AN ESSAY ON THE LOGICAL FOUNDATIONS OF SURVEY SAMPLING, PART ONE*

D. Basu
The University of New Mexico and
Indian Statistical Institute

1
An Idealization of the Survey Set-up

It is a mathematical necessity that we idealize the real state of affairs and come up with a set of concepts that are simple enough to be incorporated in a mathematical theory. We have only to be careful that the process of idealization does not distort beyond recognition the basic features of a survey set-up, which we list as follows:

(a) There exists a population—a finite collection $\mathcal{P}$ of distinguishable objects. The members of $\mathcal{P}$ are called the (sampling) units. [Outside of survey theory the term population is often used in a rather loose sense. For instance, we often talk of the infinite population of all the heads and tails that may be obtained by repeatedly tossing a particular coin. Again, in performing a Latin-square agricultural experiment the actual yield from a particular plot is conceived of as a sample from a conceptual population of yields from that plot. It is needless to mention that such populations are not real. The existence of a down-to-Earth finite population is a principal characteristic of the survey set-up.]

(b) There exists a sampling frame of reference. By this we mean that the units in $\mathcal{P}$ are not only distinguishable pairwise, but are also *observable* individually; that is, there exists a list (frame of reference) of the units in $\mathcal{P}$ and it is within the powers of the surveyor to pre-select any particular unit from the list and then observe its characteristics. Let us assume that the units in $\mathcal{P}$ are listed as

$$1, 2, 3, \ldots, N,$$

*Research supported in part by NSF Grant GP-9001.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_24, © Springer Science+Business Media, LLC 2011

where N is finite and is known to the surveyor. [We are thus excluding from our survey theory such populations as, for example, the insects of a particular species in a particular area or the set of all color-blind adult males in a particular country. Such populations as above can, of course, be the subject matter of a valid statistical inquiry but the absence of a sampling frame makes it impossible for such populations to be *surveyed* in the sense we understand the term survey here.]

(c) Corresponding to each unit $j \in \mathscr{P}$ there exists an unknown quantity Y_j in which the surveyor is interested. The unknown Y_j can be made known by observing (surveying) the unit j. The unknown state of nature is the vector quantity

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

However, the surveyor's primary interest is in some characteristic (parameter)

$$\tau = \tau(\theta)$$

of the state of nature θ. [Typically, the Y_j's are vector quantities themselves and the surveyor is seeking information about a multiplicity of τ's. However, for the sake of pinpointing our attention to the basic questions that are raised here, we restrict ourselves to the simple case where the Y_j's are scalar quantities (real numbers) and $\tau = \Sigma Y_j$.]

(d) The surveyor has prior knowledge K about the state of nature θ. This knowledge K is a multi-dimensional complex entity and is largely of a qualitative and speculative nature. We consider here the situation where K has at least the following two well-defined components. The surveyor *knows* the set Ω of all the *possible* values of the state of nature θ and, for each unit j, ($j = 1,2,\ldots,N$), he has access to a record of some known auxiliary characteristic A_j of j. [Typically, each A_j is a vector quantity. However, in our examples we shall take the A_j's to be real numbers.] The set Ω and the vector

$$\alpha = (A_1, A_2, \ldots, A_N)$$

are the principal measurable components of the surveyor's prior knowledge K. Let us denote the residual past of the knowledge by R and write

$$K = (\Omega, \alpha, R).$$

(e) The purpose of a survey is to gain further knowledge (beyond what we have described as K) about the state of nature θ and, therefore, about the parameter of interest $\tau = \tau(\theta)$. Since the surveyor is supposed to know the set Ω of all the possible values of θ, he knows the set $\mathscr{T}$ of all the possible values of τ. Initially, the surveyor's *ignorance* about τ is, therefore, *spread* over the set $\mathscr{T}$. [Later on, we shall quantify this initial *spread of ignorance* as a prior probability distribution.] In theory, the surveyor can dispel this ignorance and gain complete knowledge by making a total survey (complete enumeration) of $\mathscr{P}$. If he observes the Y-characteristic of every

unit j ($j = 1,2,\ldots,N$), then he knows the actual value of $\theta = (Y_1, Y_2, \ldots, Y_N)$ and, therefore, that of $\tau(\theta)$. We are, however, considering the case where a total survey is impracticable. By a *survey* of the population $\mathcal{P}$ we mean the selection of a (usually small) subset

$$u = (u_1, u_2, \ldots, u_n)$$

of units from $\mathcal{P}$ and then observing the corresponding Y-values

$$y = (Y_{u_1}, Y_{u_2}, \ldots, Y_{u_n})$$

of units in the subset u.

(f) We make the simplifying assumption that there are no *non-response* and *observation* errors; that is, the surveyor is able to observe every unit that is in the subset u, and when he observes a particular unit j, he finds the true value of the hitherto unknown Y_j without any error.

(g) The surveyor's blueprint for the survey is usually a very complicated affair. The survey plan must take care of myriads of details. However, in this article we idealize away most of these details and consider only two facets of the survey project, namely, the *sampling plan* and the *fieldwork*. The sampling plan is the part of the project that yields the subset u of $\mathcal{P}$ and fieldwork generates the observations y on members of u. The data (sample) generated by the survey is

$$x = (u, y) .$$

For reasons that will be made clear later, it is important to distinguish between the two parts u and y of the data x.

(h) Let $\mathcal{S}$ stand for the sampling plan of the surveyor. The plan (when set in motion) produces a subset u of the population $\mathcal{P} = (1,2,\ldots,N)$. We write $u = (u_1, u_2, \ldots, u_n)$, where $u_1 < u_2 < \ldots < u_n$ are members of $\mathcal{P}$. The fieldwork generates the vector $y = (Y_{u_1}, Y_{u_2}, \ldots, Y_{u_n})$ which we often write as $(y_1, y_2, \ldots, y_n)$. [Occasionally, we shall consider sampling plans that introduce a natural selection order among the units that are selected. For such plans it is more appropriate to think of u, not as a subset of $\mathcal{P}$, but as a finite sequence of elements $u_1, u_2, \ldots, u_n$ drawn from $\mathcal{P}$ in that selection order. In rare instances, the sampling plan may allow the possibility of a particular unit appearing repeatedly in the sequence $(u_1, u_2, \ldots, u_n)$. From the description of the sampling plan $\mathcal{S}$ it will usually be clear if we intend to treat u as a set or a sequence. In either case, we can think of u as a vector $(u_1, u_2, \ldots, u_n)$ and y as the corresponding observation vector $(y_1, y_2, \ldots, y_n)$ where $y_i = Y_{u_i}$.]

(i) *Summary*: Our idealized survey set-up consists of the following:

(i) A finite population $\mathcal{P}$ whose members are listed in a sampling frame as $1,2,\ldots,N$. Availability of each $j \in \mathcal{P}$ for observation.

(ii) The unknown state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$ and the parameter of interest $\tau = \tau(\theta)$.

(iii) The prior knowledge $K=(\Omega, \alpha, R)$.
(iv) Absence of non-response and observation errors.
(v) Choice of a sampling plan $\mathcal{S}$ as part of the survey design.
(vi) Putting the sampling plan $\mathcal{S}$ and the fieldwork into operation, thus arriving at the data (sample)

$$x=\{u=(u_1,u_2,...,u_n), \quad y=(y_1,y_2,...,y_n)\}$$

where $u_i \in \mathcal{P}$ and $y_i = Y_{u_i} (i=1,2,...,n)$.
(vii) Making a *proper* use of the data x in conjunction with the prior knowledge K to arrive at a *reasonable judgment* (or decision) related to the parameter τ.

The operational parts of the survey are its design (v), the actual survey (vi) and the data analysis (vii). In this article we are concerned only with the design and the analysis of a survey.

2
Probability in Survey Theory

We posed the survey set-up as a classic problem of inductive inference—a problem of inferring about the whole from observations on only a part. The basic questions are: Which part does one observe? Does the part (actually observed) tell us anything about the whole? and, then the main question, Exactly what does it tell? Let us now examine how probability enters into the picture.

There are three different ways in which probability theory finds its way into the mathematical theory of survey sampling. First, there is the time honored way through a probabilistic model for observation errors. Indeed, this is how probability theory first infiltrated the sacred domain of science. When we observe the Y-value Y_j of unit j, there is bound to be some observation error. In current survey theory we classify this kind of error as *non-sampling* error. In this article we have idealized away this kind of probability by assuming that there exists no observation error. We have deliberately taken this simplistic view of the survey set-up. The idea is to concentrate our attention on the other two sources of probability.

In current survey theory, the main source of probability is randomization, which is an artificial introduction (through the use of random number tables) of randomness in the sampling plan $\mathcal{S}$. Randomization makes it possible for the surveyor to consider the set (or sequence) u, and therefore the data $x=(u,y)$, as random elements. With an element of randomization incorporated in the sampling plan $\mathcal{S}$, the surveyor can consider the space U of all the possible values of the random element u and then the probability distribution p_θ of u over U. [For sampling plans usually discussed in survey textbooks, the probability distribution p_u of u is uniquely determined (by the plan) and is, therefore, independent of the state of nature θ. In part two of this article we shall take a broader view of the subject and also consider plans for which the probability distribution of u involves θ.] Now, let X be the space of all the possible values of the data (sample) $x=(u, y)$ of which

we have already recognized (thanks to randomization) the part u to be a random element. The space X is our sample space. Let P_u be the probability distribution of x over the sample space X. If $T = T(x)$ is an estimate of τ, then (prior to sampling and fieldwork) we can consider T to be a random variable and speculate about its sampling distribution and its average performance characteristics (as an estimator of τ) in an hypothetical sequence of repeated experimentations. This decision-theoretic approach is not possible unless we regard randomization as the source of probability in survey theory. From the point of view of a frequency-probabilist, there cannot be a statistical theory of surveys without some kind of randomization in the plan $\mathcal{S}$.

Apart from observation errors and randomization, the only other way that probability can sneak into the argument is through a mathematical formalization of what we have described before as the residual part R of the prior knowledge $K = (\Omega, \alpha, R)$. This is the way of a subjective (Bayesian) probabilist. The formalization of R as a prior probability distribution of θ over Ω makes sense only to those who interpret the probability of an event, not as the long range relative frequency of occurrence of the event (in an hypothetical sequence of repetitions of an experiment), but as a formal quantification of the illusive (but nevertheless very real) phenomenon of *personal belief* in the truth of the event. According to a Bayesian, probability is a mathematical theory of belief and it is with this kind of a probability theory that one should seek to develop the guidelines for inductive behavior in the presence of uncertainty. The purpose of this essay is not to examine the logical basis of Bayesian probability nor to describe how one may arrive at the actual qualification of R into a prior probability distribution of θ over Ω. [Of late, a great deal has been written on the subject. See, for instance, I. R. Savage's delightfully written new book, *Statistics: Uncertainty and Behavior.*]

Can the two kinds of probability co-exist in our survey theory? This is what we propose to find out.

3
Non-Sequential Sampling Plans and Unbiased Estimation

By a non-sequential sampling plan we mean a plan that involves no fieldwork. If the sampling plan $\mathcal{S}$ is non-sequential, then the surveyor can (in theory) make the selection of the set (or sequence) u of population units right in his office and then send his field investigators to the units selected in u and thus obtain the observation part y of the data $x = (u, y)$. A great majority of survey theoreticians have so far restricted themselves to non-sequential plans that involve an element of randomization in it. In this section we consider such plans only. The essence of non-sequentialness of a plan $\mathcal{S}$ is that the probability distribution of u does not involve the state of nature θ. Thus, the sampling plan where we continue to draw a unit at a time with equal probabilities and with replacements until we get ν distinct units is a non-sequential plan.

Given $u = (u_1, u_2, ..., u_n)$, the observation part $y = (y_1, y_2, ..., y_n)$, where $y_i = Y_{u_i} (i = 1, 2, \ldots, n)$, is obtained through the fieldwork and is uniquely determined by the state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$. The conditional

probability distribution of y given u is degenerate, the point of degeneration depending on θ. That is, for all y, u and θ

$$\text{Prob}(y \mid u, \theta) = 0 \text{ or } 1. \tag{3.1}$$

[We are taking the liberty of using the symbols u, y, x and θ both as variables and as particular values of the variables.]

For each sampling plan $\mathcal{S}$ we have the space U of all the possible values of u. The probability distribution p of u over U is θ-free, that is, is uniquely defined by the plan $\mathcal{S}$. The probability distribution p is clearly discrete. There is no loss of generality in assuming that $p(u) > 0$ for all $u \in U$. If the non-sequential plan is *purposive* (that is, the plan involves no randomization) then U is a single-point set and the distribution of u is degenerate at that point.

