RELEVANCE OF RANDOMIZATION IN DATA ANALYSIS
PART ONE

by

D. BASU

(The Florida State University)

March 1977

FSU Statistics Report No. M402
RELEVANCE OF RANDOMIZATION IN DATA ANALYSIS
PART ONE

by

D. BASU

(The Florida State University)

Abstract: Ronald A. Fisher listed randomization, replication and local control (or blocking) as the three fundamental principles of statistical experimentation. So we religiously introduce an element of pre-randomization (in the selection of survey units or in the allocation of treatments) in all our survey or experimental designs. But why do we need to randomize and what is the exact role and relevance of this artificial (fully controlled, that is) randomization at the data analysis stage? In Part One of this essay, the question is examined in the context of survey sampling. The second part will be addressed to the same question but in the context of experimental designs.

This report is prepared for presentation at a Symposium on Survey Sampling and Measurement to be held at Chapel Hill, N.C., April 14-17, 1977.
1. INTRODUCTION

This essay is a natural sequel to an earlier one (Basu, 1971) presented by me at a symposium held in Waterloo, Ontario in March/April 1970. The writing of this essay was promised and its content foreshadowed in the Waterloo essay. While presenting that essay, I had made a number of off-the-cuff remarks challenging the logic of the so-called randomization analysis of data. In this essay I propose to give a further account of my views on the question of data analysis. The time lag of over seven years between the two essays is only a measure of my diffidence and self doubt on the important question of the relevance of randomization at the data analysis stage.

Please permit me to begin on a light-hearted note by giving you an account of a brief conversation that took place between me and Professor Neyman at a breakfast table in a motel at Waterloo in April 1970. If I remember right, this took place on the morning following my spirited presentation of the Waterloo essay. The dialogue ran approximately along the following lines:

Basu: Professor Neyman, I greatly admire your mathematical theory of statistics. All your concepts are so well-defined and the theory is filled with so many beautiful theorems. But there exists a big gap in your theory.

Neyman: Well, shall we hear about it?

Basu: The theory is concerned more with the sample space than with the particular sample at hand, more with inference-making behaviors and the average performance characteristics of such behaviors than with the problem of ascertaining what particular inference is best warranted by what particular data. Your theory does not seem to recognize the fact that different samples, even though they may
be in the same sample space, may differ vastly in their information contents.

Neyman: Let us have an example.

Basu: Here is an oversimplified survey-type example:

Example 1.1: Let us suppose our population consist of 100 units. With each unit is associated an unknown number. Let \( Y_1, Y_2, \ldots, Y_{100} \) be the 100 unknown numbers and let \( Y = \Sigma Y_j \) be the parameter of interest that we have to estimate. Let us further suppose that we have the background information that one (but we do not know which one) of the 100 numbers is very large, say, of the order of \( 10^{10} \), and that all the other 99 numbers lie between 0 and 1. If \( S_n \) stands for the experiment of drawing a simple random sample of size \( n \) from the bag, then I have to agree with you that \( S_{25} \) is a more informative experiment than \( S_5 \). But suppose you have drawn a simple random sample of size 25 and you know that I have drawn another such sample of size 5. You find that all your 25 sample numbers are small and that I am beaming with pleasure. You have to concede then that your sample is worthless for the purpose of estimating \( Y \) and that I have hit the jackpot.

With a characteristic benevolent smile, Professor Neyman told me:

Neyman: Basu, you amaze me! When you make a point, you do it with so much force and such passion!

That was the end of the dialogue. Clearly, Professor Neyman wanted to enjoy his breakfast and was in no mood to be drawn into a controversy at that time. But I have often wondered about what might have been Professor Neyman's defense
had he cared to take up my challenge. Would he have lightly dismissed the example as one that has no practical import? Or, would he have gently chided me for talking about such terrible sampling plans like $S_5$ and $S_{25}$ in the context of that kind of background information?

Please note that I am concerned at the moment not with the question of how to plan a survey or experiment but with that of how to analyze the data generated by a survey or experiment. The point that I was trying to drive home with my pathological example is that the information content of the sample generated by a survey or experimental design (however well-planned the design may be) depends largely on the sample itself, that a lot of sample to sample variation in the information content is inevitable. The question of how to analyze a data can be answered only after a careful examination of the data itself.

Of course, it may be argued that, once I have clearly specified what kind of analysis would be appropriate for what data (that may be generated by a particular survey or experiment), my overall inference-making behavior vis a vis that particular survey or experiment gets well-defined as a decision function. My behavior pattern can be assessed in terms of some average performance characteristics and them compared with other hypothetical behavior patterns. There are more than one snag in this argument.

In most survey situations, the data collected would be so vast and complicated that it would take all my time and mental energy to figure out how best to analyze the particular data obtained. It would be entirely unrealistic for anyone to suggest that I could give an honest answer to the question of how
I would analyze all the different data that might be generated by the survey design. Ask an impossible question and you get an unrealistic answer. Even if I could define my inference-making behavior as a decision function, the function, as a rule, would be so complicated that it would be virtually impossible to work out its average performance characteristics. Finally, it is not clear what statistical intuition guides us through the mathematical process of averaging the "loss" over all possible data. As I have said before, a data may be good, indifferent or downright poor in its information content. When I get a good data, I thank my stars in the hope that my inference making performance will be good in that situation. Why shall I permit anyone to cast a shadow of doubt on my performance by pointing the possibility of a poor data that might have been but have not? On the other hand, when I get a relatively uninformative data, I should either try to get more information or make a clean breast of the fact and do the best I can in that situation.

Let us look back on our pathological example 1-1. How to analyze the data generated by the design $S_{25}$ (a simple random sample of size 25)? Let $y = (y_1, y_2, \cdots y_{25})$ denote the 25 sample $Y$-values. In a sense we well understand, the statistic $\tilde{T} = 100 \bar{y}$ is a "design-unbiased" estimator of $Y = \Sigma Y_j$. But is it not true that the corresponding inference-making behavior is terribly biased all the time? If $y$ contains the large $Y$-value, then we know that $\tilde{T}$ overestimates $Y$ by a factor of nearly 4. On the other hand, if $y$ fails to contain the large $Y$-value, then $\tilde{T}$ clearly underestimates $Y$ by an astronomical figure. We have no option in this case but to give up the traditional survey criterion of design-unbiasedness and to face the problem of estimation of $Y$ fairly and squarely.
Let $\Lambda$ stand for the large $Y$-value (in the population of 100 units) which to start with we knew to be a number of the order of $10^{10}$. Let $Y_\Delta = Y - \Lambda$ be the total of the 99 small $Y$-values. If the sample $y$ contains the large $Y$-value then our estimate $\hat{Y}$ of $Y$ will perhaps look like

$$\hat{Y} = \Lambda + \hat{Y}_\Delta$$

where $\hat{Y}_\Delta$ is some reasonable estimate of $Y_\Delta$ based on the 24 small $Y$-values in the sample. Perhaps the estimate $\hat{Y}_\Delta = 99 \bar{y}_\Delta$, where $\bar{y}_\Delta$ is the mean of the 24 small $Y$-values in the sample, will appear reasonable to many of us in this case. On the other hand, if $y$ fails to include $\Lambda$ then the data is of very poor quality and we should either insist on more data or make a poor job of estimation by using the formula $\hat{Y} = 10^{10} + 99\bar{y}$.

The estimator

$$\hat{Y} = \begin{cases} 
\Lambda + 99 \bar{y}_\Delta & \text{if } y \text{ includes } \Lambda \\
10^{10} + 99 \bar{y} & \text{if } y \text{ fails to include } \Lambda 
\end{cases}$$

is not design-unbiased, but who cares? And then, who wants to judge $\hat{Y}$ by averaging its performance over all possible samples each of which is either very good or very poor?

