DISCUSSION Wayne A. Fuller, Iowa State University

Dr. Koch has discussed topics that have long been of concern to statisticians. One of these, the idea of a target population was addressed by survey statisticians in the 1930's and 40's when random sampling of finite populations was being introduced. More recently discussions of "analytic surveys" again brought the topic to the surface. Most sampling texts contain some discussion of target population. On the basis of these discussions one might identify three possible objectives for the estimates constructed from a sample of a finite population.

The first would be: Estimation of a property (a parameter) of the particular finite population sampled. The parameter might be the mean, the difference between the means of two groups, or a regression coefficient. This type of inference problem is, perhaps, most natural and comfortable for the traditional survey sampler. It is the task of a number of government agencies such as the Census Bureau and the Bureau of Labor Statistics.

The second problem is the estimation of a parameter of a finite population separated by time or space from the finite population actually sampled. For example, a study of recreation activities was conducted in Iowa to predict future demand for recreational facilities. This material was requested by the State Conservation Commission as a guide for parkland acquisition, etc.

The third problem is the estimation of a parameter of an infinite population from which the finite population is a conceptual random sample. I think most will agree that scientists are often interested in inferences beyond the finite population studied. This does not mean that it is always easy to define the conceptual population of interest.

One might place the three objectives in a hierarchy, the estimation of the particular finite population parameter being the narrowest objective and the estimation of the infinite population parameter the broadest. However, a careful consideration of the problem of estimating for a second finite population seems to require a specification of the relationship between two finite populations. This in turn leads one to the infinite population concept.

When only one population is sampled it seems that the statistician can only help the subject matter specialist assemble and interpret data on which to make the judgment on comparability. On the other hand, if we have sampled a number of finite populations, for example, a number of years, we may be able to bring statistical analysis to bear on the nature of the comparability of the finite population of interest (next year). That is, one might formalize that problem by assuming that the sequence of finite populations was a realization from a common generating mechanism.

Let us consider briefly the idea of a superpopulation. One does not have to be an authority on the history of statistics or on the foundations of statistics to recognize that the ideas of superpopulation permeate the literature. For example, Fisher (1925, p. 700) in a prefatory note to his 1925 paper "Theory of Statistical Estimation" stated, "The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability." Also, little reading is required to establish the diversity of opinions statisticians hold with respect to the ideas of superpopulation. An idea of this diversity can be obtained by reading the volumes New Developments in Survey Sampling edited by Johnson and Smith (1969) and Foundations of Statistical Inference edited by Godambe and Sprott (1971).

In many of the studies of sample survey data falling within our personal experience, the investigator was interested in conclusions beyond the finite population actually sampled. As I said before, this does not mean that the investigator could perfectly specify the population of interest. If the statistician poses the question, "For what population do you wish answers?" he should be content with a rather vague answer. fact, the answer "I desire inferences as broad as possible" will be a reasonable reply in the minds of many scientists. Such an answer means that the investigator wishes a model with the potential for generalization. Given this desire, the statistician should assist in constructing models with that potential.

Treating the finite population as a sample from an infinite population is one framework which provides the potential for generalization. In fact, I believe a strong case can be made for the following position: "The objective of an analytic study of survey data is the construction and estimation of a model such that the sample data are consistent with the hypothesis that the data are a random sample from an infinite population wherein the model holds." While this statement is something of an inversion of the manner in which the traditional statistical problem is posed, it seems to be consistent with the manner in which scientific progress is made.¹

When presented with analytic survey data I believe one constructs models acting as if the data were a sample from an infinite population. (Of course one should not ignore the correlation structure of the sample data. Correlation among sample elements may arise from properties of the population or may be induced by the sample design. For example, if the sample is an area sample of clusters of households, the correlation between units in the same area cluster must be recognized in the analysis.)

A scientific investigator reports carefully the procedures, motivations, and alternative postulated models associated with the analysis. Those things considered unique in the material (the nature of the sample) are reported together with the findings for that material. The reader of the scientific report must decide if the results of the study are applicable to the reader's own problem.

