ON PARTIALLY SUFFICIENT STATISTICS

by

D. Basu

(The Florida State University)
FSU Statistics Report No. M382

July 1976
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

This essay was prepared for presentation at the annual meeting of the Institute of Mathematical Statistics at New Haven, Connecticut in August 1976.
ON PARTIALLY SUFFICIENT STATISTICS

by

D. BASU
(The Florida State University)

ABSTRACT

The notion of a sufficient statistic - a statistic that 'summarizes in itself all the relevant information in the sample x' about the universal parameter $\omega$ - is generally acclaimed as the most important discovery of R. A. Fisher in the theory of statistics. However, it is not generally recognized that the notion of a partially sufficient statistic - that isolates and exhausts all the relevant information in the sample about a sub-parameter $\theta = \theta(\omega)$ - can be very elusive if we pose the question in the classical Fisher fashion. The notion of partial sufficiency appeared in Fisher (1920) in which the parameter of interest was a sub-parameter. In this essay we try to unravel the mystery that surrounds the notion of partial sufficiency.
1. INTRODUCTION

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

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

\[
\{(X, A, P_\omega): \omega \in \Omega\}
\]

for the experiment \( E \) involves a universal parameter \( \omega \) 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 \( \theta \) that is available in the data \((E, x)\)?

Sir Ronald, being a non-Bayesian, could not face up to this question and, therefore, took refuge behind a seemingly related

**Question B:** What statistic \( T \) summarizes in itself the whole of the relevant information about \( \theta \) that is available in the sample \( x \)?

Let us understand that the two questions, though similarly phrased, are very different in their orientations. Question A is clearly addressed to the particular data \((E, x)\). But in B, Fisher is searching for a statistic \( T \) - a measurable map of the sample space \( X \) into another space \( T \) - which is such that in some meaningful sense there is no loss of information in performing the marginal experiment
$E_T$ - perform $E$ but record only $T(x)$ - instead of the full experiment $E$.

Question B, when asked in the context of the universal parameter $\omega$, led R. A. Fisher to his most significant discovery in statistical theory, namely, the notion of a sufficient statistic. But the same question, when asked in the context of a sub-parameter, turns out to be surprisingly resistant to a neat solution. No other concept in statistical theory is shrouded in greater mystery than that of a partially sufficient statistic.

It is of interest to note that in the writings of Sir Ronald the earliest mention of the notion of sufficiency can be traced to Fisher (1920) where the author was actually grappling with the problem of inferring about a sub-parameter - the $\sigma$ of a normal population with unknown mean and variance. With a sample $x = (x_1, x_2, \ldots x_n)$ from the population Fisher was concerned with the relative precisions of the two estimators of $\sigma$, namely,

$$s_1 = \sqrt{\frac{1}{2n}} \sum |x_i - \bar{x}| / n \quad \text{and}$$

$$s = [\frac{1}{n} \sum (x_i - \bar{x})^2 / n]^{\frac{1}{2}}.$$

[Fisher used the notations $\sigma_1$ and $\sigma_2$ for the above estimators, but we 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 $\sigma$. 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 $\sigma$). 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: Even in the late forties, Sir Ronald did not seem to distinguish between the notions of an estimate and that of an estimator. Furthermore, the claim that in Fisher (1920) certain facts have been 'established' or 'shown' is not quite accurate.]

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 without ever realizing how elusive the answer to question B could be when it is asked in the context of a sub-parameter.

In the sequel 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 $\sigma$?"

[Note: The name 'sufficient' is, of course, very misleading. We should never have allowed an expression like 'T is sufficient for $\theta$' 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 $\theta$'. 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
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 \( \theta \) has a 'variation independent' complement \( \phi \) - that is, the universal parameter \( \omega \) may be represented as \( \omega = (\theta, \phi) \) with the domain of variation \( \Omega \) of \( \omega \) being the Cartesian product \( \theta \times \phi \) of the respective domains of \( \theta \) and \( \phi \). In this case, we have (from Neyman-Pearson) the following

**Definition (Specific Sufficiency):** The statistic \( T: X \rightarrow T \) is specific sufficient for the parameter \( \theta \) 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

\[
\{(X, A, p_{\theta, \phi}): \theta \in \Theta, \phi \text{ fixed}\}
\]

for the experiment \( E \).

