122.

United Nations
World Population Conference
Belgrade, Yugoslavia
30 August to 10 September 1965

Nations Unies
Congrès mondial de la population
Belgrade (Yougoslavie)
30 août - 10 septembre 1965

THEORY OF SURVEYS TO ESTIMATE TOTAL POPULATION

by

W. Edwards Deming
Washington

and

Nathan Keyfitz
Chicago

THEORY OF SURVEYS TC ESTIMATE TOTAL POPULATION

By

W. Edwards Deming
Washington

and

Nathan Keyfitz
Chicago

Purpose. The purpose of this paper is to discuss some of the statistical problems encountered in estimating by sampling the total number of a population, without benefit of a previous census, and to present a device for this purpose which may have other uses as well. We will speak of two kinds of situation: (1) the population is fixed, each person being nominally attached in some recognizable manner to a fixed location, such as a dwelling unit; (2) the population is mobile--here today, somewhere else tomorrow. Always in addition to the population that is fixed, a certain number of people are not identified with any fixed location; this applies in countries that have regular modern censuses as in those that have never taken a census. We make only brief mention of the fixed population, as it is already treated in books. We introduce some theory for the moving population, in the hope that knowledge of the theory will encourage practical application.

Sampling to estimate a fixed population--the area principle. The main difficulties are about the same for the use of sampling as they are in conducting a first complete census. The basic requirement for either a complete census or a sample of a fixed population is a frame. A frame is a list of sampling units, the totality of which covers the population to be counted. For a fixed population a good set of maps constitutes an implicit list or frame.

Without a frame, there can be neither a complete census nor a sample. Without a frame there would be no way to allot work to interviewers, nor any way to know whether or when the work of taking the census is completed, nor is there any way to carry out a control.

The sampling unit for a census of a fixed population is most commonly an element of area. The enumeration district, with clearly defined boundaries and designed to contain 500 to 1000 persons, is an example. Enumeration districts may be of any shape or size--they may contain a small number of people or a large number of people. There is of course preference for uniformity insofar as the enumeration district is the work load for one enumerator, and insofar as it is a sampling unit from which an estimate of the total number of inhabitants is to be made. Pre-existing well-defined and well-established administrative units, or city blocks, may be used. We also call attention to an instance of square areas, 3 km. by 3 km., marked off by points designated as their corners, and conspicuous enough so that the enumerator might know whether he is inside the designated sample area.*

Area samples may become complex as their designers strive for efficiency, to achieve a desired precision at minimum cost. They employ stratification, ratio-estimates, regression, multiple stages, composite estimates, and a host of other ingenious devices. But underneath all of these remains the fundamental principle of the simple random sample: that out of N areal or other units defined in the population _n_ are chosen at random for enumeration. In a large-scale sample-survey, prepared in stages, it is necessary to delineate only a fraction of the N units, but there are still N units in total, in hundreds of strata. Every acre in the territory has a chance of n/N of inclusion, likewise every household has probability n/N, every store, and every individual resident.

As the probability of selection is n/N, it is possible to estimate the total number of people in the frame merely by multiplying the number of people in the sample by N/n; and similarly to estimate the total number of houses in the frame by multiplying the number of houses in the sample by N/n; and similarly to estimate the total number of acres, households, stores, or individuals. The raising factor N/n is simply the reciprocal of the probability of selection.

Point sampling. We come now to a type of sampling in which the factor N/n is not built into the design, but must be estimated. Consider a situation where the population is mobile, not attached to any areal unit. Area sampling could still be used if the people would stay in one place long enough to be counted on a sample of areas. In the circumstance that there are no maps, or that the people are too mobile, area sampling is impossible. Here we are driven to a sample of points, and it is necessary to seek an entirely different basis of design and estimation than we have just described.

Sampling to estimate a mobile population from a set of points. The sample of points takes advantage of the mobility of the population. The more people are moving, and the more rapidly they are moving, the more efficient the point sample will be. It turns an obstacle into an advantage. Suppose indeed that people are in motion in such fashion that they all have the same chance of passing every point. The first step would be to place enumerators at a number of random points; they might be put down by helicopter or make their way on foot. The techniques for ensuring randomness would be those well known in crop cutting. The enumerators would remain at their stations for a given period of time--say a day or a week. Suppose that they encounter n_1 persons in total, and that to each one they give a card or a button to show that he has been enumerated. This enumeration would constitute the first round.

*J.F. Holleman, Experiment in Swaziland, Swaziland Sample Survey 1960 (Institute for Social Research, University of Natal).