Let X be the sample space, the set of all possible samples (data) $x = (u, y)$ where u is generated by the plan $\mathcal{S}$ and y by the fieldwork. For each $\theta \in \Omega$, we have a probability distribution P_θ over X. Whatever the plan $\mathcal{S}$, the probability distribution P_θ is necessarily discrete. We write $P_\theta(x)$ or $P_\theta(u, y)$ for the probability of arriving at the data $x = (u, y)$ when θ is the true value of the state of nature. Clearly,

$$P_\theta(x) = P_\theta(u, y) = p(u) \text{ Prob } (y \mid u, \theta). \tag{3.2}$$

The surveyor takes a peep at the unknown $\theta = (Y_1, Y_2, \ldots, Y_N)$ through the sample $x = (u, y)$. Prior to the survey, the surveyor's ignorance about θ was spread over the space Ω. Once the data x is at hand, the surveyor has exact information about some coordinates of the vector θ. These are the coordinates that correspond to the distinct units that are in u. The data x rules out some points in Ω as clearly inadmissible. Let Ω_x be the subset of values of θ that are consistent with the data x. In other words, $\theta \in \Omega_x$ if $P_\theta(x) > 0$; that is, it is possible to arrive at the data x when θ is the true value of the state of nature. The subset Ω_x of Ω is well-defined for every $x \in X$. Without any loss of generality we may assume that no Ω_x is vacuous. From (3.1) and (3.2) it follows that the likelihood function $L(\theta)$ is given by the formula

$$L(\theta) = P_\theta(x) = \begin{cases} p(u) & \text{for all } \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

In other words, whatever the data x, the likelihood function $L(\theta)$ is flat (a positive constant) over the set Ω_x and is zero outside Ω_x. This remark holds true for sequential plans also (Basu, 1969). The importance of the remark will be made clear later on.

A major part of survey theory is concerned with unbiased estimation. A statistic is a characteristic of the sample x. An estimator $T = T(x)$ is a statistic that is well-defined for all $x \in X$ and is used for estimating a parameter $\tau = \tau(\theta)$. By an unbiased estimator of $\tau(\theta)$ we mean an estimator T that satisfies the identity

$$E(T \mid \theta) = \sum_x T(x) P_\theta(x) \equiv \tau(\theta), \text{ for all } \theta \in \Omega. \tag{3.4}$$

Let $w(t, \theta)$ be the loss function. That is, $w(t, \theta)$ stands for the surveyor's assessment of the magnitude of error that he commits when he estimates the parameter $\tau = \tau(\theta)$ by the number t. We assume that

$$w(t, \theta) \geq 0 \text{ for all } t \text{ and } \theta,$$

the sign of equality holding only when $t = \tau(\theta)$. The risk function $r_T(\theta)$ associated with the loss function w and the estimator T is then defined as the expected loss

$$r_T(\theta) = E[w(T, \theta) \mid \theta] = \sum_x w(T(x), \theta) P_\theta(x). \tag{3.5}$$

[If the reader is not familiar with the decision-theoretic jargons of loss and risk, he may restrict himself to the particular case where $w(t, \theta)$ is the squared error $(t - \tau(\theta))^2$ and the risk function $r_T(\theta)$ is the variance $V(T \mid \theta)$ of the unbiased estimator T.] The following theorem proves the non-existence of a uniformly minimum risk (variance) unbiased estimator of τ.

Theorem. Given an unbiased estimator T of τ and an arbitrary (but fixed) point $\theta_0 \in \Omega$, we can always find an unbiased estimator T_0 (of τ) such that $r_{T_0}(\theta_0) = 0$, that is, T_0 has zero risk at θ_0.

Proof: We find it convenient to write $T(u, y)$ and $P_\theta(u, y)$ for $T(x)$ and $P_\theta(x)$ respectively. It has been noted earlier that the conditional distribution of y given u is degenerate at a point that depends on θ. Let $y_0 = y_0(u)$ be the point of degeneration of y, for given u, when $\theta = \theta_0$. Consider the statistic

$$T_0 = T(u, y) - T(u, y_0) + \tau(\theta_0). \tag{3.6}$$

The statistic $T(u, y_0)$ is a function of u alone and so its probability distribution, and therefore its expectation are θ-free. Indeed,

$$\begin{aligned} E[T(u, y_0)] &= \sum_u T(u, y_0) p(u) \\ &= \sum_{u,y} T(u, y) P_{\theta_0}(u, y) \\ &= E[T(u, y) \mid \theta_0] \\ &= \tau(\theta_0). \end{aligned}$$

Thus, the statistic T_0 as defined in (3.6) is an unbiased estimator of τ. Now, when $\theta = \theta_0$, the statistics $T(u, y)$ and $T(u, y_0)$ are equal with probability one, and so $T_0 = \tau(\theta_0)$ with probability one. This proves the assertion that $r_{T_0}(\theta_0) = 0$.

The impossibility of the existence of a uniformly minimum risk unbiased estimator follows at once. For, if such an estimator T exists then $r_T(\theta)$ must be zero for all $\theta \in \Omega$. That is, whatever the value of the state of nature θ it should be possible to estimate $\tau(\theta)$ without any error (loss) at all. Unless the sampling plan $\mathcal{S}$ is equivalent to a total survey of the population, such a T clearly cannot exist for a parameter τ that depends on all the coordinates of θ. The following two examples will clarify the theorem further.

Example 1. Consider the case of a simple random sample of size one from the population $\mathcal{P} = (1, 2, ..., N)$. The sample is (u, y) where u has a uniform probability distribution over the N integers 1, 2, ..., N and $y = Y_u$. Let the population mean

$$\overline{Y} = \frac{1}{N}(Y_1 + ... + Y_N)$$

be the parameter to be estimated. Clearly, y is an unbiased estimator of $\overline{Y}$. Let $\theta_0 = (a_1, a_2, \ldots, a_N)$ and $\bar{a} = (\Sigma a_j)/N$. The statistic

$$T_0 = y - a_u + \bar{a} \tag{3.7}$$

is an unbiased estimator of $\overline{Y}$ with zero risk (variance) when $\theta = \theta_0$. The variance of y is $\Sigma(Y_j - \overline{Y})^2/N$ and that of T_0 is $\Sigma(Z_j - \overline{Z})^2/N$ where $Z_j = Y_j - a_j$ $(j = 1, 2, \ldots, N)$.

Example 2. Let $\mathcal{S}$ be an arbitrary non-sequential sampling plan that allots a positive selection probability to each population unit. That is, the probability Π_j that the unit j appears in the set (or sequence) u is positive for each $j (j = 1, 2, \ldots, N)$. Since $\mathcal{S}$ is non-sequential, the vector

$$\Pi = (\Pi_1, \Pi_2, \ldots, \Pi_N)$$

is θ-free. Let Y be the population total ΣY_j. A particular unbiased estimator of Y that has lately attracted a great deal of attention is the so-called Horvitz-Thompson (HT) estimator (relative to the plan $\mathcal{S}$). The HT-estimator is defined as follows. Let $u_1 < u_2 < \ldots < u_\nu$ be the distinct population units that appear in u and let $\hat{y} = (y_1, y_2, \ldots, y_\nu)$ be the corresponding observation vector. Let $p_i = \Pi u_i$ $(i = 1, 2, \ldots, \nu)$. The HT-estimator H is then defined as

$$H = \frac{y_1}{p_1} + \ldots + \frac{y_\nu}{p_\nu}. \tag{3.8}$$

That H is an unbiased estimator of Y will be clear when we rewrite (3.8) in a different form. Let $E_j (j = 1, 2, \ldots, N)$ stand for the event that the plan $\mathcal{S}$ selects unit j, and let I_j be the indicator of the event E_j. That is, $I_j = 1$ or 0 according as unit j appears in u or not. It is now easy to check that

$$H = \sum_{j=1}^{N} \Pi_j^{-1} I_j Y_j. \tag{3.9}$$

That H is an unbiased estimator of Y follows at once from the fact that

$$\begin{aligned} E(I_j) &= \text{Prob}\,(E_j) \\ &= \Pi_j \quad (j = 1, 2, \ldots, N). \end{aligned}$$

Now, let $\theta_0=(a_1, a_2, \ldots, a_N)$ be a point in Ω that is selected by the surveyor (prior to the survey) and let H_0 be defined as

$$H_0=H-\Sigma\,\Pi_j^{-1} I_j a_j+\Sigma a_j. \tag{3.10}$$

It is now clear that H_0 is an unbiased estimator of Y and that $V(H_0 \mid \theta)=0$ when $\theta=\theta_0$. Since the variances of H and H_0 are continuous functions of θ, it follows that

$$V(H_0 \mid \theta) < V(H \mid \theta) \tag{3.11}$$

for all θ in a certain neighborhood Ω_0 of the point θ_0. If the surveyor has the prior knowledge that the true value of θ lies in Ω_0, then the modified Horvitz-Thompson estimator H_0 is uniformly better than H. The estimator H_0 will look a little more reasonable if we rewrite it as

$$H_0=[\Sigma\,\Pi_j^{-1} I_j(Y_j-a_j)]+\Sigma a_j. \tag{3.12}$$

If (3.9) is a reasonable estimator of $Y=\Sigma Y_j$, then the variable part of the right hand side of (3.12) is an equally reasonable estimator of

$$\Sigma(Y_j-a_j)=Y-\Sigma a_j.$$

The strategy of a surveyor who advocates the use of (3.12) [in preference to that of (3.9)] as an estimator of $Y=\Sigma Y_j$ is quite clear. Instead of defining the state of nature as

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

he is defining it as

$$\theta'=(Y_1-a_1, Y_2-a_2, \ldots, Y_N-a_N).$$

Suppose the surveyor has enough prior information about the state of nature, so that by a proper choice of the vector $(a_1, a_2, \ldots, a_N)$ he can make the coordinates of θ' much less variable than that of θ. He is then in a better position to estimate the total of the coordinates of θ' than one who is working with θ. Consider the situation where the surveyor knows in advance that the j^{th} coordinate Y_j of θ lies in a small interval around the number a_j ($j=1, 2, \ldots, N$). In such a situation the surveyor ought to shift the origin of measurement (for θ) to the point $(a_1, a_2, \ldots, a_N)$ and represent the state of nature as

$$\theta'=(Y_1-a_1, Y_2-a_2, \ldots, Y_N-a_N).$$

If the numbers $a_1, a_2, \ldots, a_N$ has a large dispersion, then shifting the origin of measurement to $(a_1, a_2, \ldots, a_N)$ will cut down the variability in the coordinates of the state of nature to a large extent. The effect will be similar to what is usually achieved by stratification.

At this point one may very well raise the questions: Why must the surveyor choose his knowledge vector $(a_1, a_2, \ldots, a_N)$ before the survey?

Is it not more reasonable for him to wait until he has the survey data at hand and then take advantage of the additional knowledge gained thereby? Once the data is at hand, the surveyor knows the exact values of the surveyed coordinates of θ. The natural post-survey choice of a_j for any surveyed j is, therefore, Y_j. For a non-surveyed j, the surveyor's best estimate a_j of the unknown Y_j would still be of a speculative nature. If in formula (3.12) we allow the surveyor to insert a post-survey specification of the vector $(a_1, a_2, \ldots, a_N)$, then the first part of the right hand side of (3.12) will vanish and the estimator will look like

$$\begin{aligned} H_* &= \Sigma\, a_j \\ &= (Y_{u_1} + Y_{u_2} + \ldots + Y_{u_\nu}) + \sum_{j \notin u} a_j \qquad (3.13) \\ &= S + S^*, \end{aligned}$$

where S is the sum total of the Y-values of the distinct surveyed units and S^* is the surveyor's post-survey estimate of the total Y-values of the non-surveyed units.

A decision-theorist will surely object to our derivation of formula (3.13) as naive and incompetent. He will point out that we have violated a sacred cannon of inductive behavior, namely *never select the decision rule after looking at the data*. He will also point out that S^* in (3.13) is, as yet, not well-defined (as a function on the sample space X), and he will reject H_* (as an estimator of Y) with the final remark that the whole thing stinks of Bayesianism!

Nevertheless, the fact remains that formula (3.13) points to the very heart of the matter of estimating the population total Y. A survey leads to a complete specification of a part of the population total. This part is the sample total S as defined in (3.13). At the end of the survey the remainder part Y-S is still unknown to the surveyor. If the surveyor insists on putting down T as an estimate of Y, then he is in effect saying that he has reason to believe that T-S is close to Y-S. And then he should give a reasonable justification for his belief. Of course we can write any estimate T in the form

$$T = S + S^*$$

where S is the sample total and $S^* = T - S$. But then, for some T, the part S^* (of T) would appear quite preposterous as an estimate of the unknown part $Y - S$ of Y. The following two examples will make clear the point that we are driving at.

Example 3. The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the

elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take 50 y (where y is the present weight of Sambo) as an estimate of the total weight $Y=Y_1+\ldots+Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive samplings plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $100y/99$ and not $50y$." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)

Example 4. Sampling with unequal probabilities has been recommended in situations that are less frivolous than the one considered in the previous example but the recommended unbiased estimators for such plans sometimes look hardly less ridiculous than the one just considered. Let us consider the so-called pps (probability proportional to size) plans about which so many research papers have been written in the past 20 years. A pps sampling plan is usually recommended in the following kind of situation. Suppose for each population unit j we have a record of an auxiliary characteristic A_j (the size of j). Also suppose that each A_j is a positive number and that the surveyor has good reason to believe that the ratios

$$\Lambda_j = Y_j/A_j \quad (j=1,2,\ldots,N) \tag{3.14}$$

are nearly equal to each other. In this situation it is often recommended that the surveyor adopts the following without replacement pps.

