Steffen L. Lauritzen

On the Interrelationships among Sufficiency, Total Sufficiency and Some Related Concepts

Preprint
July
1974

8

Institute of Mathematical Statistics
University of Copenhagen

Steffen L. Lauritzen

ON THE INTERRELATIONSHIPS AMONG SUFFICIENCY, TOTAL SUFFICIENCY AND SOME RELATED CONCEPTS

Preprint 1974 No. 8

INSTITUTE OF MATHEMATICAL STATISTICS

UNIVERSITY OF COPENHAGEN

July 1974

This is a reprint of Technical Report No. 18 from Department of Statistics, Stanford University, prepared under the auspices of 'Office of Naval Research Contract N00014-67-A-0112-0030.'

ON THE INTERRELATIONSHIPS AMONG SUFFICIENCY,

TOTAL SUFFICIENCY AND SOME RELATED CONCEPTS

by

Steffen L. Lauritzen*

Institute of Mathematical Statistics
University of Copenhagen

Abstract

As noted by various authors, the notion of sufficiency is to weak for problems in connection with statistical inference in stochastic processes. Various attempts have been made to impose extra conditions and in the present paper we shall discuss a few of these, with the purpose of discovering in which sense the concepts so defined differ and in which sense they are alike.

Key words: adequacy, conditional independence, S-structure, statistical inference, stochastic processes, summarizing statistics, total sufficiency, transitivity, universality.

* At present visiting Stanford University.

1. Introduction.

The literature on sufficiency is extensive and it is not the aim of the present paper to give a complete survey of this. We shall discuss the relationship between a number of notions introduced by various authors with different problems in mind, but all of them being of the same nature as sufficiency. Some of these notions were defined in terms of subfields of abstract probability spaces, but we shall restate all definitions in terms of statistics and discrete probabilities as our interest is directed more towards structural properties than technical ones.

2. Sufficiency, adequacy and summarizing statistics.

In the present section we shall investigate three different properties of statistics with the same basic idea, namely that they express the intuitive statement that a statistic contains all "relevant" information.

First the classical notion of a sufficient statistic as introduced by Fisher (1920). We shall define it the following way:

Let X be a random variable on a discrete, at most denumerable space E and t a mapping from E into another discrete space F. Let $\mathcal{P}$ be a family of probabilities on E and let $Y = t(X)$.

Definition 2.1. t is said to be sufficient for $\mathcal{P}$ if there is a fixed non negative real function φ on $E \times F$ so that for all $P \in \mathcal{P}$ and all $x \in E$:

$$P\{Y = y\} > 0$$

$$\Rightarrow P\{X = x | Y = y\} = \varphi(x,y).$$

A slightly stronger notion was introduced by Freedman (1962) with the pure probabilistic motivation of generalizing de Finettis theorem for exchangeable 0-1 random variables. The notion is however closely related to sufficiency, as we shall soon see. Again let X be a random variable on a discrete space E and t a mapping from E to a discrete space F.

Definition 2.2. A probability measure P on E is said to be summarized by t if for all $x, x' \in E$

$$t(x) = t(x') \Rightarrow P\{X = x\} = P\{X = x'\}.$$

In contrast to definition 2.1, we are not dealing with a family of probabilities but only with one probability measure. To be able to see the relation between a sufficient statistic and a summarizing statistic we have to define a summarizing statistic for a family of probabilities $\mathcal{P}$. In the previous notation we define

Definition 2.3. t is said to summarize $\mathcal{P}$ if all $P \in \mathcal{P}$ are summarized by t.

Remark: Note that the term "summarizing" is essentially related to discrete random variables as opposed to other concepts dealt with in the present paper.

This is stronger than sufficiency:

Proposition 2.1. *If $\mathcal{P}$ is a family of probability measures on E and $t: E \to F$ summarizes $\mathcal{P}$, then t is sufficient for $\mathcal{P}$.*

Proof. We shall just specify the function φ in definition 1. Define φ as

$$(1) \qquad \varphi(x,y) = \begin{cases} 0 & \text{if} \quad t(x) \neq y \\ \frac{1}{N(y)} & \text{for} \quad t(x) = y \,, \end{cases}$$

where $N(y)$ is the total number of x's so that $t(x) = y$, $0 < N(y) \leq \infty$.

