Sampling Without Replacement With Probability Proportional to Size
Author(s): W. L. Stevens
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 20, No. 2 (1958), pp. 393-397
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983900

Sampling Without Replacement With Probability Proportional to Size

By W. L. Stevens

Faculdade de Filosofia, Ciências e Letras, São Paulo, Brazil

[Received December, 1957]

SUMMARY

It is shown that sampling without replacement with probability proportional to size can be achieved if the units are grouped with reference to size. When the same unit is chosen a second time, it is substituted by another unit of the same size chosen at random. The estimate of the population total is formally the same as when sampling is done with replacement. The estimate of variance differs in that from the sum of squares of deviations of the ratio, r, we subtract, for each group chosen t (> 1) times, the quantity tS/N, where S is the sum of squares of deviations within the group and N the number in the group.

1. The Method

Sampling with probability proportional to size is usually done *with replacement*, for if it is done without replacement, the probability ceases to be strictly proportional to size, unless some special device is used, such as that proposed by Yates and Grundy, which is rather complicated for $n = 2$ and hardly practicable for $n > 2$.

Unless the probability of drawing the same unit twice is negligible, the method of sampling with replacement is inefficient, the loss of information being roughly equal to the proportion of duplicates. Although the loss is not usually very serious, it is worth while inquiring if any simple method can be found for avoiding it. It is suggested that such a method is available if the values of x, the variable measuring size, are or can be grouped.

If the values of x have been rather coarsely rounded, it will often be found that these groups already exist for all or much of the population. Where they do not exist, they can be formed by replacing groups of consecutive values (when listed in descending or ascending order) by a common central value. Thus, if it is desired to have no group smaller than five, the series

$$x = \; \ldots \; 36 \quad 39 \quad 39 \quad 41 \quad 41 \; \ldots$$

can be replaced by

$$x = \; \ldots \; 39 \quad 39 \quad 39 \quad 39 \quad 39 \; \ldots$$

The technique of selecting the sample is then quite simple: if at any moment a unit is drawn a second time, it is replaced by another unit of the same size drawn at random from among the other units of the same size which have not yet been drawn.

In principle, it is therefore necessary that no group shall be smaller than n, the number of units in the sample. But this will not usually be necessary in practice. It will usually be found that when $n =$ (say) 10, if the smallest group is of size (say) 3, the probability of drawing a group more times than it has members is extremely remote. If nevertheless it did happen, we would have to draw again. To the extent that this is likely to happen, the theory of the method fails, but as we are supposing that this is very improbable, we suppose also that the theoretical results are valid in the practical situation even when the smallest number in a group is less than n.

It is also admitted that the process of grouping (if not already completed by the rounding of values of x) will entail some loss of information. We suppose that in practice this loss will be extremely small; at any rate less than the gain resulting from the elimination of multiple drawings.

The sampling plan may be formulated in a different manner. Let $X_i = N_i x_i$ represent the total size of group i. Then we select, with replacement, n groups with probability proportional to X_i. If the group i is chosen t_i times, we then select, without replacement, t_i units with equal probability from this group.

2. Estimation

We will denote groups by i and j, and units within groups by u and v. The probability of selecting any unit in group i is

$$p_i = x_i/X,$$

where X is the total size of the population. Denoting the values to be observed by y_{iu} and the number of units in group i by N_i, we put

$$\begin{aligned} N_i\mu_i &= \Sigma y_{iu} \\ N &= \Sigma N_i \\ N\mu &= \Sigma N_i\mu_i \\ &= \Sigma\Sigma y_{iu}, \end{aligned}$$

where summations are made over the units of the group, over all groups and over the population respectively. Thus μ_i is the mean of y in group i and μ the general mean in the whole population.

The probability that unit iu is chosen in a sample of n units is found by summing the probabilities that group i is chosen t times multiplied respectively by the probabilities that t units chosen out of N_i units will include the unit iu.

$$\begin{aligned} &\Sigma\left\{\frac{n!}{(n-t)!\,t!}\,(1-N_ip_i)^{n-t}\,(N_ip_i)^t\right\}\left\{\frac{t}{N_i}\right\} \\ &\qquad = np_i\,\Sigma\left\{\frac{(n-1)!}{(n-t)!\,(t-1)!}\,(1-N_ip_i)^{n-t}\,(N_ip_i)^{t-1}\right\} \\ &\qquad = np_i. \end{aligned}$$

Thus we see that the probability that the sample contains a given unit is strictly proportional to the size of that unit.

