Sankhyā : The Indian Journal of Statistics
1983, Volume 45, Series A, Pt. 3, pp. 324–337.

CONDITIONAL INDEPENDENCE IN STATISTICS

By D. BASU[1]
The Florida State University
and
CARLOS A. B. PEREIRA[2]
Universidade de São Paulo

SUMMARY. The theory of conditional independence is explained and the relations between ancillarity, sufficiency, and statistical independence are discussed in depth. Some related concepts like specific sufficiency, bounded completeness, and splitting sets are also studied in some details by using the language of conditional independence.

1. Introduction

The notion of conditional independence is a central theme in statistics. In a series of articles Dawid (1979a, 1979b; 1980), Florens and Mouchart (1977), and Mouchart and Rolin (1978) have explained at length the grammar of Conditional Independence as a language of statistics. This article is a further elucidation on the subject and is in part of an expository nature.

The statistical perspective of this article is that of a Bayesian. A problem begins with a parameter (State of Nature) θ with its prior probability model $(\theta, \mathcal{B}, \xi)$ that exists only in the mind of the investigator. There is an observable X with an associated statistics model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$. Writing $\omega = (\theta, X)$, $(\Omega, \mathcal{F}) = (\Theta \times \mathcal{X}, \mathcal{B} \times \mathcal{A})$, and Π for the joint distribution of (θ, X), there then exists a subjective model $(\Omega, \mathcal{F}, \Pi)$ for ω. Hidden behind the wings of the Bayesian probability model $(\Omega, \mathcal{F}, \Pi)$ are the four models

(i) the prior model $(\Theta, \mathcal{B}, \xi)$,

(ii) the statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$,

(iii) the posterior model $(\Theta, \mathcal{B}, \{\xi_x : x \in \mathcal{X}\})$,

and (iv) the predictive model $(\mathcal{X}, \mathcal{A}, P)$ where P is the marginal or predictive distribution of X.

---

[1]Research partially supported by NSF Grant No. 79-04693.

[2]Research supported by CNPq, CAPES and USP—Brazil.

Key Words : Conditional independence, ancillarity, sufficiency, Markov property, (strong) identification, splitting sets, measurable separability, specific sufficiency, variation independence.

AMS classification : Primary 62B05, Secondary 62B20.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_32, © Springer Science+Business Media, LLC 2011

In statistics the phenomenon of conditional independence manifests itself in a natural fashion. The statistical model that is most commonly in use is that of a sequence $\boldsymbol{X} = (X_1, X_2, \ldots)$ of observables that are independently and identically distributed (i.i.d) for each given value of θ. It was DeFinnetti (1937) who emphasized that, in view of the fact that θ is not fully known, it is appropriate to regard the sequence of X_i's not as i.i.d random variables but as an exchangeable process. The fact that the X_i's are conditionally i.i.d implies that they are positively dependent in the sense that the covariance (when exists) between any pair is non-negative. More specifically, $\text{cov}(X_i, X_j) = \text{var}(E(X_1 | \theta))$.

Consider for example the particular case where $X_1, X_2, \ldots, X_n$ are i.i.d with common distribution $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ not fully known. In almost every text book of statistics it is proved that $\bar{X} = \frac{1}{n}\Sigma X_i$ is stochastically independent of $s^2 = \frac{1}{n}\Sigma(X_i - \bar{X})^2$. Does it mean that $\bar{X}$ when observed, carries no information about s^2 ? That the answer cannot be "yes" is easily seen as follows. Suppose that the sample size $n = 25$ and that our partial knowledge about $\theta = (\mu, \sigma^2)$ is as follows, $\mu = 0$ or 1 and $\sigma^2 = 1$ or 100, (that is $\Theta = \{(0, 1), (0, 100), (1, 1), (1, 100)\}$). Suppose now that $\bar{X}$ is observed and is equal to 2.1. This observation generates the four likelihoods $L(0, 1)$, $L(1, 1)$, $L(0, 100)$ and $L(1, 100)$ where $L(0, 1) = \frac{5}{\sqrt{2\pi}} \exp\left\{-\frac{25}{2}(2.1)^2\right\}$ and so on. The relative likelihoods work out roughly as 10^{-17}, 1, $2(10)^5$ and $3(10)^5$ respectively. Thus, it is intuitive that the observation $\bar{X} = 2.1$ almost categorically rules out the points $(0, 1)$ and $(1, 1)$. Then $\bar{X}$ and s^2, even though they are conditionally independent given θ, are in effect highly dependent.

The three entities $\theta = (\mu, \sigma^2)$, $T = (\bar{X}, s^2)$ and $\boldsymbol{X} = (X_1, \ldots, X_n)$ in this order, have the Markov property in the sense that, given θ and T, the conditional distribution of $\boldsymbol{X}$ depends on (θ, T) only through T. This is the sufficiency property of T as recognized by Fisher (1920, 1922). Kolomogrov (1942) gave a Bayesian characterization of the notion of sufficiency by noting that irrespective of the choice of the prior distribution ξ for the parameter θ, the posterior distribution $\xi_{\boldsymbol{X}}$ of θ depends on $\boldsymbol{X}$ only through T. Note that the Fisher characterization of sufficiency is made only in terms of the statistical model for $\boldsymbol{X}$ whereas the Kolmogorov characterization is made in terms of a large family of Bayesian models $(\Omega, \mathcal{F}, \Pi)$ for $\omega = (\theta, X)$. (See Basu, 1977 and Cheng, 1978 for further details on these characterizations).

Fisher regarded a sufficient statistic T as one that summarizes in itself all the available relevant information in the sample X about the parameter θ. He called a statistic $Y = Y(X)$ ancillary if the conditional distribution of Y given θ does not involve θ (is the same for all values of θ). For example, the statistic $\Sigma \frac{(X_i - \bar{X})^4}{s^4}$ is ancillary. In a series of articles Basu (1955, 1958, 1959, 1964, 1967) studied the phenomena of sufficiency, ancillarity and conditional independence from various angles. In these articles, Basu's viewpoint was non-Bayesian in the sense that he did not introduce a prior distribution ξ for θ. Mouchart and Rolin (1978) studied in depth the familiar Basu theorems on sufficiency, ancillarity and conditional independence from the view point of the Bayesian model.

In this paper we too review Basu's results and also the two-parameter problem from the Bayesian perspective. Many results are stated without proof since the proofs involve standard measure theoretic arguments and can be found for instance in Loeve (1977).

2. Notation and preliminaries

Let $(\Omega, \mathcal{F}, \pi)$ be the basic probability space. By a "random object" X we mean a measurable map $\omega \to X(\omega)$ of $(\Omega, \mathcal{F})$ into another measurable space $(\mathcal{X}, \mathcal{A})$. The sub σ-algebra (to be called subfield) of X events $\{X^{-1}(A); A \in \mathcal{A}\}$ will be denoted by $\mathcal{F}_X$. The two probability spaces $(\Omega, \mathcal{F}_X, \Pi)$ and $(\mathcal{X}, \mathcal{A}, \Pi^{-1})$ are indistinguishable in a certain sense, and so we shall, as a rule, identify a random object X with the induced subfield $\mathcal{F}_X$ of $\mathcal{F}$. In that way, one could say that random objects are generators of subfields. Examples of random objects include random variables, random vectors etc.

For any two subfields $\mathcal{F}'$ and $\mathcal{F}''$ of $\mathcal{F}$, $\mathcal{F}' \vee \mathcal{F}''$ denotes the smallest subfield of $\mathcal{F}$ that contains both $\mathcal{F}'$ and $\mathcal{F}''$. The smallest subfield of $\mathcal{F}$ that contains all null sets of $\mathcal{F}$ (a set N is null if $\pi(N) = 0$) is denoted by $\bar{\mathcal{F}}_0$, and write $\mathcal{F}_0 = \{\varphi, \Omega\}$, the trivial subfield.

