On the Elimination of Nuisance Parameters

DEBABRATA BASU

Eliminating nuisance parameters from a model is universally recognized as a major problem of statistics. A surprisingly large number of elimination methods have been proposed by various writers on the topic. In this article we propose to critically review two such elimination methods. We shall be concerned with some particular cases of the marginalizing and the conditioning methods. The origin of these methods may be traced to the work of Sir Ronald A. Fisher. The contents of the marginalization and the conditionality arguments are then reexamined from the Bayesian point of view. This article should be regarded as a sequel to the author's three-part essay (Basu 1975) on statistical information and likelihood.

KEY WORDS: Marginalization and conditionality arguments; Specific and partial sufficiency; Ancillary and S-ancillary statistics; Unrelated parameters.

1. THE ELIMINATION PROBLEM AND METHODS

The problem begins with an unknown state of nature represented by the parameter of interest θ. We have some information about θ to begin with—e.g., we know that θ is a member of some well-defined parameter space Θ—but we are seeking more. Toward this end, a statistical experiment $\mathcal{E}$ is planned and performed which generates the sample observation x. Further information about θ is then obtained by a careful analysis of the data $(\mathcal{E}, x)$ in the light of all our prior information about θ and in the context of the particular inference problem related to θ. For going through the rituals of the traditional sample-space analysis of data, we must begin with the invocation of a trinity of abstractions $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$, where $\mathfrak{X}$ is the sample space, $\mathcal{A}$ is a σ algebra of events (subsets of $\mathfrak{X}$), and $\mathcal{P}$ is a family of probability measures on $\mathcal{A}$. If the model $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$ is such that we can represent the family $\mathcal{P}$ as $\{P_\theta: \theta \in \Theta\}$, where the correspondence $\theta \to P_\theta$ is one-one and (preferably) smooth, then we go about analyzing the data according to our own light and are thankful for not having to contend with any nuisance parameters.

However, instances of statistical models with $\mathcal{P}$ indexed by θ alone are very rare. Typically, we have to work with a family $\mathcal{P}$ that is indexed as

$$\mathcal{P} = \{P_{\theta,\phi}: \theta \in \Theta, \phi \in \Phi\} ,$$

where ϕ is an additional unknown parameter. If the inference problem at hand relates only to θ and if information gained on ϕ is of no direct relevance to the problem, then we classify ϕ as the nuisance parameter.

The big question in statistics is: How can we eliminate the nuisance parameter from the argument? During the past seven decades an astonishingly large amount of effort and ingenuity has gone into the search for reasonable answers to this question. Broadly speaking, this collective endeavor of the community of statisticians may be classified into the following overlapping categories:

1. To plan the experiment ε in such a fashion that the model is related to the parameter of interest and is relatively free of disturbing nuisance parameters. In this article we are not concerned with the important problems of planning experiments. Our concern is with the problem of data analysis. However, a few elimination methods, such as randomization and sequential sampling which will be discussed in a sequel, may well be classified under this heading.
2. To justify a replacement of the basic model $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$ by a related θ-oriented model $(\mathcal{T}, \mathcal{C}, \mathcal{Q})$, the family $\mathcal{Q}$ is indexed by θ alone. The marginalization and the conditionality arguments that we shall be examining in this article belong to this category.
3. To estimate the nuisance parameter away; that is, to substitute the unknown nuisance parameter ϕ by an estimated value $\hat{\phi}$. This classical method of elimination is used repeatedly in the large sample theory of statistics.
4. To Studentize in the manner of W.S. Gossett with the idea in mind to construct a reasonable looking pivotal quantity involving the sample x and the parameter of interest θ.
5. To invoke the invariance argument of Pitman-Stein-Lehmann. This particular marginalization argument will be examined in a subsequent article.
6. To delimit the argument to a small class of decision procedures, e.g., unbiased estimators, fixed size confidence intervals, similar tests, etc., whose average performance characteristics are, at least in part, free of the nuisance parameter. Mathematicians love this argument. See, e.g., Linnik (1965, 1968).
7. To eliminate the nuisance parameter from the risk function $r_\delta(\theta, \phi)$ of the decision procedure δ by the invocation of a so-called maximization (or minimax) principle. The recommendation for the choice of δ is then made on the basis of the eliminated risk function
$$R_\delta(\theta) = \sup_\phi r_\delta(\theta, \phi) .$$
In Lehmann (1959) we find this argument used quite frequently. For example, the size of a test is always understood as the maximum probability of committing an error of the first kind.
8. To invoke the fiducial argument of R.A. Fisher. With the departure of Sir Ronald from our midst, we seem to have lost our zest for this novel elimination argument.
9. To justify an elimination of the nuisance parameter directly from the likelihood function $L(\theta, \phi|x)$ generated by the particular data (ε, x). The idea is to construct a new scale $L_e(\theta, x)$ (the suffix e denotes the process of elimination of the nuisance parameter) for a direct comparison of the amount of support that the data lends to various values of θ. The maximization of likelihood with respect to ϕ is the classic example of this kind of elimination.

* Debabrata Basu is Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306. This is a revised version of an earlier work with the same title that was presented at a symposium held at the Carleton University, Ottawa, Ontario, October 24–26, 1974. The earlier version appeared in mimeographed form in Proceedings of Symposium on Statistics and Related Topics, ed., Md. E. Saleh, Carleton Math. Lecture Notes No. 52, 1975.

© Journal of the American Statistical Association
June 1977, Volume 72, Number 358
Theory and Methods Section

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_26, © Springer Science+Business Media, LLC 2011

10. To act like a Bayesian; that is, to fix a prior, compute the posterior, integrate out the nuisance parameter from the posterior to arrive at the posterior marginal distribution of the parameter of interest, and then to let the statistical argument rest on the posterior marginal distribution.

    In addition, we have the choice of a fairly large number of specialized elimination methods: the two-stage sampling plan of Stein (1945), the randomization method of Durbin (1961), the characterization argument of Prohorov (1967), the partial sufficiency argument of Hájek (1967), the M-ancillarity argument of Barndorff-Nielsen (1973), etc.

After this introduction to the problem and methods of elimination, we plunge headlong into the depths of the marginalization and the conditionality arguments and try to sort out a number of ideas related to partial sufficiency and partial ancillarity.

2. MARGINALIZATION AND CONDITIONING

The marginalization method of elimination consists of: Choosing a suitable statistic T: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{T}, \mathfrak{B})$, such that the family

$$\mathcal{P}_T = \{P_{\theta,\phi}T^{-1}: \theta \in \Theta, \phi \in \Phi\}$$

of probability measures on $(\mathcal{T}, \mathfrak{B})$ is θ-oriented, i.e., the family $\mathcal{P}_T$ is indexed by θ alone; and then recommending that the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ be given up in favor of the model $(\mathcal{T}, \mathfrak{B}, \mathcal{P}_T)$.

In effect, the method replaces the data $(\mathcal{E}, x)$ by its reduction $(\mathcal{E}_T, t)$, where $T(x) = t$. By $\mathcal{E}_T$ we mean the marginal experiment that may be operationally defined as "perform $\mathcal{E}$ but record only $T(x)$." It is not easy to justify data reduction of the above kind. A great deal of thought and mathematical expertise have gone into the many efforts made so far at such justification. Two distinct major lines of thought in this general direction are: (a) the invariance argument and (b) the partial sufficiency argument. In this article, we shall be concerned with the partial sufficiency argument only.

The conditioning method of elimination consists of: Choosing a suitable statistic Y: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{Y}, \mathfrak{C})$ such that the conditional distribution of the sample x, given $Y = y$, is θ-oriented (it depends on (θ, ϕ) only through θ) for all $y \in \mathcal{Y}$; and recommending that the data $(\mathcal{E}, x)$ be analyzed by looking at the sample x, not as a random variable with the unconditional distribution model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ but as a random variable with the θ-oriented conditional distribution model that corresponds to the condition $Y = y$, where y is the observed value of the statistic Y. In effect, the method aims at replacing the data $(\mathcal{E}, x)$ by the conditioned data $(\mathcal{E}_y{}^Y, x)$, where $\mathcal{E}_y{}^Y$ is a conceptual conditional experiment that corresponds to the observed value y of a suitable statistic Y.

For the marginalization argument, the statistic T not only needs to be θ-oriented but also has to be one that, in some sense, summarizes in itself all the relevant and usable information about θ that is contained in the data. Similarly, for the conditionality argument, it is not enough to choose just any statistic Y that will do the elimination job. The static Y needs to be such that, in some meaningful sense, we can assert that referring the observed sample x to the reference set of all possible samples x' with $Y(x')$ fixed at the present observed value $y = Y(x)$ entails no loss of information on the parameter of interest θ. The statistical literature is strewn with logicians' nightmares of the above kind. Let us see what sense we can make of such nightmares.

3. PARTIAL SUFFICIENCY AND PARTIAL ANCILLARITY

In this section we put together a number of mathematical definitions.

Definition 1 (Model): By the model (or statistical structure) of an experiment $\mathcal{E}$ we mean the usual trinity of abstractions $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$.