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 \( \mu \). The sample standard deviation \( s \) is, however, not specific sufficient for \( \sigma \). Even though \( \bar{x} \) is specific sufficient for \( \mu \), 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 \( \mu \). 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 \( \mu \). Clearly, we are looking for something more than specific sufficiency.

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

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

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

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

**Definition (\theta-oriented statistics):** The statistic $T: \mathcal{X} \to T$ is \theta-oriented if the marginal (or sampling) distribution of $T$ - that is, the measure $P^T_\omega$ on $T$ - depends on $\omega$ only through $\theta = \theta(\omega)$. In other words, $\theta(\omega_1) = \theta(\omega_2)$ implies

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

for all 'measurable' sets $B \subset T$.

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

3. PARTIALLY SUFFICIENT STATISTICS

The following interesting definition of partial sufficiency is generally attributed to Fraser (1956). But the definition appears in Olshevsky (1940) who attributed it to Neyman (1935).
Definition (p-sufficiency): The statistic \( T \) is p-sufficient for \( \theta \) if it is i) specific sufficient for \( \theta \) and ii) \( \theta \)-oriented.

In view of the requirement (i) in the definition of p-sufficiency, the parameter \( \theta \) needs to have a variation independent complement \( \phi \). The requirement (ii) is too restrictive and brings in the unpleasant consequence that \( T \) may be p-sufficient but a wider statistic \( T_1 \) may not be. Indeed, with this definition of partial sufficiency, the whole sample \( x \) is not p-sufficient for \( \theta \).

Example 1: Let \( x = (n_1, n_2, n_3) \) have a multinomial distribution with 
\[ n_1 + n_2 + n_3 = n \] (the sample size) and \( p_1, p_2, p_3 = 1 - p_1 - p_2 \) as the parameters (class probabilities). The parameters \( p_1, p_2 \) are not variation independent. However, with the re-parametrization \( \theta = p_1, \phi = p_2/(1 - p_1) \), it is easily seen that the statistic \( T \) defined as \( T(x) = n_1 \) is p-sufficient for \( \theta \).

Example 2: Let \( y \) and \( z \) be independent Poisson variables with means \( \theta \phi \) and \( \theta(1 - \phi) \) respectively, where \( 0 < \theta < \infty \) and \( 0 < \phi < 1 \). It is easy to check that the statistic \( T = y + z \) is p-sufficient for \( \theta \).

The notion of \( T \) being p-sufficient for \( \theta \) may be characterized in terms of the following factorization criterion:

\[
p(x|\theta, \phi) = g(T|\theta)h(x|T, \phi)
\]

where \( g \) and \( h \) define respectively the marginal distribution of \( T \) and the conditional distribution of \( x \) given \( T \).

The general interest in p-sufficiency is due to the following generalization of the Rao-Blackwell argument that we find in Fraser (1956). Suppose we are
considering the problem of point estimation of a real valued function \( a(\theta) \) of the parameter of interest. Let us suppose that the loss function \( W(y, \theta) \) - the loss suffered when \( a(\theta) \) is estimated by \( y \) - is convex (from below) in \( y \) for each \( \theta \in \Theta \). Let \( \mathcal{U} \) be the class of all estimators \( U \) of \( a(\theta) \) such that the 'risk function'

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

is well-defined and depends on \((\theta, \phi)\) only through \( \theta \).

**Theorem (Fraser):** If \( T \) is p-sufficient for \( \theta \), then for any \( U \in \mathcal{U} \) there exists an estimator \( U_0 = U_0(T) \) - a measurable function of \( T \) - such that

\[
r_{U_0}(\theta) \leq r_U(\theta)
\]

for all \( \theta \in \Theta \).

