Journal of Statistical Planning and Inference 2 (1978) 1-13.
© North-Holland Publishing Company

ON PARTIAL SUFFICIENCY: A REVIEW*

D. BASU**
The Florida State University, Tallahassee, Fl, U.S.A.

Received 8 September 1976; revised manuscript received 12 September 1977
Recommended by O. Kempthorne

The notion of a sufficient statistic—a statistic that summarizes in itself all the relevant information in the sample x about the universal parameter ω—is acclaimed as one of the most significant discoveries of Sir Ronald A. Fisher. It is however not well-recognized that the related notion of a partially sufficient statistic—a statistic that isolates and exhausts all the relevant and usable information in the sample about a sub-parameter $\theta=\theta(\omega)$—can be very elusive if the question is posed in sample space terms. In this review article, the author tries to unravel the mystery that surrounds the notion of partial sufficiency. For mathematical details on some of the issues raised here one may refer to Basu (1977).

AMS subject classifications: 60A05, 62A15.

Key words:
Ancillary statistics, information, nuisance parameters, specific sufficiency, θ-oriented statistics.

1. Introduction

In the beginning we have a parameter of interest—an unknown state of nature θ. With a view to gaining additional information on θ, we plan and then perform a statistical experiment $\mathscr{E}$ and thus generate the sample x. The problem of data analysis is to extract all the relevant information in the data $(\mathscr{E}, x)$ about the parameter of interest θ.

The notion of partial sufficiency arises in the context where the statistical model

$$\{(\mathscr{X}, \mathscr{A}, P_\omega): \omega \in \Omega\}$$

of the experiment $\mathscr{E}$ involves the universal parameter ω and where $\theta=\theta(\omega)$ is a sub-parameter In this case it is natural to ask:

Question A: What is the whole of the relevant information about θ that is available in the data $(\mathscr{E}, x)$?

*Based on an invited talk given by the author at the August, 1976 annual meeting of the Institute of Mathematical Statistics.

**Partially supported by NSF research grant no. MCS77 01661.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-5825-9_27, © Springer Science+Business Media, LLC 2011

It is not easy for a non-Bayesian to face up to this question. Most of us would feel more at ease when the question is rephrased in the following familiar form:

Question B: What statistic T summarizes in itself the whole of the relevant information about θ that is available in the sample x?

Let us understand that the two questions A and B, though similarly phrased, are very different in their orientations. Question A is clearly addressed to the particular data $(\mathscr{E}, x)$. But in B we are searching for a principle of data reduction. We may rephrase B in the following nearly equivalent form:

Question B* Does there exist a statistic T such that, in some meaningful sense, there is no loss of information on θ in the reduction of the data $(\mathscr{E}, x)$ to $(\mathscr{E}_T, t)$, where $\mathscr{E}_T$ is the marginal experiment—perform $\mathscr{E}$ but record only $t = T(x)$—corresponding to the statistic T?

Questions B or B*, when asked in the context of the universal parameter ω, led Fisher to the important notion of a sufficient statistic. But the same question, when asked in the context of a sub-parameter θ, turns out to be surprisingly resistant to a neat solution. The notion of partial sufficiency is indeed shrouded in a lot of mystery.

It is interesting to note that Sir Ronald introduced the notion of sufficiency into statistical literature (Fisher, 1920) first in the context of partial sufficiency. With a sample $x = (x_1, x_2, \ldots, x_n)$ from a normal population with unknown μ and σ, Fisher (1920) was concerned with the relative precisions of the two estimators

$$s_1 = (\tfrac{1}{2}\pi)^{1/2} \sum |x_i - \bar{x}|/n, \qquad s = [\sum (x_i - \bar{x})^2/n]^{1/2}$$

of the standard deviation σ. [Fisher had used the notations σ_1 and σ_2 for the above estimators, but we have opted for the more familiar s.] Introducing this paper in Fisher (1950), Sir Ronald described the main thrust of his 1920 argument in the following terms.

> "..., but the more general point is established that, for a given value of s, the (conditional) distribution of s_1 is independent of σ. Consequently, when s, the estimate based on the mean square is known, a value of s_1, the estimate based on the mean deviation, gives no additional information as to the true value (of σ). It is shown that the same proposition is true if any other estimate is substituted for s_1, and consequently the whole of the relevant information respecting the variance which a sample provides is summed up in the single estimate s".

[*Author's note:* The proposition stated in the final sentence of the above quoted paragraph was not proved in Fisher (1920). Indeed, the proposition is not true unless we limit the discussion to location invariant statistics.]

In Fisher (1922), p. 316 we find the first mention of the now famous:

Criterion of Sufficiency: That the statistic chosen should summarize the whole of the relevant information supplied by the sample.