Sampling plan. Let $A=\Sigma A_j$ and $P_j=A_j/A$ $(j=1,2,\ldots,N)$. Choose a unit (say, u_1) from the population $\mathcal{P}=(1,2,\ldots,N)$ following a plan that allots a selection probability P_j to unit j $(j=1,2,\ldots,N)$. The selected unit u_1 is then removed from the sampling frame and a second unit (say, u_2) is selected

from the remaining $N-1$ units with probabilities proportional to their sizes (the auxiliary characters A_j). This process is repeated n times so that the surveyor ends up with n distinct units

$$u_1, u_2, \ldots, u_n$$

listed in their natural selection order. After the fieldwork the surveyor has the sample

$$x = \{(u_1, y_1), \ldots, (u_n, y_n)\}$$

where $y_i = Y_{u_i} (i = 1, 2, \ldots, n)$. Let us write p_i for $P_{u_i} (i = 1, 2, \ldots, n)$ and

$$x^* = \{(u_1, y_1), \ldots, (u_{n-1}, y_{n-1})\}$$

for the vector defined by the first $n-1$ coordinates of x. It is then easy to see that (see Desraj [6] Theorem 3.13)

$$E\left(\frac{y_n}{p_n} \middle| x^*\right) = \Sigma' \frac{Y_j}{P_j} \cdot \frac{P_j}{1-p_1-p_2-\ldots-p_{n-1}}$$

$$= (\Sigma' Y_j)/(1-p_1-\ldots-p_{n-1}),$$

where the summation is carried over all j that are different from $u_1, u_2, \ldots, u_{n-1}$. Since $\Sigma' Y_j = Y-(y_1+\ldots+y_{n-1})$ it follows at once that

$$E\left(y_1+\ldots+y_{n-1}+\frac{y_n}{p_n}(1-p_1-\ldots-p_{n-1}) \mid x^*\right) = Y. \tag{3.15}$$

Therefore, the unconditional expectation of the lefthand side of (3.15) is also Y. And so we have the so-called Desraj estimator

$$D = y_1+\ldots+y_{n-1}+\frac{y_n}{p_n}(1-p_1-\ldots-p_{n-1}), \tag{3.16}$$

which is an unbiased estimator of Y. Writing S for the sample total $y_1+\ldots+y_n$ we can rewrite (3.16) as

$$D = S + S^* \tag{3.17}$$

where

$$S^* = \frac{y_n}{p_n}(1-p_1-\ldots-p_n).$$

Let us examine the face-validity of S^* as an estimate of Y^*, the total Y-values of the unobserved population units.
Writing $A = \Sigma A_j$, $a_i = A_{u_i}$ $(i = 1, 2, \ldots, n)$ and $A^* = A-a_1-\ldots-a_n$ (the total A-value of the unobserved units), we have

$$S^* = \frac{y_n}{p_n}(1-p_1-\ldots-p_n) \tag{3.18}$$

$$= \frac{y_n}{a_n} A^* \quad (\text{since } p_i = \frac{a_i}{A}).$$

Clearly, S^* would be an exact estimate of Y^* if and only if

$$\frac{y_n}{a_n}=\frac{Y^*}{A^*}=\frac{\Sigma' Y_j}{\Sigma' A_j} \tag{3.19}$$

(the summation is over the unobserved j's).

Now, if the surveyor claims that according to his belief (3.17) is a good estimate of Y, then that claim is equivalent to an assertion of belief in the near equality of the two ratios

$$\frac{y_n}{a_n} \text{ and } \frac{Y^*}{A^*}.$$

What can be the logical basis for such a belief? We started with the assumption that the surveyor has prior knowledge of near equality in the N ratios in (3.14). At the end of the survey, the surveyor has observed exactly n of these ratios and they are

$$\frac{y_1}{a_1}, \frac{y_2}{a_2}, \ldots, \frac{y_n}{a_n}. \tag{3.20}$$

The surveyor is now in a position to check on his initial supposition that the ratios in (3.14) are nearly equal. Suppose he finds that the observed ratios in (3.20) are indeed nearly equal to each other. This will certainly add to the surveyor's conviction that the unobserved ratios Λ_j (where j is different from $u_1, u_2, \ldots, u_n$) are nearly equal to each other and that they lie within the range of variations of the observed ratios in (3.20). Now, Y^*/A^* is nothing but a weighted average of the unobserved ratios (the weights being the sizes of the corresponding units). It is then natural for the surveyor to estimate Y^*/A^* by some sort of an average of the observed ratios. For instance, he may choose to estimate Y^*/A^* by $(y_1+\ldots+y_n)/(a_1+\ldots+a_n)$. This would lead to the following modification of the Desraj estimate (3.17):

$$D_1=S+\frac{y_1+\ldots.+y_n}{a_1+\ldots.+a_n}A^* \tag{3.21}$$

$$=\frac{y_1+\ldots.+y_n}{a_1+\ldots.+a_n}A$$

(and this we recognize at once as the familiar ratio estimate). Alternatively, the surveyor may choose to estimate the ratio Y^*/A^* by the simple average

$$\frac{1}{n}\left(\frac{y_1}{a_1}+\ldots.+\frac{y_n}{a_n}\right)$$

of the observed ratios. This will lead to another variation of the Desraj estimate, namely

$$D_2=(y_1+\ldots.+y_n)+\frac{1}{n}\left(\frac{y_1}{a_1}+\ldots.+\frac{y_n}{a_n}\right)A^*. \tag{3.22}$$

What we are trying to say here is the simple fact that both (3.21) and (3.22) have much greater face validity as estimates of Y than the Desraj estimate (3.17). In the Desraj estimate we are trying to evaluate Y^*/A^* by the n^{th} observed ratio $y_n \mid a_n$ and are taking no account of the other $n-1$ ratios. This is almost as preposterous as the estimate suggested by the circus statistician in the previous example. Suppose the surveyor finds that the n observed ratios $y_i \mid a_i$ $(i = 1, 2, \ldots, n)$ are nearly equal alright, but $y_n \mid a_n$ is the largest of them all. In this situation how can he have any faith in the Desraj estimate

$$D = S + \frac{y_n}{a_n} A^*$$

being nearly equal to Y? [Remember, the factor A^* will usually be a very large number.] Again, what does the surveyor do when he discovers that his initial supposition that the ratios Y_j/A_j $(j = 1, 2, \ldots, N)$ are nearly equal, was way off the mark? Will it not be ridiculous to use the Desraj estimate in this case? Here we are concerned not with the mathematical property of unbiasedness of an estimator but with the hard-to-define property of face validity of an estimate. An estimate T of the population total Y has little face validity if after we have written T in the form

$$T = S + S^*$$

we are hard put to find a reason why the part S^* should be a good estimate of Y^*.

4
The Label-Set and The Sample Core

We have noted elsewhere that, for a non-sequential sampling plan $\mathcal{P}$, the label part u of the data $x = (u, y)$ is an ancillary statistic; that is, the sampling distribution of the statistic u does not involve the state of nature θ. The sampling distribution of u is uniquely determined by the plan. It is therefore obvious that the label part of the data cannot, by itself, provide any information about θ. Knowing u, we only know the names (labels) of the population units that are selected for observation. [When u is a sequence, we also know the order and the frequency of appearance of each selected unit in u.] With a non-sequential plan $\mathcal{P}$, the knowledge of u alone cannot make the surveyor any wiser about θ. The surveyor may, and often does, incorporate his prior knowledge of the auxiliary characters $\alpha = (A_1, A_2, \ldots, A_N)$ in the plan $\mathcal{P}$. But this does not alter the situation a bit. The label part u of the data x will still be an ancillary statistic.

If the label part u is informationless, then can it be true that the observation part y of the data $x = (u, y)$ contains all the available information about θ? A little reflection will make it abundantly clear that the answer must be an emphatic, no. A great deal of information will be lost if the label part of the data is suppressed. Without the knowledge of u, the surveyor cannot relate the components of the observation vector y to the population units and so

he cannot make any use of the auxiliary characters $\alpha = (A_1, A_2, \ldots, A_N)$ and whatever other prior knowledge he may have about the relationship between θ and α.

Let us call a statistic $T = T(u, y)$ *label-free* if T is a function of y alone. So far, the only label-free estimator that we have come across is the estimator y of $\bar{Y}$ in Example 1. If in this case the surveyor has prior knowledge that the true value of θ lies in the vicinity of the point $\theta_0 = (a_1, a_2, \ldots, a_N)$, then he would naturally prefer the estimator (3.7) as an unbiased estimator of $\bar{Y}$. The surveyor can arrive at an estimate like (3.7) only if he has access to the information contained in u. In survey literature, we find several attempts at justifying label-free estimates. But a reasonable case for a label-free estimate can be made only under the assumption of a near complete ignorance in the mind of the surveyor. But, in these days of extreme specialization, who is going to entrust an expensive survey operation in the hands of a very ignorant surveyor?! To remain in survey business, the surveyor has to carefully orient himself to each particular survey situation, gather a lot of auxiliary data $A_1, A_2, \ldots, A_N$ about the population units, and then make intelligent use of such data in the planning of the survey and in the analysis of the survey data. Considerations of label-free estimates are, therefore, of only an academic interest in survey theory.

Let us denote by $\hat{u}$ the set of distinct population units that are selected (for survey) by the sampling plan. The set $\hat{u}$ is a statistic—a characteristic of the sample $x = (u, y)$. We call $\hat{u}$ the label-set and find it convenient to think of $\hat{u}$ as a vector

$$\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_\nu),$$

where $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_\nu$ are the ν distinct unit-labels that appear in u, arranged in an increasing order of their label values. The *observation-vector* $\hat{y}$ is then defined as

$$\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\nu),$$

where $\hat{y}_i = Y_{\hat{u}_i}$ $(i = 1, 2, \ldots, \nu)$.

We denote the pair $(\hat{u}, \hat{y})$ by $\hat{x}$ and call it the *sample-core*. For each sample $x = (u, y)$ we have a well-defined sample-core $\hat{x} = (\hat{u}, \hat{y})$. In the literature the sample-core has been called by other fancy names like *order statistic* or *sampley*, etc. [It should be noted that, though we can think of $\hat{u}$ as a subset of $\mathcal{P}$, we cannot think of $\hat{y}$ as a set, because the values in $\hat{y}$ need not be all different. Even if it were possible to think of $\hat{y}$ as a set, it would not be fruitful to do so. For, if $\hat{u}$ and $\hat{y}$ are both conceived as sets, then we have no way to relate a member of $\hat{y}$ to the corresponding label in $\hat{u}$. This is the reason why we prefer to think of the label-set $\hat{u}$ as a vector and of $\hat{y}$ as the corresponding observation-vector.]

The sample core $\hat{x}$ is a statistic. The mapping $x \to \hat{x}$ is usually many-one. For instance, in the pps plan of Example 4, the number ν (of distinct units selected) is the same as n, but, for each value of the label-set $\hat{u}$, there are exactly $n!$ values of u (corresponding to the $n!$ different selection-orders in which the n units might have been selected). Here the mapping $x \to \hat{x}$ is $n!$ to 1.

In part two of this essay we shall establish the fact that the sample-core $\hat{x}$ is a sufficient statistic. This means that, given $\hat{x}$, the conditional distribution of the sample x is uniquely determined (does not involve the unobserved part of the state of nature θ). The widely accepted principle of sufficiency tells us that if T be a sufficient statistic then every reasonable estimator (of every parameter τ) ought to be a function of T. Following Fisher we may call an estimator H *insufficient* if H is not a function of the sufficient statistic T. The Desraj estimator (3.17) of the population total Y is then an insufficient estimate. If we rewrite the Desraj estimate as

$$D = S + \frac{y_n}{a_n} A^*$$

where S is the sample Y- total and A^* is the total A-values of the unobserved units, then it is clear that both S and A^* are functions of the sample-core $\hat{x} = (\hat{u}, \hat{y})$. [Indeed, S is a function of $\hat{y}$ and A^* is a function of $\hat{u}$.] However, $y_n|a_n$ (the ratio corresponding to the last unit drawn in the without replacement pps plan) is not a function of $\hat{x}$. Knowing $\hat{x}$, we only know that the ratio $y_n|a_n$ may have been any one of the n ratios

$$\hat{y}_i|\hat{a}_i \quad (i = 1, 2, \ldots, n)$$

where $\hat{y}_i = Y_{\hat{u}_i}$ and $a_i = A_{\hat{u}_i}$.

If we define λ_i $(i = 1, 2, \ldots, n)$ as the conditional probability of $u_n|a_n$ being equal to $\hat{y}_i|\hat{a}_i$, then the λ_i's are well-defined (θ-free) constants, $\Sigma\lambda_i = 1$ and

$$\bar{D} = E(D \mid \hat{x}) = S + \left[\sum_1^n \lambda_i \frac{\hat{y}_i}{\hat{a}_i}\right] A^*. \tag{4.1}$$

Since D is an unbiased estimator of $Y = \Sigma Y_j$, so also is the estimator $\bar{D}$. From the Rao-Blackwell theorem it follows that (if $n > 1$) the variance of $\bar{D}$ is uniformly smaller than that of D. The estimator $\bar{D}$ has been variously called in the literature, the *symmetrized* or the *un-ordered* Desraj estimator. In view of what we explained in the previous section, the symmetrized Desraj estimator (4.1) looks much better than the original Desraj estimator on the score of face-validity. However, the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$ in (4.1) are much too complicated to make $\bar{D}$ an acceptable estimator of Y. The estimates (3.21) and (3.22) have about the same face-validity as that of (4.1) and are much simpler to compute. However, (4.1) scores over the other two estimates on the dubious criterion of unbiasedness!

The estimator (4.1) cannot be the only unbiased estimator of Y that is a function of the sample core $\hat{x}$. Consider the estimator

$$\frac{y_1}{p_1} = \frac{y_1}{a_1} A\,, \tag{4.2}$$

where y_1 is the Y-value of the first unit that was drawn (by the pps plan of Example 4) and a_1 is the corresponding A-value. Clearly, (4.2) is an *insufficient* unbiased estimator of Y. The symmetrized version of (4.2) will be

$$\left(\Sigma\mu_i \frac{\hat{y}_i}{\hat{a}_i}\right) A, \tag{4.3}$$

where

$$\mu_i = P\left(\frac{y_1}{a_1} = \frac{\hat{y}_i}{\hat{a}_i} \,\middle|\, \hat{x}\right) \quad (i = 1, 2, \ldots, n).$$

The estimator (4.3) is unbiased and is a function of $\hat{x}$.