**Proof:** Choose and fix \( \phi_0 \in \phi \) and define \( U_0 \) as follows:

\[
U_0(T) = \mathbb{E}(U|T, \theta, \phi_0).
\]

Since \( T \) by definition is specific sufficient for \( \theta \), it follows that \( U_0(T) \) is well-defined as an estimator - that is, the parameter \( \theta \) does not enter into the definition of \( U_0 \). Since \( T \) is \( \theta \)-oriented, so also must be \( U_0 \). Therefore, the risk function corresponding to \( U_0 \) depends on \((\theta, \phi)\) only through \( \theta \). By supposition so also is the risk function corresponding to \( U \). Therefore,

\[
r_U(\theta) = \mathbb{E}[W(U, \theta)|\theta, \phi_0] \geq \mathbb{E}[W(U_0, \theta)|\theta, \phi_0] \quad \text{(Jensen's inequality)}
\]

\[
= r_{U_0}(\theta)
\]

for all \( \theta \in \Theta \).
The above theorem may be generalized along the lines suggested by Hajek (1965). Let $U'$ be the class of all estimators $U$ for which the risk function $r_U(\theta, \phi)$ is well-defined but not necessarily free of $\phi$. Let

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

be the 'maximum risk' associated with $U$ for a given value of the parameter $\theta$. We then have the following

**Theorem (Hajek):** If $T$ is $p$-sufficient for $\theta$, then for any $U \in U'$ there exists an $U_0 = U_0(T)$ such that

$$R_{U_0}(\theta) \leq R_U(\theta)$$

for all $\theta \in \Theta$

**Proof:** Define $U_0$ as in the proof of the previous theorem and observe that

$$R_U(\theta) \geq r_U(\theta, \phi_0) \quad \text{(from defn. of } R_U)$$

$$\geq r_{U_0}(\theta, \phi_0) \quad \text{(Jensen's inequality)}$$

$$= r_{U_0}(\theta) \quad \text{(since } U_0 \text{ is } \theta\text{-oriented)}$$

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

for all $\theta \in \Theta$.

That the above is a generalization of the Fraser theorem will be clear once we observe that $r_U(\theta) \equiv R_U(\theta)$ for each $U$ in the class $U$ (as defined in the context of the earlier theorem).
Now, let us note the fact that the proofs of the previous two theorems rest heavily on the supposition that $T$ is $\theta$-oriented but make little use of the supposition that $T$ is specific sufficient for $\theta$. What was needed is the existence of a $\phi_0$ in $\Phi$ such that, with $\phi$ fixed at $\phi_0$, the statistic $T$ becomes sufficient for $\theta$ in the usual sense.

4. $H$-Sufficiency

The remark made at the end of the previous section suggests the following generalization — due to Hajék (1965) — of the notion of $p$-sufficiency. For each $\theta \in \Theta$, let $\Omega_\theta = \{\omega: \theta(\omega) = \theta\}$ and let $P_\theta$ be the 'convex hull' of the family

$$P_\theta = \{P_\omega: \omega \in \Omega_\theta\}$$

of measures on $(\chi, A)$. The class $P_\theta$ is the set of all $Q$ that has a representation of the form

$$Q(A) = \int_{\Omega_\theta} P_\omega(A) \, d\xi_\theta(\omega), \text{ for all } A \in A,$$

where $\xi_\theta$ is some 'mixing' probability measure on $\Omega_\theta$. [For the sake of brevity of exposition we have ridden slipshod over the usual measurability requirements.]

**Definition** ($H$-sufficiency): The statistic $T$ is $H$-sufficient (partially sufficient in the sense of Hajék) for $\theta$, if, for each $\theta \in \Theta$, there exists a choice of a measure $Q_\theta \in P_\theta$ such that

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

and

ii) $T$ is $\theta$-oriented in the full model $\{(\chi, A, P_\omega): \omega \in \Omega\}$. 
That p-sufficiency - defined in the particular context where \( \Omega = \Theta \times \Phi \) - implies H-sufficiency, follows from the fact that we can choose and fix \( \phi_0 \in \Phi \) and then define \( Q_\theta = P_{\theta, \phi_0} \), which is a mixture probability corresponding to a 'degenerate' mixing measure. Observe that the notion of H-sufficiency, unlike that of p-sufficiency, does not presuppose \( \Theta \) having a variation independent complement \( \phi \). Also observe that the condition of \( \Theta \)-orientedness (in the definition of H-sufficiency) has the unfortunate consequence (as in the case of p-sufficiency) that \( T \) may be H-sufficient but a wider statistic (for instance, the whole sample \( x \)) may not be. Hajek (1965) 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 our concern is to isolate the 'whole of the relevant information' about a sub-parameter \( \Theta \).