Writing $r_{iu} = y_{iu}/p_i$, $R = \Sigma r_{iu}$ and $\bar{r} = R/n$, where Σ denotes summation over the sample, then the expected value of R is

$$E(R) = n\Sigma\Sigma p_i(y_{iu}/p_i) \text{ (summed over population)}$$
$$= nN\mu.$$

Hence an unbiased estimate of the total, $N\mu$, is provided by $R/n = \bar{r}$, formally identical with the estimate when sampling is done with replacement.

In practice, of course, we more often calculate the ratio in relation to size x, in which case, writing $r = y/x$, we have estimate $\bar{r}X$.

3. Variance

First let us find the probability that the sample contains units iu and iv $(u \neq v)$, i.e., a given pair within group i. This is the sum of products of the probability that the group is chosen t times by the probability that t units chosen at random from the N_i units will include the given pair.

$$\Sigma \left\{\frac{n!}{(n-t)!\,t!}(1-N_ip_i)^{n-t}(N_ip_i)^t\right\}\left\{\frac{t(t-1)}{N_i(N_i-1)}\right\}$$
$$= \frac{n(n-1)\,N_ip_i^2}{N_i-1}\Sigma\left\{\frac{(n-2)!}{(n-t)!\,(t-2)!}(1-N_ip_i)^{n-t}(N_ip_i)^{t-2}\right\}$$
$$= n(n-1)\,N_ip_i^2/(N_i-1).$$

Next we find the probability that the sample contains the pair iu and jv $(i \neq v)$, i.e., a pair belonging to two groups. Supposing that the groups are chosen respectively s and t times, the required probability is

$$\Sigma\Sigma\left\{\frac{n!}{(n-s-t)!\,s!\,t!}(1-N_ip_i-N_jp_j)^{n-s-t}(N_ip_i)^s(N_jp_j)^t\right\}\left\{\frac{st}{N_iN_j}\right\}$$
$$= n(n-1)\,p_ip_j\,\Sigma\Sigma\left\{\frac{(n-2)!}{(n-s-t)!\,(s-1)!\,(t-1)!}(1-N_ip_i-N_jp_j)^{n-s-t}(N_ip_i)^{s-1}(N_jp_j)^{t-1}\right\}$$
$$= n(n-1)\,p_ip_j.$$

Now consider R^2, where $R = \Sigma y_{iu}/p_i$ over the sample. The expansion of R^2 will contain terms of three kinds:

squares like $(y_{iu}/p_i)^2$
products like $y_{iu}y_{iv}/p_i^2$
products like $y_{iu}y_{jv}/p_ip_j$ $\quad (i \neq j)$.

The contribution from terms of the first kind, to $E(R^2)$, the expected value of R^2, will be found by summing over the population, the product of the square by the probability of unit iu being included in the sample.

$$\Sigma\Sigma(y_{iu}/p_i)^2(np_i) = n\Sigma\Sigma y_{iu}^2/p_i. \tag{3.1}$$

Similarly, the contribution of terms of the second kind, $y_{iu}y_{iv}/p_i^2$, is found by summing, over all such pairs, the product of the term by the probability that the pair occurs in the sample.

$$\sum_i \sum_{u \neq v} \left\{\frac{y_{iu}y_{iv}}{p_i^2}\right\}\left\{\frac{n(n-1)\,N_i p_i^2}{N_i - 1}\right\} = n(n-1)\sum_i \left\{\frac{N_i}{N_i - 1}\sum_{u \neq v} y_{iu}y_{iv}\right\}$$

Now

$$\begin{aligned}\sum_{u \neq v} y_{iu}y_{iv} &= \sum_u y_{iu}(N_i\mu_i - y_{iu})\\ &= N_i^2\mu_i^2 - \sum_u y_{iu}^2\\ &= N_i(N_i - 1)\,\mu_i^2 - (N_i - 1)\,\sigma_i^2\end{aligned}$$

where σ_i^2 is the variance of y within group i, defined, as usual,

$$\sigma_i^2 = (\Sigma y_{iu}^2 - N_i\mu_i^2)/(N_i - 1).$$

Hence the contribution of terms like $y_{iu}y_{iv}/p_i^2$ is

$$n(n-1)(\Sigma N_i^2\mu_i^2 - \Sigma N_i\sigma_i^2). \tag{3.2}$$

Finally we find the contribution of terms like $y_{iu}y_{jv}/p_ip_j$ where $i \neq j$:

$$\begin{aligned}\sum_{i \neq j}\sum_u \sum_v \left\{\frac{y_{iu}y_{jv}}{p_ip_j}\right\}\{n(n-1)\,p_ip_j\} &= n(n-1)\sum_{i \neq j}\sum_{u,v} y_{iu}y_{jv}\\ &= n(n-1)\sum_{i \neq j} N_iN_j\mu_i\mu_j.\end{aligned} \tag{3.3}$$

