ROLE OF THE SUFFICIENCY AND LIKELIHOOD PRINCIPLES IN SAMPLE SURVEY THEORY*

By D. BASU

University of New Mexico and *Indian Statistical Institute*

SUMMARY. In this paper, the statistical model for sample surveys is first put in the conventional set-up of $(\Omega, \alpha, \mathcal{P})$, and it is shown that a maximal sufficiency reduction is always possible for a sample survey model. The corresponding minimal sufficient statistic is derived. We examine the role of the sufficiency and likelihood principles in the analysis of survey data and arrive at the revolutionary but reasonable conclusion that, once the sample has been drawn, the inference should not depend in any way on the sampling design. This poses the problem of designing a survey which will yield a good (representative) sample. The randomisation principle is examined from this view point and it is noticed that there is very little, if any, use for it in survey designs.

1. INTRODUCTION

This article was written with the object of emphasizing the following four points.

(a) The first point is only of pedagogical interest. Recently, a series of interesting papers have appeared [Pathak (1964), Godambe (1966), Hanurav (1968), Joshi (1968) to mention only a few] in which the statistical model for sample surveys has been so formulated as to confuse conventional statistical mathematicians [ordinarily incapable of speculating about anything excepting the trinity of $(X, \alpha, \mathcal{P})$!] into the belief that the analysis of survey type data falls outside the mainstream of the theory of statistical analysis. In these formulations, one sees on the surface a 'sample space' S (of possible samples s) with just one probability measure p on S. [How can there be any inference with just one measure ?!] The pair (S, p) is called the sampling design. A typical sample $s \epsilon S$ is a subset of (or a finite sequence with its members drawn from) a fixed population Π of individuals $1, 2, 3, \ldots, N$. The parameter is an unknown vector $\theta = (Y_1, Y_2, \ldots, Y_N)$. A statistic is a very special kind of a function of the sample s and the parameter θ. [How can a statistic be anything but a function defined on the sample space ?!] And so on and on it (the new formulation) goes, apparently blazing a new trail in the wilderness of statistical thought. In this article we point out that it is not really necessary to formulate the survey model in the above 'unfamiliar' manner. We need not abandon the trinity $(X, \alpha, \mathcal{P})$!

(b) The second point emphasized here is also of a purely academic nature. If we assume that the set of 'possible' values for the parameter θ is uncountable, then the family $\mathcal{P}$ in the sample survey model $(X, \alpha, \mathcal{P})$ would be typically undominated. This raises the possibility that there may not exist a maximal sufficiency reduction of the survey data (and other hair raising possibilities!). But the saving grace for the survey model is that each member of $\mathcal{P}$ is always a discrete measure. The existence

* This research was partially supported by Research Grant No. ZU-2582 of the National Science Foundation.

of the maximal sufficiency reduction of the data (the minimal sufficient statistic) is always assured if we take every set as measurable. Also it is very easy to characterize and use the minimal sufficient statistic.

(c) In this article we examine the role of the twin principles of sufficiency and likelihood in the analysis of survey data and arrive at the revolutionary but entirely reasonable conclusion that at the analysis stage the statistician should not pay any attention to the nature of the sampling design. Indeed, the analyst need not even know the sampling design that produced the data.

(d) It goes without saying that there is a great need for designing the survey very carefully. How else can we expect to get a good (representative) sample ? Currently, survey statisticians make extensive use of the random number tables. In this article, the author very briefly examines the randomization principle and comes to the conclusion that there is very little (if any) use for it in survey designs.

2. STATISTICAL MODELS AND SUFFICIENCY

The notion of a sampling (or statistical) experiment is idealized as a statistical model $(X, \alpha, \mathcal{P})$ where

(i) X is the sample space,

(ii) α is a fixed σ-field of subsets of X, called the measurable sets or the events, and

(iii) $\mathcal{P} = \{P_\theta | \theta \varepsilon \Omega\}$ is a fixed family of probability measures P_θ on α.

The family $\mathcal{P}$ is indexed by the unknown state of nature (the parameter) θ. The set of all the possible values of θ is the parameter space Ω.