A subfield of $\mathcal{F}$ is said to be completed if it contains $\bar{\mathcal{F}}_0$. For any subfield $\mathcal{F}'$ of $\mathcal{F}$ its completion $\bar{\mathcal{F}}'$ is defined by

$$\bar{\mathcal{F}}' = \mathcal{F}' \vee \bar{\mathcal{F}}_0.$$

For a random object X, the notation $X \in \mathcal{F}'$ indicates that $\mathcal{F}_X \subseteq \mathcal{F}'$ and X is said to be essentially $\mathcal{F}'$ measurable. A random variable is a random object with range $(R_1, \mathcal{B}_1)$ where R_1 is the real line and $\mathcal{B}_1$ is the Borel

σ-algebra. A random variable f is said to be bounded if there exists $a \in R_1$ such that $\pi\{\omega : |f(\omega)| \leqslant a\} = 1$. In the sequel, all random variables will be regarded as bounded unless stated otherwise and the use of small letters shall be restricted to their representation. The notation $f \subseteq X$ indicates that the random variable f is ess—$\mathscr{F}_X$ measurable. In the same spirit, for two random objects X and Y, we write $X \subseteq Y$ to indicate that $\bar{\mathscr{F}}_X \subseteq \bar{\mathscr{F}}_Y$. The class of all bounded random variables on $(\Omega, \mathscr{F}, \Pi)$ is denoted by L_∞ and $L_\infty(X)$ denotes the class of all ess—$\mathscr{F}_X$ measurable random variables. Here and for the rest of this article, equality of two random variable means essential equality; that is $f = g$ means $\{\omega : f(\omega) \neq g(\omega)\}$ is a null set.

The conditional expectation of f, given a random object X, is a random variable $f^{*X} \in L_\infty(X)$ such that

$$\int f g d\pi = \int f^{*X} g d\pi \quad \forall g \in L_\infty(X)$$

Another notation for f^{*X} is $E(f \mid X)$. When the conditioning random object X is implicit in the context, f^* is substituted for f^{*X}. The map $f \to f^*$ of L_∞ to $L_\infty(X)$ is linear, constant preserving, idempotent and is a contraction in the L_p norm if $p \geqslant 1$.

3. Conditional independence : definition, properties and the drop/add principles

In this section the definition and properties of conditional independence are briefly discussed.

Three random objects X, Y and Z are being considered and in this section * stands for $^*Z$ operator.

Definition 1. (Intuition) : The random objects X and Y are conditionally independent given Z (in symbols $X \perp\!\!\!\perp Y \mid Z$) if for any $f \in L_\infty(X)$

$$E(f \mid Y, Z) = f^{*YZ} = f^*$$

Note that to say $X \perp\!\!\!\perp Y \mid Z$ is equivalent to say that $X \mid (Y, Z)$ has the same conditional disribution as $X \mid Z$. This is the intuition behind the definition. In the case where Z is essentially a generator of $\mathscr{F}_0$, we obtain the independence of X and Y in the usual sense. In this case the notation is $X \perp\!\!\!\perp Y$.

Definition 1*a*. (Symmetric) : The random objects X and Y are conditionally independent given Z if for any $f \in L_\infty(X)$ and $g \in L_\infty(Y)$

$$(fg)^* = f^* g^*.$$

The following well known theorem gives the equivalence of the two definitions showing that $X \amalg Y \mid Z$ implies $Y \amalg X \mid Z$ which is not clear by looking at definition 1.

Theorem 1 : *Definitions* 1 *and* 1*a are equivalent.*

The concept of conditional independence (c.i.) gives rise to many questions. Among them are questions involving the drop and add (Drop/Add) principles. Suppose that X, Y, Z, W, X_1, Z_1, are random objects such that $X \amalg Y \mid Z$, $X_1 \subseteq X$, $Z_1 \subseteq Z$. What can be said about the relationship $\amalg$ if X_1 is substituted for X, Z_1 for Z, (Y, W) for Y or (Z, W) for Z ? In other words, can $\mathscr{F}_X$, $\mathscr{F}_Y$ or $\mathscr{F}_Z$ be essentially reduced or enlarged without destroying the c.i. relation ? In general the answer is no. However for certain kinds of reductions and enlargements, the relationship will be preserved. To indicate that the relationship $\amalg$ does not hold we write not $\amalg$.

It is not difficult to find examples showing that arbitrary reductions or enlargements of $\mathscr{F}_Z$ = the conditioning subfield, may destroy the c.i. property. With the example of the normal distribution presented in the introduction, we have $\bar{X} \amalg s^2 \mid \theta$ but not $\bar{X} \amalg s^2$. In yet another statistical context, suppose that θ is the unknown (real) parameter of interest and let X and Y be two i.i.d. random variables with common uniform distribution on the interval $(\theta-\frac{1}{2}, \theta+\frac{1}{2})$. Since θ is unknown, we can only say that $X \in \mathscr{R}$, $Y \in \mathscr{R}$. However, after X has been observed equal to x, we would for sure know that $x-1 \leqslant Y \leqslant x+1$. This shows clearly that $X \amalg Y \mid \theta$ but not $X \amalg Y$. To show that we can have $X \amalg Y$ and X not $\amalg Y \mid Z$, let W and Z be two i.i.d. $N(0, 1)$ variables and take $X = Z-W$ and $Y = Z+W$.

The above examples may be viewed as cases of Simpson's paradox (see Dawid, 1979a). The paradox, however, is much stronger. For instance, let W and Z be two independent normal variables with zero means. As before define $X = Z-W$ and $Y = Z+W$. The correlation between X and Y is given by $\rho(X, Y) = \dfrac{1-\delta}{1+\delta}$ where $\delta = \dfrac{\text{var}(W)}{\text{var}(Z)}$. Given Z, the conditional correlation is clearly equal to -1. On the other hand, δ may be taken very small to make $\rho(X, Y)$ close to 1. This shows that we can have cases where X and Y are strongly positive (negative) dependent but, when Z is given, X and Y turn to be strongly negative (positive) dependent. The problem of dependence inversion is discussed in depth in Lindley and Novick (1981). The following example may be of relevance for applied statisticians.

Example : Suppose that an urn contains θ (unknown) white balls in a total of N (known) balls. A sample, without replacement, of n balls is selected

from this urn. Let X be the number of white balls in the sample which implies that $Y = \theta - X$ is the number of white balls that remain in the urn. Since $\rho(X, Y \mid \theta) = -1$, we have X not $\amalg$ $Y \mid \theta$. It can be proved (see Whitt, 1979) that $X \amalg Y$ iff θ has binomal (prior) distributions with fixed parameters N and $p \in (0, 1)$. On the other hand, if a priori, $\Pr\{\theta = 0\} = \Pr\{\theta = N\} = \frac{1}{2}$, then $\rho(X, Y) = 1$ showing an extreme inverted dependence.

The essence of Drop/Add principles for c.i. is contained in the following propositions.

Proposition 1: If $X \amalg Y \mid Z$, then for every $X' \subset X$, we have:

(i) $X' \amalg Y \mid Z$

(ii) $X \amalg Y \mid (Z, X')$

(iii) $(X, Z) \amalg (Y, Z) \mid Z$.

By way of explanation, if $X \amalg Y \mid Z$, then the relation $\amalg$ is preserved when (i) X and Y are increased (Add) by any essential part of Z, (ii) Z is increased (Add) by any essential part of X or Y, and (iii) X and Y are arbitratily reduced (Drop).

To end this section we present an extreme case of Drop/Add principles for the conditioning random object. It appears in Dawid (1980) and it was originally introduced by G. Udny Yule in terms of collapsibility of contingency tables. It must clarify the problems with Simpson's paradox for contingency tables.

Proposition 2: Let X, Y and Z be three random objects such that $\mathcal{F}_Z = \{\phi, \Omega, A, A^C\}$ with $0 < \Pi(A) < 1$. If $X \amalg Y$ and $X \amalg Y \mid Z$, then either $X \amalg Z$ or $Y \amalg Z$.

The proof becomes simple when we recognize the following result.

Lemma: *If* $X \amalg Y$ *and* $X \amalg Y \mid Z$, *then for every atom* A *of* Z *with* $\Pi(A) > 0$, *we have*

$$E(I_A \mid (X, Y)) = \frac{1}{\Pi(A)} E(I_A \mid X) E(I_A \mid Y).$$

Proof of lemma: Let B, C, be two sets such that $I_B \subset X$, $I_C \subset Y$

$$\begin{aligned}
\int_{BC} E(I_A \mid X) E(I_A \mid Y) d\Pi &= \int E(I_{AB} \mid X) E(I_{AC} \mid Y) d\Pi \\
&= \int E(I_{AB} \mid X) d\Pi \int E(I_{AC} \mid Y) d\Pi = \Pi(AB)\Pi(AC) \\
&= [\Pi(A)]^2 \Pi(B \mid A)\Pi(C \mid A) = [\Pi(A)]^2 \Pi(BC \mid A) \\
&= \Pi(A)\Pi(ABC) = \Pi(A) \int_{BC} E(I_A \mid (X, Y)) d\Pi.
\end{aligned}$$

Since sets of the form BC generate $\mathcal{F}_{XYT}$ a standard argument completes the proof.