We suppose that the family $\mathcal{P}$ is indexed as $\mathcal{P} = \{P_\omega: \omega \in \Omega\}$ and call ω the *universal parameter*. Let $\theta = \theta(\omega)$ be the parameter of interest. By a statistic T we mean a measurable map of $(\mathfrak{X}, \mathfrak{A})$ into another measurable space $(\mathcal{T}, \mathfrak{B})$.

Definition 2 (Ancillarity): The statistic T is ancillary if the marginal (or sampling) distribution of T is the same for all $\omega \in \Omega$—i.e., for all $B \in \mathfrak{B}$, the function $P_\omega(T^{-1}B)$ is a constant in ω.

Definition 3 (θ-Oriented Statistic): The statistic T is θ oriented if the marginal distribution of T depends on ω only through $\theta = \theta(\omega)$. That is, $\theta(\omega_1) = \theta(\omega_2)$ implies $P_{\omega_1}(T^{-1}B) = P_{\omega_2}(T^{-1}B)$ for all $B \in \mathfrak{B}$.

Observe that every ancillary statistic is θ oriented irrespective of what θ is

Example 1: Let $x = (x_1, x_2, \ldots, x_n)$, with n fixed in advance, be a sample of n independent observations on a $N(\mu, \sigma)$. Let $D = (x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1)$ be the difference statistic. Clearly, D is σ oriented and, therefore, so is every measurable function $h(D)$ of D. That the class $\{h(D)\}$ of measurable functions of the difference statistic does not exhaust the family of σ-oriented statistics is seen as follows. Choose and fix two functions $h_1(D)$ and $h_2(D)$ that are identically distributed and also a Borel set E in R_1. Since $\bar{x}$ is stochastically independent of D for all (μ, σ), it now follows that the statistic T_E defined as

$$\begin{aligned} T_E(\bar{x}, D) &= h_1(D) \quad \text{if} \quad \bar{x} \in E \\ &= h_2(D) \quad \text{if} \quad \bar{x} \notin E \ , \end{aligned}$$

is σ oriented—indeed, T_E is identically distributed as $h_1(D)$ and $h_2(D)$. It is thus clear that D is not the maximum σ-oriented statistic. In fact no maximum σ-oriented statistic exists. (See Basu (1959) and (1967) for more information on this kind of problem.) In this case we have a plentiful supply of σ-oriented statistics. However, the notion of μ-orientedness is vacuous in the sense that no nontrivial (nonancillary) statistic can be μ oriented. This remark is generally true for the location parameter μ in a location-scale parameter setup.

Definition 4 (Variation Independence): The two functions $\omega \to a(\omega)$ and $\omega \to b(\omega)$ on the space Ω with respective ranges A and B are said to be variation independent if the range of the function $\omega \to (a(\omega), b(\omega))$ is the Cartesian product $A \times B$.

If the universal parameter ω can be represented as $\omega = (\theta, \phi)$, where θ and ϕ are variation independent in the preceding sense—that is, $\Omega = \Theta \times \Phi$ where Θ and Φ are the respective ranges of θ and ϕ—then we call ϕ a variation independent complement of θ. With θ as the parameter of interest, we may then call ϕ the nuisance parameter.

We have not come across a satisfactory definition of the notion of a nuisance parameter. It is only hoped that the above working definition will meet with little resistance. (See Barndorff-Nielsen (1973) for further details on the notion of variation independence.)

By a sufficient statistic, we mean a statistic that is sufficient in the usual sense with respect to the full model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$. The following definition of a specific sufficient statistic appears in Neyman and Pearson (1936). Let ϕ be a variation independent complement of θ.

Definition 5 (Specific Sufficiency): The statistic T is specific sufficient for θ if, for each fixed $\phi \in \Phi$, the statistic T is sufficient with respect to the restricted model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P}_\phi)$, where $\mathcal{P}_\phi = \{P_{\theta,\phi}: \theta \in \Theta, \phi \text{ fixed}\}$.

In Example 1, the sample mean $\bar{x}$ is specific sufficient for μ. In fact, $\bar{x}$ is a minimum specific sufficient statistic for μ. The sample standard deviation s is, however, not sufficient for σ for any specified value of μ. Indeed, a statistic can be specific sufficient for σ only if it is sufficient.

In the spirit of Definition 5, we then define the notion of specific ancillarity in the following terms. As before, let ϕ be a variation independent complement of θ.

Definition 6 (Specific Ancillarity): The statistic T is specific ancillary for θ if, for each fixed $\phi \in \Phi$, it is ancillary with respect to the restricted model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P}_\phi)$.

In other words, T is specific ancillary for θ if it is ϕ oriented, where ϕ is a variation independent complement of θ. It should be noted that the definition of θ-orientedness does not presuppose the existence of a variation independent complement ϕ, but the definitions of specific sufficiency and specific ancillarity (for θ) do.

In Example 1, with σ as the parameter of interest, it is tempting to marginalize to the statistic s. But can we logically justify such a marginalization? In what sense can we say that s summarizes in itself all the relevant and available information about σ in the absence of any information on μ? We shall return to the question later.

Suppose μ is the parameter of interest in Example 1. Marginalization to the statistic $\bar{x}$, which is specific sufficient for μ, will not eliminate σ as $\bar{x}$ is not μ oriented. We shall also lose valuable information on μ if we throw away the s-part of the sufficient statistic $(\bar{x}, s)$ and record only $\bar{x}$. For one thing, we shall no longer be able to speculate about the accuracy of $\bar{x}$ as a point estimate of μ. The marginalization method is of no use for the purpose of eliminating the scale parameter σ. As we have noted earlier, if T is μ oriented then it has to be an ancillary statistic. Surely, we do not want to marginalize to something that has nothing to do with μ! The conditionality argument is also of no use for eliminating σ. If conditioning with respect to Y eliminates σ, then Y has to be specific sufficient for σ. But, as we have stated earlier, every such Y has to be sufficient for (μ, σ). Hence, conditioning with respect to Y will eliminate μ as well! The problem of eliminating the scale parameter σ is not an easy one. Student's t-test and Stein's two-stage sampling plan are classical examples of statistical methodology that were developed to solve the problem in non-Bayesian terms.

The following definition of partial sufficiency is usually attributed to Fraser (1956). But we find the definition clearly laid out in Olshevsky (1940), who attributed it to Neyman (1935).

Definition 7 (p-Sufficiency): The statistic T is partially sufficient (denoted by p-sufficient) for θ if T is specific sufficient for θ and T is θ oriented. From this it is clear that the notion of p-sufficiency for θ presupposes the existence of a variation independent complement ϕ for θ. With the same presupposition, Sandved (1967) defined a notion of partial ancillarity in the following terms.

Definition 8 (S-Ancillarity): The statistic Y is a partial ancillary (S-ancillary) for θ if Y is specific ancillary for θ (Y is ϕ oriented) and Y is specific sufficient for ϕ. It should be noted that in Definitions 7 and 8, we are looking at the same concept but from two different angles. The statistic Y is S-ancillary for θ if and only if it is p-sufficient for ϕ.

The name S-ancillary (ancillary in the sense of Sandved) is due to Barndorff-Nielsen (1973) whose terminology for p-sufficiency is S-sufficiency. Barndorff-Nielsen's mathematical formalization of the twin notions of p-sufficiency and S-ancillarity as a "cut" may be defined as follows.

Definition 9 (Barndorff-Cut): A statistic T: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{T}, \mathfrak{B})$ defines a Barndorff-cut of an experiment

$$\mathcal{E} = \{(\mathfrak{X}, \mathfrak{A}, P_\omega): \omega \in \Omega\} ,$$

if there exist two variation independent and complementary subparameters $\theta = \theta(\omega)$ and $\phi = \phi(\omega)$, such that the marginal experiment $\mathcal{E}_T = \{(\mathcal{T}, \mathfrak{B}, P_\omega T^{-1}): \omega \in \Omega\}$ is θ oriented ($P_\omega T^{-1}$ depends on ω only through $\theta(\omega)$) and that each one of the family $\{\mathcal{E}_t{}^T: t \in T\}$ of conditional experiments is ϕ oriented.

The statistic T is then p-sufficient for θ and S-ancillary for ϕ. Observe that every sufficient statistic defines a Barndorff-cut and so also does every ancillary statistic. In the former case $\theta(\omega) = \omega$ and $\phi(\omega)$ is a known constant, and in the latter case it is the other way around.

In Example 1 there exists no Barndorff-cut that separates μ and σ. The following are a few other examples where the definition yields something.

Example 2: Let the random variables $x_i (i = 1, 2, \ldots, m)$ be iid $N(\theta, 1)$, and let $y_j (j = 1, 2, \ldots, n)$ be an independent set of iid $N(\phi, 1)$. Clearly, $\bar{x}$ is p-sufficient and $\bar{y}$ is S-ancillary for θ.

Example 3: Let x and y be independent Poisson variables with means μ and ν, respectively. With the reparametrization $\theta = \mu/(\mu + \nu)$ and $\phi = \mu + \nu$, it can

be checked that $Y = x + y$ is S-ancillary for θ. There does not exist a statistic T that is p-sufficient for θ.

Example 4: Let $x = (y, z, w)$ have a multinomial distribution with $y + z + w = n$ and p, q, $r(p + q + r = 1)$ as probabilities (parameters). The parameters p and q are not variation independent. However, when we reparametrize as $\theta = p$, $\phi = q/(1 - p)$, it is easy to check that the statistic y becomes p-sufficient for θ (S-ancillary for ϕ).