Of late, quite a few papers have been written in which the main idea is the above described method of *un-ordering an ordered estimate*, that is, making use of the Rao-Blackwell theorem and the sufficiency of the sample core $\hat{x}$. Whatever the sampling plan $\mathcal{J}$ is, the sample-core is always sufficient. Indeed, the sample-core is (in general) the minimum (minimal) sufficient statistic. However, for a non-sequential sampling plan $\mathcal{J}$, the sufficient statistic $\hat{x}$ is never *complete*. By the *incompleteness* of $\hat{x}$ we mean the existence of non-trivial functions of $\hat{x}$ whose expectations are identically zero for all possible values of the state of nature θ. This is because (when the plan is non-sequential) the label-set $\hat{u}$ (which is a component of $\hat{x}$) is an ancillary statistic. For every parameter of interest $\tau(\theta)$, there will exist an infinity of unbiased estimators each of which is sufficient in the sense of Fisher (that is, is a function of the minimal sufficient statistic $\hat{x}$.)

5
Linear Estimation in Survey Sampling

During the past several years a great many research papers have been written dealing exclusively with the topic of linear estimation of the population mean $\overline{Y}$ or, equivalently, the population total Y. Some confusion has, however, been created by the term *linear*. An estimator is a function on the sample space X. Unless X is a linear space we cannot, therefore, talk of a linear estimator. In our formulation, X is the space of all samples $x = (u, y)$ and so X is not a linear space. How then are we to reconcile ourselves to the classical statement that, in the case of a simple random sampling plan, the sample mean is the best unbiased linear estimate of the population mean? We have the often quoted contrary assertion from Godambe that in no realistic sampling situation (whatever the plan $\mathcal{J}$) can there exist a best estimator in the class of linear unbiased estimators of the population mean. This section is devoted entirely to the notions of the so-called linear estimates.

Consider first the case of a simple random sampling plan in which a number n (the sample size) is chosen in advance and then a subset of n units is selected from the population $\mathcal{P}$ in such a manner that all the $\binom{N}{n}$ subsets of $\mathcal{P}$ with n elements are alloted equal selection probabilities. Let us suppose that the plan calls for a selection of the n sample units one by one without replacements and with equal probabilities, so that we can list the selected units in their natural selection order as $u_1, u_2, \ldots, u_n$. The label part of the data is then the sequence $u = (u_1, u_2, \ldots, u_n)$ and the observation part is the corresponding observation vector $y = (y_1, y_2, \ldots, y_n)$. Clearly, the y_i's are identically distributed (though not mutually independent) random variables with

$$E(y_i) = (\Sigma Y_j)/N = \overline{Y} \quad (i = 1, 2, \ldots, n).$$

Now, if the surveyor chooses to ignore the label part u of the data, then he can define a linear estimator of $\bar{Y}$ as a linear function

$$T = b_0 + b_1 y_1 + b_2 y_2 + \ldots + b_n y_n \tag{5.1}$$

of the observation vector y, where the coefficients $b_0, b_1, \ldots, b_n$ are preselected constants. All estimators of the above kind are label-free estimators. Let L be the class of all unbiased estimators of $\bar{Y}$ that are of the type (5.1). In other words, L is the class of all estimators of the type (5.1) with

$$b_0 = 0 \text{ and } b_1 + \ldots + b_n = 1 . \tag{5.2}$$

The sample mean $\bar{y} = (\Sigma y_i)/n$ is a member of L. The classical assertion that we referred to before is to the effect that, in the class L, there exists a uniformly minimum variance estimator and that is the sample mean $\bar{y}$. This result is well-known and a fairly straightforward proof may be given for the particular case where we define variance as the mean square deviation from the mean. We, however, consider it appropriate to sketch a proof that ties in well with the general spirit of this article.

Consider the sample core $\hat{x} = (\hat{u}, \hat{y})$ where we write the label-set $\hat{u}$ as a sequence $(\hat{u}_1, \ldots, \hat{u}_n)$ with $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_n$ and look upon $\hat{y}$ as the corresponding observation vector $(\hat{y}_1, \ldots, \hat{y}_n)$. Note that the mapping $x \to \hat{x}$ is $n!$ to 1 and that the vector $\hat{y}$ is obtained from the vector y by rearranging its coordinates in an increasing order of their corresponding unit labels. Now, given $\hat{x}$, the conditional distribution of x is equally distributed over the $n!$ possible values of x and so it follows that

$$E(y_i | \hat{x}) = (\Sigma \hat{y}_i)/n = \bar{y} \quad (i = 1, 2, \ldots, n) . \tag{5.3}$$

Thus, if $T = \Sigma a_i y_i$, with $\Sigma a_i = 1$, is any member of L then from (5.3) it follows that

$$E(T | \hat{x}) = \Sigma (a_i \bar{y}) = \bar{y}. \tag{5.4}$$

And so from the Rao-Blackwell theorem it follows that $\bar{y}$ is better than T [and this is irrespective of the loss function $w(t, \theta)$ (see §3) as long as $w(t, \theta)$ is convex (from below) in t for each fixed value of θ]. Observe that, in the class L, the sample mean

$$\bar{y} = (\Sigma y_i)/n = (\Sigma \hat{y}_i)/n$$

is the only one that is a function of the sample core $\hat{x}$. Every other member of L is insufficient in the sense explained in the earlier section. And so it is no wonder that $\bar{y}$ beats every other member of L in its performance characteristics. The class $L - \{\bar{y}\}$ is certainly not worth any consideration at all.

At this stage one may ask: Why not consider the class of all linear functions of the vector $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$? The snag is that the variables $\hat{y}_1, \ldots, \hat{y}_n$ have very complicated distributions and their expectations are not easy to obtain. For instance, the variable $\hat{y}_1$ can take only the values $Y_1, Y_2, \ldots, Y_{N-n+1}$ and its expectation is a complicated linear function of these $N - n + 1$

values. It is, therefore, not easy to characterize the class of unbiased estimators of $\overline{Y}$ that are linear functions of the vector $\hat{y}$. In any case, our representation of $\hat{y}$ as a vector is a rather artificial one and it is difficult to see why we should consider linear functions of the vector $\hat{y}$.

Let us look at the problem from another angle. True, the sample space X is not linear, but the parameter space Ω of all the possible values of the state of nature $\theta=(Y_1, \ldots, Y_N)$ is a part of the N-dimensional linear space R_N. A linear function on Ω is a function of the type

$$B_0+B_1 Y_1+\ldots+B_N Y_N \tag{5.5}$$

where $B_0, B_1, \ldots, B_N$ are constants. But (5.5) is a linear function of the parameter θ and cannot be conceived of as a statistic. Consider, however, a modification of (5.5) where we replace the coefficient B_j by the variable $B_j I_j$ where I_j is the indicator of the event E_j that the unit j is selected by the sampling plan $\mathcal{S}(j=1,2,\ldots,N)$. For each set of coefficients $B_0, B_1, \ldots, B_N$ we then have a sort of a linear function [see formula (3.9)]

$$T=B_0+\sum_1^N B_j I_j Y_j \tag{5.6}$$

on Ω, where the coefficients $B_j i_j (j=1,2,\ldots,N)$ are random variables. The indicator I_j is a function of the label-set $\hat{u}$ [$I_j(\hat{u})=1$ or 0 according as j is a member of $\hat{u}$ or not]. It is easy to recognize T as a statistic—indeed as a function of the sample core $\hat{x}=(\hat{u}, \hat{y})$. Only observe that we may rewrite (5.6) as

$$T=B_0+\sum_1^n b_i \hat{y}_i \tag{5.7}$$

where $b_i=B_{\hat{u}_i}$ $(i=1, 2, \ldots, n)$. Let us repeat once again that T is not a linear function on the sample space X, but that we may stretch our imagination a little bit to conceive of T as a random linear function on Ω with coefficients that are determined by the label-set $\hat{u}$. If T is defined as in (5.6) then

$$E(T)=B_0+\Sigma B_j \Pi_j Y_j \tag{5.8}$$

where $\Pi_j=E(I_j)=P(E_j)$. And so T is an unbiased estimator of $\overline{Y}$ if and only if [we are assuming that each $\Pi_j>0$ and that Ω does not lie in a subspace (of R_N) of dimension lower than N]

$$B_0=0 \text{ and } B_j=(N\Pi_j)^{-1}(j=1,2,\ldots,N). \tag{5.9}$$

If we define a linear estimator as in (5.6), then it follows that the Horvitz-Thompson estimator [see (3.8) and (3.9)] is the only unbiased linear estimator of $\overline{Y}$. Following Godambe we, therefore, take one step further and define the class of linear estimators in the following manner:

Definition. Let $\beta_0, \beta_1, \ldots, \beta_N$ be well-defined functions of the label-set $\hat{u}$. By a generalized linear estimator T we mean a statistic that may be represented as

$$T=\beta_0+\sum_1^N \beta_j I_j Y_j. \tag{5.10}$$

Note that the β_j's and I_j's are functions of $\hat{u}$ and that it is only the observed Y_j's that really enter into the definition of T. We may rewrite T in the alternative form

$$T=\beta_0+\sum_1^n \beta_{\hat{u}_i}\hat{y}_i \tag{5.11}$$

and thus recognize it as a function of the sample core $\hat{x}$.

The generalized linear estimator T [as defined in (5.10)] is an unbiased estimator of $\bar{Y}$ if and only if

$$E(\beta_0)=0 \text{ and } E(\beta_j I_j)=N^{-1} \text{ for all } j. \tag{5.12}$$

Let us denote by $\mathcal{L}'$ the class of generalized linear unbiased estimators of $\bar{Y}$. If each $\Pi_j>0$, then $\mathcal{L}'$ is never vacuous, for we have already recognized the Horvitz-Thompson estimator

$$H=\Sigma(N\Pi_j)^{-1}I_jY_j \tag{5.13}$$

as a member of $\mathcal{L}'$. If $\theta_0=(a_1, a_2, \ldots, a_N)$ be a fixed point in Ω and we define H_0 as

$$H_0=\Sigma(N\Pi_j)^{-1}I_j(Y_j-a_j)+(\Sigma a_j)/N \tag{5.14}$$

then H_0 is a member of $\mathcal{L}'$ and has zero risk (variance) when $\theta=\theta_0$. It follows that in the class $\mathcal{L}'$ of generalized linear unbiased estimators of $\bar{Y}$ there cannot exist a best (uniformly minimum risk) estimator. This then is the celebrated Godambe assertion that we referred to in the opening paragraph of this section.

If we go back to the case of simple random sampling and compare the two classes L and $\mathcal{L}'$ [defined in (5.2) and (5.12) respectively] then we shall observe that the two classes have precisely one member in common, namely the sample mean

$$\bar{y}=\left(\sum_1^n y_i\right)/n=\sum_1^N (n^{-1}I_jY_j).$$

The Godambe class $\mathcal{L}'$ of generalized linear estimators is not an extension of the class L. The two classes L and $\mathcal{L}'$ are essentially different in character and scope. Thus, the classical assertion that the sample mean is the best linear unbiased estimate of the population mean and Godambe's denial that no such best linear unbiased estimate can ever exist are both true (each rather trivially) in their separate contexts.

Following Hanurav, we may extend the Godambe class of linear estimators by defining a linear estimate as

$$T^*=\beta_0^*+\Sigma\beta_j^* I_jY_j \tag{5.15}$$

where $\beta_0^*, \beta_1^*, \ldots, \beta_N^*$ are well-defined functions of u—the label part of the data $x=(u, y)$. The only difference between (5.10) and (5.15) is that in the former the β's are functions of $\hat{u}$, whereas in the latter, the β^*'s are functions of u. Once we remember that the I_j's are functions of the label-set $\hat{u}$, it

follows at once that

$$E(T^* \mid \hat{x}) = \beta_0 + \Sigma \beta_j I_j Y_j \tag{5.16}$$

where

$$\beta_j = E(\beta_j^* \mid \hat{x}) = E(\beta_j^* \mid \hat{u})$$

is a function of the label-set $\hat{u}$ $(j = 0, 1, 2, \ldots, N)$. Thus, the conditional expectation of each T^*, given the sufficient statistic (sample-core) $\hat{x}$, is a T as defined in (5.10). From the Rao-Blackwell theorem it then follows that for each estimator of type (5.15) we can find an estimator of type (5.10) with a performance characteristic that is at least as good as (uniformly) that of the former. From the decision theoretic point of view the extension of the class (5.10) by the class (5.15) is, therefore, sort of vacuous.