If $P\{Y = y\} > 0$, we have

$$(2) \qquad \begin{aligned} &P\{X = x \mid Y = y\} \\ &= \frac{P\{X = x \wedge Y = y\}}{P\{Y = y\}} = \chi_{t^{-1}(y)}(x) \cdot \frac{P\{X = x\}}{\sum_{z \in t^{-1}(y)} P\{X = z\}} \,, \end{aligned}$$

where χ_A is the indicator function of the set A:

$$\chi_A(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \,. \end{cases}$$

If $t(x) = y$, we have

$$(3) \qquad P\{X = z\} = P\{X = x\} \quad \text{for all} \quad z \in t^{-1}(y) \,,$$

since t was a summarizing statistic. Hence

$$(4) \qquad \begin{aligned} P\{Y = y\} &= \sum_{z \in t^{-1}(y)} P\{X = z\} \\ &= P\{X = x\} \cdot N(y) \ . \end{aligned}$$

So

$$(5) \qquad P\{Y = y\} > 0 \implies N(y) < \infty$$

and

$$(6) \qquad \begin{aligned} P\{X = x | Y = y\} &= \chi_{t^{-1}(y)}(x) \cdot \frac{P\{X = x\}}{P\{X = x\}N(y)} \\ &= \chi_{t^{-1}(y)}(x) \cdot \frac{1}{N(y)} = \varphi(x,y) \ , \end{aligned}$$

which was to be proved.

So the notion of a summarizing statistic is stronger than that of a sufficient statistic in the sense that not only is the conditional distribution of X given $t(X)$ supposed to be known, but this distribution is supposed to have the specific form (6), i.e. <u>uniform</u> on the set $t^{-1}(t(X))$.

Barndorff-Nielsen and Skibinsky (1963) considered the problem of how much one could reduce a data set and still have all relevant information for the <u>prediction</u> of an unobserved random variable when the joint distribution of the data and the unobserved random variable was completely known and defined the notion <u>adequacy</u>. This definition was later extended to the case, where this joint distribution was only known to be a member of a specified family of distributions by Skibinsky (1967).

Let X and Z be random variables on discrete, at most denumerable spaces E and G. Let $\mathcal{P}$ be a family of distributions on $E \times G$ and let $\mathcal{P}_E$ denote the induced family of marginal distributions of X. Let t be a mapping from E to an at most denumerable space F.

Definition 2.4. t is said to be adequate for Z if

i) t is sufficient for $\mathcal{P}_E$

ii) for all $P \in \mathcal{P}$: $P\{X = x\} > 0$

$$\Longrightarrow P\{Z = z | X = x\} = P\{Z = z | t(X) = t(x)\} .$$

This definition suggests that the classical notion of sufficiency is not satisfactory to the theory of statistical inference in stochastic processes as the prediction of unobserved random variables (the future of the process observed) in most cases will be relevant. In the next section we shall consider some extra conditions that have been imposed on a sequence of statistics by various authors.

3. Sequences of statistics.

In the present section we shall let $X_1, X_2, \ldots$ be a sequence of random variables on discrete at most denumerable spaces $E_1, E_2, \ldots$ and let $\mathcal{P}$ be a family of probability measures on $E_1 \times E_2 \times \cdots$. Let $\mathcal{P}^{(n)}$ denote the family of marginal distributions of $X_1, \ldots, X_n$ induced by $\mathcal{P}$. We shall consider a sequence $t_1, t_2, \ldots$ of mappings.

$$t_n : E_1 \times \cdots \times E_n \to F_n ,$$

where F_n are discrete, at most denumerable, and let $Y_n = t_n(X_1, \ldots, X_n)$.

Bahadur (1954) introduced the term of a sufficient and transitive sequence of statistics in connection with sequential decision theory, which can be stated as follows.

Definition 3.1. The sequence $t_1, t_2, \ldots$ is said to be sufficient and transitive if for all n, t_n is adequate for Y_{n+1}.