Proof of proposition : Let $p = \Pi(A)$. From lemma we have

$$E(I_A|(X, Y)) = \frac{E(I_A|X)E(I_A|Y)}{p}$$

and

$$E(I_{A^C}|(X, Y)) = \frac{[1-E(I_A|X)][1-E(I_A|Y)]}{1-p}$$

and consequently

$$\left(1-\frac{E(I_A|X)}{p}\right)\left(1-\frac{E(I_A|Y)}{p}\right) = 0$$

Since $X \amalg Y$, this equation holds only if $\frac{E(I_A|X)}{p} = 1$ or $\frac{E(I_A|Y)}{p} = 1$.

4. Bayesian inference : sufficiency, ancillarity and independence

As discussed in Dawid (1979a, 1980) many of the important Statistical Concepts are simply manifestations of the concept of conditional independence. In this section we use the framework of conditional independence to describe the Bayesian version of those statistical concepts and their properties. First we review some of the structures involved.

Let $(\mathcal{X}, \mathcal{A})$ be the usual Sample Space and $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on $(\mathcal{X}, \mathcal{A})$ where Θ is the usual parameter "Space". In addition the Bayesian considers a (prior) probability space $(\Theta, \mathcal{B}, \xi)$ where $\mathcal{B}$ is a σ-algebra of subsets of Θ such that $P_\theta(A)$ is a $\mathcal{B}$-measurable function for every fixed $A \in \mathcal{A}$. Clearly, the choice of the prior model is not completely arbitrary, since it has to match the statistical structure on the $\mathcal{B}$-measurability of $P_\theta(A)$.

We then consider the probability space $(\Omega, \mathcal{F}, \Pi)$, where now $\Omega = \Theta \times \mathcal{X}$, $\mathcal{F} = \mathcal{B} \times \mathcal{A}$ and Π is defined by

$$\Pi(F) = \int_\Theta P_\theta(F^\theta) d\xi(d\theta)$$

where $F^\theta = \{x : (\theta, x) \in F\}$. The marginal on $\mathcal{X}$ is defined by

$$P(A) = \Pi(\Theta \times A) \text{ for every } A \in \mathcal{A}.$$

Let X and Y be two random objects on $(\Omega, \mathcal{F})$. We say that X represents the sample and Y represents the parameter if

$$\mathcal{F}_X = \{\Theta \times A,\ A \in \mathcal{A}\}$$

and

$$\mathcal{F}_Y = \{B \times \mathcal{X},\ B \in \mathcal{B}\}.$$

In addition to X and Y defined above, consider two random objects X_1, and X_2 such that $(X_1, X_2) \subseteq X$. The Bayesian version of the concepts of sufficiency and ancillarity is contained in the following.

Definition 2 :

(a) If $X \amalg Y \mid X_1$ we say X_1 is sufficient for X with respect to Y.

(b) If $X_2 \amalg Y$ we sav that X_2 is ancillary with respect to Y.

The classical concept of statistical independence between X_1 and X_2 nas its Bayesian version as

(c) $X_1 \amalg X_2 \mid Y$.

Basu (1955, 1958) speculates under what conditions two of the three relation (a), (b) and (c) imply the third. In this section we study Basu's theorems under the Bayesian framework. The next result which is Basu's first conjecture presents conditions to have (b) and (c) implying (a).

Proposition 3 : If in addition to $X_2 \amalg Y$ and $X_1 \amalg X_2 \mid Y$ we have $X \amalg Y \mid (X_1, X_2)$ then $X \amalg Y \mid X_1$.

Proof: Note that $X_2 \amalg Y$ and $X_1 \amalg X_2 \mid Y$ implies $X_2 \amalg Y \mid X_1$, also $X_2 \amalg Y \mid X_1$ and $X \amalg Y \mid (X_1, X_2)$ implies $X \amalg Y \mid X_1$.

Arguing similarly it is easy to see that if $X_1 \amalg X_2$, then (a) implies (b) and (c). The meaning of $X_1 \amalg X_2$ in classical statistics however is void.

Note that Proposition 3 gives conditions for reducing (Drop) the conditioning random object. Actually all of Basu's theorems are cases of Drop/Add principles.

Basu (1955) stated that any statistic independent of a sufficient statistic is ancillary. Later on Basu (1958) presented a counter example and recognized the necessity of an additional condition (connectedness) on the family $\{P_\theta : \theta \in \Theta\}$ of probability measures. Koehn and Thomas (1975) strengthened this result by introducing a necessary and sufficient condition on the family. More recently Basu and Cheng (1979), generalizing results of Pathak (1975) showed the equivalence between these two conditions in coherent models.

The following theorem is a Bayesian version of the result of Koehn and Thomas (1975).

Theorem 2 : *Let* $X_1 \subset X$ *be a sufficient random object* (*i.e.* $X \amalg Y \mid X_1$). *The random object* $Y \wedge X_1$ *is essentially a constant* (*i.e.* $F_{Y \wedge X_1} = F_0$) *iff* $X_2 \amalg Y$ *whenever* $X_2 \subset X$ *and* $X_1 \amalg X_2 \mid Y$ (*i.e.* X_2 *is ancillary if* X_1 *and* X_2 *are statistically independent*).

Proof : $E(I_A \mid Y) = E(I_A \mid X_1, Y) = E(I_A \mid X_1)$ by $X \amalg Y \mid X_1$. Now since, $X_1 \wedge Y = F_0$ it follows that $E(I_A \mid Y)$ is a constant. Take X_2 such that $X_2 \equiv Y \wedge X_1$. Since $X_2 \subset Y$, $X_1 \amalg X_2 \mid Y$. Then by hypotheses $X_2 \amalg Y$, which implies that $X_2 \amalg X_2$ since $X_2 \subset Y$; that is $X_2 \equiv Y \wedge X_1$ is essentially a constant.

Remarks : The condition introduced by Koehn and Thomas (1975) is the non existence of a splitting set. A set A in the sample space is a 'splitting' set if $P_\theta(A) = 0$ or 1 for all θ, and at least for a pair $\{\theta_1, \theta_2\} \epsilon \Theta$, $P_{\theta_1}(A) = P_{\theta_2}(A^c) = 1$. In the Bayesian framework, an analogous definition is as follows: A set A such that $I_A \subset X$ is a splitting set if $0 < \Pi(A) < 1$ and $E(I_A \mid Y) = E^2(I_A \mid Y)$. It is easy to see that if A is a splitting set then $I_A \subset Y \wedge X$. We conclude that the non existence of a splitting set is equivalent to $Y \wedge X$ being essentially a constant.

Basu (1955) proved that any ancillary statistic is statistically independent of any bounded complete sufficient statistic. The Bayesian analogue of the concept of boundedly completeness is the concept of strong identifiability (Dawid, 1980 and Mouchart and Rolin, 1978). The main objective of this section is to study this concept and present Basu's result under the Bayesian framework.

Definition 3 : The random objects X and Y are said to be measurably separated conditionally on Z if $(X, Z) \wedge (Y, Z) \equiv Z$. When Z is essentially a constant we simply say that X and Y are measurably separated.

A large list of results related with this concept appears in Mouchart and Rolin (1978).

Let X and Y be two random objects. We shall study some aspects of the linear maps $L_x(Y) \xrightarrow{*} L_\infty(X)$ and $L_\infty(X) \xrightarrow{+} L_\infty(Y)$ where $*$ is for $E(\cdot \mid X)$ and $+$ for $E(\cdot \mid Y)$.

Definition 4 : We say that X is strongly identified by Y and write $X \lll Y$ if the map $L_\infty(X) \xrightarrow{+} L_\infty(Y)$ is essentially one-one.

Proposition 4 : If the map $L_\infty(Y) \xrightarrow{*} L(X)$ is essentially onto then $X \ll Y$.

Proof : Let $(f, h) \subset X$ and $f^+ = 0$. Since $*$ is essentially onto $\exists\, g \subset Y$ such that $g^* = h$. Then

$$E(fh) = E(fg^*) = E(f^+g) = 0$$

since h is arbitrary $f = 0$.

Let $X_{[Y]}$ be the random object that generates the smallest subfield that contains all functions g^* where $g \subset Y$. The following result shows that $X_{[Y]}$ may be viewed as a Bayesian minimal sufficient statistic.

Proposition 5 : (i) $X \amalg Y \mid X_{[Y]}$

(ii) If $X_1 \subset X$ is such that $X \amalg Y \mid X_1$ then $X_{[Y]} \subset X_1$.

Proof : The proof is easy and hence omitted.

