YRB 65-07

UNITED STATES DEPARTMENT OF AGRICULTURE
AGRICULTURAL ECONOMICS
STATISTICAL REPORTING SERVICE

Miscellaneous Report on Methods Research

"A Random Sample Using Limited Mail and Non-Response Interviews"

by

Charles E. Rogers

Research and Development Branch
Standards and Research Division

March 1965

Introduction

Often a simple random sample is desired from a population or substratum to produce data for a specified characteristic. In many such situations, a combination of mailed questionnaires and interviews may be used to produce the data at less cost than that by interview alone. This method is especially desirable for data which may be collected with equivalent quality by either mail or personal interview. One generally used method is to draw a random sample and interview all non-respondents to a mailed questionnaire. However, this leaves a variable number of interviews to be conducted and all must be completed.

This paper presents a method which will produce a random sample when the number of interviews must be predetermined. The method might also be adapted to assure a random sample when interviewing may be only partially completed. The sample size depends upon the number of interviews and the actual rate of voluntary returns by mail for a specific subset from a random ordering of the total list. A weighting together of respondent and non-respondent groups is not required in the estimation and analysis.

Sampling Method

A population of N elements is listed so that each element appears once and the elements are mutually exclusive for the characteristic to be estimated. If these N elements are randomly ordered and then serially numbered, any consecutive n elements from a random start will give a simple random sample of size n. Essentially the same effect might be created by selecting a random sample ordered by draw or by selecting a systematic sample from the original listing and randomly ordering the n elements selected.

Notation:

- N elements in population
- n elements are selected as potentially in the sample
- t_1 interviews can be made
- t_2 mail returns usable in the final ordered sample
- $t = t_1 + t_2$ total usable returns in final sample, and $t \leq n$
- p is the probability of a mailed questionnaire being returned
- q is the probability of a non-response from mailed questionnaires
- $q = 1 - p$

The general procedure considered is to mail a questionnaire to each of the n selected elements and to interview t_1 non-respondents to this mail survey. The number of non-respondent interviews is limited to t_1, for example, the decision is made in advance that only 50 interviews will be taken. In order to preserve randomness in the final sample, the t_1 interviews are the first t_1 non-respondents on the ordered list. The estimate will then be made using the specified characteristic for the ordered elements until a non-respondent appears and no more interviews are available. This will constitute a random sample of $t_1 + t_2$ elements.

The sample size to be initially selected (n) is dependent on the number of interviews to be taken (t_1) and the expected rate of mailed returns (p). In order to simplify the decision on initial sample size, the assumption will be made that (1) when less than t_1 non-respondents appear, potential information is lost, and (2) the required sample size to give the desired precision in the estimate is slightly smaller than n. The expected rate of mailed returns must be estimated in advance and the validity of this estimate is important in mailing enough questionnaires to secure maximum information but as few extras as possible since these may have to be discarded.

Illustration

Consider a numerical example where a sample of size 70 is needed and the maximum number of interviews is fixed at 50 by funds available.

$$N = 2000$$
$$t_1 = 50$$
$$p = 1/3$$
$$q = (1 - p) = 2/3$$
$$n = \frac{t_1}{q} = 75$$

From the 2000 elements in the population, 75 will be selected by some random process (generally a somewhat optimistic value for p is assumed to allow for variation in the rate of mailed returns since n varies inversely as 1 - p).

These 75 elements are ordered by draw if initially selected by a random number process for each element, or randomly ordered if selected systematically. Once the elements are randomly ordered and identified, this order must be maintained throughout the sampling process. After the 75 questionnaires are mailed and returns checked off, the first 50 non-respondents are interviewed. Returns usable for estimation become the first $t_1 + t_2$ questionnaires in the ordering, where t_2 is consecutive mailed returns up to a non-respondent who is not interviewed.

Consider the 75 randomly ordered elements $E_1E_2E_3$---------$E_{68}E_{69}E_{70}E_{71}E_{72}$---------$E_{75}$. If only 21 are returned by mail and 50 interviews are to be used, the first 50 of the serially numbered non-respondents would be selected and the last or 50th interview might be the 69th element in the ordering. Element 70, 71, etc., which were also returned by mail will be used as available until the next non-respondent element in the ordering is found. Any mailed returns after one non-respondent is missed must be discarded so the sample may be summarized as a random sample. The mean and its variance being estimated in the usual way:

$$\bar{X} = \sum_{i=1}^{t} X_i/t$$

$$V(\bar{X}) = \sum_{i=1}^{t} (X_i - \bar{X})^2/t\,(t - 1)$$

Discussion of Method

In order to assure validity of later proofs, the assumption must be made that the size of sample eventually used is stochastically independent of the value of the characteristic for the sample drawn and ordered.

In most sampling for agricultural characteristics, it is believed that the value of the characteristic may enter into the decision of the respondent to return or not return a questionnaire. This fact allows two sources of variation, due to characteristic value, in the sample size under this system.

(1) The random ordering may affect sample size through response rate if either large or small valued elements tend to cluster in the ordering even though the sample mean may equal the population mean.

(2) The response rate (and through it the sample size) may be affected by the difference of the mean of the particular random sample from the mean of the population.

This effect of characteristic value on sample size in the individual sample must not prevent stochastic independence if the assumption is to be fulfilled. Two characteristics of the sample must be considered.

(1) The selected sample is merely one sample from all possible samples in the population since it was randomly selected.