In other words, $t_1, t_2, \ldots$ is sufficient and transitive iff it at each step n contains all information relevant for the prediction of the value of the next statistic. This is related to but different from the notion of a totally sufficient statistic, introduced by Lauritzen (1972) in terms of abstract measure spaces and restated in terms of discrete probability spaces in Lauritzen (1974).

Definition 3.2. t_n is said to be totally sufficient if it is adequate for $X_{n+1}, \ldots, X_{n+k}$ for all $k = 1, 2, \ldots$.

That the two notions are different can be seen by the following example:

Example 1. Let X_1, X_2 be independent Poisson distributed with mean $\lambda > 0$, and let $X_n = X_2 + Z_1 + \cdots + Z_{n-2}$ for $n \geqq 3$, where $Z_1, Z_2, \ldots$ are independent of X_1, X_2 and independent identically Poisson distributed with mean 1. The sequence $t_1, t_2, \ldots$ of mappings defined by

$$t_1(x) = x \quad \text{and} \qquad (7)$$
$$t_n(x_1, \ldots, x_n) = x_1 + x_2 \quad \text{for} \quad n \geqq 2$$

is sufficient and transitive, whereas e.g. t_2 is not totally sufficient as (X_1, X_2) and $X_3 = X_2 + X_1$ are not conditionally independent.

On the other hand, the sequence $s_1, s_2, \ldots$ defined as

$$(8) \quad \begin{cases} s_1(x) = x \\ s_2(x_1,x_2) = (x_1,x_2) \\ s_3(x_1,x_2,x_3) = (x_1,x_2,x_3) \\ s_4(x_1,x_2,x_3,x_4) = (x_1,x_2,x_4) \\ s_5(x_1,x_2,x_3,x_4,x_5) = (x_1,x_2,x_3,x_5) \quad \text{and} \\ s_n(x_1,\ldots,x_n) = (x_1,x_2,x_n) \quad \text{for} \quad n \geqq 6 \end{cases}$$

is totally sufficient but not sufficient and transitive, because (X_1,X_2,X_3,X_4) and $s_5(X_1,\ldots,X_5) = (X_1,X_2,X_3,X_5)$ are not conditionally independent given $s_4(X_1,\ldots,X_4) = (X_1,X_2,X_4)$.

If one wants to insure a sequence of totally sufficient statistics to be sufficient and transitive, an extra condition has to be imposed. The following algebraic property of a sequence of statistics is a slight weakening of "S-structure" as defined by Freedman (1962).

Definition 3.3. $t_1, t_2, \ldots$ is said to have $\sum$-structure if for all m,n

$$t_n(x_1,\ldots,x_n) = t_n(y_1,\ldots,y_n)$$

$$\Longrightarrow t_{n+m}(x_1,\ldots,x_n,x_{n+1},\ldots,x_{n+m}) = t_{n+m}(y_1,\ldots,y_n,x_{n+1},\ldots,x_{n+m})$$

In other words, $t_1, t_2, \ldots$ has $\sum$-structure if for all m and n, there is a mapping ψ_{nm}

$$\psi_{nm}: F_n \times E_{n+1} \times \cdots \times E_{n+m} \to F_{n+m}$$

so that

$$t_{n+m}(x_1,\ldots,x_{n+m}) = \psi_{nm}(t_n(x_1,\ldots,x_n),x_{n+1},\ldots,x_{n+m}) .$$

The term $\sum$-structure is chosen to emphasise that t_{n+m} is a "generalized sum" of $t_n(x_1,\ldots,x_n)$ and $(x_{n+1},\ldots,x_{n+m})$, as is e.g. the case in classical exponential families, where we have

$$t_n(x_1,\ldots,x_n) = t(x_1) + \cdots + t(x_n)$$

for t being some function from E into k-dimensional Euclidean space.

We can now show the following result:

Proposition 3.1. *If for any* n, t_n *is totally sufficient and if* $t_1,t_2,\ldots$ *has* $\sum$*-structure, then* $t_1,t_2,\ldots$ *is sufficient and transitive*.

