A HISTORICAL PERSPECTIVE OF THE RECENT DEVELOPMENTS IN THE THEORY OF SAMPLING FROM ACTUAL POPULATIONS*

BY

V.P. GODAMBE

Indian Statistical Institute, New Delhi
University of Waterloo, Waterloo, Ontario

Ladies and Gentlemen,

A few days ago when Dr. Daroga Singh suggested that I deliver this year's Panse Memorial Lecture I had some hesitation to accept the honour that was being done to me. For, unfortunately I did not have the privilege of coming into contact with late Dr. Panse although I have been aware of his pioneering contributions to applications of statistical methods in agriculture and other fields. However, I am, indeed, glad to have an opportunity to associate myself with his name through this lecture. I propose to present some thoughts on historical developments of survey-sampling theory. The subsequent dicussion would reveal why for today's talk I have preferred the phrase, 'sampling from actual population' instead of the more familiar and indeed satisfactory term, 'survey-sampling'.

The popular understanding today of statistics consists of probabilistic estimates about some population characteristic like total production of a country or average income, based on some random samples. But as I would demonstrate, essentially this meshing of the probability calculus with the actual social statistics proved to be the most difficult job for the early probabilists. This was true even after the probability calculus was sufficiently well developed and sophisticated to include in it Bayes Theorem, the Law of large Numbers etc. Indeed the present day commonplace notion that estimates could be calculated on the basis of randomly drawn samples is but a result of a very slow development of human thought (far slower than the development of mathematical theory of probability).

*"Dr. Panse Memorial Lecture" delivered on 29th March, 1976 in New Delhi.

1. The Hypothetical Population Model

As is commonly understood the theory of probability originated with the study of the chance phenomena associated with gambling instruments such as roulette or dice. This was in 17th century and the names of the authors usually quoted are Fermat and Pascal. The experiment of spinning a roulette or throwing a die had a well-defined outcome and it was clearly understood from the beginning that the subject matter of the theory of probability was the results obtained by repetitions of such an experiment; a random experiment. This is not to say that formally or even informally a frequency theory of probability, propounded later on by Von Mises in this century and Venn in the preceding one, was taken for granted by the early probabilists. Actually Leibnitz with his knowledge of jurisprudence clearly proposed a view that probability was a relation between two propositions. This could be considered as the origin of the 'logical theory' of probability developed by Keynes, Jeffreys and Carnap during the present century. Also Pascal clearly put forward what in the present day terminology could be called decision theoretic, subjective approach of probability. He even went on to define utilities thus foreseeing some of the recent developments of subjective probability by Ramsey, De Finetti, Savage and others. Actually the early authors of probability theory used three aspects of the subject namely frequency, logical and subjective often without making the distinction. In spite of all this, it could be safely said that at least until the end of 19th century—(and even until this day I believe) the most important object of study for the probabilists was the results obtained by independent repetitions over and over again, perhaps indefinitely, of some chance experiment illustrated simply as before by spinning of a roulette or throwing a die. Otherwise It would not be possible to understand the most important developments in literature on the subject during this period devoted to the study of law of large numbers (Bernoulli early 18th century) and central limit theorems (Laplace, 19th century). The central aspects of these works is that they deal with the results obtained when some chance experiment is repeated independently for a large number of times. As a matter of fact until the end of 19th century most applications of probability theory required besides Bayes Theorem the law of large numbers and the central limit theorem. The errors of measurements arising out of astronomical data and otherwise were studied in this way. Thus by this time probabilists perfected a very fruitful model of an infinite-population obtained by indefinite, independent repetitions of some chance experiment. We shall call this a *hypothetical population model*, the term first explicitly used though much later by Fisher (1956).