6
Homogeneity, Necessary Bestness and Hyper-Admissibility

During the past few years, altogether much too much has been written on the subject of linear estimates of the population total Y. The original sin was that of Horvitz and Thompson who in 1952 sought to give a classification of linear estimates of Y. The tremendous paper-writing pressure of the past decade has taken care of the rest. For a plan $\mathcal{S}$ that requires that the n sample units be drawn one at a time, without replacements, and with equal or unequal probabilities, Horvitz and Thompson called an estimator T to be of T_1-type if T be of the form (5.1) with $b_0 = 0$, where $b_1, b_2, \ldots, b_n$ are pre-fixed constants and $y_1, y_2, \ldots, y_n$ are the n observed Y-values in their natural selection order. An estimator of the type (5.6) with $B_0 = 0$ (definable for an arbitrary plan $\mathcal{S}$) was classified as a T_2-type estimator. By a T_3-type estimator, Horvitz and Thompson meant an estimator $T = \beta S$, where S is the sample total and $\beta = \beta(\hat{u})$ is an arbitrary function of the label-set $\hat{u}$. That is, a T_3-type estimator is of the form (5.10) with $\beta_0 = 0$ and $\beta_1 = \beta_2 = \ldots = \beta_N$. Prabhu Ajgaonkar (1965) combined the features of the T_2 and T_3 type estimators to define his T_5-type (someone else must have defined the T_4-type!) estimators as estimators of the type (5.10) with

$$\beta_0 = 0 \text{ and } \beta_j = \beta B_j \quad (j = 1, 2, \ldots, N) \tag{6.1}$$

where β is a function of $\hat{u}$ and $B_1, B_2, \ldots, B_N$ are pre-fixed constants. With the exception of the T_1-type estimators, all the other types are subclasses of the Godambe class of linear homogeneous estimators, that is, estimators of the type (5.10) with $\beta_0 = \beta_0(\hat{u}) \equiv 0$ for all values of $\hat{u}$. Let us denote the Godambe class of linear homogeneous unbiased estimators of Y by $\mathcal{L}_0$. The rest of this section is devoted to a study of the class $\mathcal{L}_0$.

The class $\mathcal{L}_0$ is the class of all estimators of the form

$$T = \Sigma \beta_j I_j Y_j, \tag{6.2}$$

where β_j is a function of the label-set $\hat{u}$, I_j is the indicator of the event $j \in \hat{u}$ and

$$E(\beta_j I_j) = 1 \quad (j = 1, 2, \ldots, N). \tag{6.3}$$

Let us count the degrees of freedom that we have in setting up an estimator in $\mathscr{L}_0$. Let U be the set of all the possible values (given a plan $\mathscr{S}$) of the label-set $\hat{u}$ and let U_j be the subset of those $\hat{u}$'s that include the unit j. [The event E_j that $j \in \hat{u}$ is then the same as $\hat{u} \in U_j$.] Let m_j be the number of members in the set U_j. We are assuming that no U_j is vacuous; that is, no E_j is an impossible event; that is, $\Pi_j = P(E_j) > 0$ for all j. Thus,

$$m = \Sigma m_j \geq N. \tag{6.4}$$

For defining a T in $\mathscr{L}_0$ we need to define the N functions $\beta_1, \beta_2, \ldots, \beta_N$ on U. Since $I_j = I_j(\hat{u}) = 0$ for all $\hat{u} \notin U_j$, it is clear that we really need to define β_j on the set U_j only ($j = 1, 2, \ldots, N$). [The values of β_j outside the set U_j have no bearing on the statistic T as defined in (6.2).] Thus, we can think of each β_j as an m_j-dimensional vector. Now (6.3) is, in reality, a linear restriction on the m_j-dimensional vector β_j. We, therefore, have $m_j - 1$ degrees of freedom in our choice of the function (vector) β_j and so we have in all

$$\Sigma(m_j - 1) = m - N \tag{6.5}$$

degrees of freedom in our selection of a T in $\mathscr{L}_0$. We may visualize $\mathscr{L}_0$ as an $m - N$ dimensional surface (plane) in the m-dimensional Euclidean space R_m.

Let us stop for a moment to consider the extreme (and rather trivial) situation where $m = N$, that is, $m_j = 1$ for all j. This is the case of a unicluster (the terminology is Hanurav's) sampling plan, that is, a plan $\mathscr{S}$ that partitions the population $\mathscr{P}$ into a number of mutually exclusive and collectively exhaustive parts and then selects just one of these parts as the label-set $\hat{u}$. In this case we have no degree of freedom in the selection of a T; that is, the class $\mathscr{L}_0$ is a one point set consisting only of the Horvitz-Thompson estimator

$$T_0 = \Sigma \Pi_j^{-1} I_j Y_j. \tag{6.6}$$

Let us return to the non-trivial case where $m > N$. As we remarked before, a member T in $\mathscr{L}_0$ is then determined by our choice of $(\beta_1, \beta_2, \ldots, \beta_N)$ which we may look upon as an m-dimensional vector lying in an $m - N$ dimensional plane. The problem is to choose a T in $\mathscr{L}_0$ that has minimum variance. Now, if T be as in (6.2) then

$$V(T) = \sum_j V(\beta_j I_j) Y_j^2 + 2 \sum_{j<k} \mathrm{Cov}(\beta_j I_j, \beta_k I_k) Y_j Y_k \tag{6.7}$$

which depends on the state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$. For each $\theta \in \Omega$, it is then clear that $V(T)$ is a (positive semi-definite) quadratic form in the m-dimensional vector $(\beta_1, \beta_2, \ldots, \beta_N)$. For each θ in Ω, there clearly exists a choice of the vector $(\beta_1, \beta_2, \ldots, \beta_N)$ that minimizes (6.7). Except in some very special situations (with Ω a very small set), there cannot exist a choice of $(\beta_1, \beta_2, \ldots, \beta_N)$ that will minimize (6.7) uniformly for all $\theta \in \Omega$. In the class $\mathscr{L}_0$ of all linear homogeneous unbiased estimators of Y there does not exist a uniformly minimum variance unbiased estimator (Godambe, 1955).

So the search was on for some other performance criterion that would uphold some estimator as the best in the class $\mathscr{L}_0$ (or in some other smaller or

larger class). Of late two rather curious such criteria have been proposed for consideration. They are (a) Ajgaonkar's criterion of *necessary bestness* and (b) Hanurav's criterion of *hyper-admissibility*. Let us first consider necessary bestness, the curiouser of the two criteria.

"In order to choose a serviceable estimator from the practical point," writes Ajgaonkar (1965, p.638), "we propose the following criterion of the necessary best estimator."

Definition (Ajgaonkar). Between two unbiased estimators T and T' (of the population total Y) with variances

$$V(T)=\Sigma a_j Y_j^2+2\sum_{j<k} a_{jk}Y_jY_k$$

and

$$V(T')=\Sigma b_j Y_j^2+2\sum_{j<k} b_{jk}Y_jY_k$$

the estimator T is *necessary better* than T' if $a_j \leq b_j$ for all j. The estimator T (in the class C) is *necessary best in* C if it is necessary better than every other estimator in C.

From (6.7) and the above definition it then follows that the estimator $T=\Sigma\beta_j I_j Y_j$ is necessary best in the class $\mathcal{L}_0$ if and only if $V(\beta_j I_j)$ is uniformly minimum for all j. From the Schwarz inequality we have

$$V(\beta_j I_j)\,V(I_j) \geq [\mathrm{Cov}(\beta_j I_j, I_j)]^2=[E(\beta_j I_j^2)-E(\beta_j I_j)\,E(I_j)]^2. \qquad (6.8)$$

Since $I_j^2 = I_j$, $E(I_j)=\Pi_j$, $V(I_j)=\Pi_j(1-\Pi_j)$ and $E(\beta_j I_j)=1$ for all j, we at once have

$$V(\beta_j I_j) \geq (1-\Pi_j)/\Pi_j \quad (j=1,2,\ldots,N). \qquad (6.9)$$

The sign of equality holds for all j in (6.9) if we select

$$\beta_j=\Pi_j^{-1} \quad (j=1,2,\ldots,N),$$

that is, if T is the Horvitz-Thompson estimator. Thus, in $\mathcal{L}_0$ there exists a unique necessary best estimator and that is the Horvitz-Thompson estimator (5.22). [Ajgaonkar (1965) gave a very complicated looking proof of the necessary bestness of (6.6) in the subclass of T_5-type estimators as defined in (6.1), and for a particular class of sampling plans. The present proof is a simplification of a proof suggested by Hege (1967).]

But why necessary bestness? It is hard to figure out how Ajgaonkar stumbled across this curious name and definition. Let us hazard a guess. We begin with a most unrealistic assumption that the space Ω contains points of the type

$$(0,\ldots,0,Y_j,0,\ldots,0),$$

that is, vectors with only one non-zero coordinate Y_j ($j=1,2,\ldots,N$), and let

Ω_0 be the subset of all points of the above kind. For a typical $\theta \in \Omega_0$ the variance of $T = \Sigma \beta_j I_j Y_j$ is equal to

$$V(\beta_j I_j) Y_j^2 \text{ (for some } j \text{ and } Y_j).$$

Hence, if we restrict our attention to the subset Ω_0 of Ω, the necessary best estimator in $\mathcal{L}_0$ is also the uniformly minimum variance estimator. The Horvitz-Thompson estimator has uniformly minimum variance (in $\mathcal{L}_0$) over the subset Ω_0.

Let us now consider the hyper-admissibility thesis of Hanurav (1968). Hyper-admissibility as the name suggests, is a strengthening of the decision-theoretic notion of admissibility. In order not to draw the attention of the reader away from the present context, let us define admissibility in the narrow framework of unbiased point estimation (of the population total Y) with variance as the risk function. Let T_0 and T_1 be unbiased estimators of Y.

Definition. T_0 is *uniformly better* than T_1 if

$$V(T_0) \leq V(T_1) \quad \text{for all } \theta \in \Omega$$

with the strict sign of inequality holding for at least one $\theta \in \Omega$.

Let C be a class of unbiased estimators of Y. We tacitly assume that C is a convex class, that is, when T_0 and T_1 are both members of C then so also is $(T_0 + T_1)/2$. For instance, the class $\mathcal{L}_0$ is convex.

Definition. $T_0 \in C$ is *admissible* in C if there does not exist a $T_1 \in C$ that is uniformly better than T_0.

If T_0 is admissible in C, then for any alternative $T_1 \in C$ it must be true that T_1 is not uniformly better than T_0; that is, either

(a) $V(T_0) \equiv V(T_1)$ for all $\theta \in \Omega$, or

(b) $V(T_0) < V(T_1)$ for at least one $\theta \in \Omega$.

Now, in view of the admissibility of T_0 and the convexity of C, the alternative (a) is impossible. Suppose (a) holds. Consider the estimator

$$T_* = (T_0 + T_1)/2$$

and observe that

$$\begin{aligned} V(T_*) &= \tfrac{1}{4}\{V(T_0) + V(T_1)\} + \tfrac{1}{2}\rho \sqrt{V(T_0)V(T_1)} \\ &= V(T_0)(1+\rho)/2 \\ &\leq V(T_0), \end{aligned}$$

where ρ is the correlation coefficient between T_0 and T_1. Since T_0 is admissible, it follows that $V(T_*) \equiv V(T_0)$ for all θ; that is, $\rho \equiv 1$ for all θ. Therefore, $T_0 = a + bT_1$. Since, T_0 and T_1 are both unbiased estimators of Y, it follows that $a = 0$ and $b = 1$. This contradicts the initial supposition that T_0 and T_1 are different estimators.

Thus, in our present context, we may redefine admissibility as

Definition (Hanurav). $T_0 \in C$ is admissible in C if, for any other $T_1 \in C$, it is true that

$$V(T_0) < V(T_1)$$

for at least one value of θ, say θ_{01}, in Ω. [The point θ_{01} will usually depend on T_0 and T_1.}

It is clear that the admissibility of an estimator T_0 depends on two things, namely, (a) the extent of the class C that T_0 is referred to and (b) the extent of the space Ω in which θ is supposed to lie. The smaller the class C and the larger the space Ω, the easier it is to establish the admissibility of a T_0 in C. A little while ago we noted that, in the class $\mathcal{L}_0$, the Horvitz-Thompson estimator (6.6) is the only one that has uniformly minimum variance over the set Ω_0 of all points θ with only one non-zero coordinate. If we are allowed to make the unrealistic assumption that $\Omega \supset \Omega_0$, then the admissibility of (6.6) in $\mathcal{L}_0$ follows at once. Godambe and Joshi (1965) proved the admissibility of (6.6) in the wider class of all unbiased estimators of Y, under the very unrealistic assumption that $\Omega = R_N$. As we have noted earlier [see (3.10) and (3.11)], the Horvitz-Thompson estimator is no longer admissible (even in the small class $\mathcal{L}$ of all linear unbiased estimators of Y) if it is known that Ω is a small neighborhood of a point $\theta_0 = (a_1, a_2, \ldots, a_N)$.

Hanurav sought to strengthen the notion of admissibility as follows. Following Godambe, he made the unrealistic assumption that $\Omega = R_N$ and then defined a *principal hyper-surface (phs)* of Ω as a linear subspace of all points $\theta = (Y_1, Y_2, \ldots, Y_N)$ with

$$Y_{j_1} = Y_{j_2} = \ldots = Y_{j_k} = 0$$

where $0 \leq k < N$ and $(j_1, \ldots, j_k)$ is a subset of $(1, 2, \ldots, N)$. [The whole space Ω corresponds to the case $k = 0$. There are $2^N - 1$ phs's of Ω.] Let Ω^* be a typical phs in Ω. Let C be a class of unbiased estimators of Y.

Definition (Hanurav). $T_0 \in C$ is hyper-admissible in C if, for every phs $\Omega^* \in \Omega$, it is true that T_0 is admissible in C when we restrict θ to Ω^*.