Remark : From Proposition 5 it is easy to see that a Burkholder type theorem on intersection of sufficient subfields is true in the Bayesian framework. Precisely, if $X \amalg Y \mid X_1$ and $X \amalg Y \mid X_2$ then $X \amalg Y \mid X_1 \wedge X_2$.

When $X_{[Y]} \equiv X$, X is said to be identified by Y (Dawid 1980, and Mouchart and Rolin 1978). The name strong identification was motivated by the following result.

Proposition 6 : If $X \ll Y$ then $X_{[Y]} \equiv X$.

Proof : Note that $X \amalg Y \mid X_{[Y]}$. Thus $\forall f \subset X$

$$E\{E(f \mid Y, X_{[Y]}) \mid Y\} = E\{E(f \mid X_{[Y]}) \mid Y\}.$$

For $f^+ = E(f \mid X_{[Y]})$. Since $X \ll Y$ we have that

$$E\{(f - f^+) \mid Y\} = 0 \rightarrow f = f^+.$$

Then $\forall f \subset X$, $f \subset X_{[Y]}$ and $X = X_{[Y]}$.

The Bayesian version of Basu's theorem is contained in the result below.

Theorem 3 : *Let X, Y and Z be three random objects. If $X \amalg Y$, $X \amalg Y \mid Z$ and $Z \ll Y$ then $X \amalg Z \mid Y$.*

Proof : Since $X \amalg Y \mid Z$ we have, for any $f \subset X$

$$E(f \mid Y, Z) = E(f \mid Z)$$

and since

$$X \amalg Y,\ E(f \mid Y) = E(f)$$

Therefore

$$E[\{E(f \mid Z) - E(f)\} \mid Y] = 0.$$

Now since $Z \ll Y$ $E(f \mid Z) = E(f)$ we thus have $E(f \mid Y, Z) = E(f)$.

Note that to obtain Basu's theorem, we consider X as the sample, Y as the parameter, and X_0 and X_1 two random objects such that $(X_0, X_1) \subset X$, $X_0 \amalg Y$, $X \amalg Y \mid X_1$ and $X_1 \ll Y$. Clearly $X_0 \amalg Y \mid X_1$ and the result $X_0 \amalg X_1 \mid Y$ follows.

Lehman and Scheffe (1950) proved that if a sufficient statistic is boundedly complete, then it is a minimal sufficient statistic. The proposition below is a Bayesian version of this result.

Proposition 7 : Suppose $X_1 \subset X$ and $X \amalg Y \mid X_1$. If $X_1 \ll Y$ then $X_1 = X_{[Y]}$.

Proof : From Proposition 6 $X_{[Y]} \subset X_1$ and $X_1 \amalg Y \mid X_{[Y]}$.

Let $f \subset X_1$ then $E(f \mid Y) = E[E(f \mid X_{[Y]}) \mid Y]$ or

$$E(f \mid Y) = E[E(f \mid X_{[Y]}, Y) \mid Y]$$
$$= E(E(f \mid X_{[Y]}) \mid Y)$$

since $X_1 \ll Y$ we conclude that $f = E(f \mid X_{[Y]}) \subset X_{[Y]}$.

Remark : The concept of strong identifiability may be generalized as follows. X is strongly identified by Y conditionally on Z ($X \ll Y \mid Z$) if for every $f \subset (X, Z)$, $E(f \mid Y, Z) = 0$ implies $f = 0$. Analogously, X is identified by Y conditionally on Z if

$$(X, Z)_{[Y, Z]} = (X, Z).$$

All the results of this section may be easily generalized by introducing a conditioning random object Z to each relation stated. For our future work we intend to relate these general results with the work of Dawid (1979c), Ferreira (1980) and Godambe (1980).

5. The two parameter problem

We now briefly discuss sufficiency in the presence of a nuisance parameter.

Suppose that the parameter Y is such that $Y \equiv (Y_1, Y_2)$. Let X represent the sample, $X_1 \subset X$ be specific sufficient with respect to Y_2, and $X_2 \subset X$ be specific sufficient with respect to Y_1. That is, $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$. (See Basu, 1978 for details on the notion of specific sufficiency). The question here is under what conditions does the specific sufficiency of (X_1, X_2) imply the sufficiency of (X_1, X_2) ?

Proposition 8 : If $(X_1, Y_1) \wedge (X_2, Y_2) \subset (X_1, X_2)$ then $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$ imply $X \amalg Y \mid (X_1, X_2)$.

Proof: We have

$$X \amalg Y \mid (X_1, Y_1)$$

and

$$X \amalg Y \mid (X_2, Y_2)$$

and a simple argument yields

$$X \amalg Y \mid (X_1, Y_1) \wedge (X_2, Y_2).$$

The following related result may also be of interest.

Proposition 9: If $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$ then

$$X \amalg Y \mid (X_1, X_2) \text{ if and only if } X \amalg Y_1 \mid (X_1, X_2).$$

Note the condition $X \amalg Y_1 \mid (X_1, X_2)$ does not have an interpretation in classical statistics since distributions depend on both parameters Y_1 and Y_2.

The following example is again relevant. Note that the parameter space θ is variation independent (if the parameter space is the cartesian product of the domain Y_1 by the domain of Y_2; see Basu, 1977 and Barndorff-Nielsen, 1978).

Take

$$\Theta = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

Then $\Theta = \Theta_1 \times \Theta_2$ where $\Theta_1 = \Theta_2 = \{0, 1\}$

$$X = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

$$P_\theta = \delta_\theta \text{ the point mass at } \theta.$$

Then $T = I_{\{(0,1),(1,0)\}}(x)$ is specific sufficient for θ_1 and for θ_2 but is not sufficient. This example Shows that variation independence on (θ_1, θ_2) and specific sufficiency of X_1, and X_2 does not imply that (X_1, X_2) being sufficient.

Acknowledgement. Thanks are due to the referee for helping in improving presentation of the paper.

References

Ash, R. B. (1972): *Real Analysis and Probability*, Academic Press, New York.

Bahadur, R. R. (1955): Measurable subspaces and subalgebras. *Proc. Amer. Math. Soc.*, 6, 565–70.

Barndorff-Nielsen, O. (1978): *Information and Exponential Families in Statistical Theory*, John Wiley, New York.

Basu, D. (1955): On statistics independent of a complete sufficient statistics. *Sankhyā*, A, 15, 377–80

——— (1958): On statistics independent of a sufficient statistics. *Sankhyā*, A, 20, 223–26.

——— (1959): The family of ancillary statistics. *Sankhyā*, A, 21, 247–56

——— (1964): Recovery of ancillary information. *Sankhyā*, A, 26, 3–16.

——— (1967): Problems relating to the existence of maximal and minimal elements in some families of statistics (Subfields). *Proc. Fifth Berkeley Sym. Math. Statist. Prob.*, **1**, 41–50.

——— (1977): On the elimination of nuisance parameters. *Jour. Amer. Statist. Assoc.*, **72**, 355–66.

——— (1978): On partial sufficiency: A review. *J. Statist. Plan. Inf.*, **2**, 1–13.

BASU, D. and CHENG, S. C. (1979): A note on sufficiency in coherent models. *Int. J. Math. Math. Sci.* To appear.

BURKHOLDER, D. L. (1961): Sufficiency in the undominated case. *Ann. Math. Statist.*, **32**, 1191–200.

CHENG, S. C. (1978): A mathematical study of sufficiency and adequacy in statistical theory. Ph.D. Dissertation, FSU, Florida.

DAWID, A. P. (1979a): Conditional independence in statistical theory. *J. Roy. Statist. Soc.*, B, **41**, 1–31.

——— (1979b): Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc.*, B, **41**, 249–52.

——— (1979c): A Bayesian look at nuisance parameters. *Trabajos de Estadistica*, To appear.

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

DE FINETTI, B. (1937): Foresight: Its logical laws, its subjective sources. Translated edition 1964 in *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, editors). John Wiely, New York.

——— (1970): *Theory of Probability*, Vols. 1 and 2. Translated edition, 1974. John Wiley, London.

DOOB, J. L. (1953): *Stochastic Processes*, John Wiley, New York.

FERREIRA, P. E. (1980): Comments on Berkson's Paper "In dispraise of ....". Unpublished report.

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. *Mon. Not. Roy. Ast. Soc.*, **80**, 758–70.

——— (1922): On the mathematical foundations of theoretical statistics. *Phil. Trans.*, A, **222**, 309–68.

FLORENS, J. P. and MOUCHART, M. (1977): Reduction of Bayesian experiments. *CORE*, Discussion Paper 7737.