Proof. t_n is clearly sufficient for $\mathcal{P}^{(n)}$ for all n. We have to show that Y_{n+1} and $X_1,\ldots,X_n$ are conditionally independent given Y_n. We get

$$\begin{aligned} &P\{Y_{n+1} = y|X_1 = x_1,\ldots,X_n = x_n\} \\ (9) \qquad &= P\{t_{n+1}(x_1,\ldots,x_n,X_{n+1}) = y|X_1 = x_1,\ldots,X_n = x_n\} \\ &= P\{\psi_{n,n+1}(t_n(x_1,\ldots,x_n),X_{n+1}) = y|X_1 = x_1,\ldots,X_n = x_n\} ,\end{aligned}$$

where $\psi_{n,n+1}$ satisfies

$$(10) \qquad t_{n+1}(x_1,\ldots,x_{n+1}) = \psi_{n,n+1}(t_n(x_1,\ldots,x_n),x_{n+1}) .$$

As t_n is totally sufficient, X_{n+1} and $X_1,\dots,X_n$ are conditionally independent given Y_n, and we get from (9) that

$$
\begin{aligned}
&P\{Y_{n+1} = y | X_1 = x_1,\dots,X_n = x_n\} \\
(11) \qquad &= P\{\psi_{n+1}(t_n(x_1,\dots,x_n),X_{n+1}) = y | Y_n = t_n(x_1,\dots,x_n)\} \\
&= P\{Y_{n+1} = y | Y_n = t_n(x_1,\dots,x_n)\} ,
\end{aligned}
$$

which was to be proved.

Martin-Löf (1973) defined an algebraic consistency condition for a sequence of statistics which is slightly weaker than $\sum$-structure. We shall call it $\sum^*$-structure:

Definition 3.4. $t_1,t_2,\dots$ is said to have $\sum^*$-structure if for all n,m there is a function

$$N_{n,m}: F_n \times F_m \to \{0,1,\dots\}$$

so that for all $(x_1,\dots,x_n) \in E_1 \times\cdots\times E_n$

$$
\begin{aligned}
&\#\{(x_{n+1},\dots,x_{n+m}) \in E_{n+1} \times\cdots\times E_{n+m} : t_{n+m}(x_1,\dots,x_{n+m}) = y\} \\
&= N_{n,m}(t_n(x_1,\dots,x_n),y) .
\end{aligned}
$$

Remark: The discreteness of the sample space is also essential here, as the number of points corresponding to given values of the statistics occur in the definition in a fundamental manner.

It is immediate that we have

Proposition 3.2. If $t_1,t_2,\dots$ has $\sum$-structure, then it has $\sum^*$-structure.

Proof. Trivial.

$\sum^*$-structure becomes an important property when the conditional distributions given the statistics are determined by the numbers

(12) $$N_n(y) = \#\{(x_1,\ldots,x_n) \in E_1 \times\cdots\times E_n : t_n(x_1,\ldots,x_n) = y\}$$

and we have

Proposition 3.3. If for all n, $\mathcal{P}^{(n)}$ is summarized by t_n and if $t_1,t_2,\ldots$ has $\sum^*$-structure, then $t_1,t_2,\ldots$ is sufficient and transitive.

Proof. We already know that t_n is sufficient for $\mathcal{P}^{(n)}$ from proposition 2.1, As in proposition 3.1 it remains to be shown that Y_{n+1} and $X_1,\ldots,X_n$ are conditionally independent given Y_n. We have for all n

(13) $$\begin{aligned} &P\{X_1 = x_1,\ldots,X_n = x_n\} \\ &= P\{X_1 = x_1,\ldots,X_n = x_n | Y_n = t_n(x_1,\ldots,x_n)\} \cdot P\{Y_n = t_n(x_1,\ldots,x_n)\} \\ &= \frac{P\{Y_n = t_n(x_1,\ldots,x_n)\}}{N_n(t_n(x_1,\ldots,x_n))}, \end{aligned}$$

according to proposition 2.1.

