Zidek, 1973).

Suppose the parameter of interest Θ is invariant under $\bar{G}$, so that $\Theta(\lambda) \equiv \Theta(\bar{g} \circ \lambda)$, and let T be the maximal invariant function of X under G. Then Barnard proposed that, in the absence of prior information, T should be regarded as containing all the available information about Θ. Such a statistic T is termed *G-sufficient* for Θ. It can be shown (see e.g. Lehmann, 1959, p.220) that, if Θ is a *maximal* invariant function of Λ under $\bar{G}$, then a G-sufficent statistic T will be Θ-oriented.

Example 5.1. Let $\mathbf{X} = (X^i : i = 1, \ldots, n)$ be a random sample from $N(\mu, \sigma^2)$. Take G as the additive group of real numbers, a typical element a taking $\mathbf{X}$ into $\mathbf{X} + a\mathbf{1}$; then we way may take $\bar{G} = G$, with $a \circ (\mu, \sigma^2) = (\mu + a, \sigma^2)$. A maximal invariant statistic is $\mathbf{X} - \bar{X}\mathbf{1} = (X^i - \bar{X} : i = 1, \ldots, n)$ (where $n\bar{X} = \Sigma_{i=1}^{n} X^i$), which is thus G-sufficient for the invariant parameter σ^2.

Further reduction is possible using ordinary sufficiency, either in the full or in the reduced model. Either way, this yields the statistic $\Sigma(X^i\text{-}\bar{X})^2$ in the above example, as containing all the available information about σ^2 in the absence of prior knowledge (about μ, in particular).

Example 5.2 Let $\mathbf{X}^i$ ($i=1, \ldots, n$) be a random sample from the bivariate normal distribution with entirely unknown mean-vector and dispersion matrix. Let G consist of the group of location-scale transformations acting on each component separately (but identically for all i). After reduction by sufficiency, this yields the sample correlation coefficient as G-sufficient for its population counterpart.

Example 5.3. Sample Survey. Consider a sampling frame of labelled units, denoted by $i=1,2, \ldots ,m$. With unit i is associated an unknown quantity Y_i, and we take as our parameter Λ the ordered set $(Y_1, Y_2, \ldots , Y_m)$. The sampling scheme is determined by a known probability distribution P yielding S, a random subset of $\{1,2, \ldots ,m\}$, and the data consist of $X = \{(i, Y_i):i\epsilon S\}$.

Let G and $\bar{G}$ each be isomorphic to the group of permutations of $(1,2, \ldots , m)$, acting on data $x = \{(i,y_i):i\epsilon S\}$ as $g\circ x = \{(g^{-1}i, Y_i): i \epsilon S\}$, and on parameter $\lambda = (y_1, \ldots , y_m)$ as $\bar{g}\circ\lambda = (y_{g1}, \ldots , y_{gm})$. The sampling distributions are equivariant under G and G if and only if, under P, all subsets of the same size are equally probable; that is to say, for *simple random sampling* with a possibly random sample size. We have maximal invariants: $T=$the order statistic of $(Y_i:i\epsilon S)$, and Θ = the order statistic of $(Y_1, \ldots, Y_m)$, and thus, under simple random sampling, T is G-sufficient for Θ.

An ancillary statistic based on T is N, the size of sample taken, and the conditional distribution of T given (N,Θ) is a multivariate hypergeometric distribution.

Example 5.4. (Schou, 1978). Let $\mathbf{X}^i(i=1, \ldots, n)$ be a random sample of unit vectors in $\mathbf{R}^2$ drawn from the Fisher-von Mises distribution on the circle. The parameter Λ can take any value in $\mathbf{R}^2$, and the model densities have the form $c(\|\lambda\|) \exp (\lambda' \mathbf{x}) (\|\mathbf{x}\| = 1)$. A sufficient statistic is $\mathbf{S}_n = \Sigma_{i=1}^n \mathbf{X}^i$.

A typical element of G rotates each $\mathbf{X}^i$ about $\mathbf{0}$ through the same angle α, and has the same effect on both $\mathbf{S}_n$ and Λ. After a sufficiency reduction to $\mathbf{S}_n$, the maximal invariant is $\|\mathbf{S}_n\|$, which is thus G-sufficient for the maximal invariant parameter $\|\Lambda\|$.

5.2. S-sufficiency

Let T be a statistic. The experiment which yields observation of X, with model densities $f(x\,|\,\lambda)$, can be regarded as made up of two components:

(i) the *reduced experiment*, yielding observation of T, with model densities $f(t\,|\,\lambda)$ derived by marginalization from $f(x\,|\,\lambda)$; and

(ii) the *conditional* experiment, after observing $T = t$, yielding observation of X but with model densities $f(x\,|\,t,\lambda)$ derived from $f(x\,|\,\lambda)$ by conditioning on T.