2. A Random Sample

This hypothetical population model was available when the professional statistician appeared on the stage around the middle of the 19th century. It is generally believed that statistics or statistical theory originated with the investigations of biological and sociological phenomena such as inheritance and the like during the last century. Soon the statistical theory was directed towards studying relationships between different factors that influenced those phenomena and towards studying the underlying chance mechanisms. For instance, the sizes of a dozen observed human skulls were supposed to have been produced by some kind of chance mechanism operating in the background. It was but a small step to replace this chance-mechanism by a hypothetical population generated by independent repetitions of this chance mechanism. Thus the statistician adopted the hypothetical population model perfected by the probabilists as mentioned in the preceding paragraph. It is true that F. Galton, K. Pearson and their contemporary statisticians were less explicit about it. And as I said before the first clear recognition of the hypotherical population model is due to Fisher. The chance mechanism determined the frequency distribution in the corresponding hypothetical population. In as much as the chance mechanism was not fully known (and hence was being studied) so also the corresponding frequency distribution was supposed to be known only up to certain unknown parameter, like the mean or the variance. For example in the simple chance mechanism of tossing a coin, the parameter may be the 'probability of head turning up'. Now the parameter here is objective in the sense that it is determined by the chance mechanism. But it could also be called a (technically useful) fiction in the sense that the parameter could not be determined by any well defined operation in contrast to some other 'parameters' which we would discuss later in section 5. Any way an estimate of this unknown parameter could be obtained on the basis of some observations made on the chance mechanism. These observations were called a *random sample* from the corresponding hypothetical population. For instance the random sample of a dozen skulls mentioned before. Thus the *general statistical theory* could be said to have evolved out of the problem of inferring, on the basis of a 'random sample' drawn from a hypothetical population, the associated frequency function containing an unknown parameter. Most importantly the randomness here is *postulational* in contrast to *physical* randomisation which we would discuss subsequently in Section 5.

3. Actual Populations

In relation to the general satatistical theory just mentioned one must note the interesting development concerning social statistics that was taking place since late 17th century. The development was somewhat independent of the development of the theory of probability. In 1662 Graunt published in England his work on social statistics. This contained mortality tables, vital statistics, bills of birth and death etc. The tables were constructed from the data collected in an *arbitrary* or haphazard manner from the *population.* In contrast to the hypothetical population discussed previously this population was real or *actual.* It could consist of households in a country or men, women of a society. These tables on social statistics then provided the basis for fixing rates of annuities or taxes. Some calculations made in this respect would now look utterly absurd. The satistical arguments then made were very primitive and even grossly erroneous. This one could say even after the more scholarly work of De Witt from Denmark and others which followed Graunt's publication. The most striking feature of all this work on social statistics is that the authors do not seem to be at all aware of the fact that the tables on social statistics used for fixing annuities, taxes, etc. could themselves be improved if they were based on a properly drawn sample instead of some arbitarily collected data from the population. This awareness was to come much later.

4. Early Contributions to Sampling from Actual Populations

During the nineteenth century apparently it was recognised that one could make inferences concerning an actual (finite) population characteristic such as its mean on the basis of sampled observations, with the help of the calculus of probability. Laplace for instance is said to have estimated the number of French people on the basis of a sample and also the ratio of male to female births. Interesting discussion concerning some surveys conducted in Germany during 19th century, is found in Schott (1923). For other discussions on the topic around 1900 we refer to Bowley (1906). An early example of this is from the 13th century England: "Rules" of Robert Grosseteste, Bishop of Lincoln (1240-1242) as quoted by Oschinsky (1971) in which the lord or lady is advised to estimate the total amount of grain in the barn by commanding that every twentieth sheaf should be examined as it enters. The oldest instance I know of, of the suggestion that calculus of probability could be used to make inferences about an actual population total on the basis of a sample is to be found in *Mahabharat*, the great Indian epic. I am tempted to give some details. This is from Vana-Parva; Nala-Damayanti Akhyan:

> The God Kali has his eye on a beautiful princess and is dismayed when Nala wins her hand. In revenge an evil spirit enters the body of the virtuous prince. Crazed with frenzy for gambling, Nala loses his kingdom, and wanders demented for many years. Nala's change of fortune is described in a remarkable anecdote. In an alien form, he has been travelling with another king Bhangasuri. This latter, wanting to flaunt his skill in numbers, estimates the number of leaves, and the number of fruit, on two great branches of a spreading tree. There are, he avers, 2,095 fruits. Nala counts all night and is duly amazed by morning. Bhangasuri accepts his due:
>
> > I of dice possess the science, and in numbers thus am skilled.
>
> He agrees to teach this science to Nala in exchange for some classes in horsemanship, in which, despite his exile, Nala still excels. At the end of this sensational course in survey-sampling Nala vomits out the poison of Kali, and is restored his normal form. Kali, exorcised by mathematics, retires to the tree. Nala returns to his kingdom, offers his still faithful bride as his final stake and quickly recoups all his losses, and lives happily ever after.