On the same page we find it suggested that, in the case of a sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma)$, the statistic s fully satisfies the criterion of sufficiency. It is thus clear that from the very beginning Sir Ronald had been grappling with the notion of partial sufficiency.

In this article we shall be examining several definitions of partial sufficiency that have been proposed from time to time. In every case we shall look back on this original problem of Fisher and ask ourselves the question: "Does this definition make s partially sufficient for σ?"

[*Author's note*: The name "sufficient" is, of course, very misleading. We should never have allowed an expression like "T is sufficient for θ" to creep into any statistical text. It is less misleading to use expressions like "T is sufficient for the sample x" or "T isolates and exhausts all the information in x about θ". Perhaps we should agree to substitute the name "sufficient" by the more descriptive characterization "exhaustive", which also comes from Fisher. Having said all these, we are nevertheless going to use the expression "partially sufficient for θ" in the rest of this essay!]

2. Specific sufficient statistics

In Neyman and Pearson (1936) we find one of the earliest attempts at making some sense of the elusive notion of partial sufficiency. Let us suppose that the parameter of interest θ has a "variation independent" complement ϕ—that is, the universal parameter ω may be represented as $\omega = (\theta, \phi)$ with the domain of variation Ω of ω being the Cartesian product $\Theta \times \Phi$ of the respective domains of θ and ϕ. In this case, we have (from Neyman–Pearson) the following:

Definition (specific sufficiency). The statistic $T: \mathscr{X} \to \mathscr{T}$ is *specific sufficient* for the parameter θ if, for every fixed $\phi \in \Phi$, the statistic T is sufficient in the usual sense—that is, T is sufficient with respect to the restricted model

$$\{(\mathscr{X}, \mathscr{A}, P_{\theta,\phi}) : \theta \in \Theta, \phi \text{ fixed}\}$$

for the experiment $\mathscr{E}$.

With a sample $x = (x_1, x_2, \ldots x_n)$ of fixed size n from $N(\mu, \sigma)$, the sample mean $\bar{x}$ is specific sufficient for μ. The sample standard deviation s is, however, not specific sufficient for σ. Even though $\bar{x}$ is specific sufficient for μ, in no meaningful sense of the terms can we suggest that $\bar{x}$ exhaustively isolates all the relevant information in the sample x about the parameter μ. Surely, we also need to know s in order to be able to speculate about, say, the precision of $\bar{x}$ as an estimate of μ. Clearly, we are looking for something more than specific sufficiency.

The fact of T being specific sufficient for θ may be characterized in terms of the following factorization of the frequency (or density) function p on the sample space $\mathscr{X}$:

$$p(x \mid \theta, \phi) = G(T(x), \theta, \phi)\ H(x, \phi).$$

Alternatively, we may characterize the specific sufficiency of T (for θ) by saying that the conditional distribution of any other statistic T_1, given T and (θ, ϕ), depends on (θ, ϕ) only through ϕ.

Before going on to other notions of partial sufficiency, it will be useful to state the following:

Definition (θ-oriented statistics). The statistic $T:\mathscr{X}\to\mathscr{T}$ is *θ-oriented* if the marginal (or sampling) distribution of T—th:.t is, the measure $P_\omega T^{-1}$ on $\mathscr{T}$-depends on ω only through $\theta=\theta(\omega)$. In other worr.s, $\theta(\omega_1)=\theta(\omega_2)$ implies

$$P_{\omega_1}(T^{-1}B)=P_{\omega_2}(T^{-1}B)$$

for all 'measurable' sets $B\subset\mathscr{T}$.

It should be noted that the notion of θ-orientedness does not rest on the existence of a variation independent complementary parameter ϕ. In our basic example of a sample from $N(\mu,\sigma)$, observe that $\bar{x}$ is not μ-oriented but that s is σ-oriented.

3. Partial sufficiency

If we put together the two definitions of the previous section, then we have the following definition of partial sufficiency that is usually attributed to Fraser (1956).

Definition. The statistic T is *partially sufficient* for θ if it is specific sufficient for θ and is also θ-oriented.

See Basu (1977) for a number of examples of partially sufficient statistics. In the example of a sample $(x_1,x_2,\ldots,x_n)$ from $N(\mu,\sigma)$, the statistic $\bar{x}$ is not partially sufficient for μ as it is not μ-oriented and the statistic s is not partially sufficient for σ as it is not specific sufficient for σ. In view of the specific sufficiency part of the above definition, it is necessary that the parameter θ has a variation independent complement ϕ. The requirement of θ-orientedness brings in the unpleasant consequence that T may be partially sufficient for θ but a wider statistic T_1 need not be. Indeed, the whole sample x is never partially sufficient for θ.