That the two theorems of the previous section remain true when the condition of p-sufficiency is replaced by the less stringent one of H-sufficiency, is checked as follows. Let \( a(\theta) \) be a real valued parameter that we wish to estimate. Let us suppose that the loss function \( W(y, \theta) \) - the loss suffered when \( a(\theta) \) is estimated, by \( y \) - is convex from below. For any \( U \in U' \), let \( r_U(\omega) = \int_X W(U, \theta) \, dP_\omega \) be the risk function and let \( R_U(\theta) = \sup_{\omega \in \Omega_\theta} r_U(\omega) \) be the maximum risk corresponding to a particular \( \theta \in \Theta \).
Now, with a statistic $T$ that is $H$-sufficient for $\theta$, let us define $U_0$ as

$$U_0(T) = E(U|T, Q_\theta)$$

where $\{Q_\theta : \theta \in \Theta\}$ is the family of mixture probability measures that appear in the definition of $H$-sufficiency. In view of condition (i) in the definition of $H$-sufficiency, it follows that $\theta$ does not enter into the definition of $U_0$ — that is, $U_0$ is a bonafide estimator. From the Jensen's inequality we then have that, for all $\theta \in \Theta$.

\[ (*) \quad \int_X W(U_0, \theta)dQ_\theta \leq \int_X W(U, \theta)dQ_\theta. \]

Since $T$ is $\theta$-oriented, so also must be $U_0 = U_0(T)$. Therefore, the risk function $r_{U_0}(\omega)$ is a constant — let us denote this by $\overline{r}_{U_0}(\theta)$ — on each $\Omega_\theta$. Since each $Q_\theta$ is some $\xi_\theta$ — mixture of

\[ \{P_\omega : \omega \in \Omega_\theta\} \]

it follows at once that the left hand side of $(*)$ is

\[ \int_{\Omega_\theta} r_{U_0}(\omega)d\xi_\theta (\omega) = \overline{r}_{U_0}(\theta) = R_{U_0}(\theta). \]

On the other hand, the right hand side of $(*)$ is

\[ \int_{\Omega_\theta} r_U(\omega)d\xi_\theta (\omega) \leq R_U(\theta). \]

In case $U \in \mathcal{U}$ (the class of estimators whose risk function depends on $\omega$ only through $\theta = \theta(\omega)$), the right hand side of $(*)$ would be equal to

\[ \overline{r}_{U}(\theta) = R_{U}(\theta). \]
We thus have the

Theorem (Hajék): If $T$ is $H$-sufficient for $\theta$, then, for each $U \in U'$, there exists $U_0 = U_0(T)$ such that

$$R_{U_0}(\theta) \leq R_U(\theta) \text{ for all } \theta \in \Theta.$$ 

Furthermore, if $U \in U$ then $\bar{r}_{U_0}(\theta) \leq \bar{r}_U(\theta)$ for all $\theta \in \Theta$.

Before examining Hajék's complete class theorem, 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 $\mu$. This is because $T$ can be $\mu$-oriented only if it is an ancillary statistic, in which case it cannot, of course, be partially sufficient for $\mu$ - this remark holds true for a general location-scale parameter set-up with $\mu$ as the location parameter. On the other hand the statistic $s$ is $\sigma$-oriented. Let us examine whether $s$ is $H$-sufficient for $\sigma$.

The density (or likelihood) function factors as

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

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

For each $\sigma \in (0, \infty)$, let $\xi_\sigma$ be our choice of the mixing measure on the range space $R_1$ of the nuisance parameter $\mu$. The corresponding family

$\{Q_\sigma: 0 < \sigma < \infty\}$ of mixture measures - on the sample space $R_n$ - will have the density function

$$\bar{p}(x|\sigma) = \int_{-\infty}^{\infty} p(x|\mu, \sigma) d\xi_\sigma(\mu)$$

$$= A(\sigma) \exp \left[ -\frac{ns^2}{2\sigma^2} \right] \int_{-\infty}^{\infty} \exp \left[ -\frac{n(\bar{x}-\mu)^2}{2\sigma^2} \right] d\xi_\sigma(\mu)$$
We shall recognize \( s \) as \( \mathcal{H} \)-sufficient for \( \sigma \) provided we can find a family \( \{ \xi_{\sigma} \} \) of mixing measures such that

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

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

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

establishing condition (i) of the definition of \( \mathcal{H} \)-sufficiency.