By the term statistic we mean a characteristic of the sample x. A general and abstract formulation of the notion of a statistic is that of a mapping of X onto a space Y. Thus, a statistic $T = T(x)$ is an arbitrary function with X as its domain. Every statistic T defines an equivalence relation [$x \sim x'$ if $T(x) = T(x')$] on the sample space X. This leads to a partition of X into equivalent classes of sample points. As we need not distinguish between statistics that induce the same partition of X, it is convenient to think of a statistic T as a partition $\{\pi\}$ of X into a family of mutually exclusive and collectively exhaustive parts π.

The statistic (partition) $T = \{\pi\}$ is said to be wider (larger) than the statistic $T^* = \{\pi^*\}$ if every π is a subset of some π^*—in other words, if every π^* is a union of a number of π's. Given a statistic $T = \{\pi\}$, consider the class of all measurable sets (members of α) that are unions of some π's. They constitute a sub-σ-field (sub-field) of α and is denoted by α_T. We call α_T the sub-field induced by T. If T is wider than T^* then $\alpha_T \supset \alpha_{T^*}$.

An abstract and very general formulation of the notion of sufficient statistic is the following :

Definition : The statistic T is sufficient $[\alpha, \mathcal{P}]$ if, corresponding to every real-valued bounded, α-measurable function f, there exists an α_T -measurable f^* such that for all $B \in \alpha_T$ and $\theta \in \Omega$

$$\int_B f \, dP_\theta = \int_B f^* \, dP_\theta.$$

The notion of sufficiency has been studied in great details in statistical literature. In the particular case where the family $\mathcal{P}$ of probability measures is dominated by a σ-finite measure λ, we have the following factorization theorem of fundamental importance.

Theorem : *Let $p_\theta = dP_\theta/d\lambda$ be a fixed version of the Radon-Nikodym derivative of P_θ wrt λ. A necessary and sufficient condition for the sufficiency of the statistic T is that there exists, for each $\theta \in \Omega$, an α_T-measurable function g_θ and a fixed α-measurable function h such that, for each $\theta \in \Omega$,*

$$p_\theta(x) = g_\theta(x)h(x) \quad a.e.w.\,[\lambda].$$

In a dominated set-up, most of the properties of sufficient statistics flow from the above factorization theorem. For example, if T is sufficient then any statistic T^* that is wider (larger) than T is also sufficient. Again, with a separability condition on $\mathcal{P}$, it is true that there exists a sufficient statistic T which is essentially smaller (narrower) than every other sufficient statistic T^*. Such a sufficient statistic is called the minimal (or least) sufficient statistic.

That neither of the above two propositions need hold for general undominated set-ups has been exhibited by Burkholder (1961) and Pitcher (1957). Consider the following two examples.

Example 1: Let X be the real line, α the σ-field of Borel sets and $\mathcal{P}$ the class of all discrete two-point probability distributions P_θ on the line that are symmetric about the origin. [That is, the entire mass of P_θ is equally distributed over the two points $-\theta$ and θ, where $\theta > 0$.] Let E be a non-Borel set that excludes the origin but is symmetric about it. Let $T(x) = |x|$ and let

$$T^*(x) = \begin{cases} |x| & \text{if } x \in E \\ x & \text{if } x \notin E. \end{cases}$$

Clearly, T^* is wider than T. However, in this example, T is sufficient but T^* is not.

Example 2 : Let X, α and E be as in the previous example and let $\mathcal{P} = \{P_\theta\}$ be defined as follows. If $\theta \in E$ then the whole mass of P_θ is equally distributed over the points $-\theta$ and θ. If $\theta \notin E$, then P_θ is degenerate at θ. In this example, there does not exist a minimal sufficient statistic.

In each of the above two examples, we are dealing with a family of measures each member of which is discrete. In example 1, each measure has its entire mass concentrated at two points only; in example 2, each measure has its entire mass

distributed over at most two points. True, we are dealing, in each case, with an undominated family of measures. But that is not where the real trouble lies. In these examples, our difficulties stem from our artificially restricting ourselves to Borel sets only. If in the above two examples we take α to be the class of all subsets, then we do not have to face the above kind of anomalous situations. The natural domain of definition of discrete measures is the σ-field of all subsets. In sample survey theory, we need not consider non-discrete probability measures. By a discrete model we mean the following.

Definition : The statistical model (X, α, P_θ), $\theta \in \Omega$, is called a discrete model if

(i) each P_θ is a discrete measure,

(iii) α is the class of all subsets of X, and.