(Reproduced from *History and Philosophy of Science Seminar by* Ian Hacking)

It is important to see what sort of use of the calculus of probability was envisaged by the authors in the preceding paragraph. As I said before, until the end of the 19th century the most often employed tools for statistical inference apart from Bayes theorem were law of large numbers and central limit theorem. These theorems were concerned with identically and independently distributed variables; the underlying model being that of a hypothetical population of Section 11, from which unlimited numbers of observations could be made in a *sequence*. Thus the survey statistician though he was studying and sampling an actual (finite) population, assumed at least subconsciously the model of a hypothetical population. For instance he may be estimating average income per household in the country. The actual population of households of course was before him. But he was considering the hypothetical population of observations generated by repeated draws. The distribution of the variate in this hypothetical population was assumed to have certain form say normal. To draw a sample consistent with these assumptions, it was thought necessary to have a method of drawing which will not have bias concerning the variate under—study. For instance the method of drawing should not be such that it would tend to select 'richer' (or poorer) households more often. 'The average income per household in the actual population' was by means of law of large numbers or by some other reasonable manner not made explicit, was identified as the unknown parameter of the hypothetical population. Now as I said earlier in Section 2 'sampling' of a hypothetical population *postulationally* implies that the given set of observations coustitute a random sample; one may at most divide (stratify) it in groups if there is enough heterogeneity. And a survey sampling statistician working (implicitly or explicitly) with the hypothetical population model could do no better than this. This gave rise to what is now populary understood as '*representative method of sampling*'. One may for instance balance the sample by making the sample mean equal to the population mean for some known variable of which total count was available. The randomness of the sample was any way just postulational. This representative method as the foregoing discussion would suggest in the present day terminology, is more purposive than random. At any rate it envisaged no use of artificial device such as random number tables to draw a sample. A still more important point to be emphasised here is that in the inference or estimation procedure, the distribution generated by randomisation, played no important role. As we know in the later development of sampling theory and practice this 'distribution generated by randomisation' played a central role.

5. Randomisation with Artificial Devices

A new era first in sampling practice and then in sampling theory was ushered when artificial device to draw a random sample was used. An early recorded instance of this is a survey conducted in Sweden in 1912 (Dalenius, 1957) to study housing conditions. The publication and use of random number tables was still an event of a later date.

With the use of artificial randomisation devices like throwing of a die or random number tables to draw a random sample from an actual population, the concept of the distribution generated by randomisation started getting crystallized. That such a distribution must play 'an important' if not 'the central' role in inference also was gradually realised. The first theoretical formulation of this however was in relation to the problems of 'design of experiments' by Fisher in 1926. In relation to survey sampling the first well-known attempt is due to Neyman (1934) though there were not so well-known formulations previously. Our subsequent discussion of randomisation would be primarily in relation to survey-sampling.

The introduction of artificial devices of randomisation to draw a random sample posed a logically distinctive situation than the ones dealt with by the general statistical theory of Galton, Pearson and Fisher mentioned earlier in Section 2. This was the distinction: The use of artificial devices of randomisation evidently presuppose an actual population with explicitly *labelled* individuals such as for instance the population of households in a town or the population of farms in a village. Such population admits infinite modes of sampling; each subset of individuals in principle could be selected with *arbitrarily* fixed probability. In contrast as we have observed before 'sampling' hypothetical population implies simple random sampling with the possible variation of stratification. Again the labels of the individuals of an actual population provide much wider modes of estimation than those available for a hypothetical population. These new modes of estimation did not admit the traditional treatment accorded to the problem by the general statistical theory. A basic reformulation of the problem was necesary. To make the distinction between actual and hypothetical population still explicit we observe as before in Section 2 that the parameter of a hypothetical population (say probability of success in the coin tossing experiment) is also hypothetical while the parameter of an actual population (say average income per household in a town) is real in the sense that the latter could be determined in principle by a total count. In fact, inferring about some unknown parameter of a hypothetical population was equivalent to inferring about the chance mechanism which generated the population. In contrast, in the case of an actual population the only chance mechanism was the mode of randomisation of sampling which was completely known to the statistician. Irrespective of this mode of randomisation, the parameter of an actual population had its existence and could be determined by the total count. We illustrate these distinctions in our review of Neyman's work on the subject, in the next section.

6. Neyman's Approach

With the above historic perspective it seems to me that Neyman (1934) in his well-known paper on survey-sampling tacitly attempted to fit the sampling theory of actual populations within the classical model of hypothetical population. This model, as it was previously (Section 2) pointed out, underlay the general statistical theory. Now as we observed in Section 4 that while 'sampling' a hypothetical population the observations are independent and are arranged in a natural *sequence*, lst observation say x_1, 2nd x_2,...*i*th say x_i...and so on. In fact the only way to distinguish between two observations apart from their variate values (x say) is by the *order* in which they appear. In this case if the expectation of each x is some unknown θ, $E(x_i)=\theta$, $i=1,\ldots, n$ say a statement such as,

'among all unbiased linear estimates (*i.e.* of the form $\sum\limits_1^n a_i x_i$) mean $\bar{x}$, of the n observations has minimum variance' is of considerable statistical relevance. Such a statement, however, in case of sampling actual population becomes of doubtful relevance as we would subsequently show.

