ON AN IMPORTANT LIMITATION TO THE USE OF DATA FROM SAMPLES

BY

MORRIS H. HANSEN

Bureau of the Census

AND

W. EDWARDS DEMING

Bureau of the Budget, Washington

Travail présenté à la 26e Session
de l'Institut international de statistique
Berne 1949

TIRAGE À PART DU
BULLETIN DE L'INSTITUT INTERNATIONAL DE STATISTIQUE
TOME XXXII – 2e LIVRAISON
BERNE 1950

IMPRIMERIE STÆMPFLI & CIE., BERNE

ON AN IMPORTANT LIMITATION TO THE USE OF DATA FROM SAMPLES

by Morris H. Hansen

Bureau of the Census

and

W. Edwards Deming

Bureau of the Budget, Washington

The title of this article might well be «On an important limitation to the use of data from censuses», or «On an important limitation to the use of data of any kind». The central theme of the authors is built around (i) the distinction between data and the uses of data; (ii) the distinction between a standard error of a result obtained by sampling, and the risk of making a wrong forecast.

The primary purpose of conducting any survey, sample or complete census, or the purpose of carrying out any experiment, is to provide more knowledge of the past—knowledge that will help to provide a basis for rational action in the future. Action is based on a prediction of some kind. For example, a man plants potatoes because he predicts that they will bring a higher price than an alternative crop that he might grow. A certain amount of money is appropriated for some phase of a school programme and the amount is based on a prediction that a certain number of children will be in a particular age-group next year or five years hence.

Data concerning the people of a country—how many, where and how they live, their distributions by age and other personal characteristics, their education, their employment, their housing, their production both agricultural and non-agricultural, their incomes, their purchases, businesses, transportation, inventories, prices paid and received for various commodities, the condition of their crops, and a host of other types of information are today considered to be vital to intelligent administration of government and business. Current and indispensible information on essential aspects of our economy is obtained today with speed, reliability, and efficiency undreamed of a few years ago, owing to advancing statistical methodology, particularly in sampling, mechanical techniques, interviewing, and the construction of questionnaires. Two examples are furnished by the Monthly Report on the Labor Force in the United States, and by the Quarterly Report in Canada. These reports supply periodic information on characteristics of the labor force, the employed, the unemployed, and of people not in the labor force. The data in both cases are obtained from samples of only about 25 000 households, and the figures are ready for administrative use soon after the date to which the survey applies. Other examples could be drawn from many parts of the world.

It is a mistake to confuse the reliability of such data with the reliability of forecasts purporting to predict what the volume of unemployment will be six months

hence, or how many people will purchase refrigerators, or what the yield of wheat will be, or what the price of butter will be.

A survey is regarded as important only when the uses of the data are important. When a wrong decision will endanger the health or property or incomes of people, or rob them of any other values, a survey that will assist a rational decision assumes importance. Its results make headlines. A monthly collection of prices of foodstuffs is important, for example, when these prices form the basis for intelligent resolution of important problems in production, distribution, wage-adjustments, or something else.

Because of this close relationship between the importance of data and the importance of a decision that is to be based on the data, too often the standard error of a result obtained by sampling has been confused with the standard error of a forecast that is based, partially at least, on this result. Good forecasting and good luck in forecasting have sometimes wrongly been attributed entirely to good data, i. e., good sampling. Bad forecasting and bad luck in forecasting have similarly sometimes been wrongly attributed entirely to poor data, i. e., poor sampling. It is damaging to quantitative studies of science to permit this confusion to be perpetuated.

The data obtained from a survey may be reliable, and the results may be useful for the problem which the survey was expected to resolve, yet a decision based on the data may turn out to be unfortunate. One cannot see beyond the horizon, and if a decision based on a survey turns out to be unfortunate, the survey, however efficient and reliable, has, in a sense, failed, because its purpose has failed. It is therefore extremely important that the design of the questionnaire, the timing, and the precision of the results of a survey be tied closely to the needs of the forecasts that will be made and which will affect decisions.