We suppose T is Θ-oriented, and try to express the fact the conditional experiment, discarded on reduction to T, contains no useful information about Θ. One such expression is the requirement that the conditional experiment is determined entirely by nuisance parameters. Because of the arbitrary nature of nuisance parameters, this may be interpreted *either* in terms of some nominated choice of nuisance parameter, *or* as a requirement that there exist *some* choice of nuisance parameter yielding this property. For non-triviality in this latter case we must impose some restrictions (otherwise Λ itself might be regarded as a nuisance parameter!) and this motivates the insistence that Θ and the nuisance parameter Φ should be *variation-independent*: that is to say, as Λ varies over its range of possible values, Θ and Φ range over a product-space. Thus the property of S-sufficiency may be expressed as:

$$f(x\,|\,\lambda) = f(t\,|\,\theta)f(x\,|\,t,\phi) \qquad (5.1)$$

where Θ and Φ are variation-independent.
[Note that this property does *not*, in general, hold for G-sufficiency]

Example 5.5 Let X_1, X_2 have independent Poisson distributions with respective means Λ_1, Λ_2 known only to be positive. We are interested in $\Theta = \Lambda_1 + \Lambda_2$. Then $T = X_1 + X_2$ is Θ-oriented, and is in fact S-sufficient for Θ; for the conditional distribution of X given $T = t$ is Binomial $B(t;\Phi)$, where $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2)$ is variation-independent of Θ.

A trivial case of S-sufficiency arises when $X = (T,S)$, $\Lambda = (\Theta,\Phi)$ (Θ,Φ variation-independent), and

$$f(t,s\,|\,\theta,\phi) = f(t\,|\,\theta)f(s\,|\,\phi). \qquad (5.2)$$

Then the experiments producing T and S may be considered as entirely unrelated to each other.

Example 5.6. Components of variance. The data are (X_{ij}: $i = 1, \ldots, I; j = 1, \ldots, J$), generated as

$$X_{ij} = \mu + \tau Y_i + \sigma Z_{ij}, \qquad (5.3)$$

where the Y's and Z's are independent standard normal variables, and $(\mu,\ \tau^2,\ \sigma^2)$ the value of the parameter. A minimal sufficient statistic is (S_1,S_2,S_3), where

$$S_1 = X_{..},\ S_2 = J\,\Sigma^{I}_{i=1}\ (X_{i.}-X_{..})^2,\ S_3 = \Sigma^{I}_{i=1}\ \ \Sigma^{J}_{j=1}\,(X_{ij}-X_{i.})^2,$$

and where the dot operator averages over the replaced suffix.

In the sampling distribution, S_1, S_2 and S_3 are mutually independent, with $S_1 \sim N(\mu,\ \sigma_0^2/IJ)$, $S_2 \sim \sigma_0^2\ \chi^{2}_{I-1}{}'$, and $S_3 \sim \sigma^2\,\chi^2{}_{I(J-1)}$, where $\sigma_0^2 = \sigma^2 + J\tau^2$. Thus taking $\Theta = \sigma^2$, $\Phi = (\mu,\ \sigma_0^2)$, $T = S_3$, $S = (S_1,\ S_2)$, we have the factorization (5.2). However, Θ and Φ will not normally be variation-independent , since (with $\tau^2 \geq 0$) we must have $\sigma_0^2 \geq \sigma^2$, and it therefore seems that information in S may be relevant to σ^2. There are two ways in which we can get variation-independence: (1) restrict the parameter-space, for example requiring $\sigma^2 \leq a$ and $\sigma_0^2 \geq b\ (\geq a)$, μ being unrestricted; or (2) extend the parameter space to allow $\tau^2 < 0$. (This condition makes sense if interpreted in terms of the covariance structure of the (X_{ij}), rather than the synthetic representation (5.3): Dawid, 1977; we can then allow any combination of μ, $\sigma^2 > 0$, $\sigma_0^2 > 0$).

The former approach appears to distort the real problem to fit the Procrustean bed of theory, and in any case the appropriate implied parameter-space for $(\mu,\tau^2,\ \sigma^2)$ will depend on the value of J. The latter approach may or may not be regarded as appropriate, leading as it does to the possibility of negative correlations between the $(X_{i.})$, and again involving the value of J.

The above problem is the subject of Stone and Springer (1965).

6. EXAMPLES OF MARGINAL SUFFICIENCY

Marginal sufficiency may or may not go hand in hand with its various classical counterparts, as the following examples illustrate.