Furthermore

(14) $$\begin{aligned} &P\{Y_{n+1} = y \wedge X_1 = x_1,\ldots,X_n = x_n\} \\ &= \sum_{x:(x_1,\ldots,x_n,x)\in t_{n+1}^{-1}(y)} \frac{P\{Y_{n+1} = y\}}{N_{n+1}(y)} \\ &= N_{n,n+1}(t_n(x_1,\ldots,x_n)) \cdot \frac{P\{Y_{n+1} = y\}}{N_{n+1}(y)}. \end{aligned}$$

Hence from (13) and (14) we get

$$(15)\qquad \begin{aligned} &P\{Y_{n+1} = y \mid X_1 = x_1,\dots,X_n = x_n\} = \\ &\frac{P\{Y_{n+1} = y\}N_n(t_n(x_1,\dots,x_n))}{N_{n+1}(y)P\{Y_n = t_n(x_1,\dots,x_n)\}} N_{n,n+1}(t_n(x_1,\dots,x_n)). \end{aligned}$$

Since this only depends on $x_1,\dots,x_n$ through $t_n(x_1,\dots,x_n)$, we must have

$$(16)\qquad \begin{aligned} &P\{Y_{n+1} = y \mid X_1 = x_1,\dots,X_n = x_n\} \\ &= P\{Y_{n+1} = y \mid Y_n = t_n(x_1,\dots,x_n)\} , \end{aligned}$$

which was to be proved.

If we assume $\sum$-structure instead of $\sum^*$-structure, we have the even stronger result:

Proposition 3.4. *If for all* n, $\mathcal{P}$ *is summarized by* t_n *and if* $t_1, t_2, \dots$ *has* $\sum$*-structure, then* t_n *is totally sufficient for all* n.

Proof. As in the previous proof, we only have to establish that $X_1,\dots,X_n$ and $X_{n+1},\dots,X_{n+k}$ are conditionally independent given Y_n for all n and k. We have

$$P\{X_1 = x_1, \ldots, X_{n+k} = x_{n+k} | Y_n = y\}$$

$$= \frac{\chi_{t_n^{-1}(y)}(x_1, \ldots, x_n) \cdot P\{Y_{n+k} = t_{n+k}(x_1, \ldots, x_{n+k})\}}{P\{Y_n = y\} \cdot N_{n+k}(t_{n+k}(x_1, \ldots, x_{n+k}))}$$

$$(17) \qquad = \frac{\chi_{t_n^{-1}(y)}(x_1, \ldots, x_n) \cdot P\{Y_{n+k} = \psi_{n,k}(y, x_{n+1}, \ldots, x_{n+k})\}}{P\{Y_n = y\} \quad N_{n+k}(\psi_{n,k}(y, x_{n+1}, \ldots, x_{n+k}))},$$

where $\psi_{n,k}$ is given by

$$(18) \qquad \psi_{n,k}(t_n(x_1, \ldots, x_n), x_{n+1}, \ldots, x_{n+k}) = t_{n+k}(x_1, \ldots, x_{n+k}) .$$

But for fixed y, the expression (17) is a <u>product</u> in $(x_1, \ldots, x_n)$ and $(x_{n+1}, \ldots, x_{n+k})$. Hence $X_1, \ldots, X_n$ and $X_{n+1}, \ldots, X_{n+k}$ are conditionally independent given $Y_n = y$, which was to be proved.

Apparently the property of being summarizing with $\sum$- structure is very strong. Another way of strengthening total sufficiency is to assume <u>minimality</u>. As in Lauritzen (1974) we say that t_n is <u>minimal totally sufficient</u> if it is a function of any other totally sufficient statistic. We then have

<u>Proposition 3.5</u>. <u>If for all</u> n, t_n <u>is minimal totally sufficient then</u> $t_1, t_2, \ldots$ <u>has</u> $\sum$- <u>structure</u>.

<u>Proof</u>. The result follows directly from corollary 1 of Lauritzen (1974).