A forecast is an attempt to paint a picture of the future. A sample of the past may be of great help, or it may not. Much depends on the material regarding which the forecast is made. Human beings and the weather present certain problems; physical materials present others. The success of a forecast may be expected to depend not only on the type of material concerning which the forecast is made, but also on the length of time covered, and on the assumptions respecting continuing relationships or changes in relationships, the method of forecasting, and the accuracy of the data on which the forecast is based. Seriously inaccurate data may be the primary cause of a poor forecast. Obviously, however, the existence of reliable data does not ensure a reliable forecast: the reliability of the sample is only one of several factors affecting the success of the forecast.

When we deal with marbles the problems of forecasting assume simplicity. A large random sample of marbles drawn from a bowl containing black and white marbles enables us to make pretty accurate predictions concerning the percentages of black marbles that will appear in future random drawings. Our prediction is good for all time in the future: it does not depend on the weather or whether someone changes his mind.

Recent advances in statistical methodology have received widespread acceptance and respect from specialists in subject matter and even from the public. It is a fact, though, that some of this acceptance and respect is uncritical. The results of a sample are too often accepted for more than they really are. It is as important

now to reach an understanding concerning the distinction between the reliability of a set of data and the reliability of a forecast based on these data, as it was important a few years ago to spread knowledge of the fact that sampling is no longer guesswork but a science that had reached maturity.

In the hands of a specialist, sampling today can be depended on to yield, within calculable limits, the same results as a complete census would have yielded if it had been taken at the same time, with the same definitions, same questionnaire, same interviewers, given the same training, supervision and incentives, and all else held constant as well. The difficulties that are apparently encountered in trying to carry out these stipulations are avoided in the interpretation of sampling error, given further on.

In probability sampling the chief results of a sample are accompanied by a measure of precision known as the standard error of sampling. It may be well to pause and see just what a standard error of sampling of (e. g.) 2 percent means. A survey has been carried out with the aim of ascertaining how many homes in a certain area are mortgaged, and the result is stated as 75 360 (1 ± 0.02), the 0.02 being the standard error of sampling. The interpretation is this [1]: (a) if a complete census of a population were taken; and (b) if a moderately large probability sample of people, households, farms, business establishments, manufactured articles, or other units covered in the census were independently designated afterwards (as by sampling the cards punched from a census); and (c) if tabulations were carried out separately for the complete coverage, and the sample; and (d) if the procedure of independently designating a probability sample and then identifying it in the census and tabulating it were repeated again and again, then close to 95 percent of the sample ranges $X \pm 2S_x$ would overlap a particular result of the complete coverage, and more than 99 percent of the ranges $X \pm 3S_x$ would do so. The symbol X denotes an estimate made from a sample, and S_x denotes an estimate of the standard error of X. Both X and S_x vary from one sample to another.

Thus the standard error S_x provides a measure of the range of sampling error. It measures the range of error of the sample estimates that arises both from the particular kind and size of sample that is used, and the particular formula of estimation that is followed. By taking a large enough probability sample S_x can be made as small as desired.

There have been many instances when a sample has been selected subsequently from the units (people, households, farms, business establishments, manufactured articles) concerning which information was obtained by a complete coverage. One example was the 1 : 1000 sample of the schedules from the census of Japan in 1921, which was selected and tabulated after the earthquake of 1923 had removed hope of a complete tabulation. Another is the Y-sample by which the census of India of 1941 was tabulated (intended to be a 1 : 50 sample). Countless tabulations of samples of card-files provide more examples in which the sample was not designated prior to the collection of the results whence the sample was drawn. There have

[1] The authors are aware of the fact that this interpretation is only one of many ways of saying the same thing: also that some of our conditions are sufficient but not necessary. The interpretation given here is thought to be adequate and simple. Here it is assumed that X is an estimate of the total or average value of some characteristic of the finite population that is sampled.

been numerous instances in which a probability sample has been designated and carried through prior to and independently of a census operation, and the sample units later identified in the census and estimates prepared therefrom. Thus, the interpretation given by the authors may be subjected to experimental test, as it has been many times.