One way to ensure (**) is to choose for \( \xi_{\sigma} \) the uniform distribution over the whole of \( \mathbb{R}_1 \). But, with a family \( \{ Q_{\sigma} \} \) of improper mixtures, the proof of the Hajek theorem will break down. If the range of \( \sigma \) is the whole of the positive half line, then it can be shown that the factorization (**) will be achieved for no proper mixings. However, if we are willing to set a finite upper bound \( K \) for the parameter \( \sigma \) - from a practical point of view this is hardly a restriction - then it is easy to check that choice

\[
\xi_{\sigma} = N(0, \sqrt{(K^2 - \sigma^2)/n}), \quad 0 < \sigma < K
\]

as our family of mixing measures will achieve the desired factorization (**).

The above argument of Hajek (1965) establishing the \( \mathcal{H} \)-sufficiency of \( s \) for \( \sigma \) \((0 < \sigma < K)\) is very intriguing. At this point we like to contrast the approaches of Fisher and Hajek to the question of partial sufficiency of \( s \) for \( \sigma \). First, let us look at the question from the Fisher point of view.
The pair \((\bar{x}, s)\), being jointly sufficient for \((\mu, \sigma)\), contains the whole of the available information on the parameter of interest \(\sigma\). Furthermore, the two statistics \(\bar{x}\) and \(s\), being stochastically independent, yield independent (additive, that is) bits of information on \(\sigma\). If \(\mu\) were known, then we have \(n\) 'degrees of freedom' worth of information on \(\sigma\). Of these, the statistic \(s\) summarizes in itself \(n-1\) 'degrees of freedom' worth of information on \(\sigma\). If the only (prior) information about \(\mu\) 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 \(\mu\). It is in this situation of no (prior) information on \(\mu\) that Fisher would label \(s\) as exhaustive of all available and usable information on \(\sigma\). And in the event of no (prior) information on \(\sigma\) either other than \(0 < \sigma < \infty\) - Fisher would invoke his celebrated fiducial argument to declare that the status of the parameter \(\sigma\) has been altered from that of an unknown constant to that of a random variable with (fiducial) probability distribution \(\sqrt{n} \frac{s}{\chi^2_{n-1}}\). Observe that the fiducial distribution of \(\sigma\) 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 \((\mu, \sigma)\) - of the form

\[ q(\mu, \sigma) d\mu d\sigma = g(\sigma) d\mu d\sigma \]

[\(\mu\) and \(\sigma\) are independent a-priori and \(\mu\) has the (improper) uniform distribution over the whole real line], the posterior marginal distribution of \(\sigma\) depends on the sample \(x\) only through the statistic \(s\). Furthermore, the fiducial distribution of \(\sigma\) corresponds to the case where \(g(\sigma) = 1/\sigma\) \((0 < \sigma < \infty)\). Although Fisher
carefully avoided putting 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.

On the surface, Hajék's partial sufficiency argument carries a distinct Bayesian flavour. His mixing measure

$$\xi_\sigma = N(0, \sqrt{\frac{2}{(K-\sigma^2)}}/n)$$

for $\mu$ may be interpreted as the prior conditional distribution of $\mu$ given $\sigma$. With any prior $q(\mu, \sigma)$ of the form

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

the posterior marginal distribution of $\sigma$ 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 Hajék 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. It is interesting to observe that Fisher regarded $s$ as partially sufficient for $\sigma$ in the event $\mu$ is unbounded, whereas, Hajék would regard $s$ as partially sufficient when $\sigma$ is bounded.