Example 6.1. Full sufficiency. If T is sufficient for the full parameter Λ, then$T$ is marginally sufficient for Θ for *any* prior distribution on Λ. Moreover, usually the converse will hold (Håjek, 1965; Martin, Petit et Littaye, 1973).

Example 6.2. G-sufficiency. In the model of 5.1, consider the family F of prior distributions for Λ which are *invariant* under $\bar{G}$; thus if $\Pi \in \mathbf{F}$, $g \in \bar{G}$, then $\Lambda \sim \Pi \Rightarrow \bar{g} \circ \Lambda \sim \Pi$. By general results on invariance (see e.g. Dawid, 1979a, Section 8), Θ is a "sufficient statistic" in F, so that F corresponds to a bevy of Bayesians, and thus leads to an agreed marginal model for X given Θ. It may now be seen that any of these distributions for X given Θ is invariant under the action of G on X, whence T is sufficient for this marginal model, and hence marginally sufficient. Thus all Bayesians in the bevy would agree to work with the marginal model for the reduced data T, and since T is, in any case, Θ-oriented, this is equivalent to using the sampling distributions of the G-sufficient statistic T.

In the context of Example 5.4, suppose that the prior distribution for Λ is rotationally symmetric about $\mathbf{0}$ (as a particular case, Λ_1 and Λ_2, might have independent standard normal distributions); then the posterior distribution of $\Theta = \|\Lambda\|$ will be a function of $T = \|\mathrm{S}_n\|$ alone, and could be derived by combining the marginal prior of Θ with the (Θ-oriented) reduced experiment for T.

Likewise, in Example 5.3 with simple random sampling, if in the prior distribution the variables $(Y_1,\ldots,Y_m)$ are *exchangeable* (which means, simply, invariance under the group $\overline{G}$ of permutations), then the order statistic T of the data will be marginally sufficient for the order statistic Θ of the parameter, and coherent inference could be based on its multivariate hypergeometric sampling distribution (for given sample size).

The general theory developed above is of somewhat limited applicability. A proper $\overline{G}$-invariant distribution exists only when $\overline{G}$ is compact as a topological group. Usually this condition does *not* hold; it fails, for instance, in Examples 5.1 and 5.2. Then $\overline{G}$-invariant measures exist, but are improper distributions. Difficulties can now arise. For example, it is possible for the posterior distribution of Θ to depend on the data through T alone, but not to be derivable from the reduced experiment based on T. This is the *marginalization paradox* of Dawid, Stone and Zidek (1973). Such problems do not arise for proper priors.

A difficult technical problem is to discover whether a G-sufficient statistic can be marginally sufficient for a non-invariant prior distribution, and, in particular, for a proper prior in the case of a non-compact group. Case studies suggest that this will not normally be possible (Jaynes, 1980). If so, then reduction to a G-sufficient statistic, when the group is not compact, will be intrinsically incoherent, in the sense that the only prior distributions which allow such reduction are improper, and possibly paradoxical.

Example 6.3. S-sufficiency. Suppose (5.1) holds, and the prior distribution for Λ is such that Θ and Φ are independent. (Thus, so long as the parameter-space is redefined, if necessary, as the support of the prior distribution, Θ and Φ must be variation-independent). Then $\pi(\theta,\phi) = \pi(\theta)\,\pi(\phi)$, whence

$$\pi(\theta,\phi\,|\,x) \propto \pi(\theta)f(t\,|\,\theta)\;\;\pi(\phi)f(x\,|\,t,\phi). \qquad (6.1)$$

It follows that T is marginally sufficient for Θ, and the reduced experiment gives the marginal model.

In Example 5.5, suppose that we take a conjugate prior distribution: $\Lambda_i \sim \Gamma(a_i, b)$ independently. As is well known, this implies that $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2) \sim \beta\ (a_1, a_2)$, *independently* of $\Theta = \Lambda_1 + \Lambda_2 \sim \Gamma(a_1+a_2,b)$. It follows that $T = X_1 + X_2$ is marginally sufficient, so that inference for Θ follows on combining the reduced data T, having distribution $P(\Theta)$, with the marginal prior: $\Theta \sim \Gamma(a_1+a_2,\ b)$.

The above simplification is an important (but little-known) general property of conjugate inference for exponential families (Barndorff-Nielsen, 1978: Corollary 9.3). Under weak conditions, whenever a S-sufficient statistic T exists, yielding a factorization (5.1), then Θ and the nuisance parameter Φ will turn out to be independent, for any conjugate prior (where the term ''conjugate''is suitably defined). Thus conjugate Bayes inference about such a parameter Θ can always proceed in the reduced experiment.