R. A. Fisher was aware of the problem of recognizability of good and bad samples. With his theory of ancillary statistics, he attempted at a partial solution of the problem in non-Bayesian terms. [Refer to Basu (1964) for a detailed description of this method.] In this pathological example of ours, good and bad samples are easily identified in terms of the indicator $I_E$ of the
event $E$ that the sample $y$ includes the large $Y$-value $A$. Since $I_E$ is an ancillary statistic, one may want to invoke the conditionality argument of Fisher to analyze the data after conditioning it by the observed value of $I_E$. We shall return to the conditionality argument later.

The point that I was trying to make to Professor Newman seven years ago, could very well have been made in a less sensational fashion with the following realistic survey example.

**Example 1.2:** We are making an opinion survey among the community of students in a large university campus. We obtain a sampling frame from the Registrar's office and then draw a simple random sample of, say, 100 students. A large body of data is then generated by interviewing (surveying) the 100 sample students. When we carefully scrutinize the data, we discover that our sample contains a disproportionately large number of females and that the response pattern of the females to the questions related to the smoking of marijuana differs markedly from the sample group of males. This post-recognition of the fact that the sample does not truly represent the population in relation to the marijuana questions is certainly going to affect our analysis of the data with reference to the marijuana parameters.

A critic may try to point out that our trouble in this instance really stems from the fact of our initial carelessness in not stratifying the population into males and females to start with. But, it might not have been possible for us to draw samples separately from the subpopulation of the male and female students simply because we did not have separate sampling frames for them. In any case, the hindsight about the fruitful device of pre-statification of the
population into males and females may have come to us only after we examined the data relative to the marijuana questions. When we turn to the Vietnam issue we may find that the sample contains far too many foreign students, when we examine the data in relation to the student-power issue we may find that the sample contains a disproportionate number of undergraduates, and so on.

The survey design is not the only determinant of the quality of the data produced by the survey. The principal determinant of how a particular data ought to be analyzed is the data itself. The key concept in survey theory ought to be the notion of post-stratification. Example 1.2 highlights the need for post-stratification. Depending on what parameter we are trying to estimate, the data itself will usually suggest how we need to stratify it. Post-stratifications of the data in numerous ways should be recognized as an essential process of data analysis.

2. LIKELIHOOD

The likelihood principle has an inexorable logic of its own. [For a careful discussion on this topic see Basu (1975).] Yet, the principle has often been characterized as either irrelevant or inoperative in the context of survey sampling and experimental designs. Let us take a close look at the question in terms of yet another survey type example.

**Example 2.1:** An urn contains 100 tickets that are known to be numbered consecutively as $\theta + 1$, $\theta + 2$, $\vdots \theta + 100$ where $\theta$ (the parameter) is an unknown number in the set $\theta$ of all integers. Suppose we set in motion a particular sampling plan $S$, thus generating a set of 10 sample tickets bearing the
numbers \( y_1, y_2, \ldots, y_{10} \) that are recorded in their natural selection order, if any. How should we analyze the data?

Here \( x = (y_1, y_2, \ldots, y_{10}) \) is the sample. Let \( m \) and \( M \) be respectively the minimum and the maximum sample values. In order to fix our ideas, let us first consider the case where \( S \) is the traditional plan of drawing 10 tickets one by one with equal probabilities and without replacements. It is now easy to check that the likelihood function generated by the data is

\[
L(\theta \mid x) = \begin{cases} 
\text{if } M - 100 \leq \theta \leq m - 1 \\
0 \text{ otherwise}
\end{cases}
\]

(2.1)

where the constant \( p \) is equal to \( (90!)/(100!) \). Writing \( J_x \) for the set of integers in the internal \([M - 100, m - 1]\) and \( J_x(\cdot) \) for the indicator of the set \( J_x \), we can re-write (2.1) as

\[
L(\theta \mid x) = p \cdot J_x(\theta)
\]

(2.2)

The above representation of the likelihood function leads us to the following conclusions:

i) Since \( J_x(\theta) \) depends on the sample \( x \) only through the statistic \((m, M)\), it follows that \((m, M)\) is a sufficient statistic in the usual sense. Indeed \((m, M)\) is minimum sufficient.

ii) The data \( x \) rules out all values of \( \theta \) that fall outside \( J_x \). It cuts down our extent of ignorance about \( \theta \) from the parameter space \( \Theta \) to the set \( J_x \).

iii) The data \( x \) lends equal likelihood support to all the parameter points in \( J_x \).
iv) The length \( d_x = 99 - (M - m) \) of the interval \( J_x = [M - 100, m - 1] \) may be taken as a reasonable measure of the quality of the data; the smaller \( d_x \) is the better is the data \( x \). The best possible data is the one for which \( M - m = 99 \). In this case we are able to identify the value of \( \theta \) without the possibility of any error. Observe that \( S_2 \), a simple random sampling plan with sample size 2, may yield a data that is perfect in the above sense, whereas, a much more extensive survey plan, say \( S_{25} \), may very well fail to do so.

We are at last ready to make the very crucial observation that the above conclusions hold true irrespective of the nature of the survey plan \( S \) that produced the sample \( \bar{x} = (y_1, y_2, \ldots, y_{10}) \). The likelihood function \( L(\theta \mid \bar{x}) \) will have the form (2.2) for any sampling plan \( S \) that we can think of. Observe that the factor \( J_x(\theta) \) has nothing to do with the plan \( S \) and that the constant factor of proportionality \( p \) did not enter into any of the arguments (i) to (iv) above. It is this factor \( p \) that depends on \( S \) and also on \( \bar{x} \) in some cases. It will be useful to check the correctness of the above assertion for each of the following sampling plans:

\( S_{(1)} \): Continue sampling one at a time, without replacements and with equal probabilities until the sample range \( M - m \) exceeds 50.

\( S_{(2)} \): Draw a sample of size 5 one at a time, without replacements and with equal probabilities. If the sample mean exceeds 20, draw a further sample of size 5 in the same manner; otherwise stop sampling.

\( S_{(3)} \): Continue sampling one at a time with replacements and with equal
probabilities until the same number is drawn twice, with the condition
that we stop anyway if the first 10 draws are all different and the
sample range is at least 25.

What is the likelihood function generated by the sample \( x = (17, 24, 40, \\
5, 16, 37, 19, 26, 10, 62) \) under each of the above three sampling plans?

With such a sample \( x \), we have \( m = 5 \) and \( M = 62 \) and so the interval \( J_x \) is \([-38, 4] \). Irrespective of the sampling plan \( S \) (as long as \( x \) is a possible sample
for the plan \( S \)), the data would tell us unequivocally that the true value of \( \theta 
\) must lie in the interval \( J_x \) and would lend equal likelihood support to each
point in \( J_x \). The likelihood function can always be represented in the form
(2.2) where \( J_x(\theta) \) does not depend on the plan \( S \). The factor \( p \) is \((90!)/(100!)
\) in the cases of \( S_{(1)} \) and \( S_{(2)} \), but is \( 100^{-10} \) in the case of \( S_{(3)} \).

Thus, irrespective of the nature of the sampling plan \( S \). The likelihood
function is flat over the set \( J_x = [M - 100, m - 1] \). The fact that all para-
meter points in the set \( J_x \) are equally supported by the data does not mean
that we are dealing here with a case of "uninformative" likelihood as I find
it often suggested by some of my esteemed colleagues. True, we do not have a
well-defined maximum likelihood estimate of \( \theta \) (unless \( J_x \) is a one-point set).
But that does not mean that we cannot analyze the data in terms of the likeli-
hood function.