Example 5: Let $0 < \theta < 1$ and $0 < \phi < \infty$. Let X be a random variable with pdf

$$p(x|\theta, \phi) = (1 - \theta)\phi e^{\phi x} \quad \text{for} \quad x \leq 0$$
$$= \theta\phi e^{-\phi x} \quad \text{for} \quad x > 0 \ .$$

Let $x_1, x_2, \ldots, x_n$ be n independent observations on X. Let T be the number of positive x_i's, and let $Y = \sum |x_i|$. Then T and Y are respectively p-sufficient and S-ancillary for the parameter θ.

Note the similarities between Examples 2 and 5. In either case, we have for the parameter of interest θ a statistic T that is p-sufficient and a statistic Y that is S-ancillary. In each case, however, the two statistics are stochastically independent for all possible values of the universal parameter. The fact that this is not generally true is going to bother us in due course.

It will be useful to review the various definitions in terms of the corresponding factorizations of the likelihood functions. To this end let us suppose that the family $\mathcal{P} = \{P_{\theta,\phi}: \theta \in \Theta, \phi \in \Phi\}$ is dominated by a σ-finite measure μ and let $\{p(\cdot|\theta, \phi)\}$ be the corresponding family of probability density functions. To fix our ideas and to avoid all measure-theoretic difficulties let us pretend for the time being that $\mathfrak{X}$ is a countable set and that μ is the counting measure on $\mathfrak{X}$. Corresponding to any statistic $T: \mathfrak{X} \to \mathcal{T}$ we have a factorization (of p) of the form

$$p(x|\theta, \phi) = g(T|\theta, \phi)f(x|T, \theta, \phi) \ ,$$

where g defines the marginal distribution of T and f defines the conditional distribution of x given T. (Our notations are admittedly rather sloppy, but there should be no difficulty in following our meaning.) Consider now the following particular cases of the above general factorization.

Case I: $p = g(T|\theta, \phi)f(x|T)$—this corresponds to the case where T is sufficient.

Case II: $p = g(T)f(x|T, \theta, \phi)$—the statistic T is ancillary.

Case III: $p = g(T|\theta)f(x|T, \theta, \phi)$—the statistic T is θ oriented. The case where T is ϕ oriented is similar.

In the situation where θ and ϕ are variation independent parameters, the notion of θ-orientedness is the same as the notion of specific ancillarity for ϕ. Case III, therefore, also corresponds to the case where T is specific ancillary for ϕ. With θ and ϕ variation independent, we have the next case.

Case IV: $p = g(T|\theta, \phi)f(x|T, \phi)$—the statistic T is specific sufficient for θ. The case where T is specific sufficient for ϕ is similar.

Case V: $p = g(T|\theta)f(x|T, \phi)$—the statistic T is p-sufficient for θ and is S-ancillary for ϕ.

Case Va: $p = g(T|\phi)f(x|T, \theta)$—the statistic T is S-ancillary for θ and is p-sufficient for ϕ.

Instead of looking at factorizations in terms of marginal and conditional frequencies, suppose we consider factorizations of the more general form

$$p(x|\theta, \phi) = G(x, \theta, \phi)F(x, \theta, \phi) \ .$$

The very familiar

Case VI: $p = G(T, \theta, \phi)F(x)$,

when proved equivalent to Case 1, constitutes the well-known factorization theorem for sufficiency. Similarly, the factorization

Case VII: $p = G(T, \theta, \phi)F(x, \phi)$

can be shown to be equivalent to Case IV (the case of specific sufficiency for θ). Now consider

Case VIII: $p = G(T, \theta)F(x, \phi)$.

Is Case VIII equivalent to Case V? It is important to recognize that the answer is in the negative. The examples in Section 9 will clarify the matter. Finally, we have factorizations of the form

Case IX: $p = G(x, \theta)F(x, \phi)$.

It will turn out later that we really should be after factorizations of this form. Clearly, p factors in the manner of Case IX whenever we have a Barndorff-cut separating θ from ϕ (as in Cases V or Va). That the converse is not true will be variously exemplified in Section 9.

4. GENERALIZED SUFFICIENCY AND CONDITIONALITY PRINCIPLES

To understand the logic of the generalized sufficiency and conditionality principles $\mathfrak{s}^*$ and $\mathfrak{C}^*$, it is useful to consider a few hypothetical situations. (For a comprehensive discussion on the sufficiency, conditionality, invariance, and the likelihood principles refer to Basu (1975).)

(i) We have two experimental setups ε and ε', where the former provides information only on the parameter of interest θ while the latter is informative about an unrelated parameter ϕ alone—the parameter ϕ is unrelated to θ in the sense that we do not recognize the relevance of any information on ϕ for the purpose of inference making on θ. Faced with data such as $\{(\varepsilon, x), (\varepsilon', x')\}$, it makes good statistical sense to ignore the second part of the data and concentrate our attention on the relevant part (ε, x).

(ii) Let ε be an experiment whose randomness (probabilistic) characteristics depend only on θ. Having obtained the data (ε, x), suppose we choose to perform a randomization exercise $\varepsilon_{(x)}$ thus arriving at the additional data $(\varepsilon_{(x)}, y)$. If all the randomness characteristics of $\varepsilon_{(x)}$ (possibly influenced by x) are known to us, then the secondary data $(\varepsilon_{(x)}, y)$ cannot give us any additional information on θ, or

on anything for that matter. It makes good statistical sense then to suggest that the analysis of the data $\{(\mathcal{E}, x), (\mathcal{E}_{(x)}, y)\}$ ought to proceed on a total nonrecognition of the randomization exercise $\mathcal{E}_{(x)}$ and the resulting outcome y. Indeed, this is one way of looking at the sufficiency principle $\mathcal{S}$ (see Basu 1975).

(iii) If in (ii) we find that the randomness characteristics of $\mathcal{E}_{(x)}$ are fully known except for a nuisance parameter ϕ that is unrelated to θ, then we are in a situation quite analogous to (i). Conforming to the statistical intuition that told us to ignore $(\mathcal{E}', x')$ in (i), the generalized sufficiency principle $\mathcal{S}^*$ tells us to ignore $(\mathcal{E}_{(x)}, y)$ in this situation.

(iv) We have a choice of k experiments $\mathcal{E}_{(1)}, \mathcal{E}_{(2)}, \ldots, \mathcal{E}_{(k)}$. The randomness structure of $\mathcal{E}_{(y)} (y = 1, 2, \ldots, k)$ is related only to the parameter of interest. Let $\mathcal{E}$ stand for a randomization exercise that selects one of the k experiments with known (predetermined) selection probabilities $\pi_1, \pi_2, \ldots, \pi_k$. The experiment $\mathcal{E}_{(y)}$ selected by $\mathcal{E}$ is then performed resulting in the outcome x. The full data is $\{(\mathcal{E}, y), (\mathcal{E}_{(y)}, x)\}$. Since the part $(\mathcal{E}, y)$ of the data is totally uninformative, it makes good statistical sense to disregard this part of the data and focus our attention on the relevant part, i.e., $(\mathcal{E}_{(y)}, x)$. This is a version of the conditionality principle.

(v) Now, suppose in (iv) above the selection probabilities $\pi_1, \pi_2, \ldots, \pi_k$ are not fully known but depend on (are functions of) an unrelated nuisance parameter ϕ. We are now in a situation that is very similar to (i). The generalized conditionality principle $\mathcal{C}^*$ tells us to analyze the data by concentrating our whole attention on that part of the data—namely $(\mathcal{E}_{(y)}, x)$—that is related to θ.

We are now ready to state formally the two generalized principles of sufficiency and conditionality.

Principle $\mathcal{S}^*$ *(Generalized Sufficiency Principle):* If, in terms of the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ for the data $(\mathcal{E}, x)$, we recognize the statistic T as p-sufficient (partially sufficient in the sense of Definition 7) for the parameter of interest θ, then the data $(\mathcal{E}, x)$ should be reduced by marginalization to $(\mathcal{E}_T, t)$, where $\mathcal{E}_T$ is the marginal experiment corresponding to T and $t = T(x)$.

Principle $\mathcal{S}^*$ may be stated in a less severe form in the following terms.

Principle $\mathcal{S}^{**}$: If T is p-sufficient for θ, then $T(x') = T(x'')$ implies that the information content (the evidential meaning) of the data $(\mathcal{E}, x')$ and $(\mathcal{E}, x'')$ relative to the parameter θ are identical in all respects. In other words, the data $(\mathcal{E}, x')$ warrants the same inference on θ as does the data $(\mathcal{E}, x'')$.

Principle $\mathcal{C}^*$ *(Generalized Conditionality Principle):* If Y is an S-ancillary (Definition 8) for θ, then the data $(\mathcal{E}, x)$ should be analyzed by reinterpreting it as $(\mathcal{E}_{(y)}, x)$, where $\mathcal{E}_{(y)} (= \mathcal{E}_y{}^Y)$ is the conditional experiment that corresponds to the observed value $y = Y(x)$ of Y.