It follows at once that the H-T estimator $T_0 = \Sigma \Pi_j^{-1} I_i Y_j$ is the unique hyper-admissible estimator in $\mathcal{L}_0$. Suppose $T = \Sigma \beta_j I_j Y_j$ is hyper-admissible in $\mathcal{L}_0$. Consider the phs Ω_j^* of all points θ with $Y_i = 0$ for all $i \neq j$. For a typical $\theta \in \Omega_j^*$

$$V(T) = V(\beta_j I_j) Y_j^2$$

and this [as we have noted in (6.9)] is greater than

$$V(T_0) = \Pi_j^{-1}(1-\Pi_j)\, Y_j^2$$

unless $\beta_j = \beta_j(\hat{u}) = \Pi_j^{-1}$. Thus, the admissibility of T in each phs Ω_j^*implies that $T = T_0$. That T_0 is hyper-admissible, that is, is admissible on each phs, is equally trivial. Let Ω^* be a typical phs and let $T^* = \Sigma\beta_j^* I_j Y_j$ be a member of $\mathscr{L}_0$ such that

$$V(T^*) \leq V(T_0) \quad \text{for all } \theta \in \Omega^*. \tag{6.10}$$

For each one-dimensional phs $\Omega_j^* \in \Omega^*$, we must have the sign of equality in (6.10) for all $\theta \in \Omega_j^*$, and so it follows that $\beta_j^* = \Pi_j^{-1}$ for each j such that $\Omega_j^* \in \Omega^*$. Therefore, the sign of equality holds in (6.10) for all $\theta \in \Omega^*$. In other words, it is impossible to find an estimator T^* in $\mathscr{L}_0$ that is uniformly better than T_0 in the phs Ω^*; that is, T_0 is admissible (in the class $\mathscr{L}_0$) when we restrict θ to Ω^*.

In the context of the class $\mathscr{L}_0$, the twin criteria of *necessary bestness* and *hyper-admissibility* are mathematically equivalent. Before we proceed to examine the logical basis of the criterion of hyper-admissibility, let us point out a curious error committed by Hanurav (1968, p. 626). In his relation (3.2) Hanurav mistakenly asserts that T_0 is hyper-admissible (in C) if and only if, for every alternative $T_1 \in C$ and every phs $\Omega^* \subset \Omega$, we can find a point $\theta_{01} \in \Omega^*$ such that

$$V(T_0|\theta = \theta_{01}) < V(T_1|\theta = \theta_{01}). \tag{6.11}$$

We give an example to contradict the above assertion. Consider T_0 and T_1 where T_0 is as in (6.6) and

$$T_1 = \beta_1 I_1 Y_1 + \sum_{j=2}^{N} \Pi_j^{-1} I_j Y_j \text{ with } E(\beta_1 I_1) = 1.$$

In the phs Ω^* of all θ's with $Y_1 = 0$, it is clear that

$$V(T_0) \equiv V(T_1).$$

So in Ω^* we cannot find a point θ_{01} satisfying (6.11) and this in spite of T_0 being hyper-admissible in $\mathscr{L}_0$ and T_1 being an alternative member of $\mathscr{L}_0$.

The main result of Hanurav is to the effect that, for any nonunicluster sampling plan $\mathscr{L}$, the H-T estimator (6.6) is the unique hyper-admissible estimator in the class $\mathscr{U}^*$ of all polynomial unbiased estimators of Y. A quadratic estimator of Y is a statistic T of the form

$$T = \beta_0 + \Sigma\beta_j I_j Y_j + \Sigma\beta_{jk} I_{jk} Y_j Y_k \tag{6.12}$$

where $I_{jk} = I_j I_k$ is the indicator of the event that both j and k are in the label set $\hat{u}$, and the β's are functions of $\hat{u}$ with the (unbiasedness) conditions

$$E(\beta_0) = 0, \quad E(\beta_j I_j) \equiv 1, \quad E(\beta_{jk} I_j I_k) \equiv 0 \quad \text{for all } j \text{ and } k.$$

A polynomial estimator is similarly defined.

Now, let us examine the logical content of the hyper-admissibility criterion. Let $\mathscr{P}^*$ be an arbitrary but fixed subset (subpopulation) of the population $\mathscr{P}$ and let Y^* be the total Y-value of the units in $\mathscr{P}^*$; that is,

$$Y^* = \sum_{j \in \mathscr{P}^*} Y_j. \tag{6.12}$$

Suppose, along with an estimate of Y, the surveyor also needs to estimate the parameter Y^*. Once the surveyor has decided upon an estimator $T = T(\hat{u}, \hat{y})$ for Y, he may choose to derive an estimate T^* for Y^* in the following manner. Recall that $\hat{y}$ is the vector $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\eta)$ of the (observed) Y-values of the μ distinct units $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_\eta$ in the label set $\hat{u}$. Define y^* as the vector

$$y^* = (y_1^*, y_2^*, \ldots, y_\eta^*),$$

where y_i^* is y_i or zero according as $\hat{u}_i$ is or is not a member of $\mathscr{P}^*$. In other words, we derive y^* by substituting by zeros those coordinates of the observation vector $\hat{y}$ that corresponds to units that are outside the subpopulation $\mathscr{P}^*$. Now define

$$T^* = T(\hat{u}, y^*). \tag{6.13}$$

If T is an unbiased estimator of Y, then it is almost a truism that T^* is an unbiased estimator of Y^*. If T is the linear homogenous estimator $\Sigma \beta_j I_j Y_j$, then T^* is the estimator $\Sigma^* \beta_j I_j Y_j$, where the summation Σ^* extends over all j that belong to $\mathscr{P}^*$. In particular, if $\mathscr{P}^*$ is the single member subpopulation consisting of the unit j alone, then the H-T estimator $T_0 = \Sigma \Pi_j^{-1} I_j Y_j$ gives rise to the estimator

$$T_0^* = \Pi_j^{-1} I_j Y_j = \begin{cases} \Pi_j^{-1} Y_j & \text{if } j \text{ is surveyed} \\ 0 & \text{otherwise} \end{cases} \tag{6.14}$$

for the parameter $Y^* = Y_j$.

The estimate (6.14) is similar to the one considered in example 3 of Section 3 and is, of course, utterly ridiculous. But in the makebelieve world of mathematicians, we are allowed to make any supposition. Let us pretend that when a surveyor estimates Y by T, he naturally commits himself to estimating each of the $2^N - 1$ subtotals Y^* by the corresponding derived estimate T^*. Given a class $C = \{T\}$ of estimators of Y, let us consider, for each subtotal Y^*, the class $C^* = \{T^*\}$ of derived estimators of Y^*. The estimator $T_0 \in C$ is hyper-admissible in C if, for each subtotal Y^*, the derived estimator T_0^* is admissible in C^*. According to Hanurav, if the sampling plan $\mathscr{S}$ is nonunicluster, then given any linear (or polynomial) unbiased estimator T that is different from the H-T estimator, he can always find another unbiased estimator T_1 and a subtotal Y^* such that the derived estimator T_1^* (for Y^*) is uniformly better than the derived estimator T^*.

7
Linear Invariance

We have idealized away many of the mathematically intractable features of the survey operation. But even with our oversimplified mathematical framework, the dimension N of the state of nature θ will usually run into several hundred thousands. It is clear that we are dealing with a most complex inference situation. A typical survey operation is an essentially non-repeatable, once in a lifetime affair. The surveyor, who is a specialist in the particular survey area, plans the survey, collects and analyzes the huge survey data and then arrives at his estimates of the various parameters of interest. Why does he need to consult a mathematician? How can the deductive processes of mathematics be of any use to the surveyor in his purely inductive inference making efforts? The author suspects that the answer lies in the general concensus among the scientific community that the mathematicians are the true watchdogs of rationalism. It may well be argued that this great reverence for mathematicians, this identification of rationalism with deduction, this over-eagerness to put every argument (be it in the realm of economics, psychology, survey theory, even philosophy) in the mold of pure deduction have done more harm than good to the general growth of knowledge. True, a good mathematician, having sharpened his mind with constant exercises in deductive reasoning, will often be able to comb out many a tangle created by unclear thinking on the part of the scientist. But new tangles are created by our over-eagerness to force a mathematical model for a situation that is essentially non-mathematical in nature. We close Part One of our essay with one more example of such a tangle in survey theory.

When the surveyor calls upon a decision-theorist (let us abbreviate the name to DT) to audit his survey work, the DT does not attempt to evaluate the thought process by which the surveyor arrived at his estimate T (for, say, the population total Y) from the data x. Indeed, the DT denies the very existence of a rational thought process that may lead us from the particular data x to the estimate T. [So far we have been freely using the two terms *estimate* and *estimator* and did not care to distinguish between them. But the whole controversy that is now raging in survey theory may be summarized as the difference between the estimate and the estimator. To the surveyor, the parameter Y is an unknown variable and the estimate T is a constant suggested by the data x at hand. The DT thinks of Y as an unknown constant and looks upon T as a random variable—a function on the sample space X.] As the DT cannot evaluate the estimate T, he proceeds to force an estimator out of the surveyor. For this he needs the sampling plan f to be randomized and, preferably, non-sequential. Once the DT has figured out the space X of all the possible data x (that the surveyor might have obtained from the survey), he would ask the surveyor to answer the impossible question of how he would have estimated Y for each x in X. If the function $T(x)$ is very complicated (as it would usually be) then that would be the end of the DT's audit. The estimator $T(x)$ better be simple enough so that the DT can evaluate the risk function—the average performance characteristics of the estimator. But before the risk function is evaluated, the DT would like to know the surveyor's loss function which again better be a simple one. As the DT

cannot answer the question: How good (rational) is the estimate T?, he evades the issue and proceeds to answer what he thinks to be a nearly equivalent question: How good is the average performance characteristic of the estimator T?

Instead of looking at the average performance characteristics of T, the DT may try to evaluate the estimator T by examining it directly as a function on X. A criterion that is frequently used for such direct evaluation of the estimator T is the criterion of linear invariance. The DT tries to find out if the surveyor's estimate T of the population total depends in some way on the scale in which the population values (the state of nature θ) are measured. With a linear shift (change of origin and scale) in the measurement scale for the population values, the state of nature $\theta=(Y_1, Y_2, \ldots, Y_N)$ will be shifted to $\theta' = (a+bY_1, a+bY_2, \ldots, a+bY_n)$ and the parameter Y will be shifted to $Y' = Na + bY$. With the same shift in the measurement scale the data

$$x = \{(u_1, u_2, \ldots, u_n), (y_1, y_2, \ldots, y_n)\}$$

will appear as

$$x' = \{(u_1, u_2, \ldots, u_n), (a+by_1, \ldots, a+by_n)\}. \tag{7.1}$$

Since x and x' represent the same data (in two different scales) it is natural to require that they lead to the same estimate (in the two scales) of the population total. This leads us to the following

Definition. The estimator $T = T(x)$ is origin and scale invariant if

$$T(x') \equiv Na + bT(x) \tag{7.2}$$

for all x, a, and $b > 0$, where x' is defined as in (7.1). We call T scale invariant if the above identity holds with $a = 0$.

One reason why there is so much interest (see Section 6) in linear homogeneous estimators is that they are supposed to be scale invariant. [As we shall presently point out, the above supposition is true only under some qualifications.] The Horvitz-Thompson estimator $T_0 = \Sigma\Pi_j^{-1} I_j Y_j$ is clearly scale invariant. It will be origin invariant only if

$$\Sigma\Pi_j^{-1} I_j \equiv N \tag{7.3}$$

for all samples. Since the expected value of the left hand side is clearly equal to N, we may restate the identity (7.3) as $V[\Sigma\Pi_j^{-1} I_j] = 0$ or equivalently

$$\begin{aligned} N^2 &= E[\Sigma\Pi_j^{-1} I_j]^2 \\ &= \Sigma\Pi_j^{-1} + \sum_{j \neq k} [\Pi_{jk}/(\Pi_j \Pi_k)] \end{aligned} \tag{7.4}$$

where Π_{jk} is the probability that both j and k are in the sample. In the case of simple random sampling with sample size n, it is clear that $\Pi_j = n/N$ for all j and so the H-T estimator reduces to the simple origin and scale invariant estimator

$$G = (N S)/n \tag{7.5}$$

where S is the sample total.

So far mathematicians have generally avoided the non-homogeneous linear estimators of the type

$$\beta_0 + \Sigma \beta_j I_j Y_j \tag{7.6}$$

in the mistaken belief that such estimators cannot possibly be scale invariant. It is tacitly assumed that any function $\beta = \beta(\hat{u})$ of the label-set $\hat{u}$ is necessarily *scale-free*; that is, the value of $\beta(\hat{u})$ depends only on $\hat{u}$ and not on the scale in which the population values are measured. That this need not be so is seen as follows. Suppose the surveyor defines β as

$$\beta(\hat{u}) = \Sigma I_j a_j \tag{7.7}$$

where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a pre-selected fixed point in the space Ω. The function β is clearly scale invariant. That is, if the surveyor is told that, in the new measurement scale, each of the population values is to be multiplied by the scaling factor b, then he (the surveyor) will automatically represent the point θ_0 as $(ba_1, ba_2, \ldots, ba_n)$ and re-compute $\beta(\hat{u})$ as $b\beta(\hat{u})$. Let us look back on the modified H-T estimator

$$H_0 = \Sigma \Pi_j^{-1} (Y_j - a_j) + \Sigma a_j \tag{7.8}$$

that we had considered earlier in (3.12), where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a pre-selected fixed point in Ω. A surveyor using (7.8) as his estimating formula for Y can never be accused of violating the canon of linear invariance. [We are not saying that H_0 is a respectable or a reasonable estimator of Y. We are only saying that, apart from being an unbiased estimator of Y with zero variance when $\theta = \theta_0$, the estimator H_0 is origin and scale invariant.] It has been repeatedly asserted by Godambe (see either of his 1968 papers) that, in the class of all estimators that are functions of the label-set $\hat{u}$ and the sample total S, the estimator $G = (NS)/n$ (where n is the number of units in $\hat{u}$) is the unique origin and scale invariant one. However, observe that if $\beta = \beta(\hat{u})$ is any scale-free function of $\hat{u}$ and

$$\beta_0(\hat{u}) = \Sigma a_j - \beta(\hat{u}) \Sigma I_j a_j ,$$

where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a fixed point in Ω, then the estimator

$$G_0 = \beta_0 + \beta S = \Sigma a_j + \beta \Sigma I_j (Y_j - a_j) \tag{7.9}$$

is an origin and scale invariant function of $\hat{u}$ and S.