(iii) for each $x \in X$, there exists a $\theta \in \Omega$, such that, $P_\theta(\{x\}) > 0$.

[*Remark* : Condition (iii) only ensures that we do not entangle ourselves with possibilities that have zero probabilities for each possible value of the parameter θ. Condition (ii) ensures that all sets and functions are measurable.]

We, henceforth, deal with discrete models only. A discrete model is undominated if and only if X is uncountable. The Burkholder-Pitcher type pathologies cannot occur in discrete models (Basu and Ghosh, 1967).

3. Sufficiency in discrete models

Let (X, α, P_θ), $\theta \in \Omega$, be a discrete model. For each $x \in X$ let

$$\Omega_x = \{\theta \mid P_\theta(x) > 0\}.$$

[We, henceforth, write $P_\theta(x)$ for $P_\theta(\{x\})$.] The set Ω_x is the set of parameter points that are consistent with the observation (sample point) x. No Ω_x is vacuous.

For discrete models, the minimal sufficient statistic always exists and is uniquely defined as follows. Consider the binary relation on X : "$x \sim x'$ if $\Omega_x = \Omega_{x'}$ and $P_\theta(x) \mid P_\theta(x')$ is a constant in θ for all $\theta \in \Omega_x = \Omega_{x'}$".

The above is an equivalence relation on X. The partition (statistic) induced by the equivalence relation is the minimal (least) sufficient statistic. This is an easy consequence of the following factorization theorem (Basu and Ghosh, 1967).

Theorem : *If (X, α, P_θ), $\theta \in \Omega$, be a discrete model, then a necessary and sufficient condition for a statistic (partition) $T = \{\pi\}$ to be sufficient is that there exists a function g on X such that, for all $\theta \in \Omega$ and $x \in X$,*

$$P_\theta(x) = g(x) P_\theta(\pi_x),$$

where π_x is that part of the partition $\{\pi\}$ that contains x.

The above factorization theorem is a direct and easy consequence of the definitions of sufficient statistics and discrete models (as stated in Section 2).

If $T = \{\pi\}$ be a sufficient partition and if g be defined as in the previous theorem, then it follows that $g(x) > 0$ for all $x \in X$ and that, for each π,

$$\sum_{x \in \pi} g(x) = 1.$$

Each part of a sufficient partition must be countable. What the above factorization theorem is telling us is nothing but the intuitively obvious proposition that $\{\pi\}$ is sufficient if and only if, for each part π, it is true that the conditional distribution of the sample x given π is θ-free. This is indeed the original definition of sufficiency as proposed by Fisher.

Another consequence of the above theorem is that if $T = \{\pi\}$ is a sufficient statistic then any statistic T^* that is wider than T is necessarily sufficient. It also follows that (for discrete models) there exists a one-one correspondence between sufficient statistics (partitions) and sufficient sub-fields (Basu and Ghosh, 1967).

An alternative (but equivalent) way of characterizing the minimal sufficient statistic for a discrete model is the following. For each $x \in X$ let $L_x(\theta)$ stand for the likelihood function, i.e.

$$L_x(\theta) = \begin{cases} P_\theta(x) & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{for} \quad \theta \notin \Omega_x. \end{cases}$$

Let us standardize the likelihood function as follows.

$$\overline{L}_x(\theta) = \frac{L_x(\theta)}{\sup_\theta L_x(\theta)}.$$

Consider the mapping

$$x \to \overline{L}_x(\theta),$$

a mapping of X into a class of real-valued functions on Ω. This mapping is the minimal sufficient statistic. [A little reflection would show that the partition (of X) induced by the above mapping is the same as the one induced by the equivalence relation described earlier in this section.]

4. The sample survey models

The principal features of a sample survey situation are as follows. There exists a well-defined population Π — a finite set of distinguishable objects called the (sampling) units. Typically, there exists a list of these units—the so-called sampling frame. Let us list the population as

$$\Pi = (1, 2, 3, \ldots, N).$$

The unit i has an unknown characteristic Y_i. The unknown state of nature is

$$\theta = (Y_1, Y_2, \ldots, Y_N).$$

The statistician has some prior information or knowledge K about θ. This knowledge K is largely of a qualitative and speculative nature. For example, the statistician knows that θ is a member of a well-defined set Ω (the parameter space). He also knows, for each unit i, some characteristic A_i of the unit i. Let us denote this set of known auxiliary characteristics by

$$A = (A_1, A_2, \ldots A_N).$$

Thus, A is a principal component of K. In K is also embedded what the statistician thinks (knows) to be the true relationship between the unknown θ and the known A.