GODAMBE, V. P. (1979): On sufficiency and ancillarity in presence of nuisance parameter. Unpublished report.

HALL, W. J., WIJSMAN, R. A. and GHOSH, J. K. (1965): The relationship between sufficiency and invariance. *Ann. Math. Statist.*, **36**, 375–614.

KOEHN, U. and THOMAS, D. L. (1975): On statistics independent of a sufficient statistics: Basu's lemma. *Amer. Statist.*, **29**, 40–2.

KOLMOGOROV, A. N. (1942): Determination of the center of dispersion and degree of accuracy for a limited number of observations (in Russian). *Izvestija Akademii Nauk*, Ser. Mat., **6**, 3–32.

LEHMAN, E. L. and ACHEFFE, H. (1950) : Completeness, similar regions and unbiased estimation, Part I. *Sankhyā*, A, **10**, 305–40.

LINDLEY, D. V. and NOVICK, M. R. (1981) : The role of exchangeability in inference. *Ann. Statist.*, **9**, 45–58.

LOEVE, M. (1977) : *Probability Theory*, 4th ed. Springer-Verlag, NY.

MOUCHART, M. and ROLIN, J. M. (1978) : A note on conditional independence. Unpublished report.

MOY, S. T. C. (1954) : Characterization of conditional expectation as a transformation on function spaces. *Pacific J. Math.*, **4**, 47–63.

PATHAK, P. K. (1975) : Note on Basu's lemma. Unpublished report.

PICCI, G. (1977) : Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM J. Appl. Math.*, **33**, 383–98.

WHITT, W. (1979) A note on the influence of the sample on the posterior distribution. *Jour. Amer. Statist. Assoc.* **74**, 424–426.

Paper received : *March*, 1981.
Revised : *January*, 1983.

Sankhyā : The Indian Journal of Statistics
1983, Volume **45**, Series A, Pt. 1, pp. 99–104.

A NOTE ON BLACKWELL SUFFICIENCY AND A SKIBINSKY CHARACTERIZATION OF DISTRIBUTIONS

By D. BASU* and CARLOS A. B. PEREIRA**

The Florida State University, Tallahassee

SUMMARY. A Skibinsky (1970) characterization of the family of hypergeometric distributions is re-examined from the point of view of sufficient experiments and a number of other distributions similarly characterized.

1. Introduction

Consider an urn containing N balls x of which are white. If a simple random sample of n ($n \leqslant N$) balls is drawn from the urn, then the number of white balls in the sample has the hypergeometric distribution with parameters N, n, and x [denoted by $h(N, n, x)$]. Skibinsky (1970) introduced the following characterization of $h(N, n, x)$:

"A family of $N+1$ probability distributions (indexed say by $x = 0, 1, \ldots, N$), each supported on a subset of $\{0, 1, \ldots, n\}$ is the hypergeometric family having population and sample size parameters N and n respectively (the remaining parameter of the x-th member being x), if and only if for each number θ, $0 < \theta < 1$, the mixture of the family with binomial (N, θ) mixing distribution is the binomial (n, θ) distribution."

Writing $b(N, \theta)$ for the binomial distribution over $\{0, 1, \ldots, N\}$ and the symbol $\sim$ for "distributed as", we may restate Skibinsky's characterization as follows :

Let $X \sim b(N, \theta)$, $0 < \theta < 1$, and let $\{\tau_x : x = 0, 1, \ldots, N\}$ be a family of probability distributions on $\{0, 1, \ldots, n\}$, where $n \leqslant N$. Consider the random variable Y such that the conditional probability distribution of Y given $\{X = x\}$ is τ_x for all x (i.e., $Y \mid X = x \sim \tau_x$). Then $Y \sim b(n, \theta)$ for all θ in $(0, 1)$ if and only if τ_x is $h(N, n, x)$ for all x.

Skibinsky (1970) proved the above result in several interesting ways, but somehow the perspective of Blackwell sufficiency eluded him. Written for its pedagogical interest, this note is an elucidation on the notion of Blackwell sufficiency and an unification of a number of results analogous to Skibinsky's characterization of the Hypergeometric distribution.

*Research partially supported by NSF Grant No. 79-04693.

**Research supported by CAPES and USP—Brazil.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_33, © Springer Science+Business Media, LLC 2011

2. Blackwell sufficiency

A statistical experiment related to a parameter $\theta \in \Theta$ is idealized as an observable random variable (or vector), X, associated with a sample space $\mathcal{X}$ and a family $\{p_\theta : \theta \in \Theta\}$ of probability functions (distributions) on $\mathcal{X}$ indexed by θ. We avoid all measurability difficulties by restricting ourselves only to discrete sample spaces. Given two spaces $\mathcal{X}$ and $\mathcal{Y}$, a transition function τ, from $\mathcal{X}$ to $\mathcal{Y}$, is a family

$$\tau = \{\tau_x : x \in \mathcal{X}\}$$

of probability functions, τ_x, on indexed by $x \in \mathcal{X}$. Thus, the family of Hypergeometric probability functions $\{h(N, n, x) : x = 0, 1, \ldots, N\}$ is a transition function from $\{0, 1, \ldots, N\}$ to $\{0, 1, \ldots, n\}$.

Let X and Y be two experiments with models $(\mathcal{X}, \{p_\theta : \theta \in \Theta\})$ and $(\mathcal{Y}, \{q_\theta : \theta \in \Theta\})$ respectively.

Definition (Blackwell) : The experiment X is *sufficient for* (at least as informative as) the experiment Y and write $X > Y$ if there exists a transition function $\tau = \{\tau_x : x \in \mathcal{X}\}$ from $\mathcal{X}$ to $\mathcal{Y}$ such that

$$q_\theta(y) = \sum_x \tau_x(y) p_\theta(y) \qquad \ldots \quad (2.1)$$

for all $y \in \mathcal{Y}$ and $\theta \in \Theta$.

A transition function τ satisfying (2.1) is called here a Blackwell transition function. It is easy to check that the relation $>$ defines a partial order on the family of experiments related to θ.

If $T = T(X)$ is a sufficient statistic in the classical sense of Fisher (i.e., the conditional distribution of X given $\{T = t\}$ does not involve θ), then it follows at once that T is sufficient for X in the sense of Blackwell $(T > X)$. Of course, X is sufficient for T in either sense.

The intuitive content of the relation $X > Y$ is as follows :

If we perform the experiment X, note its outcome x, and finally carry out a postrandomization exercise that chooses a point $y \in \mathcal{Y}$ in accordance with the probability function τ_x, then the experiment Y^* of such a choice of y is in a sense indistinguishable from the experiment Y in that both are endowed with the same model $(\mathcal{Y}, \{q_\theta : \theta \in \Theta\})$. Any decision rule related to θ that is based on the experiment Y can therefore be perfectly matched (in terms of their average performance characteristics) by a randomized rule based on X.

For two fixed integers N and n ($n \leqslant N$), consider now the simple case where $X \sim b(N, \theta)$ and $Y \sim n(n, \theta)$, $0 < \theta < 1$. To prove that $X > Y$ we consider an experiment $W = (W_1, \ldots, W_N)$ where its components W_i, $i = 1, \ldots, N$, are i.i.d Bernoulli variables with parameter θ. Note that since $X^* = W_1 + \ldots + W_N$ is sufficient for W in the classical sense, $X^* > W$. On the other hand, for $Y^* = W_1 + \ldots + W_n$, $W > Y^*$. Therefore, $X^* > Y^*$. Since X and X^* (Y and Y^*) are indistinguishable in their models, $X > Y$.

To conclude our version of Skibinsky's characterization we note that

$$Y^* \mid X^* = x \sim h(N, n, x).$$

Then a Blackwell transition function $\{\tau_x\}$ for our problem is the family of Hypergeometric probability functions. That is,

$$P_R\{Y = y \mid \theta\} = \sum_x \tau_x(y) P_\theta\{X = x \mid \theta\} \qquad \ldots \quad (2.2)$$

for every $y \in \mathcal{Y}$ and every $\theta \in (0, 1)$ where $\tau_x(\cdot)$ is the Hypergeometric probability function with parameter (N, n, x). Finally, the uniqueness of $\{\tau_x\}$ as a Blackwell transition function follows from the fact that the family $\{b(N, \theta) : 0 < \theta < 1\}$ of probability distributions is complete. If $\{\tau'_x\}$ is another transition function satisfying (2.2), then

$$\sum_x [\tau_x(y) - \tau'_x(y)] P_\theta\{X = x \mid \theta\} = 0$$

for every $y \in \mathcal{Y}$ and therefore $\tau_x(y) = \tau'_x(y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

3. Further characterizations

Consider now an urn with N balls of k ($k \leqslant N$) different colors. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ be the vector of frequency counts for the colors; that is, x_i ($i = 1, 2, \ldots, k$) is the number of balls with the i-th color. If a simple random sample of n balls is drawn from the urn, then the sample vector of frequency counts has the multivariate Hypergeometric distribution with parameters, $N, n,$ and $\boldsymbol{x}$. This distribution is denoted by $H(N, n, \boldsymbol{x})$ and its support by

$$Z_n^k = \{(x_1, \ldots, x_k) : x_i \in Z, \ \Sigma x_i = N\}$$

where $Z = \{0, 1, \ldots\}$.