(2) This sample ordering is one from all possible orderings since it was a random process.

The joint effect of these random processes should also be random and provide logical basis for stochastic independence as desired.

From another viewpoint, the sample estimate may be considered as consisting of two parts, the mean of the interview portion and the mean of the mailed portion. These means are, in effect, weighted by the respective proportions of the usable sample falling in each. Let: $\bar{X} = \hat{W}_1\bar{X}_1 + \hat{W}_2\bar{X}_2$ where $\hat{W}_1 = \frac{t_1}{t}$ and $\hat{W}_2 = \frac{t_2}{t}$.

Further let $\hat{W}_1 = W_1$ and $\hat{W}_2 = W_2$ when the entire sample of n elements is used for estimation. When all of the elements are used, the sample may be considered as fixed in size and the mean estimator is known to be unbiased and have minimum variance. The difference in weights may be considered as adding a component of variation. The fixed size sample estimator may be written as $\bar{X}' = W_1\bar{X}_1 + W_2\bar{X}_2$.

The difference (D) is:

$(\bar{X}' - \bar{X})$ or $(W_1\bar{X}_1 + W_2\bar{X}_2) - (\hat{W}_1\bar{X}_1 + \hat{W}_2\bar{X}_2)$. Since $W_1 = 1 - W_2$ and $\hat{W}_1 = 1 - \hat{W}_2$ this difference reduces as follows:

$[\,(1 - W_2)\,\bar{X}_1 + W_2\bar{X}_2\,] - [\,(1 - \hat{W}_2)\,\bar{X}_1 + \hat{W}_2\bar{X}_2\,]$

$\bar{X}_1 - W_2\bar{X}_1 + W_2\bar{X}_2 - \bar{X}_1 + \hat{W}_2\bar{X}_1 - \hat{W}_2\bar{X}_2$

$\hat{W}_2\bar{X}_1 - W_2\bar{X}_1 + W_2\bar{X}_2 - \hat{W}_2\bar{X}_2$

$(\hat{W}_2 - W_2)\,\bar{X}_1 - (\hat{W}_2 - W_2)\,\bar{X}_2$

and $(D) = (\hat{W}_2 - W_2)\,(\bar{X}_1 - \bar{X}_2)$

With the random ordering these two quantities $(\hat{W}_2 - W_2)$ and $(\bar{X}_1 - \bar{X}_2)$ should be independent and the expected value of D equal to zero. However, variation in D may add to variation in the mean estimate by the amount of D^2. Hence, this estimator will not be a minimum variance estimator. In general, D^2 is likely to be quite small even if n is less than 100 since $(\hat{W}_2 - W_2)^2$ will usually be small.

Obviously, t (the total usable sample) may vary from t_1 to n. In the "Annals of the Royal Agricultural College of Sweden," Vol. 17, Sandelius has shown that for finite populations and non-sequential sampling the usual mean estimate and its variance are unbiased estimates of the parameters even though n is a random variable. He further shows that these unbiased estimates are probably not "best" estimates as defined in cases of fixed sample size since unique "best" estimates (in terms of minimum variance) generally do not exist for samples which vary in size. However,

they are conditional best linear unbiased estimates and provide the most logical estimation procedure. The proof of unbiasedness consists of using a given sample $(x_1 x_2 \text{----------} x_n)$ with θ a parameter to be estimated.

Let E_n be the conditional expectation for fixed n and $g = g(x_1 x_2 \text{----------} x_n)$ be a function of the sample values with the property $E_n g = \theta$ so that g is a conditional unbiased estimate of θ. Then it can be shown that $Eg = E(E_n g) = \theta$. The mean is such a parameter and its estimate has the necessary property for unbiasedness independent of the random variable n. Further, let $\text{Var}_n g = E_n(g - \theta)^2$ and let $\text{Var } g = E(g - \theta)^2$ exist; then $\text{Var } g = E E_n(g - \theta)^2 = E \text{Var}_n g$. When n is restricted to values between given integers, the existence condition will be satisfied. Now let $h = h(x_1 \text{----------} x_n)$ be (for given n) a conditional unbiased estimate of $\text{Var}_n g$. Then by the above, $E_h = {}^n(E_n h) = E \text{Var}_n g = \text{Var } g$.

It can be shown that the usual mean and variance estimator is conditionally unbiased and therefore is unbiased with random n.

Summary

The procedures and estimators considered here are for simple random sampling. However, the sample procedures may be applied to stratified sampling by considering each strata as a population and properly combining the estimates. The usual estimates for stratified mean and variance are unbiased.

$$\bar{X}_{st} = \frac{\Sigma N_h \bar{X}_h}{N}$$

$$V(\bar{X}_{st}) = \frac{\Sigma N_h^2 S^2_{\bar{x}h}}{N^2}$$

The formulas assume that the stratum sizes are known and the units are assigned to strata in advance. If stratum sizes are known but units are assigned to strata after sample information is available, post stratification estimators should be used.

Although sometimes taken for granted this same property of unbiasedness allows the use of these estimators in other cases. For example:

(1) A systematic sample drawn from a list of unknown size which permits starting the drawing of a sample from a list before the complete list is available.

(2) A sample drawn from a list covering a given area with estimates made for subareas using only that portion of the sample which falls in each subarea.

In addition, the use of limited mail and non-response interviews does not require the selection and maintaining of a large list from which a relatively large portion will neither report by mail or be interviewed.