In his 1934 paper, Neyman using the then novel Gauss-Markoff technique of least squares obtained linear unbiased least variance estimates in the sense stated in the above paragraph for mean of actual (survey) population assuming simple random sampling with or without replacement and with stratification. On the basis of his above 'optimum' estimates he compared efficiencies of different modes of stratification with appropriate sample size allocations.

The success of Neyman's work appeared so overwhelming that for several years to follow sample survey statisticians often concerned themselves with finding linear unbiased least variance estimates with the halp of Gauss-Markoff theorem, for different situations. The situation at times might have been characterised by the prior knowledge of some correlation coefficient, etc. This invariably resulted in the claims like the regression estimate, the ratio estimates (allowing for the small bias)..., etc. were linear unbiased least variance estimates for the survey-population mean, provided 'such and such' conditions were satisfied. Talk of 'most efficient estimates' become common. Ample instances of this are provided by the text books on the subject.

Again, more elaborate sampling procedures such as sampling with arbitrary probabilities and the like were put forward to reduce the variance of estimates. But essentially these elaborate sampling procedures suggested the logical inadequacy of Neyman's approach based on Gauss-Markoff theorem. It as easy to illustrate this;

Let the population consist of four individuals labelled as (1, 2, 3, 4), the corresponding variate values being (X_1, X_2, X_3, X_4). We make three draws using simple random sampling with replacement from this population. Let X-value at the first draw be x_1 that at the second be x_2 and that at the third be x_3. Now as a very special case of Gauss-Markoff theorem we can assert that of all the linear unbiased estimates $a_1x_1+a_2x_2+a_3x_3$ of $(X_1+X_2+X_3+X_4)/4$. *i.e.* $E(a_1x_1+a_2x_2+a_3x_3)=(X_1+X_2+X_3+X_4)/4$ the variance is minimised when $a_1=a_2=a_3=\frac{1}{3}$ *i.e.* for the sample mean $(x_1+x_2+x_3)/3$. But this assertion because of the existence of the labels of the individuals, (1, 2, 3, 4) unlike in the case of hypothetical population is of very doubtful statistical significance: For, making draws with some artificial device of randomisation essentially implies that at each draw we first observe the individual (label) drawn and then its X-value. With the use of these labels we can for instance calculate mean over only distinct individuals in the three draws; that is even if an individual is repeatedly drawn in three draws it is counted only once for computing the mean. Now this mean is an unbiased estimate for $(X_1+...+X_4)/4$ and can be shown to have smaller variance than that of $(x_1+x_2+x_3)/3$. But the mean over distinct individuals is an estimate clearly outside Gauss-Markoff set up and so are all the estimates which effectively utilize the labels. And in the class of all unbiased estimates which take into account individual labels none has uniformly smallest variance (Godambe 1955).

7. The New Model

The negative result mentioned at the end of the last section in fact necessitated the replacement of the hypothetical population model by a new model which is essentially based on the recognition of individual labels. In effect, for the new model a generic point in the sample-space is a subset of individuals (labels) from the population that could possibly be selected along with the corresponding variate values. (For instance in the example of a population with 4 indviduals (1, 2, 3, 4) with corresponding variate values (X_1, X_2, X_3, X_4), $(1, 2; X_1, X_2)$ is a point in the sample-space). Naturally now the parameter is the vector of all variate values associated with the different individuals in the population. In the example just referred to, the parameter is (X_1, X_2, X_3, X_4). The value of this parameter together with the sampling scheme determines the distribution on the sample-space, defined above. This new model at once provided a frame work within which different statistical concepts and tools could be brought to bear upon sampling from actual or survey populations. That is survey-sampling became an integral part of the general statistical theory. For details we refer to Godambe (1966).

The new model proved to be specially fruitful for studying the different aspects of randomisation: Bayesian non-Bayesian of various types. From a thorough going Bayesian view point it seems that the only purpose of randomisation would be to protect the assumption of exchangeability of the prior distribution*. This, at most can explain stratified simple random sampling. Another approach not quite Bayesian but close to it is to assume a class of parameterized prior distributions and to infer, about the actual population mean say, with the help of the least squares technique and the like provide by general statistical theory. This approach is not much different than the one envisaged by the authors around the turn of 19th century. (Bowley, 1906). Neither does this approach can take us far enough to interpret randomisation. Particularly unequal probability sampling (in its sufficient generality) is out of its reach. At the other extreme Neyman's thorough going frequency approach provides legitimate confidence intervals for many arbitrarily chosen modes of randomisation *i.e.*, arbitrary sampling schemes. Having many answers to a question is as bad as having no answer at all. Hence a satisfactory solution must lie somewhere in between the two views mentioned above. And already such a solution is suggested by, if not implied in, the brilliant works on randomisation by Fisher and others during 1930-1940, though these works