The notion of partial sufficiency may be characterized in terms of the following factorization criterion:

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

where g and h denote respectively the marginal probability function of T and the conditional probability function of x given T. Note that the marginal distribution is θ-oriented and the conditional distribution is ϕ-oriented.

The interest in the Fraser definition of partial sufficiency stems from the following generalization (Fraser, 1956) of the Rao–Blackwell argument. Let $a(\theta)$ be an arbitrary real valued function of θ and let $W(y,\theta)$ denote the loss sustained when $a(\theta)$ is estimated by y. Let us suppose that, for each $\theta\in\Theta$, the loss function $W(y,\theta)$ is convex in y. Finally, let $\mathscr{U}$ be the class of all estimators U such that the risk function

$$r_U(\theta)=r_U(\theta,\phi)=\mathbf{E}[W(U,\theta)|\theta,\phi]$$

is finite and depends on (θ,ϕ) only through θ.

Theorem (Fraser). *If T is partially sufficient for θ, then for any $U \in \mathscr{U}$ there exists an estimator $U_0 = U_0(T) \in \mathscr{U}$ such that $r_{U_0}(\theta) \leqq r_U(\theta)$ for all $\theta \in \Theta$.*

The proof of the theorem consists of choosing and fixing a particular value ϕ_0 of ϕ and then considering the statistic $U_0 = U_0(T) = \mathbf{E}(U \mid T, \theta, \phi_0)$ as an estimator of $a(\theta)$. That U_0 does not involve the parameter θ follows from the supposition that T is sufficient for θ when ϕ is fixed at ϕ_0. That $U_0 \in \mathscr{U}$ follows from the supposition that T is θ-oriented. The rest follows at once from Jensen's inequality.

The above theorem may be generalized along the lines suggested by Hájek (1967). Let $\mathscr{U}'$ be the class of all estimators U for which the risk function $r_U(\theta, \phi)$ is finite (but not necessarily free of ϕ). Let $\mathrm{R}_U(\theta) = \sup_\phi r_U(\theta, \phi)$ be the maximum risk associated with U for a particular θ.

Theorem (Hájek). *If T is partially sufficient for θ, then for any $U \in \mathscr{U}'$ there exists a $U_0 = U_0(T)$ such that $R_{U_0}(\theta) \leqq R_U(\theta)$ for all θ.*

The definition of U_0 is the same as in the previous theorem. The rest of the proof follows from the following chain of relations

$$R_U(\theta) \geqq r_U(\theta, \phi_0) \geqq r_{U_0}(\theta, \phi_0) = R_{U_0}(\theta).$$

If $U \in \mathscr{U}$, that is, if the risk function for U is free of ϕ, then $r_U(\theta) \equiv R_U(\theta)$ and so the above theorem is a generalization of the Fraser theorem.

Let us take note of the fact that the proofs of the previous two theorems rest heavily on the supposition that T is θ-oriented but make very little use of the supposition that T is specific sufficient for θ. What is needed is the sufficiency of T (for θ) for just one specified value ϕ_0 of ϕ. Consider the following example.

Example. Let $x = (x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n)$ be $m+n$ independent normal variables with unit variances and with $E(x_i) = \theta$ $(i = 1, 2, \ldots, m)$ and $E(y_j) = \theta\phi$ $(j = 1, 2, \ldots, n)$, where $\theta \in [a, b]$ is the parameter of interest and $\phi \in \{0, 1\}$ is the nuisance parameter. The likelihood function factors as

$$p(x \mid \theta, \phi) = A(x) \exp[-\tfrac{1}{2}m(\bar{x} - \theta)^2] \exp[-\tfrac{1}{2}n(\bar{y} - \theta\phi)^2].$$

The pair $(\bar{x}, \bar{y})$ constitute the minimal sufficient statistic. The statistic $\bar{x}$ is θ-oriented and is sufficient for θ when $\phi = 0$. Therefore, we can invoke either the Fraser or the Hájek complete class theorem and suggest a reduction of the data x to the statistic $\bar{x}$. However, such a data reduction will clearly result in a substantial loss of information in the event $\phi = 1$. Looking at the full data we should usually be able to make a good guess of the true value of ϕ. For instance, if $m = 2$, $n = 200$, $\bar{x} = 16.02$ and $\bar{y} = 17.45$ then we know for (almost) sure that $\phi = 1$ and should naturally rebel against the idea of reducing the data to $\bar{x}$.