As we have said before, corresponding to any statistic T we can conceive of a decomposition of the experiment $\mathcal{E}$ into a two-stage experimental setup in which the marginal experiment $\mathcal{E}_T$ is followed by the conditional experiment $\mathcal{E}_t{}^T$ that corresponds to the observed value $t = T(x)$ of T. The original data $(\mathcal{E}, x)$ may then be viewed as $\{(\mathcal{E}_T, t), (\mathcal{E}_t{}^T, x)\}$. If T is p-sufficient for θ then, by definition, the experiment $\mathcal{E}_T$ is θ oriented, and the experiment $\mathcal{E}_t{}^T$ is ϕ oriented. So, in view of (i) and (iii), it makes good statistical sense to invoke principle $\mathcal{S}^*$ and marginalize the data to $(\mathcal{E}_T, t)$. Conversely, if T is S-ancillary for θ then, by definition, $\mathcal{E}_T$ is ϕ oriented and $\mathcal{E}_t{}^T$ is θ oriented. So, in view of (i) and (v), it appears logical that we ought to ignore the $(\mathcal{E}_T, t)$ part of the data and analyze it as $(\mathcal{E}_t{}^T, x)$. This is the generalized conditionality principle $\mathcal{C}^*$.

5. A CHOICE DILEMMA

In the writings of R.A. Fisher we find the conditionality argument used in three different ways: to recover the ancillary information in the data when it is found that the maximum likelihood estimator is not sufficient; to eliminate the nuisance parameter as in the case of the celebrated test of independence with a 2×2 multinomial data; and to generalize the fiducial argument as in the case of multiple observations on a random variable with a location parameter in its distribution.

In Basu (1964), while studying in depth Fisher's recovery of information argument, the author discovered a disturbing inherent difficulty in the conditionality argument. The difficulty flows from the fact that, in general, there does not exist a largest ancillary statistic in the sense of the usual partial order on statistics. Even in the simplest of situations we may have two ancillary statistics Y and U such that the statistic (Y, U) is not ancillary. Indeed, the pair (Y, U) may be fully informative, i.e., sufficient. In such a situation, the conflict between which of the two ancillaries to choose for the purpose of conditioning the data remains unresolved, despite some valiant efforts by Barnard and Sprott (1971) and Cox (1971), in non-Bayesian terms. The generalized conditionality (S-ancillarity) argument founders on the same non-uniqueness rock. We reproduce here an example from Basu (1964) that has attracted a lot of attention from non-Bayesians.

Example: Let X be a random variable with range $\{1, 2, 3, 4\}$ and probability distribution

$$X: \quad 1 \qquad 2 \qquad 3 \qquad 4$$
$$\text{Prob}: \quad (1-\theta)/6 \quad (1+\theta)/6 \quad (2-\theta)/6 \quad (2+\theta)/6 \ ,$$

where $0 < \theta < 1$. We have n independent observations on X. The cell frequencies $x = (n_1, n_2, n_3, n_4)$ constitute the minimum sufficient statistic. The likelihood function is

$$L(\theta) = (1-\theta)^{n_1}(1+\theta)^{n_2}(2-\theta)^{n_3}(2+\theta)^{n_4} \ .$$

Let us write Bin (n, p) for the Binomial distribution with parameters n and p. Observe that $Y = n_1 + n_2$ is an ancillary statistic with probability distribution Bin $(n, \frac{1}{3})$ and that $U = n_1 + n_4$ is another ancillary with distribution Bin $(n, \frac{1}{2})$. If we condition x by Y then we can look upon the data as a pair of independent random variables n_1 and n_3 that are distributed as

$$\text{Bin } (Y, (1-\theta)/2) \quad \text{and} \quad \text{Bin } (n - Y, (2-\theta)/4) \ ,$$

respectively. However, if we choose to condition x by the other ancillary U, then we simplify the data to two independent variables n_1 and n_3 distributed re-

spectively as

$$\text{Bin } (U, (1-\theta)/3) \quad \text{and} \quad \text{Bin } (n-U, (2-\theta)/3) \ .$$

In either case, the sample-space analysis of the conditioned data will be fairly easy and straightforward. But can anyone give a convincing argument for the choice of either Y or U as the conditioning ancillary?

We can easily introduce a nuisance parameter into the foregoing example by incorporating into the data, say, the result z of an independent coin-tossing experiment with an unknown bias ϕ in the coin. (George Barnard once remarked that when he retires he will go into business manufacturing biased coins and selling them to people like Basu!) In this case both (Y, z) and (U, z) will be S-ancillaries and we shall be back in the choice dilemma.

In contrast to the conditionality argument, the marginalization argument, in terms of the sufficiency or the generalized sufficiency principles, does not suffer from the above kind of a choice dilemma. With the kind of models that we work with in statistics, the existence of an essentially unique minimum sufficient statistic is always assured, and if the class of statistics that are p-sufficient for θ is not vacuous, then there will exist an essentially unique minimum such statistic.

6. A CONFLICT

The two elimination methods, namely, the one that marginalizes to a statistic T that is p-sufficient for θ and the one that conditions with respect to a statistic Y that is S-ancillary for θ, owe their origin to the same statistical intuition that guided us through (i) to (v) in Section 4. However, this does not mean that the two methods can co-exist in logical harmony. The possibility of a natural conflict between the methods was pointed out to the author by Philip Dawid (1975). We give below a simple example along the lines of the dilemma example of the previous section to highlight this conflict.

Example: Let T and Y be two random variables with the same range $\{1, 2, 3, 4\}$ and a joint distribution as described in the following table.

To simplify the argument let us suppose that we have only one observation $x = (t, y)$ on the pair (T, Y)—the general case where we have n observations on (T, Y) is very similar. Observe that the statistic T, defined as $T(x) = t$, is p-sufficient for θ and that the statistic Y, defined as $Y(x) = y$, is S-ancillary for θ. The trouble is that T and Y are not stochastically independent in this example. The marginal distribution of T is very different from its conditional distribution for any given value of Y.

It is thus clear that we can have a pair of stochastically dependent statistics T and Y such that (T, Y) is sufficient for (θ, ϕ), T is p-sufficient for θ, and Y is S-ancillary for θ. The nuisance parameter ϕ can be eliminated from the argument either by marginalizing to T or by conditioning (T, Y)—that is, T—by the S-ancillary Y. The two elimination methods cannot be reconciled in such cases.

What went wrong? Should we blame the statistical intuition that guided us through (i) to (v) in Section 4? The above conflict is only a manifestation of the difficulties that we have to face when we try to interpret data in some sample-space terms.

7. RAO-BLACKWELL TYPE THEOREMS

In Section 4, our case for the Sufficiency Principle $\mathcal{S}$, the Conditionality Principle $\mathcal{C}$ and their generalizations $\mathcal{S}^*$ and $\mathcal{C}^*$ rested on the highly nonmathematical phrase, "It makes good statistical sense." The author does not know how else to argue in non-Bayesian terms for these essentially Bayesian principles of data analysis. A large majority of the statisticians belonging to the Fisher-Neyman school of thought seem to agree wholeheartedly with $\mathcal{S}$ although most of them are quite wary of $\mathcal{C}$. This almost universal faith in $\mathcal{S}$ is there, partly because it makes good statistical sense, but mainly because of the widespread belief that principle $\mathcal{S}$ has been mathematically proved in the Rao-Blackwell theorem. On p. 17 of Basu (1975) we briefly examined this mathematical proof of a statistical principle. Now, let us turn the spotlight on a similar proof of $\mathcal{S}^*$ given by Fraser (1956).

Let $a(\theta)$ be a real valued function of θ. We are looking for a reasonable point estimate of $a(\theta)$ on the basis of the data $(\mathcal{E}, x)$. Let us suppose that the loss $W(t, \theta)$, when $a(\theta)$ is estimated by t, is convex in t for each θ. Let $\mathfrak{U}$ be the class of all estimators U of $a(\theta)$ such that the risk function

$$r_U(\theta) = r_U(\theta, \phi) = E[W(U, \theta) \mid \theta, \phi]$$

is well defined and θ oriented, that is, depends on $\omega = (\theta, \phi)$ only through θ.

Theorem (Fraser): If T is p-sufficient for θ then, for each $U \in \mathfrak{U}$, there exists an estimator $U_0 = U_0(T)$ such that $r_{U_0}(\theta) \leq r_U(\theta)$ for all $\theta \in \Theta$.

Joint Distribution of T and Y

	T				
Y	1	2	3	4	Total
1	$(1-\theta)(1-\phi)/12$	$(1+\theta)(1-\phi)/12$	0	0	$(1-\phi)/6$
2	$(1-\theta)(1+\phi)/12$	$(1+\theta)(1+\phi)/12$	0	0	$(1+\phi)/6$
3	0	0	$(2-\theta)(2-\phi)/24$	$(2+\theta)(2-\phi)/24$	$(2-\phi)/6$
4	0	0	$(2-\theta)(2+\phi)/24$	$(2+\theta)(2+\phi)/24$	$(2+\phi)/6$
Total	$(1-\theta)/6$	$(1+\theta)/6$	$(2-\theta)/6$	$(2+\theta)/6$	1

Proof: The statistic T is θ oriented by definition, so the risk function generated by any function of T, if well defined, must be θ oriented. Now, choose and fix $\phi_0 \in \Phi$ and define

$$U_0 = E(U \mid T, \theta, \phi_0) \ .$$

Since T, by definition, is sufficient for θ when ϕ is fixed, it follows that U_0 is well defined as an estimator; that is, the unknown θ does not enter into the definition of U_0. From Jensen's inequality it follows that, for all $\theta \in \Theta$,

$$r_{U_0}(\theta) = r_{U_0}(\theta, \phi_0) \leq r_U(\theta, \phi_0) = r_U(\theta) \ ,$$

and thus the theorem is proved. The Rao-Blackwell theorem clearly corresponds to the particular case where ϕ is a known constant ϕ_0.