The enumerators would then be allotted to a second set of random points in the same territory, and they would again enumerate all persons that they encounter, establishing, among other items on the questionnaire, whether the person has been enumerated before. Suppose that on this second round they enumerate n_2 persons in total, including n_{12} persons that had been enumerated before. The three numbers, n_1, n_2, and n_{12}, provide an estimate $n_1 n_2/n_{12}$ of the population of the territory, along with an estimate of its variance. In fact, n_2/n_{12} is the raising factor by which we multiply n_1, just as we would use the multiplier N/n in areal sampling.

<u>Point sampling with three rounds</u>. For three rounds we should need the additional notation:

n_3 the number of inhabitants enumerated at all posts at the third round

n_{23} the number of inhabitants enumerated in the third round that were also picked up in the second round

n_{123} the number of inhabitants enumerated on all three rounds

Let N be the (unknown) number of mobile inhabitants in the territory. Then the following estimates of N are available. The estimate $\hat{N}$ in Eq. 6 combines all the information from the 3 rounds.

$$N_{12} = n_2(n_1/n_{12}) = \frac{n_1 n_2}{n_{12}} \qquad (1)$$

$$N_{13} = \frac{n_1 n_3}{n_{13}} \qquad (2)$$

$$N_{23} = \frac{n_2 n_3}{n_{23}} \qquad (3)$$

$$N_{231} = \frac{n_2 n_{23}}{n_{123}} \qquad (4)$$

$$N_{132} = \frac{n_1 n_{13}}{n_{123}} \qquad (5)$$

$$N = \frac{n_1 n_2 + n_1 n_3 + n_2 n_3 + n_1 n_{13} + n_2 n_{23}}{n_{12} + n_{13} + n_{23} + n_{123} + n_{123}} \qquad (6)$$

The uncertainty attributable to accidental independent variations, including sampling, could be ascertained by analysis of the variance in these quantities between individual points or by short-cut methods similar to those that have been developed for area sampling.

<u>Hypothetical example of point sampling</u>. In order to see how the calculations work out in practice, suppose that there is

a migratory population of about 100,000 persons, and suppose that enough points be chosen for the point sample so that about 5% or 5000 persons would be included in the first round. On the second round an expected 5000 would again be included, of whom about 250 would be repeats from the first round. On the third round, there would be 5000 in all, including 250 repeats from the second and 12 from the first. In terms of the first two rounds of the preceding model, the expected values are $n_1 = 5000$, $n_2 = 5000$, $n_{12} = 250$. Substituting these expected values in Eq. 1 gives

$$N_{12} = n_1 n_2/n_{12} = 5000 \times 5000/250 = 100,000.$$

On binomial probabilities, the coefficient of variation of the estimate of N would be about 9%. An area sample that contains 10,000 persons (equal to $n_1 + n_2$) would do better than this point sample if it were arranged in efficient clusters.

Sketch of derivation of formulae. A rough estimate of the standard error in the estimate of N could be obtained from 3 rounds by comparing N_{12}, N_{13}, N_{23}. A modicum of effort in design would provide replication, by which, in each round, there would be 2 or preferably 3 to 10 replications. Each pair of replicates provides an estimate of the standard error, by formulas that are well known.

In order to gain some insight into the formulas that might be expected under one set of possible circumstances, suppose that binomial probabilities apply, which is to say that every person in the mobile population has the chance p of being enumerated in any round. Then for the first round, the chance of encountering exactly n_1 persons would be $\binom{N}{n_1} q^{N-n_1} p^{n_1}$. For the second round, the chance of encountering $n_2 - n_{12}$ persons among the $N - n_1$ who had not been enumerated before is $\binom{N-n_1}{n_2-n_{12}} q^{N-n_1-n_2+n_{12}} p^{n_2-n_{12}}$. The chance of encountering n_{12} among the n_1 is $\binom{n_1}{n_{12}} q^{n_1-n_{12}} p^{n_{12}}$. If these three probabilites are independent, the likelihood of the sample consisting in all of $n_1 + n_2 = n_1 + (n_2 - n_{12}) + n_{12}$ persons is the product of the three separate probabilities.