It is within the powers of the statistician to find out or "observe" the characteristic Y_i for any chosen unit i. A survey problem arises when the statistician plans to gain further "information" about some function $\tau = \tau(\theta)$ of the parameter θ by observing the Y-characteristics of a set (sequence)

$$i = (i_1, i_2, \ldots, i_n)$$

of units selected from Π.

Let us denote the observed Y-characteristics by

$$y = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}).$$

The problem is to make a "suitable" choice of i and then to make a "proper" use of the observations $x = (i, y)$ in conjunction with the prior "knowledge" K to arrive at a "reasonable" "judgement" about τ.

Now, let us examine how probability theory comes into the picture. If we ignore observation errors, then there is no discernable source of randomness in the above general formulation of a survey problem (excepting some very intangible quantities like "belief", "knowledge" etc. which the Bayesians try to formalize as probability.) In any survey situation there are bound to be some observation errors (the so-called non-sampling errors). Unfortunately, in current sample survey research it is not often that we find mention of this source of randomness. It is tacitly assumed that the observation errors are negligible in comparison with the so-called "sampling error". This sampling error is the distinguishing feature of the current sample survey theory. Here is a phenomenon of randomness that is not inherent to the problem but is artificially injected into the problem by the statistician himself. The survey statistician does not lean on probability theory for the purpose of understanding and controlling the mess created by an unavoidable source of randomness or uncertainty (observation errors). He uses his knowledge of probability theory to introduce into the problem a well-understood (fully controlled) element of randomness and seems to derive all his strength (intellectual conviction) from that.

The "sampling error" is the randomness that the statistician injects into the problem by selecting the set (sequence) $i = (i_1, i_2, \ldots, i_n)$ in a random manner. Given a sampling plan $\mathfrak{S}$, for each possible i there exists a number $p(i)$ which is the probability of ending up with i. Usually, this $p(i)$ does not depend on the parameter θ, although quite often it is made to depend on the auxiliary information A. [However, one may consider sequential sampling plans for which $p(i)$ depends on θ. For instance, consider the sampling plan—"Choose unit 1 and observe Y_1. If Y_1 (which we suppose is real valued) is larger than b then choose unit 2, otherwise choose unit N". For this plan i is either $(1,2)$ or $(1, N)$ and $p(i)$ depends on θ through Y_1.] Typically, the random choice of i is made in the statistical laboratory well in advance of the time that the observation job is in progress. For such typical sampling plans, the probability $p(i)$ for any possible i does not depend on θ at all. However, even if we agree to consider sequential sampling plans of the type described within the parenthesis before, it is clear that $p(i)$ for such plans can depend on $\theta = (Y_1, Y_2, \ldots, Y_N)$ only through $y = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$. As we shall presently see, this remark is important. In the sequel we write $p(i \mid \theta)$ for $p(i)$.

The sample is $x = (i, y)$, the set i together with the observation y. [For some sampling plans—like sampling with replacements—it is more natural to think of i as a finite sequence of units with repetitions allowed.] The sample space X is the set of all possible samples x.

Now, each x, when observed, tells us the exact Y-value of some population units, i.e., tells us about some coordinates of the vector θ. Let Ω_x be the set of parameter points θ that are consistent with a given sample x. If $P_\theta(x)$ be the probability that the sampling plan ends up with sample $x = (i, y)$, then it is clear that

$$P_\theta(x) = \begin{cases} p(i \mid \theta) & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases}$$

Thus, Ω_x is also the set of all parameter points that allot non-zero probabilities to x.

As we have said before, in typical sampling plans $p(i \mid \theta)$ does not depend on θ. In sequential sampling plans (where the choice of a population unit at any stage is made to depend on the observed Y-values of the previously selected units) we have noted before that $p(i \mid \theta)$ depends on θ through y. Thus, we make the following important observation that, for any sampling plan,

$$P_\theta(x) = \begin{cases} \text{constant} & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{for} \quad \theta \notin \Omega_x. \end{cases}$$

This leads us to the following general characterization of a sample survey model.