In the present case, a Bayesian analysis of the data is quite straight
forward and simple. The Bayesian will match the likelihood function \( L \) with
his prior probability distribution \( \xi \) on \( \theta 
\), arrive at his posterior distribution.
and then use this posterior distribution to justify his inference making on \( \theta \). Since the likelihood function \( L \) is flat over the set \( J_X \) and is zero outside, it is clear that \( \xi^*_X \) is the normalized restriction of \( \xi \) to \( J_X \).

Since \( J_X \) is defined entirely in terms of the statistic \((m, M)\) it is clear that the Bayesian analysis of the data does not depend on the nature of the sampling plan \( S \). At the data analysis stage the Bayesian can as well forget about the sampling plan!

As I have said elsewhere (Basu, 1969), a typical feature of the survey sampling problem is that we invariably end up with a likelihood function that is flat and design-free. Example 2.1 beautifully brings out this fact and that is why we have labored with this example at such length. In one important respect, however, this example is not symptomatic of our idealized survey set-up. Being an example of an urn model, it is typical of a survey situation where the population is essentially unlabeled like, for example, the population of blue whales in the antarctic region. We do not have a sampling frame of reference for the 100 tickets in the urn. As the tickets do not have any pre-survey identities, the question of selecting a particular ticket for observing its \( Y \)-value simply does not arise. This precludes the possibility of using any of the more sophisticated survey designs like stratified, multistage, systematic pps, \( \pi \)ps design, etc. Faced with an unlabeled urn problem like this, we can at best shake the urn (in the case of the blue whales even that is beyond us!), blindfold ourselves, pull out several units from the urn and pretend that we have in effect got a simple random sample from the population. With urn models like this, our choice of a sampling plan is essentially restricted.
to equal probability sampling, with or without replacements and with some simple or fancy stopping rules as illustrated earlier.

3. A SURVEY SAMPLING MODEL

In the Waterloo essay (Basu, 1971), I had idealized away most of the troublesome and mathematically intractable features of a large-scale survey operation by characterizing a typical survey set-up in the following simplistic terms:

i) There exists a well-defined population, a finite collection $P$ of distinguishable objects called the units.

ii) The units in $P$ are not only distinguishable pairwise but are also identifiable individually. This means that we have a sampling frame of reference (a list of the units in $P$) which enables us to pre-select any particular unit in $P$ for the purpose of observing (surveying) its characteristics. Let us suppose that $P$ is listed as 1, 2, 3, \ldots, $N$.

iii) Corresponding to each $j \in P$, there exists an unknown quantity (possibly vector-valued) $Y_j$. Let us write $\omega = (Y_1, Y_2, \ldots, Y_N)$ and call $\omega$ the universal parameter. The set $\Omega$ of all the possible (a-priori, that is) values of $\omega$ is then the parameter space.

iv) Our main concern is with survey set-ups when we have a lot of prior information about the universal parameter $\omega = (Y_1, Y_2, \ldots, Y_N)$. Let us suppose that the main ingredient of this large body of prior information comes in the form of a knowledge vector $A = (X_1, X_2, \ldots, X_N)$, where $X_j$ is a known auxiliary characteristic (possibly vector-valued) of unit $j \in P$. All our characterizations
of the prior distribution or super-population model \( \xi \) for the parameter \( \omega \) will be made in terms of such a knowledge vector \( A \).

v) We make the simplifying assumption that there are no non-response or non-sampling (observational, that is) errors. This means that when we choose to select a particular unit \( j \) for the determination of its Y-value \( Y_j \), we are always able to locate that particular \( j \) and then determine the corresponding \( Y_j \) without any observational error.

vi) By a sampling plan or design \( S \) we mean a well-defined body of rules following which we can arrive at a subset \( \delta \) of the population \( P = \{1, 2, \ldots, N\} \). We shall call \( \delta \) the sample label-set and shall often characterize it as \( \delta = (i_1, i_2, \ldots, i_n) \), where \( i_1 < i_2 < \cdots < i_n \) are a set of distinct elements of \( P \) that is listed in increasing order of their unit indices or labels. We call \( n \) the sample size.

vii) Following a sampling plan \( S \), the surveyor selects the label-set \( \delta \in P \). This selection of \( \delta \) can usually (though, not always) be carried out before the fieldwork. By fieldwork we mean that part of the survey operation that determines the Y-value \( Y_i \) for each unit \( i \in \delta \). Clearly, fieldwork is the major part of a typical survey operation.

While I recognize the immense complexity of a time-to-life survey problem, I believe it would be quite an achievement if we can sort out the major controversial issues related to survey sampling in terms of the above simplistic model.

### 4. WHY RANDOMIZE?

By randomization we mean the injection of a fully controlled element
of randomness in the selection process of the label-set \( \delta \). Why is it necessary to randomize?

It is easy to overwhelm the question by a series of counter questions. What is the alternative to a randomized choice of \( \delta \)? Purposive selection? How can you justify a purposive selection of \( \delta \)? How can you claim to be objective in your scientific methods if you refuse to randomize? With purposive selection, how can you have a sample space? Without a sample space how can you make a statistical analysis of data? Without randomization, how can you have an unbiased estimate of any population characteristic? And so on.

At this stage, it will be useful if we ask the above questions and the counter-questions in the context of an extremely simple survey type example. For the sake of pinpointing our whole attention on the theoretical issues, I find it imperative to make all my examples as simple as possible.

**Example 4.1:** The population is \( \{1, 2, \ldots, N\} \). About the universal parameter \( \omega = (Y_1, Y_2, \ldots, Y_N) \), we have the background information that each \( Y_j \) is either 0 or 1 and that \( Y_1 \leq Y_2 \leq \ldots \leq Y_N \). The parameter of interest is \( \theta = \Sigma Y_j \), that is, the number of ones among the \( Y_j \)'s. Let us imagine a mechanical device that produces \( N \) items (units) on a particular day. The item \( j \) produced by the machine is either defective \( (Y_j = 1) \) or non-defective \( (Y_j = 0) \). Once the machine produces a defective item, it continues to do so for the rest of the day. At the end of the day, we need to estimate the number of defective items in the lot of \( N \) items produced by the machine. The unit \( j \) is the \( j^{th} \) item produced by the machine.
In this example, the universal parameter \( \omega \) is completely determined by the parameter of interest \( \theta \). In other words, we do not have any nuisance parameter. Let \( S \) be an arbitrary sampling plan. Let \( \delta = (i_1, i_2, \cdots, i_n) \), with \( i_1 < i_2 < \cdots < i_n \), be the selected label-set and let \( y = (y_1, y_2, \cdots, y_n) \), where \( y_k = Y_{i_k} \), be the observation vector. Our sample is then \( x = (\delta, y) \). What does our sample \( x \) tell us about the parameter \( \theta \)?

Observe that, irrespective of the plan \( S \), some samples \( x = (\delta, y) \) are perfect in the sense that they give us full information about \( \theta \). For instance, if \( i_1 = 1 \) and \( y_1 = 1 \), that is, if we discover that \( Y_1 = 1 \), then we know for sure that every \( Y_j = 1 \) and so \( \theta = N \) for sure. Generally speaking, if \( \delta \) contains two consecutive labels, say \( v \) and \( v + 1 \) such that \( Y_v = 0 \) and \( Y_{v+1} = 1 \), then we shall know for sure that \( \theta = N-v \).

Let us define by \( v = v(x) \) to be the largest \( i \in \delta \) such that \( Y_i = 0 \), if \( Y_i = 1 \) for all \( i \in \delta \) then define \( v = 0 \). Likewise, define \( w = w(x) \) to be the smallest \( i \in \delta \) such that \( Y_i = 1 \), if \( Y_i = 0 \) for all \( i \in \delta \) then define \( w = N+1 \). Observe that, irrespective of the survey plan, \( v \leq w - 1 \) for all possible samples \( x \), the sign of equality holding only if the sample \( x \) is perfect in the sense of being fully informative about \( \theta \). Writing \( J_x \) for the set of integers in the closed interval \([N-w+1, N-v]\), it is easy to check that the data \( x \) tells us for sure that \( \theta \) must lie in the set \( J_x \).