For Example 5.6, S_3 will be marginally sufficient for σ^2 (and $(S_1,\ S_2)$ for $(\mu,\ \sigma_0^2)$) if σ^2 and (μ,σ_0^2) are *a priori* independent. Again, interpreted in terms of $(\mu,\ \tau^2,\ \sigma^2)$, this requirement cannot hold for more than one value of J, and so appears quite artificial.

Example 6.4. Complex sampling (Sugden, 1978). Suppose a sample survey is conducted as in Example 5.3, but with a complex sampling scheme which is not equivalent to simple random sampling. Consider again the family of exchangeable prior distributions, which constitute a bevy for inference about the order-statistic Θ of Λ, and hence yield an agreed marginal model for X given Θ. Once again, the order-statistic T of X is marginally sufficient for Θ; this follows because the posterior distribution does not depend on the sampling scheme, and since, for the particular case of simple random sampling, the posterior for Θ with an exchangeable prior depends on T alone, this must hold for any sampling scheme. Consequently, the bevy can confine itself to the reduced experiment for T.

Now in general T will not be Θ-oriented, and it would therefore seem that reduction of the data to T does not afford much simplification. However, it may be seen that simplicity returns if we work with the marginal model for T given Θ, as follows. Firstly, since sample-size N (a function of T) is ancillary in the full model, it is ancillary in the marginal model; and now a symmetry argument shows that, conditional on N, the marginal model for T will be multivariate hypergeometric, exactly as for simple random sampling.

Example 6.5. L-independence. (Barndorff-Nielsen, 1978, Example 3.8). Consider a birth and death process, with birth and death intensities Λ and M, observed continuously from time 0 to time T, in which there are initially ℓ individuals. Let B, D and Z denote respectively the number of births, the number of deaths, and the total time lived by all individuals. Then $(B,\ D,\ Z)$ is sufficient for (Λ, M), and the likelihood based on data (b, d, z) is proportional to

$$\lambda^b \mu^d e^{-(\lambda+\mu)z}. \tag{6.1}$$

Since this factorizes as a function of λ and μ, we call Λ and M *L-independent*, although (6.1) can *not* be produced by *S*-sufficiency, and is not of the form (5.1).

Suppose that Λ and M are *a priori* independent. Then $\pi(\lambda \mid \text{data}) \propto \pi(\lambda)$. $\lambda^b e^{-\lambda z}$, a very straightforward calculation. For inference about Λ, all the Bayesian has to do is to store the relevant factor of his likelihood and combine it with his prior.

Here $T = (B, Z)$ is marginally sufficient for Λ, but it would not be quite so straightforward to make inference about Λ from the reduced experiment, since (B, Z) is *not* Λ-oriented and has a complicated distribution. In this case, it does not help to derive the marginal model for (B, Z) given Λ, which is also complicated and depends on the distribution assigned to M.

The lesson here is that, even when a marginally sufficient statistic exists, it may not be most profitable to the Bayesian to work with its sampling distributions (in either the full or the marginal model); other uses may be more appropriate. The same moral is pointed by the next example.

Example 6.6. Optional stopping. Consider again the Fisher-von Mises distribution of Example 5.4, but with sequential observation of $\mathbf{X}^1$, $\mathbf{X}^2$, ..., stopping according to the following rule: if $\mathbf{X}^1$, $\mathbf{X}^2$, ... , $\mathbf{X}^r$ have been observed with values $\mathbf{x}^1$, ..., $\mathbf{x}^r$, then observations terminates if the first component x_1^r of $\mathbf{x}^r$ is negative; otherwise $\mathbf{X}^{r+1}$ is observed. This rule leads, with probability one for all Λ, to termination of observation at some random finite stage N.

The data may be expressed as $(n, \mathbf{x}^1, \ldots, \mathbf{x}^n)$, the observed values of $(N, \mathbf{X}^1, \ldots, \mathbf{X}^N)$. By a standard result on optional stopping, the posterior distribution for Λ will be identical with that based on observing values $(\mathbf{x}^1, \ldots, \mathbf{x}^n)$ for $(\mathbf{X}^1, \ldots, \mathbf{X}^n)$ in the non-sequential set-up of Example 5.4, for the appropriate value of n.

In particular, consider the bevy of prior distributions for Λ which are rotationally symmetric about $\mathbf{0}$. Then, by the results of Example 6.2, the posterior distribution of $\Theta = \| \Lambda \|$ will depend only in the value of $(N, \|\mathbf{S}_N\|)$ (the value of N, taken for granted as known earlier, must now be specified). As in the last two examples, the marginally sufficient statistic $(N, \|\mathbf{S}_N\|)$ will not in general be Θ-oriented (the non-invariant stopping rule destroys that property), and so the bevy might wish to focus attention on the marginal model for $(N, \|\mathbf{S}_N\|)$. It might be conjectured, in analogy with Example 6.4, that N is ancillary in this marginal model, and that conditioning on it produces the same distribution for $\|\mathbf{S}_N\|$ as in Example 5.4. However, N is not ancillary. For example, it may easily be seen that, for $\Theta = \mathbf{0}$ (which gives an uniform distribution on the circle), the distribution for N given Θ is geometric with probability parameter 1/2; while for Θ very large, corresponding to the $(\mathbf{X}^i)$ being tightly concentrated about the same random unit vector $\mathbf{e} = \Lambda/\Theta$,$N$ will tend to be either 1 (if $e_1 < 0$) or otherwise very large (if $e_1 > 0$); each extreme holding with probability about 1/2.