This example highlights the inherent weakness of the Fraser–Hájek argument. Fraser limited his discussion to the class $\mathscr{U}$ of estimators U whose risk functions involve only θ. It is not at all clear why we have to limit our universe of discourse to such a limited class. [It is true that the statistical literature is so full of Fraser-type limited complete class theorems. Familiar examples of such theorems abound in the theories of best unbiased estimates, best similar region tests, best invariant procedures, etc.] In this example, the class of estimators of θ that are functions of $(\bar{x}, \bar{y})$ is complete in the class $\mathscr{U}'$ of all estimators, provided the loss function is convex. But the only functions of $(\bar{x}, \bar{y})$ that belong to $\mathscr{U}$ are those that do not involve $\bar{y}$. Thus, Fraser's requirement that we limit the discussion to $\mathscr{U}$ sort of forces $\bar{y}$ out of the picture even though it contains a lot of information on θ.

Hájek considered the wider class $\mathscr{U}'$ but eliminated the nuisance parameter from the argument by redefining the risk function as

$$R_U(\theta)=\sup_{\phi} r_U(\theta,\phi).$$

This method of eliminating the nuisance parameter from the risk function has been made popular by Lehmann (1959) in his famous text on tests of statistical hypotheses. A generalized version of the Minimax Principle is being invoked in this elimination argument. The author is not at all clear in his mind about the statistical content of this generalized principle. The example of this section is clearly in conflict with the principle.

4. *H*-sufficiency

Hájek (1967) pushed Fraser's notion of partial sufficiency to its natural boundary in the following manner. For each $\theta\in\Theta$ let $\Omega_\theta=\{\omega:\theta(\omega)=\theta\}$ and let $\bar{\mathscr{P}}_\theta$ be the convex hull of the family $\mathscr{P}_\theta=\{P_\omega:\omega\in\Omega_\theta\}$ of the probability measures on the sample space $\mathscr{X}$. The class $\bar{\mathscr{P}}_\theta$ is the class of all probability measures Q_θ on $\mathscr{X}$ that has the representation

$$Q_\theta(A)=\int_{\Omega_\theta}P_\omega(A)\mathrm{d}\xi_\theta(\omega)$$

for all measurable sets A, where ξ_θ is some 'mixing' probability measure on Ω_θ. [Note that we are riding slipshod over the usual measurability requirements.]

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 $Q_\theta\in\bar{\mathscr{P}}_\theta$ such that

(i) T is sufficient with respect to the model $\{(\mathscr{X},\mathscr{A},Q_\theta):\theta\in\Theta\}$ and

(ii) T is θ-oriented in the model $\{(\mathscr{X},\mathscr{A},P_\omega):\omega\in\Omega\}$.

Observe that the notion of H-sufficiency (unlike the Fraser definition of partial sufficiency) does not require θ to have a variation independent complement ϕ. If T is partially sufficient in the sense of Fraser, then it is also H-sufficient. In order to see this, we have only to choose and fix $\phi_0\in\Phi$ and then take $Q_\theta=P_{\theta,\phi_0}$ which is a mixture probability corresponding to a degenerate mixing measure.

Also observe that the requirement of θ-orientedness in the definition of H-sufficiency has the same unfortunate consequence (as in the case of Fraser's definition) that T may be H-sufficient but a wider statistic (e.g., the whole sample x) need not be so. Hájek (1967) sought to remedy this fault in his definition by putting in the additional clause (almost as an afterthought) that any statistic T_1 wider than an H-sufficient T should be regarded as H-sufficient. But such a wide definition of partial sufficiency cannot be admitted when we are concerned with the problem of isolating the whole of the relevant information about a subparameter.

The two theorems of the previous section may now be consolidated in the following complete class theorem. [For a proof refer to p. 361 of Basu (1977).]

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

Let us look back on the classical problem of a sample $x = (x_1, x_2, \ldots, x_n)$ of fixed size n from $N(\mu, \sigma)$. No statistic T can be H-sufficient for μ. This is because T can be μ-oriented only if it is an ancillary statistic, in which case it cannot, of course, be partially sufficient for μ. [This remark holds true for a general location-scale parameter set-up with μ as the location parameter.] On the other hand the statistic s is σ-oriented. Let us examine whether s is H-sufficient for σ.