The above theorem may be generalized further along the following lines suggested by Hájek (1965). Let $\mathfrak{U}'$ be the class of all estimators U such that the risk function $r_U(\theta, \phi)$ is well defined (but not necessarily θ oriented). Using the so-called minimax principle (see paragraph 7 of Section 1) let us define

$$R_U(\theta) = \sup_\phi r_U(\theta, \phi)$$

as the eliminated risk function associated with U, if $U \in \mathfrak{U}$ then $r_U(\theta) = r_U(\theta, \phi)$ is θ oriented and thus $R_U(\theta) = r_U(\theta)$. Now, if we define U_0 as in the Fraser theorem, then it follows (in view of the fact that U_0 is θ oriented) that $R_{U_0}(\theta) = r_{U_0}(\theta, \phi_0) \leq r_U(\theta, \phi_0) \leq R_U(\theta)$. This generalizes the Fraser theorem to the following result:

Theorem (Hájek): If T is p-sufficient for θ, then for each $U \in \mathfrak{U}'$ there exists an $U_0 = U_0(T)$ such that $R_{U_0}(\theta) \leq R_U(\theta)$ for all $\theta \in \Theta$.

The proofs of the preceding two theorems do not make full use of the supposition that T is p-sufficient for θ. They rest heavily on the supposition that T is θ oriented but require T to be sufficient for θ for just one specific value ϕ_0 of ϕ. This suggests the following generalization of the notion of partial sufficiency. For each $\theta \in \Theta$, let us define $\bar{\mathcal{P}}_\theta$ to be the convex hull of the family $\mathcal{P}_\theta = \{P_{\theta,\phi} : \theta \text{ fixed}, \phi \in \Phi\}$ of measures on $(\mathfrak{X}, \mathfrak{A})$. In other words, $\bar{\mathcal{P}}_\theta$ is the family of all measures Q of the form:

$$Q(A) = \int_\Phi P_{\theta,\phi}(A) d\xi(\phi) \quad \text{for all} \quad A \in \mathfrak{A} \ , \tag{7.1}$$

where ξ is an arbitrary probability measure on Φ. The following definition is due to Hájek (1965).

Definition (H-Sufficiency): The statistic T is H-sufficient (partially sufficient in the sense of Hájek) for θ if, for each $\theta \in \Theta$, there exists a choice of a measure $Q_\theta \in \bar{\mathcal{P}}_\theta$ such that, with $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$, T is sufficient in the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{Q})$, and T is θ oriented in the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$.

It should be noted that for the definition of H-sufficiency it is not necessary for θ and ϕ to be variation independent. Clearly, p-sufficiency implies H-sufficiency. We have only to choose and fix $\phi_0 \in \Phi$ and then define $Q_\theta = P_{\theta,\phi_0}$. Let us check now that the Fraser-Hájek theorems remain true even if we replace the requirement of p-sufficiency for the statistic T by the less stringent requirement of H-sufficiency. For any U, we define U_0 as $E(U \mid T, Q_\theta)$. Observe that U_0 is an estimator in view of the definition of H-sufficiency. We then invoke Jensen's inequality to prove that, for all $\theta \in \Theta$,

$$\int_{\mathfrak{X}} W(U_0, \theta) dQ_\theta \leq \int_{\mathfrak{X}} W(U, \theta) dQ_\theta \ .$$

Now, if we look back on the supposition that Q_θ is in the form (7.1) above, then, from the fact that U_0—being a function of T—is θ oriented, it follows at once that the left side of the above inequality is equal to

$$r_{U_0}(\theta) = R_{U_0}(\theta)$$

for all θ. Similarly, the right side is equal to $r_U(\theta)$ if $U \in \mathfrak{U}$ and is clearly not greater than $R_U(\theta)$ if $U \in \mathfrak{U}'$. Thus the two preceding theorems may be finally restated as:

Theorem (Fraser-Hájek): If T is H-sufficient for θ, then for any $U \in \mathfrak{U}$ there exists a $U_0 = U_0(T)$ such that $r_{U_0}(\theta) \leq r_U(\theta)$ for all θ. Furthermore, for any $U \in \mathfrak{U}'$ it is true that $R_{U_0}(\theta) \leq R_U(\theta)$ for all θ.

How much comfort can an advocate of the generalized sufficiency principle S* derive from the Fraser-Hájek theorem? Before answering this question, let us take a brief look at the question of how and where the notion of H-sufficiency fits into the ten-fold factorization scheme of the likelihood that we laid out in Section 4.

In order for T to be H-sufficient for θ it is necessary that T is θ-oriented; that is, we have a factorization of the form

$$p(x \mid \theta, \phi) = g(T \mid \theta) f(x \mid T, \theta, \phi) \ . \tag{7.2}$$

It is also necessary (in view of the sufficiency condition for T) that there exists a family $\{\xi_\theta : \theta \in \Theta\}$ of probability measures on Φ such that the "mixed" frequency function

$$q(x \mid \theta) = \int_\Phi p(x \mid \theta, \phi) d\xi_\theta(\phi)$$

factors as

$$q(x \mid \theta) = G(T, \theta) F(x) \ . \tag{7.3}$$

Let us look back at the classical problem where the sample $x = (x_1, x_2, \ldots, x_n)$ consists of n independent observations on an $N(\mu, \sigma)$. Clearly $\bar{x}$ is not H-sufficient for μ—indeed, no T can be H-sufficient for μ. But is $s^2 = \sum (x_i - \bar{x})^2$ H-sufficient for σ? Can we find a family $\{\xi_\sigma : 0 < \sigma < \infty\}$ of "mixing measures" on R_1 that will lead to a factorization of the type (7.3) above with $T = s^2$? Observe that

$$p(x \mid \mu, \sigma) = A(\sigma) \exp\left(-\frac{s^2}{2\sigma^2}\right) \exp\left[-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right],$$

where $A(\sigma) = ((2\pi)^{\frac{1}{2}}\sigma)^{-n}$.

We, therefore, need a family of mixing measures ξ_σ such that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu) = B(\bar{x})C(\sigma) \quad . \quad (7.4)$$

The above factorization clearly holds if we choose for ξ_σ the uniform distribution (the Lebesgue measure) over the whole real line. But, with such improper mixings, it is easily seen that the Fraser-Hájek theorem will fall to pieces. If the range of σ is the whole of the positive half line, then there cannot exist a family of proper mixing measures ξ_σ for which the factorization (7.4) will hold.

So how are we going to prove that we ought to marginalize to s when the parameter of interest is σ? Hájek (1965) came up with the following ingenious mathematical argument. In any particular situation, we should always be able to limit (on a priori considerations) the parameter σ to some finite interval $(0, k)$. With σ restricted to such a finite interval, the statistic s becomes H-sufficient for σ. Just check that the factorization (7.4) holds if we choose for ξ_σ the Normal measure with mean zero and variance $(k^2 - \sigma^2)/n$.

Hájek's definition of partial sufficiency is intriguing and full of mathematical possibilities. But, what are the statistical contents of Hájek's definition of partial sufficiency and his generalized Rao-Blackwell theorem? Hájek's 'proof,' that we should marginalize to s when we do not know μ, certainly does not scandalize our statistical intuition. In the language of R.A. Fisher, if we throw away $\bar{x}$ and marginalize to s, then our loss of information on σ has the measure of only one degree of freedom in the worst possible case (when μ is fully known). Of the total information available on σ, the fraction of information summarized in s is at least $(n-1)/n$. Let us now look at the following celebrated example due to Neyman and Scott (1948):

Example (Neyman & Scott): The data x consists of $2n$ observations $x_1, x_1', x_2, x_2', \ldots, x_n, x_n'$. The statistical model here corresponds to $2n$ independent normal variables with equal variances σ^2 and with x_i, x_i' having common mean $\mu_i (i = 1, 2, \ldots, n)$. The parameter of interest is σ, the nuisance parameter is the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$.

With $S^2 = \sum (x_i - x_i')^2$, $\bar{x}_i = (x_i + x_i')/2$ and $A(\sigma)$ $((2\pi)^{\frac{1}{2}}\sigma)^{-2n}$, we then have

$$p(x \mid \boldsymbol{\mu}, \sigma) = A(\sigma) \exp\left(-\frac{S^2}{4\sigma^2}\right) \exp\left[-\frac{\sum (\bar{x}_i - \mu_i)^2}{\sigma^2}\right].$$

The statistic S^2 is clearly σ oriented. Is it H-sufficient for σ? Again the answer is no if σ is unrestricted, but it is yes if we restrict σ to a finite interval (0,k). For the mixing measure ξ_σ on R_n, we now choose the one for which $\mu_1, \mu_2, \ldots, \mu_n$ are iid normal variables with means zero and variances $(k^2 - \sigma^2)/2$.