Consequently, conditioning the marginal model on N is inappropriate, and we do not recover the same reduced marginal model as for Example 5.4. It seems that our bevy cannot shortcut the complicated task of calculating the reduced marginal model.

However, this is, in reality, quite unnecessary. We know that posterior distributions will be identical with those for Example 5.4, which are easily found, so that use of the marginal model may be completely by-passed. Alternatively, we might say that it is in order to use an entirely fictitious model, in which $N = n$ is regarded as fixed and $(\mathbf{X}^1, ..., \mathbf{X}^n)$ drawn as a random sample of size n. Once again, we have a simple marginally sufficient statistic leading to a simple Bayesian inference, but it is not all helpful to work with sampling distributions.

7. D-SUFFICIENCY AND D-ANCILLARITY

The examples of Section 6 demonstrate that, even when a simple marginally sufficient statistic T exists, leading to a simple marginal posterior distribution for Θ, it may well not be fruitful for the Bayesian to concern himself with the sampling distribution of T. In particular, whether or not T is Θ-oriented will depend on irrelevant properties of the sample-space (compare Examples 6.4 and 6.6 with Example 6.2). Consequently, our next definition may be of little interest to the whole-hearted Bayesian.

Consider a model for data X, with parameter Λ, and a Bayesian B with prior distribution Π for Λ

Definition 7.1. A statistic T is *D-sufficient* for Θ (for B, or Π) if T is (i) marginally sufficient for Θ, for B, and (ii) Θ-oriented.

This definition is important for purposes of comparing Bayesian and classical concepts. In particular, we shall be examining the classical prescriptions for reduction to T, which do depend on the sample space and do, usually, have T Θ-oriented, to discover when they can be given a Bayesian justification.

From the classical viewpoint, there is another common way of eliminating nuisance parameters, namely by *conditioning*. This involves replacing the original experiment for X by the conditional experiment for X, given a statistic T. In parallel with reduction, this is motivated by the possibility of achieving the following simplification.

Definition 7.2. A statistic T is Θ-*inducing* if, for any t, the conditional experiment for X given $T = t$ is determined entirely by the value of Θ.

Using only the conditional experiment involves discarding the reduced experiment for T, and we therefore require criteria which allow us to do so without losing "useful information". These are entirely analogous to the criteria involved in discarding a conditional experiment, as already considered, and the two problems are in effect two faces of the same coin, labelled "nonformation" by Barndorff-Nielsen (1976, 1978).

We shall specifically consider the following criterion.

Definition 7.3. A (Θ-inducing) statistic T is *S-ancillary* for Θ if there exists a nuisance parameter Φ, variation-independent of Θ, which determines the reduced experiment for T (which is to say that T is Φ-oriented).

A S-ancillary statistic T gives rise to the factorization

$$f(x|\lambda) = f(x|t,\theta) f(t|\phi). \tag{7.1}$$

Comparing this with (5.1), we see that T is S-ancillary for Θ if and only if T is S-sufficient for Φ. Thus, in Example 5.5, $T = X_1 + X_2$ is S-ancillary for $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2)$, and this might justify basing inference about Φ on the conditional (binomial) model for X given T.

For the Bayesian, a generalized ancillarity criterion, which would allow him to work with a conditional model,rather than the full model , seems even less worthy of attention than generalized sufficiency, since he is not normally concerned with sampling models anyway, and in this case does not even gain, in general, by being able to discard data. One again, the following definition is of most importance for purposes of comparison between Bayesian and classical ideas.

Definition 7.4. A statistic T is *D-ancillary* for Θ (for B, or Π) if it is (i) marginally ancillary for Θ, for B, and (ii) Θ-inducing.

(Recall that "T is marginally ancillary for Θ" means that T is an ancillary statistic in the marginal model, so that $\bar{f}(t|\theta)$ does not depend on θ).