Writing $M(N, \boldsymbol{\theta})$ for the Multinomial distribution over Z_N^k, we state the natural extension of Skibinsky's characterization. Here the parameter space is the simplex

$$\mathcal{S} = \{(p_1, \ldots, p_k) : p_i \geqslant 0, \Sigma p_i = 1\}.$$

Proposition 1 : Let $X \sim M(N, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathcal{S}$, and let $\{\tau_{\boldsymbol{x}} : \boldsymbol{x} \in Z_N^k\}$ be a family of probability distributions on Z_n^k where $n \leqslant N$. Consider a random vector $Y \mid X = \boldsymbol{x} \sim \tau_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in Z_N^k$. Then $Y \sim M(n, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{S}$ if and only if $\tau_{\boldsymbol{x}}$ is $H(N, n, \boldsymbol{x})$ for all $\boldsymbol{x} \in Z_N^k$.

The proof follows the same steps of the univariate case discussed in Section 2. Here, we consider the experiment $W = (W_1, \ldots, W_N)$ where the components are i.i.d. with the common distribution being $M(1, \boldsymbol{\theta})$.

We write $X \sim \text{Poi}(\theta)$, $\theta > 0$, to indicate that X has Poisson distribution with parameter $\theta \in (0, \infty)$. Consider an additional experiment Y such that for a known number $r \in (0, 1)$, $Y \sim \text{Poi}(r\theta)$, $\theta > 0$. To prove that $X > Y$ we consider an experiment $W = (W_1, W_2)$ where its components W_1 and W_2 are independent with distributions $\text{Poi}(r\theta)$ and $\text{Poi}((1-r)\theta)$ respectively. Since $W > W_1$, $X^* = W_1 + W_2$ is sufficient for W in the classical sense, and X and X^* (Y and W_1) are indistinguishable in their models, it follows that $X > Y$.

Since $W_1 \mid X^* = x \sim b(x, r)$ for all $x \in Z$, a Blackwell transition function $\{\tau_x\}$ is the family of Binomial probability functions. That is,

$$\frac{e^{-\theta r}(\theta r)^y}{y\,!} = \sum_x \tau_x(y) \frac{e^{-\theta}\theta^x}{x\,!}$$

for every $y \in Z$ and all $\theta \in (0, \infty)$ where $\tau_x(\cdot)$ now is the Binomial probability function with parameter (x, r). The uniqueness of this family $\{\tau_x\}$ of Binomials as a Blackwell transition function follows from the completeness of the family $\{\text{Poi}(\theta) : \theta \in (0, \infty)\}$ of probability distributions on Z.

The above result in its extended form may be summarized as :

Proposition 2 : Let $X \sim \text{Poi}(\theta)$, $\theta > 0$, and let $\{\tau_x : x \in Z\}$ be a family of probability distributions on the set $Z^k = \{(y_1, \ldots, y_k) : y_i \in Z\}$. Consider a random vector $Y = (Y_1, \ldots, Y_k)$ such that $Y \mid X = x \sim \tau_x$ for all $x \in Z$. For a known vector $\boldsymbol{r} = (r_1, \ldots, r_k) \in \mathcal{S}$, the components, Y_i $(i = 1, \ldots, k)$, of Y are mutually independent and $Y_i \sim \text{Poi}(r_i\theta)$, $\theta > 0$ if and only if τ_x is $M(x, \boldsymbol{r})$ for all $x \in Z$.

We end this section with a parallel characterization of the Dirichlet-Multinomial distribution. In Basu and Pereira (1980) we studied in details this distribution and indicated its use in statistics. We define the Dirichlet-Multinomial $DM(N; \alpha_1, \ldots, \alpha_k)$ on $\boldsymbol{Z}_N^k$ as the mixture of the Multinomial family $\{M(N, \boldsymbol{p}); \boldsymbol{p} \in \mathcal{S}\}$ with $\boldsymbol{p}$ distributed as Dirichlet (on $\mathcal{S}$) with parameter $(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Its probability function is given by

$$f(z_1, \ldots, z_k) = \frac{N!\,\Gamma(\alpha)}{\Gamma(\alpha+N)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+z_i)}{z_i!\,\Gamma(\alpha_i)}$$

for all $(z_1, \ldots, z_k) \in \boldsymbol{Z}_n^k$ where $\alpha = \sum_1^k \alpha_i$. When $k = 2$, in place of $(Z_1, Z_2, \sim DM(N, \alpha_1, \alpha_2)$, we write $Z_1 \sim Bb(N; \alpha_1, \alpha_2)$ to indicate that Z_1 is distributed as Beta-Binomial with parameter $(N; \alpha_1, \alpha_2)$.

Consider a sequence of Bernoulli trials with probability of success $\theta \in (0, 1)$. If $X+\alpha$ is the number of trials needed to obtain a fixed number α of success, then X is said to be a Negative Binomial experiment with parameter $(\alpha; \theta)$ and we write $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$. Its probability function is

$$g(x|\theta) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\,\theta^{\alpha}(1-\theta)^{x} \qquad \ldots \quad (3.1)$$

for every $x \in \boldsymbol{Z}$ and all $\theta \in (0, 1)$. Note that

$$\sum_{x=0}^{\infty} \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\,(1-\theta)^{x} = \theta^{-\alpha} \text{ for every } \alpha \in (0, \infty) \text{ and all } \theta \in (0, 1).$$

Then the following results hold not only for $\alpha \in Z$ but in general for any $\alpha \in (0, \infty)$. In this case, we still write $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$, to indicate that the family of probability functions associated with the experiment X is (3.1). It is easy to check that this family is complete.

For $\alpha \geqslant \alpha_1 > 0$, let X and Y be two experiments such that $X \sim nb(\alpha; \theta)$ and $Y \sim nb(\alpha_1; \theta)$, $0 < \theta < 1$. To prove that $X > Y$, consider the experiment $W = (W_1, W_2)$ where now W_1 and W_2 are independent with distributions $nb(\alpha_1; \theta)$ and $nb(\alpha-\alpha_1; \theta)$ respectively. Following our previous chain of arguments, one can easily check that (i) $W_1+W_2 \sim X$, (ii) $W_1+W_2 > W > W_1$, (iii) $X > Y$, (iv) $W_1 | W_1+W_2 = x \sim Bb(x; \alpha_1, \alpha-\alpha_1)$, and (v) the family $\{Bb(x; \alpha_1, \alpha-\alpha_1) : x \in Z\}$ of probability functions is the unique Blackwell transition function, and thus arrive at a Skibinsky type characterization of the Beta-Binomial.

The following is a summary of an extended version of the above results.

Proposition 3 : Let $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$, and let $\{\tau_x : x \in Z\}$ be a family of probability distributions on the set Z^k. Consider a random vector $Y = (Y_1, \ldots, Y_k)$ such that $Y \mid X = x \sim \tau_x$ for all $x \in Z$. For a fixed vector $(\alpha_1, \ldots, \alpha_k)$ where $0 < \alpha_i < \infty$, $i = 1, 2, \ldots, k$, and $\alpha = \sum_1^k \alpha_i$, the components Y_i $(i = 1, \ldots, k)$ of Y are mutually independent with $Y_i \sim nb(\alpha_i, \theta)$, $0 < \theta < 1$, if and only if τ_x is $DM(x; \alpha_1, \ldots, \alpha_k)$ for all $x \in Z$.

References

Basu, D. and Pereira, C. A. B. (1982) : On the Bayesian analysis of categorical data : The problem of nonresponse. *Journal of Statistical Planning and Inference*, **6**, 345-362.

Blackwell, D. and Girshick, M. A. (1954) : *Theory of Games and Statistical Experiments*, John Wiley, N. Y.

Lehmann, E. L. (1959) : *Testing Statistical Hypothesis*, John Wiley, N. Y.

Skibinsky, M. (1970) : A characterization of hypergeometric distributions. *JASA*, **65**, 926-929.