Of course, we are prepared to assume that $0 < \sigma < k$ for some k. The Hájek proof notwithstanding, how secure do we really feel about marginalizing to S without taking a hard look at $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$? If $\boldsymbol{\mu}$ were known, then the sample would contain $2n$ units (degrees of freedom) of information on σ, out of which S summarizes in itself only n units. Are we really prepared to sacrifice n degrees of freedom at the altar of ignorance on $\boldsymbol{\mu}$? The issues raised in this example of Neyman and Scott are all very complex and we shall return to them again in a subsequent article.

We close this section with one more jab at the notion of H-sufficiency and the Rao-Blackwell type proof of our generalized sufficiency principle in sample-space terms.

Example:[1] Let $x = (x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n)$ be $m + n$ independent normal variables all with unit variances. It is known that $Ex_i = \theta (i = 1, 2, \ldots, m)$ and $Ey_i = \theta\phi (j = 1, 2, \ldots, n)$, where $-\infty < \theta < \infty$ is the parameter of interest and $\phi (= 0 \text{ or } 1)$ is the nuisance parameter.

The likelihood function neatly factors as

$$p(x \mid \theta, \phi) = A(x) \exp\left[-m(\bar{x} - \theta)^2/2\right] \cdot \exp\left[-n(\bar{y} - \theta\phi)^2/2\right] .$$

Clearly, the pair $(\bar{x}, \bar{y})$ constitutes the minimal sufficient statistic. The statistic $\bar{x}$ is θ oriented. It is also sufficient (for θ) when ϕ is fixed at the value zero. Therefore, $\bar{x}$ is H-sufficient for θ and so the Fraser-Hájek theorems proved earlier recommend marginalization to $\bar{x}$. However, the reduction of the data from $(\bar{x}, \bar{y})$ to $\bar{x}$ will mean a substantial loss of information on θ in the event $\phi = 1$. From the full data we should be able to tell (with a reasonable amount of certainty if m and n are large) whether $\phi = 0$ or 1. (If E stands for the event $m(\bar{x} - \bar{y})^2 > (m+n)\bar{y}^2$, then it is easy to check that the maximum likelihood (ML) estimator $\hat{\phi}$ of ϕ is the indicator of E and that the ML estimator of θ is $\hat{\theta} = (1 - \hat{\phi})\bar{x} + \hat{\phi}(m\bar{x} + n\bar{y})/(m + n)$.)

This example does not contradict the good statistical sense that led us to the generalized (or partial) sufficiency principle S*, but only tells us not to be unduly impressed with Fraser's mathematical proof of the principle. The statistical literature is full of this kind of proof (see for instance Lehmann (1959)) where we start on the wrong foot either by delimiting the discussion to a conveniently small (and nice) class of decision procedures or by simplifying the hypothetical risk function by an ad hoc maximization process. The author is very skeptical about the relevance of this kind of statistical mathematics in theoretical statistics.

8. THE BAYESIAN WAY

After a long journey through a whole forest of confusing ideas and examples, we seem to have lost our way. Let us now see if our Bayesian guide can find a way out of this wilderness for us.

According to a Bayesian, the role of the data $(\mathcal{E}, x)$ is to act as an operator on the experimenter's prior

[1] A referee has pointed out that a similar example appears in Barndorff-Nielsen (1973).

opinion q (a probability measure on Ω) and to transform it into a posterior opinion $q_x{}^*$. This transformation is effected through a formal use of the Bayes Theorem and the likelihood function $L(\omega) = p(x|\omega)$ generated by the data.

With $\omega = (\theta, \phi)$, where θ is the parameter of interest and ϕ is the nuisance parameter, the Bayesian analysis of data is always firmly anchored to the posterior marginal distribution $q_x{}^\dagger$ on Θ defined as

$$q_x{}^\dagger(\theta) = \sum_\phi q_x{}^*(\theta, \phi) \ ,$$

where $q_x{}^*(\omega) = L(\omega)q(\omega)/\sum_\omega L(\omega)q(\omega)$. As we said in paragraph (10) of Section 1, the Bayesian way of eliminating the nuisance parameter from the argument is to integrate it out from the posterior distribution of (θ, ϕ).

In 1942, A.N. Kolmogorov defined the notion of a sufficient statistic in the following Bayesian terms:

Definition: The statistic T is sufficient if, for every prior q on Ω, the posterior $q_x{}^*$ depends on x only through T; that is, $T(x) = T(x')$ implies that $q_x{}^* = q_{x'}{}^*$.

In the discrete setup, there is no difficulty in proving the equivalence of the above definition and the classical Fisher definition of sufficiency. In the same 1942 paper, we find Kolmogorov suggesting the following definition of partial sufficiency.

Definition (K-Sufficiency): The statistic T is partially sufficient for θ if, for all prior q on Ω, the posterior marginal distribution $q_x{}^\dagger$ on Θ depends on x only through T. (Let us call such a statistic K-sufficient for θ.)

At last we seem to have something for which we have been looking for so long. However, it was demonstrated by Hájek (1965) that the definition of K-sufficiency is vacuous in the following sense:

Theorem (Hájek): If the parameter θ is not a constant in ω, then every T that is K-sufficient for θ is sufficient (in the usual sense).

Proof: Pretending as always that we are dealing with a discrete model, we first recall that if T is not sufficient then there must exist x, x' such that $T(x) = T(x')$, but the likelihood ratio $p(x|\omega)/p(x'|\omega)$ is not a constant in ω. Therefore, if T is not sufficient, then we must have x, x' and ω_1, ω_2 such that $T(x) = T(x')$ and

$$p(x|\omega_1)/p(x'|\omega_1) \neq p(x|\omega_2)/p(x'|\omega_2) \ . \qquad (8.1)$$

Let $\omega_1 = (\theta_1, \phi_1)$ and $\omega_2 = (\theta_2, \phi_2)$. There is no loss of generality in supposing that $\theta_1 \neq \theta_2$. (Otherwise, we choose $\omega_3 = (\theta_3, \phi_3)$, with $\theta_3 \neq \theta_1 = \theta_2$, and consider the ratio $p(x|\omega_3)/p(x'|\omega_3)$ along with any one of the two ratios in (8.1) that differs from it.) Now, consider the prior q whose entire mass is equally distributed over the two points ω_1 and ω_2. Observe that

$$q_x{}^\dagger(\theta_1) = q_x{}^*(\omega_1) = p(x|\omega_1)/\sum_{i=1}^{2} p(x|\omega_i) \ ,$$

and that a similar expression holds true for $q_{x'}{}^\dagger(\theta_1)$. In view of (8.1) it follows that

$$q_x{}^\dagger(\theta_1) \neq q_{x'}{}^\dagger(\theta_1) \text{ even though } T(x) = T(x') \ .$$

Thus, T not-sufficient implies T not K-sufficient. This proves the theorem. (Observe that we do not require θ and ϕ to be variation independent either in the definition of K-sufficiency or in the proof of the above theorem.)

The fault in Kolmogorov's definition of partial sufficiency is easily detected. We may try to correct this by restricting the discussion to a relatively small class Q of prior measures q on Ω. We find the following definition in Raiffa and Schlaifer (1961):

Definition (Q-Sufficiency): The statistic T is Q-sufficient for θ if, for all $q \in Q$, the posterior marginal distribution $q_x{}^\dagger$ on Θ depends on x only through T. (In the language of Raiffa and Schlaifer, such a T is called marginally sufficient with respect to Q.)

From the beginning, we have been concerned with the problem of eliminating a parameter ϕ that is "unrelated" to the parameter of interest θ. However, we have not as yet clearly stated what we mean by two unrelated parameters. Is it enough to say that θ and ϕ are unrelated if they are variation independent and if the loss depends only on the terminal decision and the parameter θ? Clearly not, but what else can a non-Bayesian say? Just ask a non-Bayesian what he means when he agrees that the unknown true height ϕ of Mount Everest is unrelated to the unknown number θ of civilians who lost their lives in the Vietnam war! A Bayesian has no problem in defining the term. He calls θ and ϕ unrelated parameters if, apart from the condition on the loss function, his prior q for $\omega = (\theta, \phi)$ is of the form

$$q(\theta, \phi) = q_1(\theta)q_2(\phi) \ .$$

Let Q_0 be the class of all (independent) priors q of the form $q(\theta, \phi) = q_1(\theta)q_2(\phi)$. When is a statistic T going to be Q_0-sufficient for θ in the sense of our modified Kolmogorov definition of partial sufficiency? We find the following result in Raiffa and Schlaifer (1961):

Theorem (Raiffa and Schlaifer): If, for all $x \in \mathfrak{X}$, the likelihood function factors as

$$p(x|\theta, \phi) = G(T, \theta)F(x, \phi) \ ,$$

then T is Q_0-sufficient.

Proof: If the prior distribution is $q(\theta, \phi) = q_1(\theta)q_2(\phi)$, then

$$\begin{aligned} q_x{}^\dagger(\theta) &= \sum_\phi q_x{}^*(\theta, \phi) \\ &= G(T, \theta)q_1(\theta)/\sum_\theta G(T, \theta)q_1(\theta) \end{aligned}$$

depends on x only through T.

The above theorem suggests the following definition.