Consequently when we carry out a probability sample according to the rules required by theory, and prepare therefrom an estimate X, we are entitled to assert with 95 percent probability that the range $X \pm 2S_x$ would cover the result of a complete census carried out at the same date as the sample, with the same questionnaire, same interviewers, same supervision, and all else equal.

It should be borne in mind that complete censuses are not always ideal. Complete coverage is difficult to accomplish, and there are a host of nonsampling errors present in all surveys, sample and complete. Only one example need be cited to illustrate the point. The 1940 census of population in the United States included a question asking whether each person was employed in emergency (relief) work, the aim being to learn something about the characteristics of people so employed. In this case, the census was not the only source of information concerning the total number of people on the payrolls of emergency work. This number was known from the accounting records, and the census was found to be deficient by approximately 25 percent. The 1 : 20 sample was likewise deficient by approximately 25 percent.

Thus, even in the collection of «factual» information, it will often appear that the results collected by two different methods, in two different surveys, both being adequately large samples, will differ significantly even though we have attempted to make the conditions essentially the same. It will often be true, also, that an estimate made from a sample survey and a complete coverage, when the two are carried out under supposedly the same conditions, will give results that differ by more than the sampling error even when the sampling has been carried out precisely in accordance with the probability design. Conditions can never be exactly duplicated from one survey to another, and this is why the authors prefer to interpret sampling errors in the manner explained above.

It is a fact, moreover, borne out in experience, that sometimes when a probability sample of households or other units is designated as a portion of a complete coverage, and covered simultaneously with the complete coverage, estimates made from the sample show differences when compared with the complete coverage. Usually these differences are very small. Even when small, they are sometimes nevertheless well beyond the bounds of sampling error. Sometimes they are uncomfortably large. Differences may arise either because of a sampling bias, or because of differences in response that arise from a differential or preferential treatment of the households in the sample.

One example will be cited. In the census of population in the United States in 1940, certain lines on the schedules on which the names of the population were written were designated as sample lines by ruling them in heavy black lines. Thus 1 person in 20, on the average, was in the sample. This sample broadened the scope of the census, as additional information was recorded for every name in the sample. As a matter of fact, this sample provided the basis for tabulating at low cost many of the results that could have been tabulated from the complete coverage. On comparing the results it was found that negligibly small but definitely identifiable

differences appeared between the sample and the complete coverage. These differences may be attributed both to small biases in the method of sample selection and to differential treatment of the people in the sample.

Sometimes such differential treatment reflects a more careful job on the sample than on the complete coverage. Sometimes it reflects less careful treatment. At other times it simply reflects an interaction between the main survey and additional questions asked only in the sampling units. A sample survey may be carried out with a relatively small number of skilled interviewers, who, with proper training and supervision, should produce results of higher quality than could be expected from a complete census. In fact, no one would think of carrying out a detailed investigation of family or farm expenditures in a large population except by a sample. Leaving aside considerations of cost and speed, a complete coverage on such subjects would be tremendously expensive, difficult to control, and consequently subject to excessive amounts of human errors of reporting, nonresponse, incomplete coverage, etc. However, these statements are not meant to imply that every sample survey is carried out carefully: it is done so only by effort and proper circumstances.

The distinction between probability sampling and judgment sampling should perhaps be made clear here. A probability sample may be adjusted by its size so that the results will agree as closely as desired with the results of a complete coverage taken under the same conditions. Moreover, the actual range of sampling error can be calculated after the results are in, and it can be interpreted with confidence. With judgment sampling, however, there is no certainty that the results will approach the census values with increasing size of sample, and there is no objective way of calculating any ranges of error. Ranges of error can only be conjectured on the basis of expert judgment. Experience indicates that often such judgments are good, but that sometimes the best judgments turn out to be bad. Probability sampling reduces the errors arising from sampling to a question of mathematics rather than of judgment. In probability sampling, expert judgment is indeed used but it is used in those stages where it can be most effective, e. g., in deciding the most efficient and practicable sampling units, the probable effectiveness of any proposed method of trying to trim nonresponse. It is to be noted, though, that in probability sampling, judgment is not used to select the sampling units that are to be interviewed or tested.