*In the context of the foregoing discussion this prior distribution and those mentioned subsequently would relate to some hypothetical population than an actual one. But to avoid confusion one may keep them unclassified.

were in relation to design of experiments. Fisher (1936) asserted that the inferences based on the assumption of Normal distribution were valid only because they approximately agreed with the inferences based on the distribution generated by randomisation†. This assertion can be explicated as follows:

Let us suppose that the prior knowledge (often unformalised) concerning the actual population under study suggests that the variate values in the population are approximately normally distributed. That is if these variate values are plotted on a graph paper, the graph would be approximately normal. This may be because of the following more definite assumption on the part of the statistician.

> *Assumption.* The actual population under study is a random sample from a prior distribution (or super population) which is normal.

From the above Assumption it follows that the variate values associated with any arbitrarily chosen subset of individuals from the actual population is a random sample from the super population. With this one can make inference about the actual population, its mean say, with traditional techniques provided by the general statistical theory. This inference is not based on any distribution generated by randomisation. And as I said before the Assumption implies that the inference would be valid even if one draws the subset of individuals from the population in any arbitrary manner. But the inference would be invalid if the Assumption goes *wrong* in some sense. On the other hand the same inference (or approximately so) could be validated in terms of the distribution generated by randomisation if the subset of individuals were drawn at random with some suitable device. This later validation is of course independent of the above Assumption. Thus, the randomisation has protected the inference against the possibility of the Assumption being wrong. Let me emphasise that *strictly*, when the Assumption obtains, to use the distribution given by randomisation in inference is against all canons of statistical thinking. But seldom, if ever, in practice, the statistician would be strictly unerring while making assumptions. To err is human. Yet it is absurd not to make 'what look like plausible' assumptions. Then wisely enough, one should randomise to protect the inference if the assumptions go wrong. Curiously enough the

†Here and subsequently when we say 'randomisation' without any qualification we imply 'simple random sampling'.

distribution generated by the randomisation can have inferential value *strictly* when the Assumption does not obtain and is not replaced by any alternative assumption. For then logically the situation is equivalent to that of "no prior knowledge", implying no possible conditioning of frequencies. Therefore the frequencies given by randomisation are inferentially valid.

In the above discussion I have considered a rather extreme and also a very simple situation. It is extreme in the sense that in practice a statistician would consider possibility of his assumptions going wrong not totally, but in *some respects.* He could then randomise to protect only the corresponding aspects of his inference. The situation considered above is simple in the sense that only simple random sampling is considered. But a large part of survey-practice employing very sophisticated modes of randomisation, inclusion probabilities proportional to size and the like, could also be interpreted similarly. For details a reference is Godamble and Thompson (1975).

REFERENCES

1. Bowley, A.L.(1906) : Address to Economic Science and Statistics Section of British Association for the Advancement of Science. *J.R. Statist. Soc.* 69, 548-557.

2. Dalenius, T. (1957) : *Sampling in Sweden.* John Wiley and Sonse, New Yoik.

3. Fisher, R. A. (1926) : The arrangement of field experiments. *J. Min. of Agric.* 33, 503.

4. Fisher, R.A. (1936) : The coefficient of racial likeness and future of cramiometry. *J. R. Anthropo Inst.* 66, 57-63.

5. Fisher, R. A. (1956) : *Statisrical Methods and Scientlfic Inference.*

6. Godambe, V.P. (1955) : A unified theory of sampling from finite populations. *J. R. Statist. Soc.* B, 17, 268-278.

7. Godambe, V. P. (1966) : A new approach to sampling from finite populations. *J. R. Statist. Soc.* B, 28, 310-328.

8. Godambe V. P. and Thompson, M.E. (1975) : A philosophy of survey-sampling practice. '*Foundations and Philosophy of Statistical Theories in Physical Sciences*'. II, D. Reidel Publ. Co., Dordrecht, Holland

9. Neyman, J. (1934) : On two aspects of representative method. *J. R. Statist. Soc.* 97, 513-600.

10. Oschinsky, D. (1971) : *Walter of Henley and other treaties on estate management and accounting.* Oxford.

11. Schott, S. (1923) : *Statistik.* Teubner, Leipzig-Berlin (p. 43-44)