If T is D-ancillary for Θ, then $f(x|\lambda) = f(x|t, \theta) f(t|\lambda)$, whence $\bar{f}(x|\theta) = \int f(x|\lambda)\, \pi(\lambda|\theta)d\lambda = f(x|t,\theta) \int f(t|\lambda)\, \pi(\lambda|\theta)d\lambda = f(x|t,\theta) \bar{f}(t|\theta) \propto f(x|t,\theta)$. It follows that the posterior distribution for Θ satisfies $\pi(\theta|x) \propto f(x|t,\theta)\, \pi(\theta)$, and so can be found by combining the prior marginal distribution for Θ with the conditional model given T. Conversely, when T is Θ-inducing, marginal ancillarity of T is necessary for this property to hold. Thus D-ancillarity may be regarded as a Bayesian justification for working with the conditional model. Note once again that the definition involves only the *conditional* prior distribution for Λ given Θ, and so is relevant for the whole bevy of Bayesians sharing this conditional distribution, the marginal prior distribution for Θ being arbitrary.

Suppose T is S-ancillary for Θ, so that (7.1) holds. Trivially, if Θ and Φ are *a priori* independent, then $\pi(\phi \mid x) \propto \pi(\phi) f(x \mid t, \phi)$, so that T is marginally ancillary. The use of the conditional model is thereby justified if the prior independence holds. Again, it will normally hold for conjugate inference in exponential families.

The following example (from Dawid and Dickey, 1977) shows that prior independence is *not* necessary for a S-ancillary statistic to be D-ancillary.

Example 7.1. Suppose $f(x \mid \lambda) = f(x \mid t, \theta) f(t \mid \phi)$, where (Θ, Φ) takes values in $[-1,1] \times [-1,\frac{1}{2}]$. We need not specify $f(x \mid t, \theta)$, but suppose $f(t \mid \phi) = 2t^{-3}(1 + \phi t)/g(\phi)$ for $t \geq 1$, $-1 \leq \phi t \leq \frac{1}{2}$; 0 otherwise. The normalizing constant is

$$\begin{aligned} g(\phi) = {} & (1+\phi)^2 \quad (-1 \leq \phi \leq 0) \\ & (1+2\phi-8\phi^2) \quad (0 \leq \phi \leq \tfrac{1}{2}). \end{aligned}$$

In the prior, Θ and Φ are *not* independent, and in fact

$$\pi(\phi \mid \theta) = (4/3)(1-\theta\phi)\, g(\phi) \ (-1 \leq \phi \leq \tfrac{1}{2}).$$

(This does define a density, for any $\theta \epsilon [-1,1]$.)
We find $\bar{f}(t \mid \theta) = \int f(t \mid \phi)\, \pi(\phi \mid \theta)\, d\phi = 3t^{-4} \ (t \geq 1)$
$\qquad 0 \quad$ (otherwise)

so that T is both S- and D-ancillary for Θ.

(A similar example may be constructed to show that statistic T may be both S- and D-sufficient for Θ, although Θ and Φ are not *a priori* independent).

In the next Section we examine in more detail the connexions between S- and D-sufficiency and ancillarity.

8. S- AND D-NONFORMATION

The material of this Section draws heavily on Dawid and Dickey (1977).

The concepts of sufficiency and ancillarity being considered may usefully be expressed in the general framework of *conditional independence* (Dawid, 1979a), and our theorems below are applications of general properties of conditional independence to our specific problems. For further technical background and rigorous proofs, see Dawid (1980).

8.1. Ancillarity

Suppose we have S-ancillarity: $f(x|\lambda) = f(x|t,\theta)\,f(t|\phi)$. Suppose further that *T strongly identifies* Φ, as defined in Dawid (1980): that is to say, if we consider the marginal distribution of T induced by assigning a prior distribution to Φ, two different priors will induce distinct marginal distributions. This property is commonly known as "identification of mixtures" (Teicher, 1960, 1961, 1967; Barndorff-Nielsen, 1965; Chandra 1977). Clearly, strong identification implies ordinary identification.

Theorem 8.1. Under the above conditions, T is D-ancillary for $\Theta \Leftrightarrow \Theta$ and Φ are *a priori* independent.

Proof. We have already shown "$\Leftarrow$". For "$\Rightarrow$", we note that $\bar{f}\,(t|\theta) = \int f(t|\phi)\ \pi(\phi|\theta)d\phi$, and marginal ancillarity gives that $\bar{f}(t|\theta_1) = \bar{f}(t|\theta_2)$ for any θ_1, θ_2. Since $\bar{f}(t|\theta)$ is a mixture of $f(t|\phi)$ with mixing measure $\pi(\phi|\theta)$, strong identification implies that $\pi(\phi|\theta_1) = \pi(\phi|\theta_2)$, so that we have independence.