In summary, in spite of all the methodological advances that have been made in sampling, a sample survey can at best only give a picture of the past, not of the future. If the survey is well designed and carefully carried out, it is an accurate picture of the past—i. e., accurate enough for the purpose intended. Its precision will be calculable from the returns: and the possible effects of alternative questionnaires and methods of interviewing may be studied by means of auxiliary samples. Nevertheless, a sample survey is not a picture of the future, but only of the past.

To cite an example [1], the three principal polling organizations in the United States missed the election last November. In the mind of many people, this failure was a failure of sampling. Even had the best methods of sampling been used, it should be recognized that no sample could do any more than reflect, as of a particular date, what proportion of the adult population had certain stated attitudes, had

[1] The authors use this example, not to be critical of polling organizations, but because the example, although recent, has been well publicized.

behaved in specified ways in earlier elections, etc. A forecast, on the other hand, makes a statement about the net effect of people's actions on some future date. In our present state of knowledge about people, no sample—not even a complete coverage—can do this with certainty. Nevertheless, good samples can help. With advancing knowledge of psychology and of questioning and interviewing, and with advancing knowledge about the reasons why people change their minds, samples will be of much greater help in making accurate forecasts in the future than in the past.

Résumé

Sur une importante limitation de la méthode des sondages

Les relevés statistiques ont pour but de déterminer les caractères des masses statistiques. Le sondage est un mode particulier de relevé, qui consiste à choisir au hasard un certain nombre d'unités au sein d'une population. Chaque sondage est exposé à deux sortes d'inexactitudes. Les erreurs systématiques, qui surviennent aussi dans les relevés complets, ont leur origine dans la disposition même de la statistique. Les erreurs aléatoires forment la deuxième catégorie d'inexactitudes; celles-là tiennent au fait que le sondage ne touche qu'une partie de la population. Elles se distinguent toutefois des sources d'erreurs de la première catégorie par le fait qu'elles suivent certaines règles. Les fluctuations fortuites sont mesurables; elles dépendent de l'ampleur du relevé partiel et de la variabilité des valeurs particulières. Si X désigne une estimation fondée sur un sondage et S_x l'écart-type de la valeur estimée, il y a une probabilité de 95 % que la valeur tirée du relevé complet soit située dans l'intervalle $X \pm 2S_x$. Si le sondage est assez étendu, la mesure S_x des fluctuations fortuites deviendra d'autant plus petite; autrement dit, le chiffre calculé par sondage peut se rapprocher, avec l'approximation désirable, de la valeur correspondante obtenue par un relevé complet effectué dans les mêmes conditions.

Les auteurs relèvent que les nombres tirés d'un sondage ne peuvent servir directement de base à des prévisions. On souligne qu'un relevé statistique complet ou partiel ne s'attache qu'aux observations, aux phénomènes ou aux expériences se rapportant au passé ou au moment même du relevé. Il en est tout autrement pour l'avenir, qui subit l'influence de facteurs imprévisibles. De bons sondages donnent évidemment de précieux points de repère sur l'avenir, mais il faut se garder d'exiger du sondage plus qu'il ne peut fournir. Le sondage, utilisé à la place d'un relevé complet, ne fournit qu'une image schématique de la population. Le relevé partiel, comme le relevé complet reste entaché d'inexactitudes du genre des erreurs systématiques, provenant de la non-réponse, de la malfaçon du questionnaire ou des insuffisances de l'enquête statistique. Les fluctuations fortuites auxquelles un sondage est sujet ne doivent pas être confondues avec les insuffisances inhérentes à une prévision; ces dernières ne sont pas mesurables et subissent les influences les plus diverses.