Paper received : March, 1981.

13

Learning Statistics from Counter Examples: Ancillary Statistics

D. Basu [1]

Abstract

Bayesian objection to the analysis of data in frequency theory terms is amplified through several counter examples in which an ancillary statistic exists and there is a temptation to choose a reference set after looking at the data. It is argued that Fisher insisted on conditioning by an ancillary statistic, because conditioning the data $\mathbf{x}$ by an ancillary Y does not change the likelihood. In this sense Fisher discovered the supremacy of the likelihood function.

Key words and Phrases: Ancillary statistics; Conditional frequentist inference; Information; Liklihood principle; Reference set; Sufficiency principle.

1. INTRODUCTION

This paper is especially addressed to the statisticians who have not yet fully grasped the Bayesian objection to the analysis of data in repeated sampling terms. Let $\mathbf{x}$ be the sample, $f(\mathbf{x}|\theta)$ the model and θ the parameter. A statistic $Y = Y(\mathbf{x})$ is *ancillary* if the sampling distribution of Y, given θ, is θ-free (is the same for all values of θ). A statistic $T = T(\mathbf{x})$ is *sufficient* if the distribution of the sample $\mathbf{x}$, given T and θ, is θ-free. An ancillary statistic Y by itself contains no information about the parameter, whereas a sufficient statistic T is fully informative in a sense. R.A. Fisher's attempt to make sense of the notion of *information in the data* led him to these two important concepts in Statistics.

Let $L(\theta) = f(\mathbf{x}|\theta)$ be the *likelihood function* determined by the sample $\mathbf{x}$ and let $\hat{\theta}$ be the *maximum likelihood* (ML) estimate of θ. If $\hat{\theta}$ is a sufficient statistic then, according to Fisher, there would be no loss of information if the performance characteristics of $\hat{\theta}$ as an estimate of θ is sought to be evaluated in terms of the sampling distribution of $\hat{\theta}$. We shall repudiate this in the end with an example.

---

[1] Indian Statistical Institute and Florida State University

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_34,

If the ML estimate $\hat{\theta}$ is not a sufficient statistic then Fisher sought to recover the information lost in the sampling distribution of $\hat{\theta}$ with the help of an *ancillary complement* Y to the estimator $\hat{\theta}$. The ancillary statistic Y has to complement $\hat{\theta}$ in the sense that the pair $(\hat{\theta}, Y)$ is jointly sufficient. The Fisher Information $I_{\hat{\theta},Y}(\theta)$ in the sufficient statistic $(\hat{\theta}, Y)$ is then the same as the full information

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} \log L(\theta)]$$

in the sample $\mathbf{x}$. (Note that $I(\theta)$ does not relate to the particular sample $\mathbf{x}$ but is obtained by averaging the quantity $-\frac{\partial^2}{\partial\theta^2} \log L(\theta)$ over the sample space.) The Fisher Information in the statistic $\hat{\theta}$ is less than the full information $I(\theta)$. The cornerstone of the Fisher argument lies in the identity

$$I(\theta) = I_{\hat{\theta},Y}(\theta) = E[I_{\hat{\theta}}(\theta|Y)],$$

where $I_{\hat{\theta}}(\theta|Y)$ is the conditional information in the statistic $\hat{\theta}$, given Y, and the expectation on the right hand side is with respect to the ancillary statistic Y. Thus, the conditional information in $\hat{\theta}$, given Y, depends on Y and can be, for a particular value of the statistic Y, much less or much greater than the full information $I(\theta)$. The *conditionality argument* of R.A. Fisher rests on the proposition that the performance characteristics of the estimator $\hat{\theta}$ ought to be evaluated conditionally, holding the ancillary statistic Y fixed at its observed value y. As Fisher argued, the event $Y = y$, even though uninformative by itself, has a lot of latent information about θ in the sense that it helps us discern how good or bad the estimate $\hat{\theta}$ is in the present instance. The set $S(y) = \{\mathbf{x} : Y(\mathbf{x}) = y\}$ defines what Fisher called the *reference set*. Sir Ronald was trying to cut down the *sample space* S to size. We illustrate the conditionality argument with several examples.

2. EXAMPLES

Example 1: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be iid observations on a random variable that is uniformly distributed over the interval $[\theta, 2\theta]$, where $\theta > 0$ is the unknown scale parameter. With

$$m = \min x_i \text{ and } M = \max x_i,$$

the likelihood function $L(\theta)$ equals $1/\theta^n$ over the interval $[M/2, m]$ and zero outside the interval. The ML estimator $\hat{\theta} = M/2$ is not sufficient, the minimal sufficient statistic being the pair (m, M). Since the two statistics m and M are stochastically independent in an asymptomatic sense, it is clear that there will be a substantial loss of information if we marginalize the data to the ML estimator $M/2$. Comparing the mean squared error (MSE) of $M/2$ with that of m as estimators of θ, we find that the former is exactly four times better than the latter. Consider, therefore, the estimator

$T = (2M+m)/5$ which is the weighted average of $M/2$ and m with weights 4 and 1 respectively. Both $M/2$ and T are equivariant estimators of the scale parameter θ, and so their MSE's are constant multiples of θ^2. It works out that the ratio of the two MSE's tends to 25/12 as the sample size n tends to infinity. The ML estimator $\hat{\theta}$ can hardly be called an efficient estimate of θ in the usual sense of the term. Over thirty-six years ago, when I came upon this counterexample, it was pointed out to me by C.R. Rao that the ML estimator $\hat{\theta}$ ought to be judged conditionally after holding fixed its ancillary complement $Y = M/m$ at its observed value. That Y is an ancillary statistic follows from the facts that Y is scale invariant and that θ is a scale parameter. As we noted before, the likelihood mass is spread over the interval $[M/2, m]$ pinpointing the parameter θ within that interval. The statistic $Y = M/m$ varies over the range $[1, 2]$ and is indeed a measure of how good the sample is – the nearer Y is to 2 the better the sample is. While evaluating the ML estimate $\hat{\theta}$ we ought to take note of the observed value y of the statistic Y. That is, instead of referring $\hat{\theta}$ to the full sample space S, we ought to refer it to the *reference set* $S(y)$.

In terms of the full sample space S the ML estimator $M/2$ is not sufficient. But when it is conditioned by Y it suddenly becomes fully informative (sufficient, that is). Note that the other two estimators m and T also become fully informative when they are referred to the set $S(y)$. Indeed, the three statistics $M/2$, m and T become functionally related when conditioned by Y.

This example beautifully illustrates what Fisher meant by *recovery of ancillary information*. The next example illustrates how a weak pivotal quantity can be strengthened by proper conditioning with an ancillary statistic.

Example 2: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be n iid observations on a random variable with pdf $f(x-\theta)$, where f is known but θ (the location parameter) is unknown. Consider the statistic x_1 and its ancillary complement $D = (x_2 - x_1, x_3, -x_1, \ldots, x_n - x_1)$. The statistic x_1 by itself carries very little information about θ, but it becomes fully informative (sufficient) when conditioned by D. The conditional pdf of x_1, given D, has θ embedded in it as a location parameter. Fisher derived the *fiducial distribution* of the parameter θ by inverting the pivotal quantity $x_1 - \theta$ after conditioning it by the ancillary statistic D.

The previous example raises many questions. Some sample questions and answers are listed below.

Question: What is the status of the ancillary statistic D? Is it the *maximum ancillary* in the sense that every other ancillary statistic is a function of D?

Answer: No. D is never the maximum ancillary. However, in some situations D will be a *maximal ancillary* in the sense that no larger (with respect to the partial order of functional relationship) ancillary statistic exists. A

multiplicity of maximal ancillaries is a fact of life in this situation.

Question: Is the fiducial distribution of θ in Example 2 critically dependent on the choice of the pivotal quantity $x_1 - \theta$?

Answer: No. Another pivotal quantity like, say, $\mathbf{x} - \theta$, when conditioned by D, will result in the same fiducial distribution of θ. This is because $\bar{x} = x_1 + (\bar{x} - x_1)$ and $\bar{x} - x_1$ is a function of D.

Question: Can we interpret the fiducial distribution of θ probabilistically?

Answer: It was pointed out by Harold Jeffreys that the fiducial distribution of the location parameter (as derived by Fisher) coincides with the posterior distribution of θ corresponding to the uniform prior (over the entire real line) for the parameter.

In the presence of multiple ancillaries, the choice of the proper reference set is a problem. The dilemma is best exemplified by the following example.