If \( J_x \) is a one-point set, then we have complete information about \( \theta \). Otherwise, we are still left with a measure of doubt about the whereabouts of \( \theta \). What is this measure? Has the sampling plan \( S \) got anything to do with it?

It would be useful to do a little exercise. Suppose \( N = 100 \). Think of
the data $x = \{(17, 24, 40, 73), (0, 0, 1, 1)\}$. Now think of any sampling plan $S$ (let it be sequential, stratified, pp&l;=, purposive, whatever you like) that could possibly give rise to the above sample. Verify that the likelihood function

$$L(\theta) = \text{Prob}(x \mid \theta, S)$$

is flat over the set $J_x = \{61, 62, \ldots, 76\}$ and is zero outside. It will then be easy to recognize the fact that the above is true for all $N$, $x$ and $S$.

As I have said earlier in the context of example 2·1, a Bayesian will be very pleased with the simple look of the likelihood function and compute his posterior measure of doubt about $\theta$ as the normalized restriction of his prior measure to the set $J_x$. Since the set $J_x$ is defined entirely in terms of the data $x$, the Bayesian analysis of the data will be independent of the nature of the sampling plan $S$.

As in the case of example 2·1, a statistician schooled in the theory of randomization analysis of data will be usually lost with our data $x = \{(17, 24, 40, 73); (0, 0, 1, 1)\}$. This is because he is not used to think of $\delta = (17, 24, 40, 73)$ as the unique label-set generated by the experiment $S$. He is trained to look upon $\delta$ as a variable point in a set $\delta$ that is determined by $S$, and so he needs to find out the probability distribution of $\delta$ over $S$. Without an element of randomization in the plan $S$, our conventional survey sampler will not have a roomy space $S$ (well furnished with a probability distribution) to make him feel comfortable. On the other hand, if the plan $S$ is too complicated then that will make him uncomfortable again because he will not be able to
figure out the probability distribution of $\delta$ over $S$. And even when the plan
$S$ is simple enough (say, a simple random sampling plan with sample size 4)
his analysis of the data $x$ will suffer from his non-recognition of the fact
of whether the particular data is of good, moderate or poor information con-
tent.

In the present case, how should we plan to choose a sample of size 4
from the population of $N$ units? To me the simple random sampling plan
looks utterly ridiculous. If you are very ignorant about $\theta$, that is, if
all the values of $\theta$ in the parameter space $\Theta = \{0, 1, 2, \ldots, N\}$ look about
equally plausible to you, then why not sample in the following sequential
and purposive manner? First, select the unit label $[N/2]$ and observe its
$Y$-value $y_1$. Depending on whether $y_1 = 1$ or 0, select the second sample
label as $[N/4]$ or $[3N/4]$. If $y_1 = 1$ and $y_2 = 1$, then your third sample
label will be $[7N/8]$ and so on. This four step sequential sampling plan
will surely cut down the extent of your ignorance about $\theta$ from the interval
$[0,N]$ to an interval of about one-sixteenth its original size. Although this
sampling plan is purposive, in the beginning you would not know what label-
set $\delta$ you are going to end up with. The space $S$ over which $\delta$ varies is
fairly extensive. However, relative to a fixed $\theta$, there is only one label-
set $\delta$ that is possible; that is, the sampling distribution of $\delta$ is degen-
erate for each $\theta$.

In the case of this example, it is easy to see that, relative to any
given prior distribution $\xi$ of $\theta$, the optimum sampling plan would be sequen-
tial and non-randomized (purposive). Admittedly, it is not easy to justify
randomization in this case. However, it would be unfair to deny the usefulness of randomization in survey and experimental designs on the basis of a single pathological counter-example. Let us re-examine the question of randomization.

The counter-question, "How can you justify purposive sampling?", has a lot of force in it. It is only in transparently simple cases, like the one above, that one can give a clear-cut argument in favor of a particular purposive plan. In a true-to-life survey situation, it is very difficult to sell the idea of a fully purposive plan: The very purpose of a purposive plan is rooted in the scientific intuition and knowledge of a surveyor. No two surveyors are likely to agree on the choice of their survey plans. The choice of a purposive plan will make a scientist vulnerable to all kinds open and veiled criticisms. A way out of the dilemma is to make the plan very purposive, but leave a tiny weeny bit of randomization in the plan: for instance, draw a systematic sample with a random start or make a very extensive stratification of the population and then draw a sample of size 1 from each stratum!

The rationale of a fully purposive survey plan will usually be so involved that it would be almost impossible for the surveyor to clearly spell it out for the benefit of anyone other than himself. As a result, hardly anyone else (other than the surveyor) will have a clear understanding of all the factors that contributed to the selection of the label-set δ and, therefore, of the data x = (δ,y) itself. Without such an understanding how can anyone check on the validity of the conclusions drawn by the surveyor from his data?
How can I disagree with a scientist if I cannot analyze his data myself?

It is thus a clear imperative that the surveyor fully describes his survey plan and carefully explains all the considerations that led to the particular plan. And this inhibits the choice of a purposive plan. The possible criticism that the surveyor's chosen plan was not the optimum one (even with respect to his own background information) may not cast any doubt on his conclusions as long as the critic can analyze his (the surveyor's) data. No wonder, therefore, that all of us choose the path of least resistance and try to incorporate an element of randomness in the survey plan.

There are situations where the surveyor has to protect himself against unknown human frailties and biases by incorporating an element of randomness in his selection process. As an example, consider the case of the huge socio-economic survey of rural households that is continually carried out by the National Sample Survey Organization (NSSO) of India. There are over half a million villages in India. The NSSO does not have a sampling frame of the rural households (about 70 million of them) in India. An investigator is sent to a sample village to get a list of the households from the village Chowkidar (watchman), to verify the correctness of the list, to select a sample of, say, five households from the list and then to spend a great deal of time eliciting a lot of information from each of the five sample households. The NSSO cannot trust its investigators in the matter of selecting a representative sample of five households from each sample village. What if an investigator chooses the five most prosperous looking households, or the five
nearest to the village Chowkidar's residence? So each investigator is given a sealed envelope containing a page of random number table and is instructed to use the page in a particular fashion for sample selection after listing the households in alphabetical order of the names of the household heads. A sample selection process of this kind is easy to comprehend and describe, and it is not so easy to criticize a data so generated on the pretext that it cannot be analyzed.

This is about as far as I am willing to go along with the principle of randomization. However, I find that many of my esteemed colleagues have wholeheartedly committed themselves to a much stronger version of the principle. It is with this principle of randomization analysis of data that we are going to be concerned with in the rest of the essay.

5. RANDOMIZATION ANALYSIS OF DATA

Let us argue within the framework of the simplistic survey model that we described in Section 3. Indeed, we are going to restrict our attention to the simpler situation where the unknown \( y_j \)'s and the known \( x_j \)'s are real numbers \( j = 1, 2, \ldots, N \) and where the parameter of interest is \( Y = \Sigma y_j \). Let \( S \) be the survey design and \( x = (x, y) \) the sample, where \( x = \{i_1, i_2, \ldots, i_n\} \), with \( i_1 < i_2 < \ldots < i_n \), is the label set, and \( y = (y_1, y_2, \ldots, y_n) \), with \( y_k = y_{i_k} \), is the observation vector. The set \( X \) of all the values of \( x \) that were possible a priori (that is, before \( S \) was set in motion) is the sample space. An estimator \( T \) is a map of the space \( X \) into the range space \( V \) of \( Y \).