The density (or likelihood) function factors as

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

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

For each $\sigma \in (0, \infty)$, let ξ_σ be our choice of the mixing measure on the range space R_1 of the nuisance parameter μ. The corresponding family $\{Q_\sigma : 0 < \sigma < \infty\}$ of mixture measures on the sample space R_n will have the density function

$$\begin{aligned}\bar{p}(x \mid \sigma) &= \int_{-\infty}^{\infty} p(x \mid \mu, \sigma)\, d\xi_\sigma(\mu)\\ &= A(\sigma) \exp\left[-\frac{ns^2}{2\sigma^2}\right] \int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu).\end{aligned}$$

We shall recognize s as H-sufficient for σ provided we can find a family $\{\xi_\sigma\}$ of mixing measures such that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu) = B(\bar{x}) C(\sigma) \tag{1}$$

because in that case $\bar{p}(x \mid \sigma)$ will factor as

$$\bar{p}(x \mid \sigma) = A(\sigma) \exp\left[-\frac{ns^2}{2\sigma^2}\right] B(\bar{x}) C(\sigma)$$

establishing condition (i) of the definition of H-sufficiency.

One way to ensure (1) is to choose for ξ_σ the uniform distribution over the whole of R_1. But, with a family $\{Q_\sigma\}$ of improper mixtures, the proof of the Hájek theorem will break down. If the range of σ is the whole of the positive half line, then it can be shown that the factorization (1) can be achieved with no proper mixing. However, if we are willing to set a finite upper bound K for the parameter σ—from a practical point of view this is hardly a restriction—then it is easy to check that the choice of ξ_σ as the normal distribution with mean zero and variance $(K^2 - \sigma^2)/n$ will achieve the desired factorization (1). The above argument of Hájek (1967) establishing the H-sufficiency of s for σ $(0 < \sigma < K)$ is very intriguing. At this point we like to contrast the approaches of Fisher and Hájek to the question of partial sufficiency of s for σ. First, let us look at the question from the:

Fisher Angle: The pair $(\bar{x}, s)$ being jointly sufficient for (μ, σ), contains the whole of the available information on the parameter of interest σ. Furthermore, the two statistics $\bar{x}$ and s, being stochastically independent, yield independent (additive, that is) bits of information on σ. If μ were known, then we have n 'degrees of freedom' worth of information on σ. Of these, the statistic s summarizes in itself $n-1$ 'degrees of freedom' worth of information on σ. If the only (prior) information about μ that we have is $-\infty < \mu < \infty$, then there is no way that we can recover any part of the (at most one 'degree of freedom' worth of) information contained in $\bar{x}$ about μ. It is in this situation of no (prior) information on μ that Fisher would label s as exhaustive of all available and usable information on σ. And in the event of no (prior) information on σ either (other than $0 < \sigma < \infty$) Fisher would invoke his celebrated fiducial argument to declare that the status of the parameter σ has been altered from that of an unknown constant to that of a random variable with (fiducial) probability distribution $\sqrt{n}s/\chi_{n-1}$. Observe that the fiducial distribution of σ depends on the sample only through the statistic s.

A sort of improper Bayesian justification for the Fisher intuition on the problem at hand can be given by suggesting that, for every prior $q(\mu, \sigma)$ for the parameter (μ, σ) that is of the form

$$q(\mu, \sigma)\,\mathrm{d}\mu\mathrm{d}\sigma = g(\sigma)\,\mathrm{d}\mu\mathrm{d}\sigma$$

[μ and σ are independent a-priori and μ has the (improper) uniform distribution over the whole real line], the posterior marginal distribution of σ depends on the sample x only through the statistic s. Furthermore, the fiducial distribution of σ corresponds to the case where $g(\sigma) = 1/\sigma$ $(0 < \sigma < \infty)$. Although Fisher never put his arguments in the above straightforward Bayesian framework, the fact remains that Fisher's thinking on the problem of inference had a distinct Bayesian orientation.

Hájek Angle: On the surface, Hájek's partial sufficiency argument carries a distinct Bayesian flavour. His mixing measure ξ_σ—normal with zero mean and $(K^2 - \sigma^2)/n$ as variance for μ may be interpreted as the prior conditional distribution of μ given σ. With any prior $q(\mu, \sigma)$ of the form

$$q(\mu, \sigma)\,\mathrm{d}\mu\mathrm{d}\sigma = [\mathrm{d}\xi_\sigma(\mu)]g(\sigma)\,\mathrm{d}\sigma$$

the posterior marginal distribution of σ will depend on the sample x only through the statistic s. It will, however, be very hard to make any Bayesian interested in a prior $q(\mu, \sigma)$ of the above form. Apart from the fact that q depends on the sample size (which it should not), it is not possible to make any sense of q as a measure of prior belief pattern. The main thrust of the Hájek argument is, however, not Bayesian at all. He was using the Bayesian device (of averaging over the parameter space) only as a mathematical artifact to prove a complete class theorem in the fashion of Abraham Wald.