Definition : The model (X, α, P_θ), $\theta \in \Omega$ is called an SS-model if the model is discrete and if $P_\theta(x)$ is a constant for all $\theta \in \Omega_x$, where

$$\Omega_x = \{\theta \mid P_\theta(x) > 0\}.$$

From what we have said in Section 3, it then follows that

Theorem : *If* (X, α, P_θ), $\theta \in \Omega$ *be an SS-model, then the minimal (least) sufficient statistic is the mapping** $x \to \Omega_x$.

The distinguishing feature of an SS-model is that for every possible sample the likelihood function is flat. That is, for every $x \in X$ the likelihood function $L_x(\theta)$ is zero for all θ outside a set Ω_x and is a constant for $\theta \in \Omega_x$. The following is an example of a non-discrete model with the above feature.

Example 3 : Let $x = (x_1, x_2, \ldots, x_n)$ be n independent observations on a random variable that has a continuous and uniform distribution over the interval $(\theta - 1/2, \theta + 1/2)$, where θ is the parameter $(-\infty < \theta < \infty)$. Let Ω_x be the interval $(m(x), M(x))$ where $m(x) = \max x_i - 1/2$ and $M(x) = \min x_i + 1/2$. Here, $L_x(\theta) = 1$ for all $\theta \in \Omega_x$ and is zero for $\theta \notin \Omega_x$. The mapping $x \to (m(x), M(x)) = \Omega_x$ is the minimal sufficient statistic.

5. THE SUFFICIENCY AND LIKELIHOOD PRINCIPLES

The twin principles of sufficiency and likelihood both attempt to answer the same question. The likelihood principle, however, goes a great deal further in its assertion.

The question is : "What characteristic of the sample x is relevant for making an inference about the parameter θ?" In general, the sample x is a very complex entity. Must we take into account the sample x in all its detail ? Could it be that some characteritics of x are totally irrelevent for making any inference about the state of nature θ ? For instance, if in the observation x we have incorporated the outcome u from a number of tosses of a symmetric coin, then it seems very reasonable to argue that the characteristic u of x is totally irrelevent and must be ignored.

The sufficiency principle is the following. If $T = T(x)$ be a sufficient statistic, then only the characteristic $T(x)$ of x is relevant for inference making. That is, if $T(x) = T(x')$ then the inference about θ should be the same whether the sample is x or x'. The relevant information core of x is then the statistic $T_0(x)$, we T_0 is the minimal sufficient statistic.

The sufficiency principle has gained rather wide acceptance. The Neyman-Pearson school of statisticians tend to justify the principle by proving some complete class theorem that tells us that it is not necessary to consider decision rules (inference procedures) that do not depend on x through $T_0(x)$. On the other hand the Bayesians have no objection to the sufficiency principle as they point out that the posterior distribution for the parameter θ—whatever be its prior distribution—depends on x only through the minimal sufficient statistic $T_0(x)$.

As we have stated in Section 3, the mapping $x \to \overline{L}_x(\theta)$, where $\overline{L}_x(\theta)$ is the standardized (modified) likelihood function, is the minimal sufficient statistic. Thus,

*In typical survey situations, the minimal sufficient statistic (the information core of the sample) is the set of (distinct) population unit-labels that are drawn in the sample together with the corresponding Y-values.

according to the sufficiency principle, two sample points x and x' are equally informative if

$$\bar{L}_x(\theta) \equiv \bar{L}_{x'}(\theta) \text{ for all } \theta.$$

Note that the sufficiency principle does not tell us anything about the nature of the information supplied by x. The likelihood principle takes a big step forward and asserts that the information supplied by x *is* the likelihood function $\bar{L}_x(\theta)$. Whereas the sufficiency principle can compare two possible samples x and x' only when they are points in the same sample space, the likelihood principle can compare them even when they are points in different sample spaces. Consider the following example.

Example 4 : Let θ be the unknown probability of head for a given coin. The following is a list of three different experiments (among the many that one can think of) that one may perform for the purpose of eliciting information about the unknown θ.

$\mathcal{E}_1$: Toss the coin 5 times

$\mathcal{E}_2$: Toss the coin until there are 3 heads

$\mathcal{E}_3$: Toss the coin until there are 2 consecutive heads.