We now take the logarithm of the likelihood, but put it into a more tractable form by replacing all factorials by the Stirling approximation, in which N! is proportional to $N^{N+\frac{1}{2}}/e^N$, and hence $\ln N! = K + (N + \frac{1}{2}) \ln N - N$. As we have no interest in the likelihood itself, but only in its derivatives with respect to N and p, we need not concern ourselves with terms not involving these variables. Hence the logarithm of the likelihood, ln L, is

$$\ln L = K + (N + \tfrac{1}{2}) \ln N - (N - n_1 - n_2 + n_{12} + \tfrac{1}{2}) \ln (N - n_1 - n_2 + n_{12}) + (2N - n_1 - n_2) \ln q + (n_1 + n_2) \ln p. \qquad (7)$$

The derivative of ln L with regard to N and to p when equated to zero gives

$$\frac{\partial \ln L}{\partial N} = \qquad \ln N - \ln(N - n_1 - n_2 + n_{12}) + 2 \ln q = 0 \qquad (8)$$

$$\frac{\partial \ln L}{\partial q} = \frac{2N - n_1 - n_2}{q} - \frac{n_1 + n_2}{p} = 0 \qquad (9)$$

terms in 1/N being omitted. It follows from Eq. 8 that

$$q^2 = 1 - \frac{n_1 + n_2 - n_{12}}{N} \qquad (10)$$

and from Eq. 9 that

$$\frac{p}{q} = \frac{n_1 + n_2}{2N - n_1 - n_2} \qquad (11)$$

or

$$p = \frac{n_1 + n_2}{2N} \qquad (12)$$

Substitution of p from Eq. 12 into Eq. 10 gives

$$\left[1 - \frac{n_1 + n_2}{2N}\right]^2 = 1 - \frac{n_1 + n_2 - n_{12}}{N}$$

or

$$\hat{N} = \frac{(n_1 + n_2)^2}{4n_{12}} \qquad (13)$$

The maximum likelihood estimate of N, which we designate N, is thus the square of the total of the sample in the two rounds divided by four times the common portion, $\frac{(n_1 + n_2)^2}{4n_{12}}$, and of p is $\hat{p} = \frac{2n_{12}}{n_1 + n_2}$. The relative merits of Eq. 13 and Eq. 1 depend on the degree to which the binomial conditions are fulfilled.

Point sampling as a residual method. If one were in possession of a good frame for an area sample, he would use point sampling only for that portion of the population that is mobile and elusive. In many circumstances, one would combine the two methods. In fact, they are combined in the United States Census, which can be thought of in principle as a 100% sample of areas, plus a 100% sample of points for the transient population. The area sample consists in the division of the territory of the entire country into enumeration districts which are assigned one to each enumerator who is instructed to cover it solidly in respect of all residents. The point sample is enumerated on a single night, called T-night; enumerators are stationed in the lobbies of known hotels, flophouses, and other points which transients pass. In the United States the mobile population is a small enough proportion of the whole that it ordinarily is dealt with as a residual in censuses and disregarded in sample surveys.

Combination of a sample of areas and a sample of points. The combination of area and point sampling would require an initial division of the whole territory (say of a country) into suitable areas and the usual random selection from these, perhaps at several stages. In each one of the areas an enumerator takes his post and counts the people that reside in or pass through the area over a stated period of time. These areas would be situated in the various geographic zones in the country whose population is to be counted, and where the enumerators happened to be located at any moment as they go their rounds would constitute random points. The persons found within the sample areas during the time of enumeration would be treated differently according as they are (a) fixed residents of the area or (b) transients encountered by the enumerator as he makes his rounds.

The enumeration at a sample of points must be repeated another day to make at least two rounds. The procedure in Round 1 is to enumerate (in addition to the regular inhabitants of the designated areas) those that happen to be passing through, and to provide, in some manner, each person enumerated as a transient with identification so that he can be recognized if he is picked up in the second round, or in the third round.

Theory needed for the combined sample. Although the theory of area sampling is now expounded in books, and there is also a considerable literature on point-sampling, combination of the two has so far, to our knowledge, not been suggested. The advantage of the combination is not in the reduction of variance so much as in cancellation of biases. The area sample gives too small a chance of inclusion to persons that are mobile, while the sample of points gives to any person a probability of inclusion proportional to his movement. The combination suggested in the preceding section, by which the enumerator counts not only people resident in selected areas, but also outsiders that happen to be in his area while he is on his rounds, might be designed to give approximately equal probabilities to all. Any combination of probabilities would serve the purpose, if the probabilities could be evaluated. The present paper may thus be of interest even in countries that have highly developed census methods for the stable population, but which place less emphasis, with possible bias from undercount, on mobile elements.

The drawback that point sampling suffers from is the need to include a sufficient proportion of the population in each of two or more rounds. Point sampling is to be used only in cases where area sampling alone is impracticable. As there is a mobile population in every country, point sampling offers a general contribution to statistical method.