Example 3: Let $(x_1, y_1), i = 1, 2, \ldots, n$, be n iid observations on (X, Y) whose joint distribution is Bivariate Normal with zero means, unit variances and covariance θ, which is the parameter of interest. In this case both $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are ancillary statistics. Note that the pair $(\mathbf{x}, \mathbf{y})$ is the entire data and therefore is sufficient. Holding the ancillary $\mathbf{x}$ as fixed and regarding $\mathbf{y}$ as the variable, we may want to estimate θ by $\sum x_i y_i / \sum x_i^2$ and then regard the estimate as unbiased with variance $(1 - \theta^2)/\sum x_1^2$. But how about holding $\mathbf{y}$ fixed and reporting that $\sum x_i y_i / \sum y_i^2$ is an unbiased estimate with variance $(1 - \theta^2)/\sum y_1^2$? It is tempting to opt for the ancillary with the larger sum of squares. But would it not be a statistical heresy to choose the reference set after looking at the data?!

3. COX ON ANCILLARIES

D.R. Cox (1971) suggested a way to deal with the problem of multiple ancillaries. Looking back at the Fisher identity $I(\theta) = EI(\theta|Y)$, Cox argued that the basic role of the conditioning ancillary Y is to discriminate between samples with varying degrees of information. So in the presence of multiple ancillaries we should choose that Y for which $I(\theta|Y)$ is most variable in Y. So opt for the Y for which Var $I(\theta|Y)$ is maximum. One snag in the Cox argument is that Var $I(\theta|Y)$ is a function of θ and so there may not exist a Y that maximizes the function uniformly in θ. Also note that in our Example 3 the Cox method fails because, in view of the perfect symmetry between $\mathbf{x}$ and $\mathbf{y}$, Var $I(\theta|\mathbf{x}) =$ Var $I(\theta|\mathbf{y})$.

But the real snag in the Cox argument is the meaninglessness of the notion of Fisher Information as a measure of the evidential meaning of the particular data at hand. Fisher's preoccupation with the elusive notion of information in the data led him to the likelihood function which he recognized as the carrier of all the information in the data. The likelihood was then partially summarized in the two statistics $\hat{\theta}$, the ML estimate, and $Z(\hat{\theta})$, the second derivative of $-\log L(\theta)$ at $\theta = \hat{\theta}$. Note that $Z(\hat{\theta})$ is the reciprocal of the radius of curvature of the log likelihood at its mode, the

larger the value of $Z(\hat{\theta})$ the sharper is the fall of the likelihood function as θ moves away from $\hat{\theta}$. We have to stretch our minds a little to regard $Z(\hat{\theta})$ as a rough measure of the concentration of the likelihood mass around $\hat{\theta}$. The greater the concentration the more informative is the likelihood. The Fisher Information $I(\theta)$ is obtained from $Z(\hat{\theta})$ by first replacing $\hat{\theta}$ by θ and then taking the average value of $Z(\theta)$ over the whole sample space S. But how can we regard $I(\theta)$ as information in the data?

Why did Fisher insist that the conditioning statistic Y has to be ancillary? Because, conditoning the data $\mathbf{x}$ by an ancillary Y does not change the likelihood. Fisher discovered the supremacy of the likelihood but got carried away by his amazing craftmanship with sample space mathematics.

4. E.L. LEHMANN ON ANCILLARIES

Eric Lehmann (1981) finally recognized the conditionality argument. And now he has to cope with the disturbing presence of ancillary statistics. Invoking the *Sufficiency Principle*, Eric would reduce the data $\mathbf{x}$ to the minimal sufficient statistic $T = T(\mathbf{x})$. Since T is sufficient, all reasonable inference procedures ought to depend on $\mathbf{x}$ only through $T(\mathbf{x})$. This data reduction sweeps away much of the ancillary dust under the rug. But, as in Example 1, some functions of the minimal sufficient statistic T may still be recognized as ancillary statistics. Eric has yet to come out openly on the question of how to deal with such persistent ancillaries.

From what Eric writes in his 1981 article, it seems that he feels quite comfortable with statistical models for which the minimal sufficient statistic T is complete. In such cases no nontrivial function of T can be ancillary. Furthermore, thanks to the so called Basu Theorem, every ancillary statistic Y is stochastically independent (conditionally on θ) of T. Therefore, no T- based decision procedure can be altered by conditioning with an ancillary Y. So who needs to think of the conditionality argument when we have a complete sufficient statistic? Remember, Fisher looked for an ancillary complement to the ML estimate $\hat{\theta}$ only when the statistic $\hat{\theta}$ was not sufficient. So in the most favorable set up where $\hat{\theta}$ is a complete sufficient statistic, can anyone object if we evaluate the estimate $\hat{\theta}$ in terms of the sampling distribution of the estimator? We give an example to prove both Fisher and Lehmann wrong on this question.

Example 4: Consider a sequence of Bernoulli trials with parameter p that results in a finite sequence $w = SFFS \ldots FS$ of successes S and failures F. Let $X(w)$ and $Y(w)$ denote, respectively, the number of S's and the number of F's in the sample sequence w. We picture w as a sample path, the locus of a point that begins its journey at the original and travels through the lattice points of the positive quadrant, moving one step to the right for each S and one step up for each F. The lattice point with coordinates $X(w)$ and $Y(w)$ is where the sample path w ends. Our example relates to a particular

sampling (stopping) rule **R**. Writing (X, Y) for the location of the moving point, the rule is described as:

Rule **R**: Continue sampling as long as (I) $Y < 2X + 1$, (II) $Y > X - 2$, and (III) $X + Y < 100$. Alternatively, the rule may be defined as: Stop sampling as soon as the sample path hits one of the three boundary lines (i) $y = 2x + 1$, (ii) $y = x - 2$, and (iii) $x + y = 100$.

As always, the likelihood does not recognize the stopping rule and comes out as

$$L(p) = f(w|p) = p^{X(w)}q^{Y(w)}$$

where $q = 1 - p$. The pair $X(w), Y(w)$ constitute the minimal sufficient statistic. The ML estimate is $\hat{p} = X/(X + Y)$. The range of the sufficient statistic (X, Y) consists of the boundary points

$$\begin{array}{llll} (0,1), & (1,3), & \ldots, & (33,67) \quad \text{on line} \quad (i), \\ (34,66), & (35,65), & \ldots, & (50,50) \quad \text{on line} \quad (iii), \textit{and} \\ (51,49), & (50,48), & \ldots, & (2,0) \quad \text{on line} \quad (ii). \end{array}$$

The ML estimator $\hat{p} = X/(X + Y)$ monotonically increases from zero to unity as (X, Y) moves through the above set of boundary points. Hence $\hat{p}$ itself is minimal sufficient. Let us assert here without proof that $\hat{p}$ is a complete sufficient statistic in this case and that no nontrivial ancillary statistic exists.

Sir Ronald is no longer with us. So let me address the following questions to my good friend Eric Lehmann who is a living legend among us for his unparalleled erudition in Statistical Mathematics. The questions relate to Example 4.

Question: What should be our criterion for the choice of an estimate of p?

(The unbiasedness criterion is sort of vacuous in this case. There is only one unbiased estimator, which is zero or unity depending on whether the first trial results in an F or an S.)

Question: If ML is the chosen criterion, then how should we evaluate the estimate $\hat{p} = X/(X + Y)$? Does it make sense to evaluate $\hat{p}$ in terms of some average performance characteristics?

Question: Are all sample paths w equally informative?

(Even though there are no ancillary statistics in this case, we can still detect major qualitative differences between different sample paths. For instance, short sample paths like F or SS have very little to say about the parameter, whereas long paths that end on line (iii) are clearly much more informative.)

Question: Why do we need to decipher what the sample w has to say about the parameter p in terms of a sample space? Does the sample F

obtained following the rule **R** say anything different from the statement: A single Bernoulli trial has resulted in a failure?

Question: Do sample space ideas like bias, variance, risk function, etc., make any sense in this case?

Question: Why not act like a Bayesian and analyze the particular likelihood function generated by the data? Isn't is quite clear in this case that all that the data has to say about the parameter is summarized in the likelihood?

REFERENCES

Basu, D. (1988), *Statistical Information and Likelihood: A Collection of Critical Essays* by D. Basu, ed. J.K. Ghosh, Springer Verlag, New York.

Cox, D. R. (1971), The Choice Between Alternative Ancillary Statistics. *Jour. Royal Statist. Soc.* **(B)33**, 251-255.

Lehmann, E.L. (1981), An Interpretation of Completeness and Basu's Theorem. *Jour. Amer. Stat. Assoc.*, **76**, 335-340.