We give below an example x_i of a possible sample point for each experiment $\mathcal{E}_i$ $(i = 1, 2, 3)$. [H = head, T = tail]

x_1 : $H\ T\ H\ H\ T$

x_2 : $T\ T\ H\ H\ H$

x_3 : $H\ T\ T\ H\ H$

[Note that the sample spaces for the three experiments are different from one another. Also note that x_1 cannot be a sample point for either $\mathcal{E}_2$ or $\mathcal{E}_3$. Similarly, x_2 cannot be a sample point for $\mathcal{E}_3$.]

It is easy to check that the likelihood function for x_i (when it is referred to experiment $\mathcal{E}_i$) is $\theta^3(1-\theta)^2$ and this is irrespective of whether i is 1, 2, or 3. The principle of likelihood tells us that sample x_1 for experiment $\mathcal{E}_1$ gives the same information about θ as does sample x_2 from $\mathcal{E}_2$ and sample x_3 from $\mathcal{E}_3$.

From the Bayesian point of view the likelihood principle is almost a truism. The starting point for a Bayesian (in his inference making effort) is a prior probabiltiy distribution over the parameter space Ω. Let $q = q(\theta)$ be the prior probability frequency function. Having observed the sample x, the Bayesian uses the likelihood function $\bar{L}_x(\theta)$ to arrive at the posterior distribution

$$q_x^*(\theta) = \frac{q(\theta)\bar{L}_x(\theta)}{\Sigma_\theta\, q(\theta)\,\bar{L}_x(\theta)}\,.$$

To a Bayesian, the role of the sample x is only to change his prior scale of preference (probability distribution) $q = q(\theta)$, for various possible values of θ, to the posterior scale $q^* = q_x^*(\theta)$. And this change is effected through the likelihood function $\overline{L}_x(\theta)$. Possible sample points x and x' (whatever sampling experiments might generate them) are equivalent as long as they induce identical (modified) likelihood functions. The likelihood principle is essentially a Bayesian principle. It is hard to justify the principle under the Neyman-Pearson set-up.

6. Role and choice of the sampling plan

Let $\mathcal{S}$ be the chosen sampling plan and let $x = (i, y)$ be the data (sample) generated by $\mathcal{S}$. In the matter of analyzing the data, how relevant is the plan $\mathcal{S}$?

If $i = (i_1, i_2, \ldots, i_n)$ and $y = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$, then Ω_x is the set of all $\theta \in \Omega$ whose j-th co-ordinate is Y_j $(j = i_1, i_2, \ldots i_n)$—the set of θ's that are consistent with the data. Note that Ω_x depends only on x and Ω, it has nothing to do with the plan $\mathcal{S}$ The minimal sufficient statistic is the mapping $x \to \Omega_x$ and the likelihood function $\overline{L}_x(\theta)$ is

$$\overline{L}_x(\theta) = \begin{cases} 1 & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases}$$

If $q = q(\theta)$ be the Bayesian prior distribution over Ω, then the posterior distribution is

$$q_x^*(\theta) = \frac{q(\theta)\overline{L}_x(\theta)}{\Sigma_\theta q(\theta)\,\overline{L}_x(\theta)} = \begin{cases} c(x)\,q(\theta) & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases}$$

The posterior distribution $q_x^*(\theta)$ is nothing but the restriction of q to the set Ω_x. And the plan $\mathcal{S}$ does not enter into the definition of Ω_x. Thus, from the Bayesian (and the likelihood principle) point of view, once the data x is before the statistician, he has nothing to do with the plan $\mathcal{S}$. He does not even need to know what the plan $\mathcal{S}$ was. [This is because, in sample survey situations, the plan $\mathcal{S}$ is an artificial source of randomness. In other statistical situations, where randomness is unavoidable and is an inherent part of the observation process, the statistician has to "understand" the process well enough to be able to arrive at his likelihood function.]

In the Neyman-Pearson type of analysis of the data, the statistician considers not only the data x in hand but also pays a great deal of attention to what other data x' he might have obtained. In other words, he needs to know the model $(X, \alpha, \mathcal{P})$ as well as the sample x. The Bayesian needs to know only the likelihood function $\overline{L}_x(\theta)$, which, in a sample survey situation, is entirely independent of the model (the sampling plan $\mathcal{S}$). The author does not think that any reconcilliation between the two approaches to data analysis is possible.