We can summarize this result as saying that, with the strong identification property, use of S-ancillarity to allow inference from the conditional model is coherent (i.e. has a Bayesian justification) if and only if Θ and Φ are *a priori* independent; more informally, it is necessary and sufficient that Θ and Φ each carry no information about the other.

The next result gives conditions on the prior distribution, not involving the model, under which S- and D-ancillarity can *never* co-exist.

Theorem 8.2. Suppose that the prior conditional distributions of Φ given Θ, considered as a parametric family, are *boundedly complete*; that is, if $h(\Phi)$ is bounded with $\mathrm{E}[h(\Phi)|\Theta] = 0$ a.s., then $h(\Phi) = 0$ a.s. If (7.1) holds, then T is *not* D-ancillary for Θ.

Proof. Suppose the contrary, and let $k(T)$ be a bounded function. Since T is Φ-oriented, $\mathrm{E}[k(T)|\Theta,\Phi] = \mathrm{E}[k(T)|\Phi] = h(\Phi)$ say. Then $\mathrm{E}[h(\Phi)|\Theta] = = \mathrm{E}[k(T)|\Theta] = \mathrm{E}[k(T)]$ a.s. since T is marginally ancillary. So by bounded completeness $h(\Phi) = \mathrm{E}[k(T)]$ a.s., i.e. $\mathrm{E}[k(T)|\Theta,\Phi] =$ constant a.s. As this holds for any k, T must be independent of (Θ,Φ), so that T is in fact ancillary, and so cannot be Θ-inducing (barring the trivial case that X is Θ-oriented).

8.2. Sufficiency

Suppose we have S-sufficiency: $f(x|\lambda) = f(t|\theta)f(x|t,\phi)$. We look for a result analogous to Theorem 8.1.

Theorem 8.3. If, for each value of t, the distributions of X given $T = t$ strongly identify their parameter Φ, then T marginally sufficient for $\Theta \Rightarrow \Theta$ and Φ are independent in their distribution posterior to observing T.

The proof parallels that of Theorem 8.1.

Under the strong identification condition of Theorem 8.3, the distribution of Φ given (T,Θ) does not depend on Θ. Also, since T is Θ-oriented, T is independent of Φ given Θ, so that the distribution of Φ given (T,Θ) does not depend on T. We appear to have shown that Φ is independent of (T,Θ), and thus that Θ and Φ must be independent *a priori*. However, this reasoning is fallacious without further conditions (Dawid, 1979b).

Example 8.1. The parameter is (Θ,Φ) with $\Phi > 0$, $\Theta \neq 0$. The data are $(S,T) = (Y/\Phi, Z/\Theta)$, where Y and Z have independent standard exponential distributions, with density $f(y) = e^{-y}(y>0)$. We thus have "unrelated problems". Given T, the data X reduce to S, with distribution unchanged, and S strongly identifies Φ, by the uniqueness property of the Laplace transform.

Suppose the prior distribution has

$$\begin{aligned}\pi(\phi|\theta) &= e^{-\phi}\ (\phi>0) \text{ when } \theta > 0\\ &\quad 2e^{-2\phi}(\phi>0) \text{ when } \theta < 0.\end{aligned}$$

Then T is D-sufficient for Θ; indeed, we may take

$$\begin{aligned}\bar{f}(s|t,\theta) &= (1+s)^{-2}\ (s>0) \text{ when } t > 0\\ &\quad 2(2+s)^{-2}\ (s>0) \text{ when } t < 0\end{aligned}$$

independently of θ. However, Θ and Φ are *not* independent in the prior distribution.

The further condition needed to ensure the validity of our informal argument above is the non-existence of a set A for which $P(T \in A|\theta)$ is always 0 or 1, both values being taken as θ varies. Such a set is called a *splitting set* for T given Θ (Koehn and Thomas, 1975). In Example 8.1, the positive half-line is such a splitting set.

We thus have the following result.

Theorem 8.4. Suppose T is S-sufficient for Θ and that there does not exist a splitting set for T given Θ. Suppose further that, for each value of t, the distributions of X given $T = t$ strongly identify the nuisance parameter Φ. Then T is D-sufficient for Θ if and only if Θ and Φ are *a priori* independent.

Thus, under appropriate conditions on the model, reduction by S-sufficiency is "coherent" if and only if Θ and Φ are *a priori* independent. (Note that this result, in common with Theorem 8.1, does not use the property that Θ and Φ be variation-independent).

Example 8.2. In the components of variance problem of Example 5.6, take $\Theta = (\mu,\sigma_o^2)$, $\Phi = \sigma^2$, $T = (S_1, S_2)$, $S = S_3$. Then the conditions of Theorem 8.4 hold, so that inference for (μ,σ_o^2) based on (S_1,S_2) alone is coherent if and only if (μ,σ_o^2) is *a priori* independent of σ^2: a condition which, as indicated earlier, is unrealistic. The same condition is necessary and sufficient for coherent inference about σ^2 based on S_3 alone.