Our traditional survey theory (hallowed with the names of such authorities as...
as Fisher, Neyman, Mahalanobis and Yates) cannot get off the ground unless an
element of artificial (fully controlled, that is) randomization is injected
into the design $S$. In this theory, the sample $x$ is regarded as a random ele-
ment that varies over the sample space $X$ in accordance with a probability
law $P_\omega$ that is well-defined in terms of the universal parameter $\omega = (Y_1,
Y_2 \ldots Y_N)$ and, of course, the design $S$. An estimator $T = T(x)$ of $Y$ is regarded
as a random variable with range $Y$. If $E(T|\omega) = Y$ for all $\omega \in \Omega$, then $T$ is
a design-unbiased (that is, unbiased relative to a design $S$) estimator of $Y$.
The particular estimate $T(x)$ of $Y$, that corresponds to the data $x$ at hand,
is then regarded as free of any sampling bias. If the variance $V(T|\omega)$ of the
unbiased estimator $T$ can be shown to be "small", then the estimate $T(x)$ is
supposed to be "close" to the true value of $Y$. Finally, if the estimator can
be shown to be, in some sense, optimum among a class $T$ of unbiased estimators
of $Y$, then that fact is proudly put forward as a sort of a proof of the objec-
tivity of the surveyor in the choice of his estimator $T$.

The statistical intuition behind the traditional theory of survey sampling
is rooted in the value-loaded notion of unbiased estimation. Along with the
concept of significance test and that of confidence interval, the concept of
unbiased estimate is one of the three most widely used, most controversial and,
according to me, most misleading notions of statistics.

The traditional survey strategy is to so design the survey that there
exists a "good" design-unbiased estimate of $Y$. With a purposive sampling plan
it is not possible to attain the unbiasedness objective. So the surveyor
carefully introduces an element of randomization in the survey plan $S$ and then finds his good unbiased estimate $\hat{Y}$ of $Y$. Next he finds a good (and, if possible, unbiased) estimate $\hat{\sigma}^2$ of the variance of $\hat{Y}$ and calls $\hat{\sigma}$ the standard error (another value loaded expression!) of $\hat{Y}$. He then puts forward the pair $(\hat{Y}, \hat{\sigma})$ as the end-product of his data analysis with the suggestion that almost all of the available information in the data $x$ about the parameter of interest $Y$ is summarized in the estimate $\hat{Y}$ and its standard error $\hat{\sigma}$.

The most attractive feature of the above line of argument is that it is not based on any speculative probabilistic supposition like, for example, $u_1, u_2, \ldots, u_n$ are iid normal variables or the regression of $u$ on $v$ is linear, etc. The statistical argument is "Oh! so non-parametric". Indeed, the only way probability enters into the argument is through the surveyor's well-planned and fully controlled use of the randomization artifact. It is this non-parametric quality of the randomization analysis of data that many statisticians find most irresistible.

The randomization argument is charming, but is it relevant? Inherent in the argument is the supposition that the analysis of a data $x = (\delta, y)$ must be firmly based on the sampling design $S$ that produced it. We propose to re-examine this major premise of the randomization argument.
6. RANDOMIZATION AND INFORMATION

With the survey data \( x = (s, y) \) before us, let us write

\[
Y_{\text{obs}} = \sum_{i \in S} y_i = \sum_{j \in S} Y_j \quad \text{and} \quad Y^* = Y - Y_{\text{obs}} = \sum_{j \notin S} Y_j.
\]

The post-survey analysis of the data must begin with the obvious fact that, irrespective of the survey plan \( S \), the data gives us the exact value of the observed part \( Y_{\text{obs}} \) of the population total \( Y \). The whole purpose of the data analysis is to extract the whole of the relevant information in the data about the yet unobserved part \( Y^* \) of \( Y \). But is the data informative about the part of the population that is yet unobserved? To answer the question in the negative is to repudiate the basic premise of all survey theory. Yet, I have heard it said that, if the plan \( S \) is purposive, that is, if there is only one possible label set \( \Delta \), then the data \((s, y)\) cannot possibly give us any information about any unobserved \( Y_j \). So naturally I ask the

**Question:** How come the artifact of randomization makes the data informative about every \( Y_j \)?

**Answer:** With randomization allowed, we can so plan the survey as to endow every unit \( j \in P \) endowed with a non-zero probability of being included in the sample label-set \( \Delta \). So even though a particular \( j \) may fail to be selected, the fact could never be denied that it might have been. The statistical purpose of randomization is to make the survey experiment informative about every \( Y_j \) and, therefore, about \( Y = \Sigma Y_j \).

**Question:** Could you be a little more explicit about what you mean by your experiment being informative about every \( Y_j \).

**Answer:** For one thing, we can find an unbiased estimator for each \( Y_j \). We have
only to find the selection probability $\pi_j$ for a given $j$ and then define $T_j$ as follows:

$$T_j = \begin{cases} \frac{Y_j}{\pi_j} & \text{if } j \in \Delta \\ 0 & \text{if } j \notin \Delta \end{cases}$$

Actually, we need to find $\pi_j$ only if $j \in \Delta$. Now observe that $T_j$ is an unbiased estimator of $Y_j$.

**Question**: Is $T_j$ a respectable estimator of $Y_j$?

**Answer**: Yes; estimators of this kind were first considered by Horvitz and Thompson (1952). The total of $T_j$ over all $j$ is the famous Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{j \in \Delta} \frac{Y_j}{\pi_j}. \quad (6.1)$$

Several optimum properties of $\hat{Y}_{HT}$ as an estimator of $Y$ have been established by V. P. Godambe (1955), T. V. Hanurav (1968), C. R. Rao (1971) and others.

**Question**: Generally speaking, what is your definition of a statistical experiment $E$ that is informative about an unknown state of nature $\theta$?

**Answer**: An experiment $E$ is informative about $\theta$ if we can recognize a statistic $T$ (a map of the sample space of $E$ into another space) whose sampling distribution manifestly depends on $\theta$ and, maybe, on some other parameters as well.

**Question**: Are all the possible outcomes of an informative experiment equally informative?

**Answer**: Now you have pulled a really fast one. I do not know what you are really talking about. Maybe, you should check with Professor Neyman about that.

So we are back on square one! Before taking up the all-important question
of information in the data, let us briefly look at $T_j$ as an informant on $Y_j$. In a technical sense, $T_j$ is indeed a design-unbiased estimator of $Y_j$. But with the data $x = (\alpha, y)$ before us, is it not obvious that $T_j(x)$ must be a terribly biased estimate of $Y_j$ irrespective of what $x$ is? When $j \notin \delta$, the $T_j$-estimate of $Y_j$ is zero. And when $j \in \delta$, that is, when we have full information on $Y_j$, the $T_j$-estimate of $Y_j$ is $Y_j/\pi_j$. Remember, $\pi_j$ is usually a very small number!

7. INFORMATION IN DATA

At the data analysis stage, notions like sampling distribution or design-unbiasedness are hardly relevant. What is really relevant is the likelihood function

$$L(\omega) = L(\omega|X,S) = \text{Prob } (x|\omega,S)$$

generated by the data—the sample $x$ produced by the survey design $S$. Every Bayesian will wholeheartedly agree with the above proposition. Even an ardent non-Bayesian like R. A. Fisher found the proposition almost self-evident and used to think of the likelihood function as the only bridge that links the observed sample to the unknown parameter.