A majority of statisticians of the Neyman-Pearson school would readily agree to the proposition that the Bayesian analysis of the data is sensible (acceptable) when the following condition holds :

Condition B : It is reasonable to think of the parameter θ as a random variable, and the random process governing θ is at least partially discernable.

However, it is hard to understand how such statisticians reconcile themselves to the contrary positions: (a) Only the data x (the likelihood function) is relevant (for inference making) when condition B holds and (b) the whole sample space X(the model) is relevant when B does not hold. There exists a continuous spectrum of conditions between the extremes of B and not-B. But the shift of emphasis from the sample x to the sample space X is not continuous. [Fisher with his theory of ancillary statistics and choice of reference sets, made a bold but unsuccessful (see Basu, 1964) attempt to bridge the gap between the above polarities in statistical theory.]

It seems to the author that the Bayesian analysis of the data x is very appropriate in sample survey situation. Given the data $x = (i, y)$ the sampling plan $\mathcal{S}$—the model $(X, \alpha, \mathcal{P})$—ceases to be of any relevance for inference making about the parameter θ. Given the data x the statistician arrives at his posterior preference scale $q_x^*(\theta)$ for the parameter θ. If $\tau = \tau(\theta)$ be the parameter of interest, then the statistician can compute the marginal posterior distribution $q_x^*(\tau)$ of the variable τ. The question, "Given x, how much information we have about τ ?", can then be answered by first agreeing upon a suitable definition of information. [For example, we may agree to work with the Shannon definition of information or with the posterior variance (or its reciprocal) of τ.]

Given a sample x, we can now tell how good (informative) the sample is. The object of planning a survey should be to end up with a good sample The term "representative sample" has often been used in sample survey terminology. But no one has cared to give a precise definition of the term. It is implicitly taken for granted that the statistician with his biased mind is unable to select a representative sample. So a simplistic solution is sought by turning to an unbiased die (the random number tables). Thus, a deaf and dumb die is supposed to do the job of selecting a "representative sample" better than a trained statistician. It is, however, true that we do not really train our statisticians for the job of selecting and observing survey type data. [In contrast, the medical practitioner is given a much more meaningful training in understanding the many variables and their interrelations in his chosen field of specialization.]

In a Bayesian plan for selecting the sample x, there is no place for the symmetric die. Very little attention has so far been paid to the problem of devising suitable sampling strategies from this point of view. In a later document the author would elaborate some of his own ideas on the problem. We end this section by describing

a Bayesian sampling strategy for the very simple case where the statistician wants to select and observe only one unit. Suppose his prior probability distribution is $g(\theta)$. If he selects unit i and observes Y_i then his posterior (marginal) distribution for τ would be, say, $g^*(\tau \mid i, Y_i)$. Once a suitable definition of "information" is agreed upon, he can use the above distribution to compute the quantity $I(i, Y_i)$—the information about τ gained from the sample (i, Y_i). At the planning stage of the experiment, the statistician does not know the value of Y_i that he is going to observe for the unit i. Let $J(i)$ be the average value of $I(i, Y_i)$ when the averaging is done over all possible values of Y_i (weighted by the prior distribution of Y_i). Thus, $J(i)$ is the "expected" information to be gained from observing unit i. Faced with the problem of deciding which unit i to select (and then observe), the statistician would not be acting unreasonably if he selects the unit i that has maximum $J(i)$. [What if the $J(i)$'s are all equal ? Such would be the case if the prior distribution of $\theta = (Y_1, Y_2, \ldots, Y_N)$ is symmetric in the coordinates. In this situation the statistician is indifferent as to which i is selected for observation. In principle, he cannot object now to a random (with equal or unequal probabilities) selection prodedure for i. However, this does not mean that he will be willing to let another person (say, a field investigator) make the choice for him. If for nothing else, a scientist ought to be always on his guard against letting an unknown element enter into the picture.]

Of course, a non-Bayesian would sneer at the arbitrariness inherent in the difinition of $J(i)$. But the procedure described above is certainly more justifiable than our current naive reliance on the symmetric die. Any reasonable Bayesian sampling strategy would have the following characteristics. (a) The sampling plan would usually be sequential. The statistician would continue sampling (one or a few units at a time) until he is satisfied with the information thus obtained or until he reaches the end of his rope (time and cost). His decision to select the units for a particular sampling stage would depend (non-randomly) on the sample obtained in the previous stages. (b) The probability that the statistician would end up observing the units $i = (i_1, i_2, \ldots, i_n)$ in this order, would depend on i and the state of nature θ. This probability would be degenerate, i.e., zero for some values of θ and unity for the rest of the values of θ.