We have already pinpointed the flaw in Hájek's definition of partial sufficiency through our example of the previous section. In that example, $\bar{x}$ is H sufficient for θ even though marginalization to $\bar{x}$ will entail a substantial loss of information on θ in the case of the (easily discernable) event $\phi=1$. Many such examples can be easily constructed. [See Barndorff-Nielsen (1973) and Basu (1977) for other such examples.]

5. Invariantly sufficient statistics

In this section we briefly review George Barnard's thoughts on the knotty question of partial sufficiency of s for σ. The following quotation is from p. 113 of Barnard (1963).

> "The definition of sufficiency which has become universally accepted required that the distribution of any function of the observations, conditional on a fixed value of the sufficient statistic, should be independent of the parameter in question, and there is no doubt that with, this definition, s fails to be sufficient for σ. However, as was usual for him, Fisher's definition of sufficiency was designed to embody a logical notion, that of providing the whole of the available relevant information for a given parameter and the definition just referred to does not altogether succeed in this object.
>
> The availability or otherwise of information is critically dependent on knowledge or lack of knowledge. Obviously if σ is already known, s provides us with no information whatsoever. The failure of s to satisfy the definition given above for sufficiency arises from the fact that the distribution of $\bar{x}-\mu$ (with the usual notations) depends also on σ. However,... μ is given as unknown, and so the information in $\bar{x}-\mu$ is unavailable.
>
> As already remarked, Fisher was very much concerned, up to the end of his life, with the difficulty of expressing in precise mathematical form, the notions corresponding to 'known' and 'unknown'. The present writer several times suggested to him, in connection with parameters such as μ in the case of the normal distribution, ..., that these parameters correspond to groups under which the problems considered are invariant, and the notion of ignorance of μ can be represented in terms of group invariance properties".

Barnard's thoughts on the problem are best understood in the context of the simple example of a sample $x=(x_1,x_2,\ldots x_n)$ of fixed size n from $N(\mu,\sigma)$. The group $G=\{g_a : a\in R_1\}$ of transformations

$$g_a(x_1,x_2,\ldots x_n)=(x_1+a,x_2+a,\ldots x_n+a)$$

of the sample space R_n onto itself is associated with the group $\bar{G}=\{\bar{g}_a : a\in R_1\}$ of transformations

$$\bar{g}_a(\mu,\sigma)=(\mu+a,\sigma)$$

of the parameter space onto itself. The group $\bar{G}$ leaves the parameter of interest σ invariant but acts transitively on (traces a single orbit on the domain of) the nuisance parameter μ.

The problem of estimating the parameter σ is invariant with respect to the group G of transformations $g_a: \mathscr{X} \to \mathscr{X}$. The maximal invariant is the difference statistic

$$D = (x_2 - x_1, x_3 - x_1, \ldots x_n - x_1)$$

The statistic s is invariantly sufficient for σ in the sense that

(i) s is a function of D and is, therefore, σ-oriented, and

(ii) the conditional distribution of any other invariant statistics $s_1 = s_1(D)$, given s, is the same for all possible values of σ (and, of course, of μ as well).

[The notion of invariantly sufficient statistic is due to Charles Stein. See Hall, Wijsman and Ghosh (1965), and Basu (1969) for further discussion on the subject.]

We are now ready for the following

Question. What is the logical necessity for restricting our attention to only G-invariant estimators of σ?

The standard argument for restricting attention to only such T that satisfies the identity

$$T(x_1 + a, x_2 + a, \ldots x_n + a) = T(x_1, x_2, \ldots x_n)$$

for all samples $x \in R_n$ and all $a \in R_1$—that is, to measurable functions of the maximal invariant $D = (x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1)$ – runs along the following lines:

Argument. The sample $(x_1, x_2, \ldots, x_n)$ consist of n i.i.d. $N(\mu, \sigma)$'s with $\mu(-\infty < \mu < \infty)$ 'unknown' and with σ as the parameter of interest. If we shift the origin of measurement to $-a$. then the sample will take on the new look $(x_1 + a, x_2 + a, \ldots, x_n + a)$. The new model for the new-look sample will then correspond to n i.i.d. $N(\mu + a, \sigma)$'s.

Note that the new mean $\mu + a$ is 'equally unknown' as μ and that σ remains unaltered. The problem of estimating σ (with μ unknown), therefore, remains invariant with any shift in the origin of measurement. Now, an estimator T is a formula for arriving at an estimate $T(x_1, x_2, \ldots x_n)$ based on the sample $x = (x_1, x_2, \ldots, x_n)$. With the same sample represented differently as $(x_1 + a, x_2 + a, \ldots x_n + a)$, but with the problem (of estimating σ) unaltered, the same formula T will yield the estimate $T(x_1 + a, x_2 + a \ldots x_n + a)$. Clearly, the formula T will look rather ridiculous if $T(x_1 + a, x_2 + a, \ldots x_n + a)$ is not equal to $T(x_1, x_2, \ldots x_n)$ for some x and a.