The claim that $s$ is partially sufficient for $\sigma$ do not jolt our statistical intuition. After all, $s$ contains most, if not all, of the available information in the sample on $\sigma$. But let us take a close look at the following celebrated
example due to Neyman and Scott (1948):

**Example:** (Neyman and Scott): The sample \( x \) consist of \( 2n \) observations \((x_i, x'_i)\), \( i = 1, 2, \ldots, n \). The statistical model is that we have observed \( 2n \) independent normal variables with common standard deviation \( \sigma \) — the parameter of interest — and that \((x_i, x'_i)\) have common mean \( \mu_i (i = 1, 2, \ldots, n) \) — the vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_n) \) being the nuisance parameter. Let \( S^2 = \sum (x_i - x'_i)^2 \) and \( \bar{x}_i = (x_i + x'_i)/2, \)

\( i = 1, 2, \ldots, n. \) The density function is

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

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

The statistic \( S \) is clearly \( \sigma \)-oriented. But does it isolate and exhaust all the relevant and usable information about \( \sigma \)? How secure do we feel about marginalizing to \( S \) without taking a hard look at \((\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)\)? Suppose we find that the \( n \) sample means are nearly equal. Then we have the information that \( \mu_1, \mu_2, \ldots, \mu_n \) must be nearly equal and also the information that \( \sigma \) must be small. If \( \mu \) were known then (in the language of Sir Ronald) the sample would contain \( 2n \) units (degrees of freedom) of information on \( \sigma \). The statistic \( S \) summarizes in itself only \( n \) units of information. Are we really prepared to pretend complete ignorance (no prior information) on \( \mu \) and thereby sacrifice \( n \) degrees of freedom?

If we are prepared to suppose that \( \sigma \) has a finite upper bound \( n \), then following Hajek we can demonstrate the fact that \( S \) is H-sufficient for \( \sigma \). For the mixing measure \( \xi_\sigma \) on the range \( R_n \) of \( \mu \), we have only to choose the one
for which \( \mu_1, \mu_2, \ldots, \mu_n \) are iid normal variables with mean zero and variance \((\mu^2 - \sigma^2)/2\).

The following example from Basu (1950b) will pinpoint the fault in the notion of H-sufficiency.

**Example:** 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. About the means we are told that \( \text{Ex}_i = \theta \) (\( i = 1, 2, \ldots, m \)) and \( \text{Ey}_j = \theta \phi \) (\( j = 1, 2, \ldots, n \)), where \( \theta (\omega < \theta < \infty) \) is the parameter of interest and \( \phi (\phi = 0 \text{ or } 1) \) is the nuisance parameter. The likelihood function factors as

\[
p(x|\theta, \phi) = A(x) \exp\left(-\frac{m(x-\theta)^2}{2}\right) \exp\left(-\frac{n(y-\theta \phi)^2}{2}\right).
\]

The pair \((\bar{x}, \bar{y})\) constitute the minimal sufficient statistic for \((\theta, \phi)\). Now observe that \( \bar{x} \) is \( \theta \)-oriented and that it is sufficient (for \( \theta \)) when \( \phi \) is fixed at the value zero. Therefore, \( \bar{x} \) is H-sufficient for \( \theta \) - for each \( \theta \) choose the mixing measure \( \xi_\theta \) to be degenerate at \( \phi = 0 \). The complete class theorem of Hajek holds. But the reduction of the data to the H-sufficient statistic \( \bar{x} \) will entail a substantial loss of information on \( \theta \) if \( \phi = 1 \). 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. In the latter event, who would care to marginalize to \( \bar{x} \)?

The above example highlights the inherent weakness of the Fraser-Hajek type of complete class theorems. Fraser limited his discourse to the class \( U \) of estimators \( U \) whose risk functions are \( \theta \)-oriented - depends on \( \omega = (\theta, \phi) \) only through \( \theta \). Hajek considered the wider class \( U' \) but eliminated the nuisance
element from the risk function by considering only the maximum risk

\[ R_\theta(\theta) = \sup_{\theta(\omega) = \theta} R_\theta(\omega). \]

The statistical literature is full of the above kind of limited complete class theorems. According to the author, such theorems are of dubious value in the theory of statistical inference.

5. INVARIANTLY SUFFICIENT STATISTICS

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

"The definition of sufficiency which has become universally accepted requires 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, \( s \) fail to be sufficient for \( \sigma \). 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 \( \sigma \) 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 \( \sigma \). However \( \mu \) 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 $\mu$ 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 $\mu$ can be represented in terms of group invariance properties.