7. SOME CONCLUDING REMARKS

(a) Godambe (1966a) noted that the application of the likelihood principle in the sampling situation would mean that the sampling design is irrelevant for data analysis. On page 317 he writes, "One implication of this, as can be seen from (4), is that the inference about θ must not depend on the sampling design even through the probability $p(i)$ of the i that has actually been drawn. In particular, the estimator of τ should not depend on $p(i)$ or the sampling design". [In this and in the following quotation the author has taken the liberty of changing some of the notations. This was done for the sake of bringing them in line with the notations used in this article]. It is interesting to observe that Godambe immediately shies away from the revolu-

tionary implication of his remark and tries to find some excuses for not applying the likelihood principle in the sampling situation. He writes (p 317, Godambe, 1966a), "In connection with the likelihood principle, it may be further noted that here θ is the parameter and (i, y) is the sample. Thus, possibly there is some kind of relationship between the parametric space and the sample space (when the sample is observed, the parameter cannot remain completely unknown) which forbids the use of the likelihood principle. The relationships between parametric and sample spaces restricting the use of the likelihood principle are referred to by Barnard, Jenkins and Winsten (1962)". In the two 1966 papers referred to here, Godambe tries very hard to justify a particular linear estimator as the only reasonable one for the population total. Godambe's estimator depends on the sampling design. The author finds Godambe's arguments very obscure.

(b) Let us repeat once again that the posterior distribution of τ depends only on the prior distribution q (on Ω) and the sample $x = (i, y)$. It does not depend on the sampling design $\mathcal{S}$. Thus, any fixed q on Ω would give rise to a Bayes 'estimation procedure' B_q that would tell us how to estimate τ for each possible sample x —no matter what design $\mathcal{S}$ is used to arrive at x. [Note that B_q is well-defined as a function on the union of sample spaces for all designs $\mathcal{S}$.] Now, if we consider B_q in relation to a fixed design $\mathcal{S}$, then it would be classified as an admissible estimator in the sense of Wald. The findings of Godambe (1960) and Joshi (1965) therefore appear to the author as rather obvious in nature. It is so easy to reel off any number of such universally admissible estimation procedures.

(c) The mathematical content of this article is summarized in the theorem of Section 4. This result has been known (but never explicitly proved) to the author (and among others to Godambe, Hájek, Hanurav and Pathak) for the past eleven years or so. In an yet unpublished article, entitled "Classical sufficiency and its application to sampling theory", Pathak has proved this result for a particular (non-sequential) case.

References

Basu, D. (1958): On sampling with and without replacements. *Sankhyā*, 20, 287-294.

Basu, D. (1964): Recovery of ancillary information. *Sankhyā*, 25, 3-16.

Basu, D. and Ghosh, J. K. (1967): Sufficient statistics in sampling from a finite universe. *Bull. Int. Stat. Inst.*, 42, BK. 2, 850-859.

Burkholder, D. L. (1961): Sufficiency in the undominated case. *Ann. Math. Statist.*, 32, 1191-1200.

Godambe, V. P. (1960): An admissible estimate for any sampling design. *Sankhyā*, 22, 285-288.

——— (1966a): A new approach to sampling from finite populations, I: Sufficiency and linear estimation. *JRSS* (series B), 28, 310-319.

——— (1966b): A new approach to sampling from finite populations, II: Distribution-free sufficiency. *JRSS* (series B), 28, 320-328.

Hanurav, T. V. (1964): Hyper-admissibility and optimum estimators for sampling finite populations. *Ann. Math. Statist.*, 39, 621-642.

JOSHI, V. M. (1968) : Admissibility of the sample mean as estimate of the mean of a finite population. *Ann. Math. Statist.*, 39, 606-620.

PATHAK, P. K. (1964) : Sufficiency in sampling theory. *Ann. Math. Statist.*, 35, 795-809.

PITCHER, T. S. (1957) : Sets of measures not admitting necessary and sufficient statistics or subfields. *Ann. Math. Statist.*, 28, 267-268.

Paper received ; November, 1968.
Revised : May, 1969.