A Random Sample Using Limited Mail Questionnaires and Nonresponse Interviews

By Charles E. Rogers

OFTEN A 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 by interview alone. This method is especially desirable for data which may be collected with equivalent quality by either mail or personal interview. One method is to draw a random sample and interview all nonrespondents 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.

The method is relatively simple. A population of N elements, which are mutually exclusive for the characteristic to be estimated, is listed so that each element appears once and once only. 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 result may be achieved by selecting from the original listing a random sample ordered by draw or by selecting a systematic sample and randomly ordering the n elements selected.

Notation:

N = elements in population

n = elements selected as potentially in the sample

t_1 = interviews to 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 = the probability of a mailed questionnaire being returned

q = the probability of a nonresponse from mailed questionnaires

q = 1 - p

The survey procedure is to mail a questionnaire to each of the n selected elements and to interview t_1 nonrespondents to this mail survey. The number of nonrespondent interviews is limited to t_1 and the decision as to the number of interviews must be made in advance. To preserve randomness in the final sample the first t_1 nonrespondents on the randomly ordered list are interviewed. The sample will then consist of the ordered elements from the n selected until the first is reached for which no interview is available. This will constitute a random sample of $t_1 + t_2$ elements. The sample estimates of the specified characteristics are computed from this sample.

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 the determination of initial sample size (n), the expected rate of mail return is critical. When there are less than t_1 nonrespondents, the size of the constituted sample is restricted and precision is correspondingly reduced. The expected rate of mailed returns must be estimated in advance and the accuracy of this estimate is important because enough questionnaires must be mailed to obtain the number of nonrespondents for which interviews are planned, but as few extras as possible since these may have to be discarded.

There is some evidence that the value of the characteristic may affect the decision of the respondent to return or not return a questionnaire. This may, under this system, affect the sample size because of (1) clustering of either large or small-valued elements in the ordered list, or (2) difference of the mean of the ordered list from the mean of the population.

This effect of the value of the characteristic on the sample size must not prevent stochastic independence of the mean of the sample and the size of the sample if the conditions of random sampling are to be fulfilled. Consider that:

(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.

In repeated trials these random processes obviously result in equally likely combinations of size n, and each such combination contains a specified subset of elements which will produce an effective sample of t elements that is independent from those of other combinations.

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:

$$\overline{X} = \hat{W}_1 \overline{X}_1 + \hat{W}_2 \overline{X}_2 \text{ where } \hat{W}_1 = \frac{t_1}{t} \text{ and } \hat{W}_2 = \frac{t_2}{t}.$$

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 $\overline{X}' = W_1\overline{X}_1 + W_2\overline{X}_2$.

The difference (D) is:

$$(\overline{X}' - \overline{X}) \text{ or } (W_1\overline{X}_1 + W_2\overline{X}_2) - (\hat{W}_1\overline{X}_1 + \hat{W}_2\overline{X}_2).$$

Since $W_1 = 1 - W_2$ and $\hat{W}_1 = 1 - \hat{W}_2$ this difference reduces as follows:

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

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

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

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

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

With the random ordering, the two quantities $(\hat{W}_2 - W_2)$ and $(\overline{X}_1 - \overline{X}_2)$ should be independent and the expected value of D should equal zero. However, variation in D may add to variation of 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. Sandelius has shown that for finite populations and nonsequential sampling the usual mean estimate and its variance are unbiased estimates of the parameters even though sample size is a random variable.[1] 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 $(x_1 x_2 . . . x_t)$ with θ a parameter to be estimated. Let E_t be the conditional expectation for fixed t and $g = g(x_1 x_2 . . . x_t)$ be a function of the sample values with the property

[1] Martin Sandelius, "On Non-sequential Estimation when the Sample Size is a Random Variable," reprinted from Ann. Roy. Agr. Col. Sweden, Vol. 17, pp. 400-406, 1950.

$E_t g = \theta$ so that g is a conditional unbiased estimate of θ. Then it can be shown that $Eg = E(E_t g) = \theta$. Further, let $\text{Var}_t g = E_t(g - \theta)^2$ and let $\text{Var } g = E\,(g - \theta)^2$ exist; then $\text{Var } g = E\,E_t(g - \theta)^2 = E\,\text{Var}_t\,g$. When t is restricted to values between given integers, the existence condition will be satisfied. Now let $h = h(x_1 \ldots x_t)$ be (for given t) a conditional unbiased estimate of $\text{Var}_t g$. Then by the above, $Eh = (E_t h) = E\ \text{Var}_t\,g = \text{Var } g$. Since the usual mean and variance estimators are conditionally unbiased, it follows that they are unbiased with random sample size.

As an illustration of the method 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.

$$N = 2{,}000$$

$$t_1 = 50$$

$$p = 1/3$$

$$q = (1 - p) = 2/3$$

$$n = \frac{t_1}{q} = 75$$

From the 2,000 elements in the population, 75 will be selected by a random process (generally a somewhat optimistic value, for p is assumed to insure a sufficiently large n since there is variation in the rate of mailed returns and n varies inversely with 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 nonrespondents 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 nonrespondent who is not interviewed.

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

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

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

The procedures and estimators for simple random sampling may easily be extended to stratified sampling by considering each stratum as a population and properly combining the estimates for the different strata. When stratum sizes are known and the units are assigned to strata before drawing the sample, the usual formulas obtain. When stratum sizes are known but units are assigned to strata after sample information is available, post-stratification estimators should be used.