Bernard's thoughts on the problem is 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 $\mathcal{N}(\mu, \sigma)$. The group $G = \{g_a : a \in \mathbb{R}_1\}$ 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 $\mathbb{R}_n$ onto itself is associated with the group $\bar{G} = \{\bar{g}_a : a \in \mathbb{R}_1\}$ 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 $\sigma$ invariant but acts transitively on (traces a single orbit on the domain of) the nuisance parameter $\mu$.

The problem of estimating the parameter $\sigma$ is invariant with respect to the group $G$ of transformations $g_a : X \to X$. 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 \( \sigma \) in the sense that

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

ii) the conditional distribution of any other invariant statistic \( s_1 = s_1(D) \), given \( s \), is the same for all possible values of \( \sigma \) (and, of course, of \( \mu \) as well).

The notion of invariantly sufficient statistic is due to Charles Stein. See Hall, Wijisman and Ghosh (1965) and Basu (1969) for further discussions 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 \( \sigma \)?

The standard argument for restricting attention to only such \( T \) that satisfy the identity

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

for all samples \( x \in \mathbb{R}^n \) and all \( a \in \mathbb{R}^1 \) — 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 \( n \) iid \( N(\mu, \sigma)'s \) with \( \mu \) 'unknown' and with \( \sigma \) 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 be that we have observed \( n \) iid \( N(\mu + a, \sigma)'s \) — note that the new mean \( \mu + a \) is 'equally unknown' as \( \mu \) and that \( \sigma \) remains unaltered. The problem of estimating \( \sigma \) (with \( \mu \) '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 \(\sigma\)) 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 \(T(x_1 + a, x_2 + a, \ldots, x_n + a) \neq T(x_1, x_2, \ldots, x_n)\) 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. [The author can never cease to marvel at the great success story of the invariance argument in the statistical research of the last three decades.] 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 dubious proposition that \(\mu + a\) is 'equally unknown' as \(\mu\). [Only an improper Bayesian with uniform prior (over the whole real line) for \(\mu\) can make any sense of 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 \(\sigma\) that correspond to the prior distribution \(q\) for \((\mu, \sigma)\) - for the sake of this argument let us imagine \(T_q(x)\) to be
the posterior mean of $\sigma$ for a given sample $x$ and the prior $q$. In $T_q$, we thus have a well-defined formula for estimating $\sigma$. 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 $(\mu, \sigma)$ 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 of Pitman-Stein-Lehmann.

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 (improper) priors $q$ of the form

$$q(\mu, \sigma) d\mu d\sigma = g(\sigma) d\mu d\sigma$$

the posterior marginal distribution of $\sigma$ will depend on $x$ only through $s$ and so the Bayes estimator $T_q(x)$ for $\sigma$ 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.
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 our 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 \( \omega \) 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 \( E \) that is already endowed with an assumed statistical model.

**Proposition 1:** To reduce (or marginalize) the data \( x \) to the statistic \( T = T(x) \) will entail a total loss of all available information on the (universal) parameter \( \omega \), if the marginal distribution of \( T \) is the same for all possible values of \( \omega \).
Such a statistic $T$ may be called totally uninformative about $\omega$.

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

Isn't it 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 $\omega$, then should we not speculate about it in terms of the parameter space $\Omega$ rather than in terms of the sample space $X$?!

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 $\Omega$, the posterior distribution $q(\cdot|x)$ on $\Omega$ 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. It was Fisher who recognized (for the first time in writing) 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 $\omega$ 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, Hajék and Barnard all seem to have tacitly assumed that $T$ can isolate information on $\theta$ only if it is $\theta$-oriented. This sample space requirement of $\theta$-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 $\theta$-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 $\theta$-oriented statistic? Can't we then isolate the information on $\theta$? What is information on $\theta$? How can we isolate something that we have not even cared to define?

Barnard said" "...the notion of ignorance on $\omega$ can be represented in terms of group invariance properties." [See p. 19]. 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 $\omega$ in terms of group invariance properties of the model? Who is ignorant? 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."

[See the last two sections of Basu (1975b) for a discussion on partial sufficiency from the Bayesian point of view.]
REFERENCES