Definition (L-Sufficiency): The statistic T is L-sufficient for θ if, for all $x \in \mathfrak{X}$, the likelihood factors as in the statement of the previous theorem.

We just proved that L-sufficiency implies Q_0-sufficiency. Is the converse true? The answer is, of course, no. If T is sufficient, in the sense of Fisher or Kolmogorov, then it is Q-sufficient for every Q and in particular for Q_0. Raiffa and Schlaifer's definition of Q_0-sufficiency for θ suffers from a defect very similar to that of Kolmogorov's definition of partial sufficiency (K-sufficiency). The definition is too wide and fails to pinpoint the exact notion of partial sufficiency we are after. Perhaps an example will make clear the point we are driving at.

Example: Let $x_1, x_2, \ldots, x_n$ be iid $N(\theta\phi, 1)$ where $0 < \theta < \infty$ is the parameter of interest and $\phi(=-1 \text{ or } 1)$ is the nuisance parameter. Just imagine $\mu(-\infty < \mu < \infty)$ to be the common mean and then let $\theta = |\mu|$ and $\phi = \text{Sgn } \mu$.

The statistic $T = |\bar{x}|$ is θ oriented. It is a reasonable estimator of θ, but is it, in some sense, partially sufficient for θ? Check that T is not Q_0-sufficient for θ. Indeed, the notion of Q_0-sufficiency leads us to $\bar{x}$ which is sufficient. If, however, we agree to restrict our discussion to the smaller class $Q_0' \subset Q_0$ of (independent) priors q of the form $q(\theta, \phi) = q_1(\theta)q_2(\phi)$ such that q_2 is the uniform prior on $\Phi = \{-1, 1\}$, then it is easy to check that $T = |\bar{x}|$ is Q_0'-sufficient for $\theta = |\mu|$.

If we look back on the proof of the one-way implication theorem above, then it will be clear that L-sufficiency takes us far beyond Q_0-sufficiency. If T is L-sufficient for θ then the posterior marginal $q_x^\dagger$ on Θ depends on the sample x only through T and on the prior $q = q_1q_2$ only through q_1. In Bayesian terms, we may redefine the notion of L-sufficiency as follows:

Definition (*B-Sufficiency*): The statistic T is B-sufficient (partially sufficient in a restricted Bayes sense) for θ if, for $q = q_1q_2 \in Q_0$ and $x \in \mathfrak{X}$, the posterior marginal $q_x^\dagger$ on Θ depends on x only through T and on q only through q_1. (Indeed, one may try to further generalize the above notion of partial sufficiency by restricting q to an arbitrary but fixed class Q of priors on $\Omega = \Theta \times \Phi$ and calling q_1 the prior marginal on Θ. In the present context we have, however, no use for such a generalization.) In the next section we develop the theme of B-sufficiency to its natural conclusion.

9. UNRELATED PARAMETERS

Let us consider a rather loosely formulated question: Under what circumstances can we recognize the nuisance parameter ϕ to be so *unrelated* to the parameter of interest θ that we can meaningfully isolate the whole of the relevant information about the parameter θ contained in the data $(\mathcal{E}, x)$?

This is a good test question with which we can try to classify a statistician into one or another of the numerous feuding groups (or mutual admiration societies) that divide the current community of statisticians. For instance, a pucca (fully baked) Bayesian will probably dismiss the question out of hand as naive, incompetent, and unnecessarily argumentative. This is because a pucca-Bayesian has no use for the notion of "information in the data." According to him the natural dwelling place for information is the head of a homo sapien, and he recognizes only two kinds of statistical information—prior and posterior. Being a pucca-Bayesian, he always knows his prior q as a well-defined probability measure on $\Omega = \Theta \times \Phi$. Given the data he can, therefore, compute the posterior information q_x^* and then isolate the information $q_x^\dagger$ on θ by integration.

In the pucca-Bayesian statistical theory of Bruno de Finetti and L.J. Savage, there is no room for a family Q of prior distributions. However, having examined the question from various angles, the author has come to recognize the merit of Kolmogorov's half-baked[2] Bayesian approach to the problem at hand. In the spirit of Kolmogorov, Raiffa, and Schlaifer, let us put down the following definition for unrelated parameters. Let $\theta \in \Theta$ and $\phi \in \Phi$ be two parameters that enter into the statistical structure or model of an experiment $\mathcal{E}$, and let Q_0 be the class of all product probability distributions $q = q_1q_2$ on $\Omega = \Theta \times \Phi$.

Definition (*Unrelatedness*): The parameters θ, ϕ are unrelated relative to a model of the experiment $\mathcal{E}$ if, for all prior $q \in Q_0$ and all sample outcomes x of $\mathcal{E}$, the posterior distributions q_x^* also belong to the class Q_0.

If the likelihood function $L(\theta, \phi|x) = p(x|\theta, \phi)$ factors as

$$p(x|\theta, \phi) = A(\theta, x)B(\phi, y) \ , \tag{9.1}$$

then, for any prior $q(\theta, \phi) = q_1(\theta)q_2(\phi)$, it is easily seen that the posterior factors as

$$q_x^*(\theta, \phi) = q_x^\dagger(\theta)q_x'(\phi) \ ,$$

where

$$q_x^\dagger(\theta) = A(\theta, x)q_1(\theta)/\sum_\theta A(\theta, x)q_1(\theta) \ ,$$

with a similar expression holding true for $q_x'(\phi)$. Conversely, if

$$q_x^*(\theta, \phi) = p(x|\theta, \phi)q_1(\theta)q_2(\phi)/\sum_{\theta,\phi} pq_1q_2$$

belongs to Q_0 then it is equally clear that $p(x|\theta, \phi)$ must factor in the manner of (9.1) above. We thus have the

Theorem: The parameters θ, ϕ are unrelated relative to a model of the experiment $\mathcal{E}$ if and only if the likelihood function factors in the manner of (9.1).

It is then easy to recognize whether the parameter of interest is unrelated (in the preceding sense) to the nuisance parameter or not. With such a recognition of unrelatedness, (and, of course, with the further condition that the nuisance parameter has nothing to do with the hazards of incorrect decisions) the Bayesian will not waste his time in figuring out his prior q_2 for ϕ as long as he is satisfied that his prior q for (θ, ϕ) must be in the class Q_0. He will carefully figure out his prior q_1 for θ and then work out his posterior for θ as

$$q_x^\dagger(\theta) = A(\theta, x)q_1(\theta)/\sum_\theta A(\theta, x)q_1(\theta) \ .$$

[2] The Hindi antonym of pucca is so hard to spell in English!

In Basu (1975) we examined in depth the question of information in the data. Our conclusion was that, relative to a particular statistical model for the experiment $\mathcal{E}$ in question, Fisher's notion of the "whole of the relevant information" about $\omega = (\theta, \phi)$ that is contained in the data $(\mathcal{E}, x)$ may be identified with the likelihood function

$$L(\theta, \phi \mid x) = p(x \mid \theta, \phi) \ .$$

What we are saying now is that when the likelihood comes factored as in (9.1), when, on a priori considerations, we are willing to regard θ and ϕ as independent entities, and when information gained on ϕ is of no direct relevance to the decision problem on hand (i.e., ϕ does not enter into the loss function), then we may regard the function

$$L^{\dagger}(\theta \mid x) = A(\theta, x)$$

as the "whole of the relevant information" on θ that is supplied by the data $(\mathcal{E}, x)$. This may be regarded as a generalized likelihood principle.

The generalized sufficiency principle $\mathcal{S}^*$ and the generalized conditionality principle $\mathcal{C}^*$ are in conformity with the above principle. The existence of a statistic T that is p-sufficient for θ or of a statistic Y that is S-ancillary for θ presupposes a factorization of the likelihood as in (9.1). The principles $\mathcal{S}^*$ and $\mathcal{C}^*$ are indirectly advising us to concern ourselves with the factor of $L(\theta, \phi \mid x)$ that involves only θ. This is precisely why the p-sufficiency and the S-ancillarity arguments do not lead us astray.

Also observe that we can have a statistic T that is L-sufficient (B-sufficient) for θ if and only if θ and ϕ are unrelated in the sense of the likelihood factoring as in (9.1). If and when the likelihood factors in the above manner, we can always fashion a statistic T that is minimal L-sufficient for θ and a statistic Y that is minimal L-sufficient for ϕ. For example, T will be defined in terms of the equivalence relation: $x' \sim x''$ if $A(\theta, x') = C(x', x'')A(\theta, x'')$ for all $\theta \in \Theta$. In general, such a T will fail to be θ-oriented; that is, T will not be p-sufficient for θ. Similarly, Y will, in general, fail to be S-ancillary for θ. Indeed, we shall give an example where T and Y are the same. In such an example the same statistic T is in some sense isolating all the relevant information about θ and also all the information about the unrelated parameter ϕ.

A major source of our confusion on the important question of when and how we can isolate the information on the parameter of interest, is the fact of our arguing (in the manner of Sir Ronald) in terms of statistics. The notion of a statistic as a measurable map has hardly any relevance at the data analysis stage. It was Sir Ronald who distorted the question "what is information?" to the question "what (statistic) has all the information?" He taught us that a statistic is sufficient if and only if it summarizes in itself all the relevant information in the data. In the same spirit, we have been looking for a statistic T that is partially sufficient for θ—a statistic that summarizes in itself all the relevant and usable information about θ in the event of ignorance about the nusiance parameter ϕ.