The above invariance argument of Pitman–Stein–Lehmann has been sold in many different packages to a vast community of statisticians. However, a close look at the present package will immediately reveal the fact that the argument does not really add up to anything that is logically compelling.

For one thing, the part of the argument that asserts that the problem remains invariant with any shift of the origin of measurement is questionable. The argument rests heavily on the supposition that $\mu + a$ is 'equally unknown' as μ. Only an improper Bayesian with uniform prior (over the whole real line) for μ can make a case for such a statement.

Secondly, implicit in the argument lies the supposition that the choice of the estimator (estimating formula) T as a function on the sample space may depend on the statistical model (which, in this case, does not change with any shift in the origin of measurement) and the kind of 'average performance characteristics' that we find satisfactory but must not (repeat not) depend on any pre-conceived notions that we may have on the parameters in the model. This, of course, is not a tenable supposition (as all Bayesians will readily agree).

Let T_q be a typical Bayes estimator of σ that corresponds to the prior distribution q for (μ, σ)—for the sake of this argument let us imagine $T_q(x)$ to be the posterior mean of σ for a given sample x and the prior q. In T_q we thus have a well-defined formula for estimating σ. Every such formula T_q is invariant for every shift in the origin of measurement. This is because when the origin is shifted to $-a$, the sample $(x_1, x_2, \ldots, x_n)$ shifts to $(x_1+a, x_2+a, \ldots, x_n+a)$, the parameters (μ, σ) move to $(\mu+a, \sigma)$ and the prior q changes itself to the corresponding prior q_a for $(\mu+a, \sigma)$. It is easy to see then that

$$T_q(x_1, x_2, \ldots x_n) = T_{q_a}(x_1+a, x_2+a, \ldots x_n+a)$$

for all q, x and a. Thus, no Bayes rule violates the essence of the invariance argument.

However, if for a particular q, we look upon $T_q(x)$ as a function on the sample space, then we shall find that the function will typically depend on x through both $\bar{x}$ and s. [As we have noted in the previous section, for all (improper) priors q of the form $q(\mu, \sigma)\,\mathrm{d}\mu\mathrm{d}\sigma = g(\sigma)\,\mathrm{d}\mu\mathrm{d}\sigma$ and also for some curious looking proper priors of the Hájek kind. the posterior marginal distribution of σ will depend on x only through s and so with such a choice of the prior q, the Bayes estimator $T_q(x)$ for σ will be G-invariant as a function on the sample space.]

There is no logical necessity for restricting our attention to only G-invariant estimators as long as we take care to avoid using estimating procedures that do not recognize the arbitrariness that is inherent in the choice of the origin of measurement, etc. As we have noted earlier, all Bayes estimation procedures are invariant in a sense.

6. Final remarks

Sir Ronald was deeply concerned with the notion of information (about a parameter) in the data, but never directly faced up to such basic questions as: What is information? How informative is this data? Have we obtained enough information on the parameter of interest? etc.

The mathematical definition of information that we got from Fisher is a most curious one. The definition does not relate to the concept of information in the data but is supposed to bring out the notion of information in (the statistical model of) an experiment and the associated family of marginal experiments. Even then, the Fisher information $I(\omega)$ can hardly be interpreted in terms of the average (or expected) amount of knowledge gained (or uncertainties removed) about the universal parameter ω when the experiment is performed. And we get no prescription from Sir Ronald about how to 'marginalize' his information function (or matrix) to a sub-parameter. We must reject the notion of Fisher information on the ground of irrelevance in the present context.

The Fisher criterion of sufficiency—that the statistic chosen should summarize the whole of the relevant information supplied by the data—should be looked upon only as a principle of data reduction relative to a particular statistical model of the experiment. The earliest thoughts of Fisher on the subject of sufficiency crystalized around the following two propositions that are stated here relative to a fixed experiment $\mathscr{E}$ that is already endowed with an assumed statistical model.

Proposition 1. *To reduct (or marginalize) the data x to the statistic $T = T(x)$ will entail a total loss of all available information on the (universal) parameter ω if the marginal distribution of T is the same for all possible values of ω. Any such statistic T may be regarded as 'marginally uninformative' about ω.*

Proposition 2. *To reduce the data x to the statistic T will entail no loss of available information on ω if the conditional distribution of every other statistic T_1 given T is the same for all possible values of ω. Such a statistic T may be called sufficient, fully informative, or exhaustive of all available information on ω.*

Is it not remarkable that we now have the notions of 'no information' and 'full information' (meaning, exhaustive of all available information) without ever mentioning what we mean by information?! If by information we mean the state of our knowledge about the parameter ω, then should we not speculate about it in terms of the parameter space Ω rather than in terms of the sample space $\mathscr{X}$?!