If we include the notion of a minimal sufficient statistic in our considerations, we can "summarize" the results in the following diagram:

$$
\begin{tikzcd}
\text{minimal totally sufficient} \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \text{totally sufficient} \arrow[ddr, Rightarrow] & \text{minimal sufficient} \arrow[dd, Rightarrow] \\
\text{totally sufficient with } \Sigma\text{- structure} \arrow[dr, Rightarrow] & & \\
& \text{sufficient and transitive} \arrow[r, Rightarrow] & \text{sufficient} \\
\text{summarizing with } \Sigma\text{- structure} \arrow[uu, Rightarrow] \arrow[r, Rightarrow] & \text{summarizing with } \Sigma^{*}\text{-structure} \arrow[u, Rightarrow] \arrow[r, Rightarrow] & \text{summarizing} \arrow[u, Rightarrow]
\end{tikzcd}
$$

(The implications that are not proved in the previous are trivial).

At this point the author feels uncomfortable as a statistician. Is it really so that all these various notions are relevant? It is certainly true that in many examples at least some of the notions coincide. So far we have dealt with $\mathcal{P}$ being an arbitrary family of probability measures which in some sense is unreasonable from a statistical point of view. In the last section we shall impose regularity conditions on $\mathcal{P}$ and see how many of the implications in the diagram turn into equivalences.

4. Independence and universality.

Let us assume that for all $P \in \mathcal{P}$, $X_1, X_2, \ldots$ are <u>independent</u> random variables. It is then immediate that total sufficiency and sufficiency coincide and the same is of course true for minimal total sufficiency and minimal sufficiency. Hence it appears from the diagram that e.g. "<u>minimal sufficiency" implies everything but "summarizing</u>" and is thus a very strong property.

Barndorff-Nielsen (1973) discussed the notion of a <u>universal</u> family of probability measures in connection with the notion of M-ancillarity. Let X be a random variable on a discrete, at most denumerable set E and $\mathcal{P}$ a family of probability measures on E.

<u>Definition 4.1</u>. $\mathcal{P}$ is said to be <u>universal</u> if for all $x \in E$ there is a $P \in \mathcal{P}$ so that

$$P\{X = x\} \geqq P\{X = y\}$$

for all $y \in E$.

The following result, given in e.g. Barndorff-Nielsen (1973) shows a relation to the discussion in the preceding sections:

<u>Proposition 4.1</u>. <u>If</u> $\mathcal{P}$ <u>is universal and</u> t <u>is sufficient for</u> $\mathcal{P}$, <u>then</u> t <u>summarizes</u> $\mathcal{P}$.

<u>Proof</u>. The proof is exactly as in Barndorff-Nielsen (1973), theorem 2.1. Although E is assumed to be finite in that paper, this assumption is irrelevant for the validity of the proof.

Hence, if $\mathcal{P}^{(n)}$ in the previous section is assumed to be universal for all n, "sufficient" implies "summarizing" and "totally sufficient with Σ - structure" implies "summarizing with Σ - structure". Hence from the diagram it appears that "minimal totally sufficient" implies everything but minimal sufficient. Finally, if $X_1, X_2, \ldots$ are all assumed to be independent and at the same time $\mathcal{P}^{(n)}$ assumed to be universal for all n, "minimal sufficient" implies everything else.

References

Bahadur, R.R. (1954): Sufficiency and statistical decision functions. Ann. Math. Statist., 25, 423-462.

Barndorff-Nielsen, O. (1973): On M-ancillarity. Biometrika, 60, 447-455.

Barndorff-Nielsen, O. and M. Skibinsky (1963): Adequate subfields and almost sufficiency. Appl. Math. Publ. No. 329, Brookhaven national laboratory, 39 pp.

Fisher, R.A. (1920): A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Monthly Notices R. Astronomical Soc., 80, 758-770. Reprinted (1950) in Fisher's Contributions to Mathematical Statistics. Wiley, New York.

Freedman, D.A. (1962): Invariants under mixing which generalize de Finetti's theorem. Ann. Math. Statist., 33, 916-923.

Lauritzen, S.L. (1972): Sufficiency and Time Series Analysis. Preprint No. 11. Inst. of Math. Statistics, University of Copenhagen.

Lauritzen, S.L. (1974): Sufficiency, prediction and extreme models. Scand. J. Statist. 1. (to appear.)

Martin-Löf, P. (1973): Repetetive structures. Inst. of Math. Statist. Univ. of Stockholm.

Skibinsky, M. (1967): Adequate subfields and sufficiency. Ann. Math. Statist., 38, 155-161.