References

1. Ajgaonkar, S.G. Prabhu, "On a Class of Linear Estimates in Sampling with Varying Probabilities without Replacements," *Journal of the American Statistical Association*, 60, 637-642, 1965.
2. Basu, D., "On Sampling With and Without Replacements," *Sankhyā*, 20, 287-294, 1958.

3. Basu, D., "Recovery of Ancillary Information," *Sankhyā*, 26, 3-16, 1964.
4. Basu, D. and Ghosh, J.K., "Sufficient Statistics in Sampling from a Finite Universe," *Proceedings of the 36th Session of Int. Stat. Inst.*, 850-859, 1967.
5. Basu, D., "Role of the Sufficiency and Likelihood Principles in Sample Survey Theory," *Sankhyā*, 31, 441-454, 1969.
6. Desraj, *Sampling Theory*, McGraw-Hill, 1968.
7. Godambe, V.P., "A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society, B*, 17, 269-278, 1955.
8. Godambe, V.P., "An Admissible Estimate for Any Sampling Design," *Sankhyā*, 22, 285-288, 1960.
9. Godambe, V.P. and Joshi, V.M., "Admissibility and Bayes Estimation in Sampling Finite Populations – Part I, II, and III," *Annals of Mathematical Statistics*, 36, 1707-1742, 1965.
10. Godambe, V.P., "Contributions to the United Theory of Sampling," *Rev. Int. Stat. Inst.*, 33, 242-258, 1965.
11. Godambe, V.P., "A New Approach to Sampling From a Finite Universe, Part I and II," *J. Roy. Statist. Soc.*, 28, 310-328, 1966.
12. Godambe, V.P., "Bayesian Sufficiency in Survey-Sampling," *Ann. Inst. Stat. Math.* (Japan), 20, 363-373, 1968.
13. Godambe, V.P., "Some Aspects of the Theoretical Developments in Survey Sampling," *New Developments in Survey Sampling*, Wiley-Interscience, 27-53, 1968-69.
14. Hájek, J., "Optimum Strategy and Other Problems in Probability Sampling," *Casopis Pest. Math.*, 84, 387-423, 1959.
15. Hanurav, T.V., "On Horvitz and Thompson Estimator," *Sankhyā, A*, 24, 429-436, 1962.
16. Hanurav, T.V., "Hyper-Admissibility and Optimum Estimators for Sampling Finite Populations," *Ann. Math. Statist.*, 39, 621-642, 1968.
17. Hege, V.S., "An Optimum Property of the Horvitz-Thompson Estimate," *Journal of the American Statistical Association*, 62, 1013-1017, 1967.
18. Horvitz, D.G. and Thompson, D.J., "A Generalization of Sampling Without Replacements from a Finite Universe," *J. Am. Stat. Ass.*, 47, 663-685, 1952.
19. Joshi, V.M., "Admissibility of the Sample Mean as Estimate of the Mean of a Finite Population," *Ann. Math. Statist.*, 39, 606-620, 1968.
20. Midzuno, H., "On the Sampling System with Probability Proportionate to Sum of Sizes," *Ann. Inst. Stat. Math.* (Japan), 3, 99-107, 1952.
21. Murthy, M.N., "On Ordered and Unordered Estimators," *Sankhyā, A*, 20, 254-262, 1958.
22. Pathak, P.N., "Sufficiency in Sampling Theory," *Ann. Math. Statist.*, 35, 795-808, 1964.
23. Raiffa, H. and Schlaifer, R.O., *Applied Statistical Decision Theory*, Boston Division of Research, Graduate School of Business Administration, Harvard University, 1961.
24. Roy, J. and Chakravarti, I.M., "Estimating the Mean of a Finite Population," *Ann. Math. Statist.*, 31, 392-398, 1960.
25. Zacks, S., "Bayes Sequential Designs for Sampling Finite Populations," *J. Am. Stat. Ass.*, 64, 1969.

[D. Basu presented a very brief summary of his paper and then went on to make some additional remarks. The following is an abstract of these additional remarks.

The sample core $\hat{x} = (\hat{u}, \hat{y})$ is always a sufficient statistic. In general, it is minimal sufficient. The sufficiency principle tells us to ignore as irrelevant all details of the sample $x = (u, y)$ that are not contained in the sample core $\hat{x}$. The likelihood principle tells us much more. The (normalized) likelihood function is the indicator of the set Ω_x of all parameter points θ that are consistent with the sample x. The set Ω_x depends on the sample x only through the sample-core $\hat{x}$. Given $\hat{x}$, the set Ω_x has nothing to do with the sampling plan $\mathcal{J}$. And this is true even for sequential sampling plans. For one who believes in the likelihood principle (as all Bayesians do) the sampling plan is no longer relevant at the data analysis stage.

A major part of the current survey sampling theory was dismissed by D. Basu as totally irrelevant. He stressed the need for science oriented, down to earth data analysis. The survey problem was posed as a problem of extrapolation from the observed part of the population to the unobserved part. An analogy was drawn between the problem of estimating the population total ΣY_j and the classical problem of numerical integration. In the latter the problem is to 'estimate' the value of the integral $\int_a^b Y(u)du$ by 'surveying' the function $Y(u)$ at a number of 'selected points' $u_1, u_2, \ldots, u_n$. Which points to select and how many of them, are problems of 'design'. Which integration formula to use and how to assess the 'error' of estimation, are problems of 'analysis'. True, it is possible to set up a statistical theory of numerical integration by forcing an element of randomness in the choice of the points. But, how many numerical analysts will be willing to go along with such a theory?

The mere artifact of randomization cannot generate any information that is not there already. However, in survey practice, situations will occasionally arise where it will be necessary to insist upon a random sample. But this will be only to safeguard against some unknown biases. In no situation, is it possible to make any sense of unequal probability sampling.

The inner consistency of the Bayesian point of view is granted. However, the analysis of the survey data need not be Bayesian. Indeed, who can be a true Bayesian and live with thousands of parameters? According to the author, survey statistics is more an art than a science.]

COMMENTS

G. A. Barnard:

First, a point of detail. Dr. Basu suggested, in the unwritten part II of his paper, a method of estimation using an assistant and dividing the data into two parts, D_1 and D_2. He said that no estimate of error would be available. I simply want to point out that an error estimate could be obtained, in an obvious way, if Dr. Basu has a twin, Basu[1], and his assistant also has a twin, assistant[1]. Then the data are divided into four sets, D_1, D_1', D_2, and D_2'.

Second, a general point. Dr. Basu and others here are concerned particularly with the problems which arise when it is necessary to make use of the additional information or prior knowledge α. In many sample survey situations

this prior knowledge of individuals is negligible (at least for a large part of the population under discussion) and in this case the classical procedures, in particular the Horvitz-Thompson estimators, apply in a sensible manner. It is important that we should not appear in this conference to be casting doubt on procedures which experience has shown to be highly effective in many practical situations.

The problem of combining external information with that from the sample is in general difficult to solve. For instance, in ordinary (distribution-free) least squares theory, the additional information that one or more of the unknown parameters has a bounded range makes the usual *justifications* inapplicable and no general theory appears to be possible, though it is easy to see what we should do in some particular cases.

With Dr. Basu's elephants, a realistic procedure (on the data he has given) would seem to be to think of the measurement to be made on one elephant as providing some estimate of how the animals have put on weight, or lost it, during the past three years. The circus owner should be able to give good advice on how elephants grow, but in the absence of this it would seem plausible to assume that the heaviest elephant, being fully mature, will have gained nothing, and that the percentage growth will be a linear function of weight three years ago. It would then be wise to select an elephant somewhat lighter than Sambo for weighing. The estimation procedure is clear.

Evidently, as Dr. Basu suggests, no purely mathematical theory is ever likely to be able to account for an estimation procedure such as that suggested. But I do not think this implies that all mathematical theories are time wasting in this context. A judicious balance is necessary. In particular, as I have said, we should not throw overboard the classical theory, or the work of Godambe, Horvitz, Thompson and others, just because we can envisage situations where these results would clearly not be applicable.

V. P. Godambe:

Professor Basu has given a very interesting presentation of some ideas in survey sampling theory which many of us have been contemplating for some time. I find it difficult however to agree with him in one respect. The likelihood principle, which does not permit the use of the sampling distribution generated by randomization for inference purposes, is unacceptable to me in relation to survey sampling. It seems as though the likelihood principle has different implications for two intrinsically similar situations: for the coin tossing experiment the likelihood principle allows the use of binomial distributions while infering about the binomial parameter but if the experiment is replaced by one of the drawing balls from a bag containing black and white balls, the likelihood principle does not allow the use of corresponding binomial (or hypergeometric if sampling is without replacement) distribution to infer the unknown proportion of white balls in the bag.

Professor Basu's comments on HT-estimator (example of weighing elephant and so on) are humourous and I wonder if he wants us to take them at all seriously. The comments fail to take into account the fact that the inclusion probabilities involved in HT-estimator are inseparably tied to the prior knowledge represented or approximated by a *class* of (indeed a very very wide one)

prior distributions (Godambe, *J. Roy. Statist. Soc.*, 1955) on the parametric space. I believe the only way of making sense of sampling practice and theory is through studying the frequency properties implied by the distributions generated by different modes of randomization (that is, different sampling designs) of the estimators obtained on the basis of the considerations of prior knowledge; of course one should also study the implications of reversing the role of *frequency properties* and *prior knowledge* (reference: section 7, Godambe's and Thompson's Symposium paper).

At the end of his paper Professor Basu comments on "origin and scale invariant estimator" in my paper, "Bayesian Sufficiency in Survey-Sampling", *Ann. Inst. Stat. Math.*, 1968. My assertion about the uniqueness in the paper is certainly true. Basu's comments suggest a different type of invariance which is already discussed in our (Godambe and Thompson) symposium paper (section 3).

J. Hájek:

Professor Basu and myself both like the likelihood function connected with sampling from finite populations, but for opposite reasons. He likes it to support the likelihood principle in sample surveys, and I like it to discredit this principle by showing its consequences in the same area. We both are wrong, because the probabilities of selection of samples are in a vague sense dependent on the unknown parameter, because they depend on the same prior facts (prior means and expectations, etc.) that have influenced the values under issue. Consequently, we do not have exactly the situation assumed in applications of the likelihood and conditionality principles. Of course this dependence of parameter and sample strategy is hard to formalize mathematically. My recognition of this dependence is due to a discussion I had recently with Professor Rubin on the conditionality principle.

As to the Horvitz-Thompson estimate, its usefulness is increased in connection with ratio estimation. For example, if the probabilities of inclusion are π_i and we expect the Y_i's to be proportionate to A_i, then we should use the estimate

$$\left(\sum_{i=1}^{N} A_i\right) \frac{\sum_{i \in s} Y_i/\pi_i}{\sum_{i \in s} A_i/\pi_i},$$

which would save the statistician's circus job. This estimate is not unbiased but the bias is small, and the idea of unbiasedness is useful only to the extent that greatly biased estimates are poor no matter what other properties they have.

J.C. Koop:

Professor Basu's essay is very stimulating and sometimes also provocative.

Regarding Sambo, I find the choice of selection probability for him (equal to 99/100) rather unwise in the face of the existence of a list of elephants' weights taken three years ago in the owner's possession. Sambo, we are told, was a middle-sized elephant, and knowing the existence of Jumbo in the herd, it might have been wiser to choose the selection probabilities directly propor-

tional to the respective weights of the elephants according to the available records. The reason being that if the elephants grew such that their present weights are directly proportional to their weights three years ago, then the variance of the estimate (equal to the selected elephant's weight divided by its selection probability), is zero. The circus statistician ought to have known better, and one should not be surprised that he was fired!

I am in complete agreement with him that the label of each unit in a sample (or in my terminology, the identity of a unit) cannot be discarded on the ground that it does not provide information. His discussion on this important point is very clear and can be read with profit.