In 1967, being somewhat bewildered and bemused by a long list of mostly unreadable papers on survey sampling that appeared in the Annals of Mathematical Statistics and the Sankhya, I resolved to settle to my own satisfaction the following two related questions: In a survey set-up, what is the smallest statistic that summarizes in itself all the information in the full data? And, what is all the information in the data? The results of the above investigation are carefully laid out in Basu & Ghosh (1967) and Basu (1969). In short
they are as follows:

Let $S$ be an arbitrary sampling plan. The plan may be randomized or
purposive, sequential or non-sequential. For any well-defined sampling
plan leading to a sample outcome, we should always be able to work out the
corresponding likelihood function. The outcome of $S$ need not be recorded in
the summary form of $x = (\delta, y)$. Let $x'$ be the form in which the sample out-
come of $S$ is actually recorded. We only suppose that $x'$ is recorded in suf-
ficient details, so that we can reduce $x'$ to the summary form $x$ if we wish.
We call $x$ the sample core.

Let $\omega = (Y_1, Y_2, \ldots, Y_N)$ be the universal parameter and let $\Omega$ (a subset
of $\mathbb{R}_N^*$) be the parameter space. The parameter space $\Omega$ can be quite-arbitrary.
For a given outcome $x'$ of $S$ let $\Omega_{x'}$ be the set of parameter points in $\Omega$ that
are consistent with the observation $x'$. In other words, if $x = (\delta, y)$,
where $\delta = (i_1, i_2, \ldots, i_n)$ and $y = (y_1, y_2, \ldots, y_n)$, is the sample core of $x'$, then

$$
\Omega_{x'} = \{ \omega: \omega \in \Omega \text{ and } Y_{i_k} = y_k \ (k = 1, 2, \ldots, n) \}.
$$

It is clear that the subset $\Omega_{x'}$ of $\Omega$ depends on the sample $x'$ only through
its core $x$. Let us, therefore, designate $\Omega_{x'}$ by $\Omega_x$. Observe that once the
sample core $x$ is before us, we can determine $\Omega_x$ without any reference to the
survey design $S$. It is easy to deduce [see Basu (1969)] that, irrespective
of the nature of the design $S$, the probability that $S$ gives rise to the sam-
ple $x'$ is

$$
\text{Prob}(x'|\omega, S) = \begin{cases} 
q & \text{if } \omega \in \Omega_x \\
0 & \text{if } \omega \notin \Omega_x
\end{cases}
$$
where \( q = q(x^\ast, S) \) depends only on \( x^\ast \) and the design \( S \). Writing \( I_x \) for the indicator of the set \( \Omega_x \), we can now re-write the above in the following manner:

\[
L(\omega|x^\ast, S) = q(x^\ast, S) I_x(\omega)
\]

(7.1)

where \( L \) is the likelihood function determined by the data.

The equation (7.1) holds the key to both the questions that I asked earlier. Firstly, it follows at once that, in the context of a fixed sampling plan \( S \), however complex the plan may be and no matter how the sample \( x^\ast \) is recorded, the sample core \( x \) is always sufficient. There can never be any loss of information if we insist on recording the outcome of the survey operation as \( x = (\delta, y) \), where \( \delta \) the subset of population units that are sampled and \( y \) is the corresponding observation vector. [That is why we have been representing the sample in this form from the very beginning.] Usually, the sample core \( x \) will be the minimum sufficient statistic. But if we look back on Example 4.1, it would be apparent that \( x \) may contain some redundant information. The minimum sufficient statistic is the map \( x^\ast \rightarrow \Omega_x \rightarrow \Omega_x^* \) from the sample space \( \Omega \) of \( S \) to a class of subsets of the parameter space \( \Omega \). [See Basu & Ghosh (1967) for some measure-theoretic difficulties that may arise when \( \Omega \) and \( \Omega \) are both supposed to be uncountable and how to overcome such difficulties.]

Secondly, if we look back on (7.1) and ask the question, "What is all the information in the data?", then we see at once that the relevant part of the likelihood function \( L(\omega|x^\ast, S) \) is the factor
Thus, given the data \((x,S)\), all parameter points in the set \(\Omega_x\) have equal likelihood support and all points outside \(\Omega_x\) have zero likelihood support (that is, are rejected outright by the data). As we have noted earlier, given the sample \(x\), we can determine \(\Omega_x\) without any reference to the design \(S\). Thus, at the data analysis stage, we have no need to concern ourselves with the exact nature of the design \(S\).

As I see it, the question "What does the data tell us?" is a misguided one. How can an inanimate object tell us anything? The data \((x,S)\) is a representation of an experience. It is for us to interpret that experience in the light of our other experiences. The right question to ask is "How should we make a beginning with the complex process of interpretation of a particular data?" My Bayesian answer to the question is: "Begin with the likelihood function generated by the data".

The likelihood function is only the beginning and must not be regarded as an end in itself. In our survey situation the function is flat over the set \(\Omega_x\) and is zero outside. This flatness of the likelihood appears to be a matter of great concern to many non-Bayesians. But this very flatness makes a Bayesian happy, because the mathematics of his data analysis becomes very simple in theory. The Bayesian's posterior distribution for \(\omega\) is now the normalized restriction of his prior to the set \(\Omega_x\). Since \(\Omega_x\) does not depend on \(S\), the Bayesian can take in his stride any sampling plan \(S\), however complex it may be. Indeed, at the data analysis stage he need not even care to ascertain the exact nature of the plan.
Many eyebrows were raised when I made the last remark in the opening section of Basu(1969). The remark was misinterpreted to mean that I can analyse a data \((x, \cdot)\) without knowing what kind of \(S\) filled the blank spot. Of course, that cannot be true. As I have said earlier, no one can analyse a sample \(x\) without a clear understanding of how that sample was generated. That is why I refuse to analyse a sample \(x\) that was purposively selected by another person unless I fully understand the rationale of that purposive selection. If, however, I know that the plan \(S\) is one of a kind \(\{S_1, S_2, \ldots, S_k\}\) that I fully understand, then my Bayesian analysis of the data \((x, S)\) will not depend on the exact nature of \(S\). In this case I can summarize the data further to \((x, \cdot)\).

When do I "fully understand" a plan \(S\)? When I can be persuaded to set down the likelihood function in the form (7.1). In other words, when I can blissfully forget about the plan at the data analysis stage!

8. A CRITICAL REVIEW

Over the years, I have heard many critical remarks on this neo-Bayesian thesis on survey sampling and have indeed suffered from a lot of self-doubt myself. We close this part one of the essay with a partial list of such remarks and self-doubts. Immediately following remark i, the paragraphs marked as i(a), i(b), etc. will summarize my current thinking on the subject.

Remark 1: It is a mistake to regard the individual \(Y\)-values as parameters. The fact that the \(Y_j\)'s are unknown (before the survey) does not mean that they can be regarded as parameters. Is it not a fact that in most textbooks on statistics, a parameter is defined as an unknown characteristic of an unknown (probability) distribution? If \(F(t)\) is the proportion of \(Y\)-
values that do not exceed \( t \), then in \( F \) we have an unknown c.d.f. The population mean \( \bar{Y} \) is a bonafide parameter because it is the mean of the distribution defined by \( F \). Similarly, the maximum \( Y \)-value is a parameter. But how can you regard, say, \( Y_1 \) as a characteristic of \( F \)?

1(a): The remark has a lot of force in it when we are dealing with an unlabeled population like, say, the blue whales in the antarctic region. But here we are talking about a case where each unit \( j \) has a pre-survey identity. Therefore, the \( Y \)-value \( Y_j \) of the \( j \)th unit is a well-defined but unknown state of nature and so is a parameter according to my book.

1(b): Just before the commencement of a football (or cricket) match, an important parameter is the outcome of the impending coin-tossing experiment. Once the experiment is over, all the uncertainties (in the mind of a typical spectator) about this parameter are removed, but the spectator is still uncertain about his parameter of interest, namely, the outcome of the game. In my survey set-up, the situation is exactly analogous. In the beginning, I was uncertain about the \( N \) parameters \( Y_1, Y_2, \ldots, Y_N \). After the survey, I know for sure the \( Y_j \)'s that correspond to \( j \in \Delta \). The state of my uncertainty about the parameter of interest \( Y = \Sigma Y_j \) is altered but not removed.