In this example, interest may well centre on μ alone, so that reduction to (S_1,S_2) would not eliminate all nuisance parameters. It seems likely that (S_1,S_2) will be marginally sufficient for μ (although not, of course, μ-oriented) only under the above prior independence; however, I do not have a proof of this.

Stone and Springer (1965, Rider) prove a theorem very similar to Theorem 8.4 and apply it to the variance-components model. However, they omit the splitting-set condition.

Example 8.3. We show that the strong identification condition of Theorem 8.4 may not always be required for the result to hold. Consider again Example 5.5, and suppose T is D-sufficient for Θ. We have

$$\bar{f}(x_1 | t,\theta) = \binom{t}{x_1} \int_0^1 \phi^{x_1} (1-\phi)^{t-x_1} \pi(\phi | t,\theta) d\phi$$

and $\pi(\phi | t,\theta)$ may be replaced by $\pi(\phi | \theta)$, since T is Θ-oriented, so that T and Φ are independent given Θ. Thus $\bar{f}(x_1 | t,\theta)$ will be determined by the first t moments of $\pi(\phi | \theta)$, and, so long as these are the same for every value of θ, $\bar{f}(x_1 | t,\theta)$ will not involve θ. Here the distributions of X given $T = t$ do *not* strongly identify Φ, for any t. However, the marginal sufficiency requirement that $f(x_1 | t,\theta)$ should not involve θ *for all* t ensures that *all* moments of $\pi(\phi | \theta)$ are constant, whence $\pi(\phi | \theta)$ is itself constant, so that we must have Θ and Φ independent.

REFERENCES

BARNARD, G.A. (1963). Some logical aspects of the fiducial argument. *J. Roy. Statist. Soc., B.* **25**, 111-114.

BARNDORFF-NIELSEN, O. (1965). Identifiability of mixtures of exponential families. *J. Math. Anal. Appl.*, **12**, 115-21.

— (1976). Nonformation. *Biometrika* **63**, 567-571.

— (1978). *Information and Exponential Families in Statistical Theory.* Wiley: Chichester - New York - Brisbane.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Ass.* **72**, 355-366.

— (1978). On partial sufficiency: a review. *J. Stat. Plann. Inference,* **2**, 1-13.

CHANDRA, S. (1977). On the mixtures of probability distributions. *Scand. J. Statist.* **4**, 105-112.

DAWID, A.P. (1975). On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. Roy. Statist. Soc. B.* **37**, 248-258.

— (1977). Invariant distributions and analysis of variance models. *Biometrika* **64**, 291-7.

— (1979a). Conditional independence in statistical theory (with Discussion). *J. Roy. Statist. Soc. B* **41**, 1-31.

— (1979b). Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc. B* **41**, 249-252.

— (1980). Conditional independence for statistical operations. *Ann. Statist.* **8**, 598-617.

DAWID, A.P. & DICKEY, J.M. (1977). Problems with nuisance parameters-traditional and Bayesian concepts. *Tech. Report.,* University College London.

DAWID, A.P., STONE, M. & ZIDEK, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference (with Discussion). *J. Roy. Statist. Soc. B,* **35**, 189-233.

HÁJEK, J. (1965). On basic concepts of statistics. *Fifth Berkeley Symposium on Mathematical Statistic and Probability* **1**, 139-162.

JAYNES, E.T. (1980). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys,* (A. Zellner, ed.). Amsterdam: North Holland.

KOEHN, U. & THOMAS, D.L. (1975). On statistics independent of a sufficient statistic: Basu's lemma. *American Statistician* **29**, 40-42.

LEHMANN, E.L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.

LINDLEY, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference.* Cambridge: University Press.

MARTIN, F., PETIT, J.L. & LITTAYE, M. (1973). Indépendance conditionelle dans le modèle statistique bayésien. *Ann. Inst. Henri Poincaré, B,* **9**, 19-40.

RAIFFA, H.A., & SCHLAIFER, R.S. (1961). *Applied Statistical Decision Theory.* Boston: Harvard University.

SCHOU, G. (1978). Estimation of the concentration parameter in von Mises-Fisher distributions. *Biometrika* **65**, 369-377.

STONE, M. & SPRINGER, B.G.F. (1965). A paradox involving quasi prior distributions. *Biometrika* **52**, 623-627.

SUDGEN, R.A. (1978). *Exchangeability and the foundations of survey sampling.* Ph. D. Thesis, University of Southampton.

TEICHER, H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31**, 55-73.

— (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244-248.

— (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.* **38**, 1300-2.