However, I am somewhat surprised at his lack of appreciation for the basic ideas contained in Horvitz and Thompson's path-breaking paper of 1952 as evidenced by the following statement in section 6 of his paper: "During the past few years, altogether too much has been written on the subject of linear estimators of the population total Y. The original sin was that of Horvitz and Thompson who in 1952 sought to give a classification of linear estimates of Y. The tremendous paper-writing pressure of the past decade has taken care of the rest." These two writers constructed three linear estimators, each depending on one of the following three basic features of what I subsequently termed as the axioms sample formation in selecting units one at a time, namely, (i) the order of appearance of a unit in a sample, (ii) the presence or absence of a unit in the sample and (iii) the identity of the sample itself. Sample survey theorists have since benefited from their work. I for one felt in 1956 that the various types of estimators in the literature of that time needed classification and starting with these three features of sample formation, showed that $2^3-1=7$ types or classes of linear estimators, $T_1, T_2, \ldots, T_7$ were possible for one-stage sampling, three of which were those of Horvitz and Thompson. Godambe in his fundamental paper of 1955 found what I subsequently classified as the T_5-type of estimator, which should certainly not be attributed to Ajgaonkar, whose work began much later. In the process of this classification, it was found that an estimator given in the early pages of Sukhatme's text book of 1954 is of the T_4-type, that is, an estimator where the coefficients attached to the variate-values (observations in Basu's terminology) depended on the identity of the unit (label) and the order of appearance of the unit. Among other things, all this work was described in a thesis accepted by the North Carolina State University in 1957 and published in its *Institute of Statistics Mimeo Series* as No. 296, in 1961. Subsequently in 1963 I revised some of this work and amplified some of its ramifications in a paper in *Metrika*, Vol. 7(2) and (3).

One may ask what is the use of recognizing the three features of sample formation? In the context of the real world of sample surveys it must be said that they have physical meaning, which has some bearing on how an estimator may be constructed. Equally important, they point to the information supplied by the sample even before (field) observations on its members are made. In discussing Dr. C.R. Rao's excellent paper, I constructed a class of estimators where two of the features of sample formation were used, viz., (ii) which is equivalent to recognizing the identity of the distinct units (labels) and (iii) the identity of the sample itself, to show that an estimator of this class can have smaller M.S.E. than the U.M.V. estimator, derived through an appeal to the

principles of maximum likelihood, sufficiency and completeness, carried over almost bodily from classical estimation theory, thus bringing into question the extent of relevance of these principles in estimation theory for sample surveys of a finite universe. (It must be stressed that this does not detract from Dr. Rao's valuable paper which I interpret as a probe to uncover the difficulties of the subject.)

R. Royall:

Although I agree with much of what is said in this paper, I must take exception to one fundamental point. In section 2 Professor Basu states that: "From the point of view of a frequency probabilist, there cannot be a statistical theory of surveys without some kind of randomization in the plan S."

"Apart from observation errors and randomization, the only other way that probability can sneak into the argument is through a mathematical formalization of what we have described before as the residual part R of the prior knowledge, $K = (\Omega, \alpha, R)$. This is the way of a subjective (Bayesian) probabilist. The formalization of R as a prior probability distribution of θ over Ω makes sense only to those who interpret the probability of an event, not as the long range relative frequency of occurrence of the event (in a hypothetical sequence of repetitions of an experiment), but as a formal quantification of the ... phenomenon of *personal belief* in the truth of the event."

It seems frequently to be true that at some time before the values $y_1, y_2, \ldots, y_N$ are fixed it is natural and generally acceptable to consider these numbers as values, to be realized, of random variables $Y_1, Y_2, \ldots, Y_N$. For instance, these might be the numbers of babies born in each of the N hospitals in the state during the next month. What particular values will appear is uncertain, and this uncertainty can be described probabilistically. Although subjectivists would presumably accept these statements, in many finite populations such models are precisely as *objective* as those used everyday by frequentists. If such a model is appropriate before the y's are realized, it seems to be equally appropriate after they are fixed but unobserved. If a fair coin is flipped, the probability that it will fall heads is one half; if the coin was flipped five minutes ago, but the outcome has not yet been observed, my statement that the probability of heads is one half is no less objective now than it was six minutes ago. The state of uncertainty is not transformed from objective to subjective by the single fact that the event which determines the outcome has already occurred.

It can be argued that since the event has already occurred, the outcome should be treated as a fixed but unknown constant (so that now the probability of heads is one if the fixed but unknown outcome *is* heads and otherwise is zero). Such an argument leads back to the conventional model but rests on an unduly restrictive notion of the scope of objective probability theory.

The probability of one half for heads arises from my failure to notice that the coin is slightly warped. It can be argued that all probability models for real phenomena are likewise conditioned on personal knowledge and should therefore be called subjective. Be that as it may, (i) many statisticians do not consider themselves to be subjectivists and (ii) *super-population* models are frequently as objective as any other probability models used in applied sta-

tistics. Since such models, in conjunction with non-Bayesian statistical tools, can be extremely useful in practice as well as in theory, it seems to me to be a mistake to insist that they are available only to subjective Bayesians without pointing out that in this context the term applies to essentially all practicing statisticians.

REPLY

Professor Koop and Professor Godambe seem to think that the real difficulty in the elephant problem lies in the 'unrealistic' sampling plan—a plan that is 'not related' to the background knowledge. I always thought that the real purpose of a sampling plan is to get a good representative sample. If the owner knows how to relate the present weight of the representative elephant Sambo to the total weight of his fifty elephants, then he ought to go ahead and select Sambo. Why does he need a randomized sampling plan? Professor Koop wants to allot larger selection probability to Jumbo, the large elephant. Does he really prefer to have Jumbo rather than Sambo in his sample? I think Professor Koop is actually indifferent as to which elephant he selects for weighing. He *knows* more about the circus elephants than the circus owner. He 'knows' that the 50 ratios of the present and past weights of the elephants are nearly equal. Therefore, he has made up his mind that the ratio estimate is a good one irrespective of which elephant is selected. But he is not prepared to go all the way with me and assert the goodness of the ratio estimate irrespective of the selection plan. Professor Koop needs to allot unequal selection probabilities (proportional to their known past weights) to the 50 elephants so that he can mystify his non-statistical customers with the assertion that his estimate is then an unbiased one. As a scientist he has been trained to make a show of objectivity. May I ask what Professor Koop would do if the elephant trainer informs him that Jumbo (the big elephant) is on hunger strike for the past 10 days? Should he not try to avoid selecting Jumbo? He should, because now he does not know how to relate the present weight of Jumbo to the total weight of the 50 elephants.

In survey literature, we often come across the term *representative sample.* But to my knowledge the term has never been properly defined. At one time it used to be generally believed that the simple random sampling plan yields a representative sample. However, the difficulty with this naive sampling plan was soon recognized and so surveyors turned to stratification and other devices (like ratio and regression estimation) to exploit their background information about a specific survey problem. It is not easy to understand how surveyors got messed up with the idea of unequal probability sampling. I think it started with the idea of making the ratio estimate look unbiased. Thus Lahiri devised his method of using the random number tables in such a manner that the probability of selecting a particular sample set of units is proportional to the total 'size' of the units. This plan made the ratio estimate look 'good'. The flood-gate of unequal probability sampling was then opened and a surprisingly large number of learned papers has been published on the subject. What is even more surprising is that no one seems to worry about the fact that the surveyor can allot only one set of selection probabilities $\pi_1, \pi_2, \ldots, \pi_N$, but

that he has usually to estimate a vast number of different population totals. For each particular population total the surveyor may be able to find an appropriate ratio (or regression) estimate. But how can he possibly make all these different ratio estimates look 'good'?

Of late, a great deal has been written about the Horvitz-Thompson estimate. A little while ago Professor Rao proved an optimum property of the method. But to me the H-T estimate looks particularly curious. Here is a method of estimation that sort of contradicts itself by alloting weights to the selected units that are inversely proportional to their selection probabilities. The smaller the selection probability of a unit, that is, the greater the desire to avoid selecting the unit, the larger the weight that it carries when selected.

The question that Professor Hájek raised in the first part of his comments is exceedingly important and is one that, at one time, had given me a great deal of trouble. As Professor Hájek admitted, the question is hard to formulate and is even harder to answer. In the second part of my essay, I shall discuss the problem in greater detail. To-day, let us try to understand the difficulty in the context of the circus elephants. Suppose the surveyor (the owner) selects three elephants u_1, u_2, and u_3 with probabilities proportional to their past weights (and, say, with replacements) so that the data is $x = [(u_1,y_1), (u_2,y_2), (u_3,y_3)]$. In this case, the selection probability of the labels $u = (u_1,u_2,u_3)$ depends on the past weights of the 50 elephants and, therefore, also *depends* on their present weights—the state of nature $\theta = (Y_1, Y_2,..., Y_{50})$. If the selection probability of u depends on θ, then the very fact of its selection gives the surveyor some information about θ. Should the surveyor ignore this fact and act as if he always wanted to select this set of labels u and analyze the data x on that basis? This is precisely what I am advising the surveyor to do and this is what Professor Hájek thinks to be an error. But let us stop and think for a moment. Does the information that u is selected tell the surveyor anything (about θ) that the surveyor did not know already? When the question is phrased this way, one will be forced to admit that there is no real difference between the above plan and a simple random sampling plan. Indeed, the important point that I am trying to make is this, that even when the sampling plan is sequential, the relevant thing is the data generated by the plan and the likelihood function (which depends only on the data and has nothing whatsoever to do with the plan).

However, contrast the above sampling plan with a plan where the owner asks the elephant trainer to give him the names of three elephants that come first to his mind. If (u_1,u_2,u_3) are the three elephants that are selected by the above plan, then the surveyor does not really know how he got the labels (u_1,u_2,u_3) and so he cannot analyze the data x. Could it be that the three elephants were refusing to eat for some time and that is why they were on the trainer's mind at the time? If the owner must depend on the trainer for the names and present weights of three sample elephants, and if he does not have the sampling frame (so that he cannot select the labels himself), then he may be well advised to instruct the trainer to select the three sample labels at random. Randomness is a devil no doubt, but this is a devil that we understand and have learnt to live with. It is easier to trust a known devil than an unknown saint!

The second point raised by Professor Hájek is easier to deal with. If the

surveyor *knows* that the ratios of the present and past weights of the elephants are nearly equal, then why does he not use the ratio estimate itself? I do not see any particular merit in the estimate suggested by Professor Hájek.

Now, let us turn to Professor Godambe's objection to the likelihood principle in the context of survey sampling. It will be easier for us to understand Godambe's point if we examine the following example. In a class there are 100 students. An unknown number τ of these students have visited the musical show *Hair*. Suppose we draw a simple random sample of 20 students and record for each student, not his (or her) name, but only whether he has seen *Hair*. The likelihood is then a neat (hypergeometric) function involving only the parameter of interest τ. Godambe likes this likelihood function. However, if we had also recorded the name of each of the selected students, then the likelihood function would have been a lot messier. It would no longer have been a direct function of τ, but would have been a function of the state of nature $\theta = (Y_1, Y_2, ..., Y_N)$, where Y_j is 1 or 0 according as the student j has or has not seen *Hair*. Godambe does not know how to make any sense of this likelihood function. My advice to Professor Godambe will be this: "If the names (labels) are 'not informative', if there is no way that you can relate the labels to the state of nature θ, then do not make trouble for yourself by incorporating the labels in your data". After all, isn't this what we are doing all the time? When we toss a coin several times to determine the extent of its bias, do we record for each toss the exact time of the day or the face that was up when the coin was stationary on the thumb? We throw out such details from our data in the belief that they are not relevant (informative). Statistics is both a science and an art. It is impossible to rationalize everything that we do in statistics. These days we are hearing a lot of a new expression—rationality of type II. It is this second kind of rationality that will guide a surveyor in the matter of selection of his sample and the recording of his data.

The final remark of Professor Godambe seems to suggest that he has not quite understood what I said in the last paragraph of my essay. It is simply this that the constants in the estimating formula of the surveyor need not be (indeed, they should not be) pure numbers like π and e. The estimating formula (estimator) that the surveyor chooses surely depends on the particular inference situation. If the mathematician wishes to find out how the estimator behaves in the altered situation where the population values are measured in a different scale, he should first ascertain from the surveyor whether he (the surveyor) would like to adjust the constants in his formula to fit the new scale. When the surveyor is given this freedom, then it is no longer true that $G = NS/n$ is the only linearly invariant estimator in the class of all estimators that depend only on the label-set and the sample total. The Godambe assertion holds true only in the context of a severely restricted choice.

If I have understood Professor Royall correctly, then he claims that his super-population models for the parameter $\theta = (Y_1, Y_2, ..., Y_N)$ are non-Bayesian in the sense that such models do have objective frequency interpretations. His contention about the tossed coin in the closed palm is somewhat misleading. Let us examine a typical super-population model in which the Y_j's are assumed to be independent random variables with means αA_j and variances βA_j^{γ}, where A_j is a known auxiliary character of unit j and α, β, γ are known (or unknown) constants ($j = 1,2,...,N$). To me, such a model looks

exactly like a Bayesian formalization of the surveyor's background knowledge or information. Certainly, there is nothing objective about the above model. Indeed, is any probability model objective? When a scientist makes a probability assumption about the observable X, he is supposed to be very objective about it. But as soon as he makes a similar statement about the state of nature θ he is charged with the unmentionable crime of subjectivity. Mr. Chairman, you have always been telling us that the ultimate decision is an 'act of will' on the part of the decision (inference) maker. Isn't it equally true that the choice of the probability model for the observable X is also an act of will on the part of the statistician? Equally subjective is the choice of the 'performance characteristics'. A true scientist has to be subjective. Indeed, he is expected to draw on all his accumulated wisdom in the field of his specialization. My own subjective assessment of the present day controversy on objectivity in science and statistics is this that the whole thing is only a matter of semantics.

If we define mathematics as the art and science of deductive reasoning—an effort at deducing theorems from a set of basic postulates, using only the three laws of logic—then statistics (the art and science of induction) is essentially anti-mathematics. A mathematical theory of statistics is, therefore, a logical impossibility!