Remark 2: It is misleading to regard \( Y_1, Y_2, \ldots, Y_N \) as parameters. It is much more realistic to think of them as already realized but yet unobserved values of \( N \) random variables \( \eta_1, \eta_2, \ldots, \eta_N \). This will allow us to build a realistic and useful statistical model \( \xi \) (the, so called, superpopulation model) for \( \eta_1, \eta_2, \ldots, \eta_N \). Relative to such a model \( \xi \), the problem of survey design is to make an optimum choice of the label-set \( \Delta \). Once such a choice
is made, the problem of estimating $Y = \sum Y_j$ becomes a problem of predicting the unobserved part

$$Y^* = \sum_{j \notin S} Y_j$$

of $Y$ in terms of the observed $Y$-values $\{Y_j : j \in S\}$.

2(a): There is a large measure of agreement between the Bayesian and the above Prediction approach of Richard Royall (1971). Royall's analysis of survey data is free of the survey design $S$. Like a Bayesian, Royall would analyse the data $(x,S)$, where $x = (\delta,y)$, by ignoring $S$ and then looking upon the label-set $\delta$ as the one and only one that he needs to concern himself with. However, Royall would look upon $y = \{Y_j : j \in \delta\}$ as a random vector and would try to find a $\delta$-unbiased predictor of $Y^*$ that is optimum in some reasonable sense. Royall has built for himself what he considers to be a comfortable half-way-house between the Mahalanobis-Yates and the full-fledged Bayesian approach to survey sampling. I propose to evaluate this Prediction approach in another article.

2(b): From my remark in 1(b) it should be clear that a Bayesian does not make a distinction between a random variable and a parameter. A parameter (random variable) is any unknown entity that is within reach of human speculation. Before the survey, all the $Y$-values were parameters or random variables. After the survey, some of the parameters become known, but the rest are still random variables. It is not necessary to make a distinction between a random variable $\eta_1$ and its realized but still unobserved value $Y_1$. Thus, it is redundant to hypothesize the existence of the $N$ random variables $\eta_1$,
\( \eta_2, \ldots, \eta_N \). The \( Y_j \)'s themselves are random variables. A Bayesian would readily agree that the problem of analysing a survey data is a prediction problem.

**Remark 3:** J. Hájek as a discussant on Basu (1971) made in essence the following remark: Let \( S \) be a typical randomized survey design like \( pp\), \( \pi p\), etc. where the design-probability \( p(\delta) \) of the label-set \( \delta \) is made to depend on some auxiliary characters \( \kappa_1, \kappa_2, \ldots, \kappa_N \) that are related to the universal parameter \( \omega = (Y_1, Y_2, \ldots, Y_N) \). Hence, the knowledge that a particular set \( \delta \) is selected by the design \( S \) will give us some indirect information on \( \omega \). Therefore, Basu as a discussant on Rao (1971) committed a logical error when he invoked the Conditionality Principle to suggest that, at the data analysis stage, Rao ought to hold the particular chosen \( \delta \) as fixed and not take into account other possible label-sets that might have been.

3(a): It is true that \( p(\delta) \) depends directly on \( \kappa_1, \kappa_2, \ldots, \kappa_N \) and, therefore, indirectly on \( \omega \). However, since we already know \( \kappa_1, \kappa_2, \ldots, \kappa_N \), we cannot say that from the knowledge of \( \delta \) we get some information on \( \omega \) via \( (\kappa_1, \kappa_2, \ldots, \kappa_N) \). In the case of a non-sequential survey design like \( pp\), \( \pi p\), etc., the label-part \( \delta \) of the sample \( x = (\delta, y) \) is truly an auxiliary statistic.

**Remark 4:** For the Likelihood Principle to be operative, we need a sample space \( X \), a parameter space \( \Omega \) and a kernel function \( p = p(x|\omega) \) mapping \( X \times \Omega \) into the half-line \( [0, \infty) \). In problems of survey sampling and experimental designs, we have just one probability distribution defined by the randomization scheme of the survey or experimental design. How can we have the Trinity \( (X, \Omega, p) \) of abstractions in such cases?
4(a): This can be done, and this is precisely what I achieved in Basu (1969) in the context of survey designs. The analogous case of experimental designs will be discussed in part two of the essay.

4(b): Consider, for example, the very simple case where the population \( P = \{1, 2, 3\} \) and where \( \omega = (Y_1, Y_2, Y_3) \) denotes the unknown \( Y \)-values. Let \( S \) be the simple random sampling plan of size 1. The sample is \( x = (i, y) \) where \( i \) is the selected unit and \( y \) the corresponding \( Y \)-value. What are \( X, \Omega \) and \( p \) in this case?

4(c): We must define the parameter space \( \Omega \) first. Now define \( X_1 \) to be the set of all possible samples of the type \((1, y)\). That is, \( X_1 \) is the set of all pairs \((1, y)\) such that there exists an \( \omega \) in \( \Omega \) with \( Y_1 = y \). The sets \( X_2 \) and \( X_3 \) are similarly defined. The full sample space is \( X = X_1 \cup X_2 \cup X_3 \).

4(d): For a fixed \( x = (i, y) \) define \( \Omega_x \) to be the set of all \( \omega \) in \( \Omega \) such that \( Y_i = y \). Let \( I_x \omega \) denote the indicator of the set \( \Omega_x \). For our simple random sampling plan \( S \), the kernel function \( p \) is

\[
p(x|\omega) = \begin{cases} 
1/3 & \text{if } I_x(\omega) = 1 \\
0 & \text{otherwise}
\end{cases}
\]

For a more complicated sampling plan, say, a \( pps \) plan, the factor 1/3 will be replaced by a factor like \( q(i) \).

Remark 5: Let us go back to the simple example in 4(b). For each \( \omega = (Y_1, Y_2, Y_3) \) let us define \( X_\omega \) as the set \( \{(1, Y_1), (2, Y_2), (3, Y_3)\} \) and call it the (conditional) sample space, given \( \omega \). Instead of defining, Basu-fashion, the sample space as \( X = \cup \{X_\omega : \omega \in \Omega\} \), it is better to define the sample space \( X_\omega \) for each \( \omega \) separately. [Gudambe in personal communication].

5(a): There is no real advantage in defining the sample space as a
collection \( \{ X_\omega \} \) of conditional sample spaces. In my representation of the sample space as \( X = \cup X_\omega \), the set \( X_\omega \) is the carrier of the measure \( P_\omega \) (on \( X \)) that is indexed by \( \omega \). The only outcome of representing the sample space as the collection \( \{ X_\omega \} \) will be that the theory of surveys would appear to fall outside of the mainstream of statistical theory, which it does not.

**Remark 6:** Kolmogorov (1933) made the correct fundamental distinction between a "zero probability event" (null-set) and "logically impossible event" (empty-set). Kolmogorov's set-up had only one probability measure. However, the same fundamental distinction must be observed, when a single measure is replaced by a family of measures, to resolve some difficulties of statistical logic that otherwise arise. This resolution enabled Godambe and Thompson (1971) to establish some fiducial distributions on sound statistical footing. [Godambe in personal communication].

6(a): What Godambe is saying in effect is that anomalies and paradoxes would arise if logically impossible events are included in the class of events (measurable sets) as events with zero probability. Kolmogorov (1933) was interested on discoursing on 'non-atomic' measures on uncountable sets. Therefore, he had to make the logical distinction between impossible events and zero-probability events. In survey theory all our probability distributions are discrete. Therefore, we need not distinguish between the two notions. Whenever a subset \( E \) of \( X \) has zero \( P_\omega \) - measure we should regard \( E \) as logically impossible with respect to \( \omega \) and vice versa.