Let me give a preface to my next remarks. When the originally scheduled third discussant was unavailable, it was decided to replace him with a biometrician, in order to add balance to the group of discussants. Time was short and biometricians were in even shorter supply. I was tapped for the position by a biometrician who is not attending the meetings. Hence, I feel a certain obligation to biometricians in general, if not to the absent member of that group.

Therefore, in my role as a biometrician, I would like to emphasize the importance of the knowledge of "biology" (or other subject matter fields) in model construction. Let me do this with an illustration. I have never used stepwise procedures in constructing models for empirical data. I have always felt that the subject matter person and I should actually specify an array of possible models at every step of the process. I feel that we should be better able to specify a model than a machine. This does not mean that we do not try alternative models or that we are blind to the data. Preliminary summaries, plots, and residual analyses are used. But I feel that it is important to think about the material using all available knowledge. intuition, and common sense at every step of the model building process. It seems to me that real effort is often required to persuade a subject matter person to share his knowledge with his statistical consultant. Perhaps it is because his knowledge is vague, based on analogy and conjecture. But it is precisely the kind of knowledge that should be fed into the model building process. Working together in specifying models often brings this kind of information to the surface. As Leslie Kish said last night, statisticians and statistical methods are powerful tools available to the scientist. They are not substitutes. The really successful consultant never forgets this fact. The first question, the last question, and the question at all steps between is: Does it make sense?

Dr. Koch mentioned that the variables we observe are often imperfect representations of the concepts that interest us. There are at least two levels to the problem. The first level is the failure to obtain the same value for a particular variable in different attempts to measure it. This kind of error is called response error in survey methodology and measurement area in the physical and biological sciences. If the independent variable in a simple regression is measured with error, the coefficient is biased towards zero. In the multiple independent variable case, the effects of measurement error are pervasive, but not easily described. If the error variances are known (or estimated from independent sources) there are techniques available for introducing that knowledge into the estimation procedure. I feel that this is an area that deserves more emphasis in

the "statistical methods" literature.

The second level of the problem is more subtle. Consider an IQ test. The repeatability of such tests is fairly well established and the reliability (a measure of the relative error variance) is often published with the test. Yet we realize that the mean of an individual's test scores is not perfectly correlated with that illusive concept we call intelligence. It may not even be linearly related (the scale problem). Thus, we must always be on guard against drawing incorrect conclusions by treating a variable as if it is perfectly (or even linearly) related to our concept. My colleague, Leroy Wolins, has collected a file of applied papers that he believes contain errors of the second kind.

I close, believing that the items we have been discussing will be of concern to statisticians and scientists for years to come.

FOOTNOTES

1/I believe that Kempthorne and Folks (1971, p. 507) come to this position in their discussion of Pierce.

REFERENCES

- [1] Cochran, W. G. (1946), Relative accuracy of systematic and stratified random samples for a certain class of populations. <u>Ann.</u> <u>Math. Statist. 17</u>, 164-177.
- [2] Cochran, W. G. (1963), <u>Sampling Techniques</u>. Wiley, New York.
- [3] Deming, W. E. (1950), Some Theory of Sampling. Wiley, New York.
- [4] Deming, W. E. and Stephan, F. F. (1941), On the interpretation of censuses as samples. J. Amer. Statist. Assoc. <u>36</u>, 45-59.
- [5] Fisher, R. A. (1925), Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22, 700-725.
- [6] Fisher, R. A. (1928), Book review, <u>Nature</u> 156-196.
- [7] Godambe, V. P. and Sprott, D. A. (1971), Foundations of Statistical Inference. Holt Rinehart and Winston, Tronto.
- [8] Johnson, N. L. and Smith, H. (1969), New Developments in Survey Sampling. Wiley, New York.
- [9] Kempthorne, O. and Folks, L. (1971), Probability, Statistics, and Data Analysis. Iowa State University Press, Ames, Iowa.
- [10] Madow, W. G. (1948), On the limiting distribution of estimates based on samples from finite universes. <u>Ann. Math. Statist</u>. 19, 535-545.