It so happens that Fisher's 'sample space' definition of sufficient (information-full, that is) statistic agrees with the following Bayesian definition of sufficiency due to A. N. Kolmogorov (1942):

Definition. The statistic T is *sufficient* if, for every prior $q(\cdot)$ on Ω, the posterior distribution $q(\cdot|x)$ on Ω depends on x only through $T(x)$.

It is to the lasting credit of Sir Ronald that, having discovered the 'sample space' definition of sufficiency, he was able to put the notion in the correct perspective by characterizing a sufficient statistic as that characteristic of the sample knowing which we can determine the likelihood function up to a multiplicative factor. Fisher recognized that, relative to a given model, the whole of the relevant information in the data is summarized in the corresponding likelihood function. This is only a short step away from the Bayesian insight on the knowledge business.

The 'sample space' definition of sufficiency for the universal parameter ω is all right. But the weakness and inadequacy of this approach becomes apparent when we try the sample space way to 'isolate' all the 'available' relevant information on a sub-parameter. Note that we now have to deal with the new term 'isolate' and that the term 'available' suddenly springs to life with a new meaning. Fraser, Hajek and Barnard all seem to have tacitly assumed that T can isolate information on θ only if it is θ-oriented. This sample space requirement of θ-orientedness for the partially sufficient T has been a major source of our trouble with the notion of partial sufficiency. The statistical insight that leads to θ-orientedness as a prime requirement for partial sufficiency, cannot be reconciled with any Bayesian insight on the subject. What if there are no non-trivial θ-oriented statistic? Can't we then isolate the information on θ? What is information on θ? How can we isolate something that we have not even cared to define?

Barnard (1963) said "..., the notion of ignorance on μ can be represented in terms of group invariance properties." What is ignorance? Lack of prior information? How can we talk about lack of information when we have not even attempted to define what we mean by information? In any case, how can we possibly characterize ignorance on μ in terms of group invariance properties of the model? Who is ignora ? The scientist or the model?!

In September 1967 the author had asked the late Professor Renyi the question: "Why are you a Bayesian?" Promptly came back the answer: "Because I am interested in the notion of information. I can make sense of the notion in no other way".

Acknowledgement

The author wishes to thank Professor Oscar Kempthorne for carefully going over an earlier draft of the paper.

References

Barnard, G.A. (1963). Some logical aspects of the fiducial argument. *J. R. Statist. Soc.* B25, 111-114.

Barndorff, Nielsen O. (1973). Exponential families and conditioning. Sc.D. thesis, Dept. of Maths., Univ. of Copenhagen, Denmark.

Basu, D. (1969). On sufficiency and invariance. In: *Essays in Probability and Statistics*. Univ. of North Carolina and Indian Statistical Institute. 61-84.

Basu, D. (1975a). Statistical information and likelihood (with discussions). *Sankhyā* A37, 1-71.

Basu, D. (1977). On the elimination of nuisance parameters. *Jl. Am. Stat. Assoc.* 72, 355-366.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and the mean square error. *Monthly Notices of the Royal Astronomical Soc.* 80, 758-770. [Also reproduced in Fisher (1950).]

Fisher. R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London* A222, 309-368. [Also reproduced in Fisher (1950).]

Fisher. R.A. (1950). *Contributions to Mathematical Statistics*. Wiley, New York.

Fraser, D.A.S. (1956). Sufficient statistics with nuisance parameters. *Ann. Math. Statist.* 27, 838-842.

Hájek, J. (1967). On basic concepts of statistics. In: *Proc. Fifth Berkeley Symp.* 1, 139-162.

Hall, W.J., R.A. Wijsman and J.K. Ghosh (1965). The relationship between sufficiency and invariance. *Ann. Math. Statist.* 36, 575-614.

Kempthorne, Oscar and Leroy Folks (1971). *Probability Statistics and Data Analysis*. The Iowa State University Press, Ames, IA.

Kolmogorov, A.N. (1942). Determination of the centre of dispersion and degree of accuracy for a limited number of observations. *Izv. Akad. Nauk. USSR. Ser. Mat.* 6, 3-32. [In Russian.]

Neyman, J. (1935). On a theorem concerning the concept of sufficient statistic. *Giorn. Ist. Ital. Attuari* 6, 320-334. [In Italian.]

Neyman, J. and E.S. Pearson (1936). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Stat. Res. Memoirs* 1, 133-137.