6(b): I have come to the conclusion that the logical difficulties mentioned by Godambe are illusory. And I find the Godambe-Thompson (1971) thesis
on fiducial distributions in survey theory utterly incomprehensible.

Remark 7: In survey sampling, some of the co-ordinates of the parameter \( \omega \) are observed but the rest are not. Essentially based on the fact of non-observance of some co-ordinates of \( \omega \), it is now asserted that the data \( x \) makes all parameter points in the set \( \Omega_x \) equally likely. There must be a logical fallacy in the argument. The assertion that two parameter points are equally likely can be made on the basis of knowledge but never on account of ignorance or lack of observation.

7(a): The trouble lies in some of us taking the value-loaded expression "likelihood" too seriously. The statement that the likelihood function is flat over the set \( \Omega_x \) should be interpreted to mean that the (inanimate and unintelligent) data \( x \) equally supports all parameter points in \( \Omega_x \). It does not mean that the surveyor should consider all points in \( \Omega_x \) to be equally likely.

Remark 8: The fact of a flat likelihood renders any discussion of the maximum of the likelihood null and void. This, also, is the graveyard of the likelihood principle, because if any statistician claims that there is not partial evidence on the set of unobserved \( y \)-values, he should visit a psychiatrist or leave statistics. [This is Oscar Kempthorne (in his inimitable style) as a discussant on Rao (1971).]

8(a): The length that even reputable statisticians are sometimes willing to travel in search of a maximum likelihood estimate is really amazing. In the case of the simple example considered in 4(b) above, if we ignore the label-part \( i \) of the data \( x = (i, y) \) and reduce the data \( x \) to \( y \), then it is
clear that Prob \( (y|y) \) is equal to 1 for \( \omega = (y, y, y) \). Thus, if \( (y, y, y) \in \Omega \) for all \( y \), then the maximum likelihood estimate of the population mean \( \bar{Y} \) is well-defined if we reduce the data to \( y \) and is \( y \) itself. [The case of a simple random sample of size \( n \) is analogous]. It is, however, not generally recognized that \( y \) will still be the ML estimate of \( \bar{Y} \), in the above restricted sense, even when the plan \( S \) do not allot equal selection probabilities to the three population units!

8(b): During the past several years a lot of debate took place on the relevance and/or informativeness of the label-set \( \delta \) as part of the data. In my simplistic survey set-up, where I have assumed the existence of a set of auxiliary values \( x_1, x_2, \ldots, x_N \), the labels are of course always informative and can never be suppressed from the data.

8(c): I have heard it said that the full likelihood function, being flat over the set \( \Omega_x \), is uninformative about the unobserved \( Y \)-values. Indeed, it is this kind of assertion that Kempthorne was so violently reacting against. Once he recognizes that the likelihood is only an intermediate step and not the end product of data analysis, Kempthorne's discomfiture with the likelihood principle will perhaps disappear. Or, will it? [The following is an estimate of what my good friend Kemp might say at this stage.]

Remark 9: The Likelihood Principle is only the thin end of your wedge. You are trying to sell me the whole Bayesian package. How can I act like a Bayesian when I do not recognize subjective speculations on uncertainties and utilities as valid scientific methods. There are three kinds of probabilities that statisticians are writing about these days,
a) Personal Probabilities: These are unalloyed and unashamed subjective speculations on uncertainties;

b) Model Probabilities: These conditional probability speculations on the observables involve some subjective elements and a lot of mathematical opportunism; and

c) Randomization Probabilities: Ah! these are the only kind of probability that really exist. Do not ask me to buy a new theory of statistics based on the non-existent probabilities of the kind (a) and (b). I have enough trouble already with the Neyman-Pearson-Wald kind of statistics that is based on (b). By the way, can I make you interested in a theory of statistics that is based on (c) alone?

9(a): Let us quote directly from Bruno de Finetti's own preface to his two-volume treatise on the Theory of Probability:

"My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:

PROBABILITY DOES NOT EXIST.

The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time, ... or Fairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs."

9(b): What about randomization probabilities? I believe, when De Finetti made the above stirring remark, he was not thinking of randomization probabil-
ities. If pressed hard on the matter, De Finetti would probably give a grudging recognition of objectivity to this kind of probability, but would perhaps insist on joining in the rider that this kind of probability has no relevance at the data analysis stage.

9(c): How can probabilities of type (a) and (b) be useful if they are not real? Similar questions were asked about irrational, negative, and imaginary numbers when they were first introduced into mathematics. Data analysis is a speculative process, a mind interacting with a given data. Probability theory should be looked upon as a guideline for the modes of thought and behavior of a human mind when faced with uncertainty. I strongly suspect that the Bayesian guideline, in terms of subjective probabilities of the type (a) and (b), is more reliable than the objective randomization probabilities of Oscar Kempthorne. Whether a mode of thinking that is manifestly superior to the subjective probability mode can be devised for data analysis, is a question that can be answered only with a large measure of uncertainty at this moment.

Remark 10: The inner consistency of the Bayesian point of view is granted. However, the analysis of the survey data need not be fully Bayesian. Indeed, who can be a true Bayesian and live with thousands of parameters? [Basu (1971), p. 234.]

10(a): In a typical large-scale survey situation, the population size \( N \) runs into hundreds of thousands, the dimension of the unknown \( Y \)-character \( Y_j \) for unit \( j \) runs into scores, and the dimension of the available auxiliary vector \( X_j \) for unit \( j \) may also run into dozens. How can we then entertain the thought of calling \( \omega = (Y_1, Y_2, \ldots, Y_N) \) the universal parameter
and then making a full-fledged Bayesian analysis of the data in terms of a prior \( q \) and the likelihood function (7.1)? Our representation of the likelihood function in the pretty, flat, and design-free form of (7.1) is viable only in some simplistic survey-type problems that I have exemplified in this and in some earlier essays. The primary purpose of this representation was to highlight the simple fact that the information content of the data \((x, S)\) depends only on the sample \( x \).

10(b): The Bayesian as a surveyor must make all kinds of compromises with his theory. In the beginning, he may not have a clear perception of how the unknown multi-dimensional \( Y_j \)'s are related to the known multi-dimensional \( X_j \)'s. His choice of the sample units must, therefore, be based on some ad hoc decisions taken on the basis of some vague, ill-defined suppositions. To avoid public criticisms and to substitute some unknown possible selection biases by a well-understood random selection process, he may even agree to introduce an element of randomization in the plan \( S \).

10(c): Once the data is generated by a large-scale survey plan, the speculative process of exploratory data analysis begins. I cannot put this enormous speculative process into the straight jacket of a theory. I happen to believe that data analysis is more an art form than a scientific method. In these days of powerful computers, it is possible to analyze the same data over and over again in many different ways. It is this repeated process that will lead us to the underlying relationships between the \( Y_j \) and the \( X_j \) vectors, if any. We may need to post-stratify the particular data in many different ways in order to find out how best to extrapolate
from the observed part of the population to the yet unobserved part.

10(d): I have heard it said: "Bayesianism is like a bar of soap. It is a good cleansing agent for the Fisher-Neyman theory, but in the process of cleaning up, it will disappear itself". I do not know if that is going to happen or not. But if the accumulated knowledge and technology of mankind finally sweeps away the Bayesian methods from large-scale survey theory, I strongly believe that the Bayesian wisdom -- at the data analysis stage, hold the sample fixed and speculate about the parameters--will linger on.
REFERENCES


REFERENCES (con'd)


[A fuller list of references will appear at the end of Part Two of this essay.]