We end this marathon discussion with three examples of statistical models where the parameters come naturally separated in the manner of (9.1), and yet we cannot take advantage of the fact (and isolate the information on the parameter of interest) in terms of either the generalized sufficiency or the conditionallity principle.

Example 1: We have a multinomial distribution with three categories and with probabilities

$$\theta\phi, \ (1 - \theta)(1 + \phi)/2 \quad \text{and} \quad (1 + \theta)(1 - \phi)/2 \ ,$$

where $0 < \theta < 1$ and $0 < \phi < 1$. With n observations, the three frequencies (n_1, n_2, n_3) constitute the minimal sufficient statistic, and the likelihood factors as

$$2^{-(n_2+n_3)}[\theta^{n_1}(1 - \theta)^{n_2}(1 + \theta)^{n_3}][\phi^{n_1}(1 + \phi)^{n_2}(1 - \phi)^{n_3}] \ .$$

We do not have any statistic that is p-sufficient, H-sufficient or S-ancillary for θ. The statistic $T = (n_1, n_2, n_3)$ is minimal L-sufficient (B-sufficient) for θ and also for ϕ. The (likelihood) information in the data on the parameter of interest θ is crying to be isolated as

$$L^{\dagger}(\theta) = \theta^{n_1}(1 - \theta)^{n_2}(1 + \theta)^{n_3} \ .$$

If θ and ϕ are independent a priori and if ϕ does not enter into the loss function, then a Bayesian will analyze the data in the same manner as he would have done in the hypothetical case when ϕ were known to be equal to $\frac{1}{2}$, say. Can anyone suggest a reasonable sample-space analysis of the data?

Example 2: Let $0 < \theta < \infty$ and $0 < \phi < \infty$. Let X and Y be two random variables with probability density functions

$$\theta e^{-\theta(x-\phi)}I(x - \phi) \quad \text{and} \quad \phi e^{-\phi(y+\theta)}I(y + \theta) \ ,$$

respectively, where $I(\cdot)$ stands for the indicator of the positive half of the real line. The sample consists of n independent observations $x_1, x_2, \ldots, x_n$ on X together with an independent set $y_1, y_2, \ldots, y_n$ of n independent observations on Y. Observe that the likelihood neatly factors as

$$[\theta^n \exp(-n\theta\bar{x})I(y_{(1)} + \theta)] \cdot [\phi^n \exp(-n\phi\bar{y})I(x_{(1)} - \phi)] \ ,$$

where $x_{(1)} = \min x_i$ and $y_{(1)} = \min y_i$. Clearly, the two parameters θ, ϕ are unrelated relative to the model. The statistic $(\bar{x}, y_{(1)})$ is B-sufficient (L-sufficient) for θ. The Bayesian analysis of the data is very simple as ϕ gets eliminated almost by itself. Can anyone suggest how to deal with the nuisance parameter in non-Bayesian terms?

Anyone who would sneer at the last two examples, on the grounds that they are not apparently related to any real life problem, is advised to take a hard look at the next example.

Example 3: The experiment consists of the observation, for each of n week days in a large metropolitan area, of the number of accidents involving one or more auto-

mobiles and also the corresponding number of such accidents involving one or more fatalities. The parameter of interest is the proportion θ of automobile accidents that result in death. The mean number ϕ of auto accidents per working day is the nuisance parameter. The statistical model for our record,

$$x = \{(x_1, y_1), \ldots, (x_n, y_n)\} ,$$

of the number of accidents x_i and the corresponding number of fatal accidents y_i on the ith day ($i = 1, 2, \ldots, n$) is that we have a set of n independent observations on a pair of random variables (X, Y) such that X is a Poisson variable with mean ϕ and Y, given X, is a Binomial variable Bin (X, θ). Now with $N = \sum x_i$ and $T = \sum y_i$, the likelihood neatly factors as

$$p(x|\theta, \phi) = A(x)\{\phi^N \exp(-n\phi)\}\{\theta^T(1-\theta)^{N-T}\} . \quad (9.2)$$

If n were a preselected constant, then the statistic N, distributed as a Poisson variable with mean $n\phi$, would qualify as an S-ancillary for θ. In this case the generalized conditionality principle will eliminate ϕ and will permit us to argue in some sample-space terms. Sir Ronald would have advised us to reduce the data to the minimal sufficient statistic (N, T), hold the ancillary N as fixed (at its observed value), and then look upon T as an observation on a Binomial variable with parameters N (known) and θ (unknown).

What happens if we do not preselect n but let it be determined by the very system that was under observation? Suppose we continue our observations until $T = \sum y_i$ exceeds a preselected number, say 10. How should we analyze the data then? Observe that our stopping rule has no effect on the likelihood function which comes factored in the same form as (9.2) above. Now the triple (n, N, T) constitutes the minimal sufficient statistic—the statistic T is nearly a constant but not quite. The statistics (N, T) and (n, N) are B-sufficient (L-sufficient) for θ and ϕ, respectively, but the notions of p-sufficiency and S-ancillarity are vacuous in this instance.

In a subsequent article, we shall study in depth various conditionality and marginalization arguments which have been put forward for the purpose of eliminating a nuisance parameter that is *not* unrelated to the parameter of interest in the present sense of separated (factored) likelihood.

[*Received January 1976. Revised December 1976.*]

REFERENCES

Anderson, E.B. (1967), "On Partial Sufficiency and Partial Ancillarity," *Skandinavisk Aktuarietidskrift*, 50, 137–52.

Barnard, G.A., Jenkins, G.M., and Winsten, C.B. (1962), "Likelihood Inference and Time Series (with Discussions)," *Journal of the Royal Statistical Society*, Ser. A, 125, 321–75.

———, and Sprott, D.A. (1971), "A Note on Basu's Examples of Anomalous Ancillaries (with Discussions)," *Foundations of Statistical Inference*, Toronto: Holt, Rinehart & Winston, 163–76.

Barndorff-Nielsen, O. (1973), "Exponential Families and Conditioning," published Sc.D. thesis, Department of Mathematics, University of Copenhagen.

Basu, D. (1959), "The Family of Ancillary Statistics," *Sankhyā*, 21, 247–56.

——— (1964), "Recovery of Ancillary Information," *Sankhyā*, A 26, 3–16.

——— (1965), "Problems Relating to the Existence of Maximal and Minimal Elements in Some Families of Statistics (Subfields)," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 41–50.

——— (1975), "Statistical Information and Likelihood (with Discussions)," *Sankhyā*, A, 37, 1–71.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference (with Discussions)," *Journal of the American Statistical Association*, 57, 269–326.

Cox, D.R. (1958), "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–72.

——— (1971), "The Choice Between Alternative Ancillary Statistics," *Journal of the Royal Statistical Society*, Ser. B, 33, 251–5.

Dawid, A.P. (1975), "On the Concepts of Sufficiency and Ancillarity in the Presence of Nuisance Parameters," *Journal of the Royal Statistical Society*, Ser. B, 37, 248–58.

Durbin, J. (1961), "Some Methods of Constructing Exact Tests," *Biometrika*, 48, 41–55.

Fisher, R.A. (1934), *Statistical Methods for Research Workers*, 5th ed., Edinburgh: Oliver and Boyd.

——— (1956), *Statistical Methods and Scientific Inference*, 1st ed., Edinburgh: Oliver and Boyd.

Fraser, D.A.S. (1956), "Sufficient Statistics with Nuisance Parameters," *Annals of Mathematical Statistics*, 27, 838–42.

Hájek, J. (1965), "On Basic Concepts of Statistics," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 139–62.

Kolmogorov, A.N. (1942), "Determination of the Center of Dispersion and Degree of Accuracy for a Limited Number of Observations," (in Russian) *Izvestija Akademii Nauk* SSSR, Ser. Mat. 6, 3–32.

Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, New York: John Wiley & Sons.

Linnik, Yu.V. (1965), "On the Elimination of Nuisance Parameters in Statistical Problems," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 267–80.

——— (1968), *Statistical Problems with Nuisance Parameters*, Translations of Math. Monographs, Vol. 20, Providence, Rhode Island: American Mathematical Society.

Neyman, J. (1935), "On a Theorem Concerning the Concept of Sufficient Statistic," (in Italian), *Giornale dell' Instituto Italiano degli Atturai*, 6, 320–34.

———, and Pearson, E.S. (1936), "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses," *Statistical Research Memoirs of the University of London*, 1, 133–37.

———, and Scott, E.L. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Olshevsky, L. (1940), "Two Properties of Sufficient Statistics," *Annals of Mathematical Statistics*, 11, 104–6.

Prohorov, Yu.V. (1965), "Some Characterization Problems in Statistics," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 341–50.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge: Harvard University Press.

Sandved, E. (1966), "A Principle for Conditioning on an Ancillary Statistic," *Skandinavisk Aktuarietidskrift*, 50, 39–47.

——— (1972), "Ancillary Statistics in Models Without and With Nuisance Parameters," *Skandinavisk Aktuarietidskrift*, 55, 81–91.

Stein, Charles (1945), "A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance," *Annals of Mathematical Statistics*, 16, 243–58.