The contribution of the two kinds of product terms, from (3.2) and (3.3), is

$$\begin{aligned}n(n-1)\{\Sigma N_i^2\mu_i^2 + \sum_{i \neq j} N_iN_j\mu_i\mu_j - \Sigma N_i\sigma_i^2\}\\ &= n(n-1)\{(\Sigma N_i\mu_i)^2 - \Sigma N_i\sigma_i^2\}\\ &= n(n-1)(N^2\mu^2 - \Sigma N_i\sigma_i^2).\end{aligned} \tag{3.4}$$

Adding in the contribution of the square terms from (3.1), we obtain

$$E(R^2) = n\{\Sigma\Sigma y_{iu}^2/p_i + (n-1)(N^2\mu^2 - \Sigma N_i\sigma_i^2)\}$$

whence we obtain the variance of R

$$\begin{aligned}\sigma^2(R) &= E(R^2) - nN^2\mu^2\\ &= n\{\Sigma\Sigma y_{iu}^2/p_i - N^2\mu^2 - (n-1)\,\Sigma N_i\sigma_i^2\}.\end{aligned} \tag{3.5}$$

4. ESTIMATE OF VARIANCE

Consider first $\Sigma(r - \bar{r})^2 = \Sigma r^2 - R^2/n$, the sum of squares of deviations of the ratios $r = y/p$. The expected value of Σr^2 has already been found (3.1)

$$E(\Sigma r^2) = n\Sigma\Sigma y_{iu}^2/p_i.$$

Hence, subtracting $E(R^2)/n$, we find the expected value

$$E\{\Sigma(r - \bar{r})^2\} = (n-1)\{\Sigma\Sigma y_{iu}^2 - N^2\mu^2 + \Sigma N_i\sigma_i^2\}. \tag{4.1}$$

It is seen, comparing with (3.5), that the mean square deviation agrees with that required for an unbiased estimate of $\sigma^2(R)$ up to the second term, but disagrees in the third, having $+\,\Sigma N_i\sigma_i^2$ instead of $-\,(n-1)\,\Sigma N_i\sigma_i^2$.

Consider therefore in a group i, chosen t_i times, $t_i > 1$, the quantity

$$t_i S_i / N_i$$

where

$$S_i = \Sigma(r_{iu} - \bar{r}_i)^2$$
$$\bar{r}_i = \Sigma r_{iu}/t_i$$

summation being over units of group i in the sample.

With t_i fixed,

$$E(S_i) = E\{\Sigma(y_{iu} - \bar{y}_i)^2\}/p_i^2$$
$$= (t_i - 1)\, \sigma_i^2/p_i^2.$$

Thus

$$E(t_i S_i/N_i) = E\{t_i(t_i - 1)\}\, \sigma_i^2/N_i p_i^2.$$

The expected value of $t_i(t_i - 1)$, when t_i is in a binomial distribution, is the well known result

$$E\{t_i(t_i - 1)\} = n(n - 1)\, N_i^2 p_i^2$$

Hence

$$E(t_i S_i/N_i) = n(n - 1)\, N_i \sigma_i^2.$$

Summing over all groups which have $t_i > 1$, we find

$$E(\Sigma t_i S_i/N_i) = n(n - 1)\, \Sigma N_i \sigma_i^2.$$

Hence

$$E\{\Sigma(r - \bar{r})^2 - \Sigma t_i S_i/N_i\} = (n - 1)\{\Sigma\Sigma y_{iu}^2/p_i - N^2\mu^2 + \Sigma N_i\sigma_i^2 - n\Sigma N_i\sigma_i^2\}$$
$$= (n - 1)\{\Sigma\Sigma y_{iu}^2/p_i - N^2\mu^2 - (n - 1)\, \Sigma N_i\sigma_i^2\}.$$

By comparison with (3.5), we conclude that an unbiased estimate of variance of R is ns^2 and of the estimated total is s^2/n, where

$$s^2 = \frac{\Sigma(r - \bar{r})^2 - \Sigma t_i S_i/N_i}{n - 1}.$$

In spite of the complexity of the analysis, the final result is very simple. From the sum of squares of deviations, $\Sigma(r - \bar{r})^2$, used in the analysis when sampling is done with replacement, we merely subtract, for each group chosen t_i (> 1) times, the quantity

$$t_i S_i/N_i$$

where N_i is the number in the group and S_i the sum of squares of deviations of r within the group.

As most groups chosen more than once will, in fact, be chosen twice, we note that for $t = 2$, the correction to be subtracted is

$$(r_{i1} - r_{i2})^2/N_i$$

where r_{i1} and r_{i2} are the two ratios observed.

When ratios are calculated with reference to the size x_i, the formula for the variance will contain the additional factor X^2.

Reference

Yates, F. & Grundy, P. M. (1953), "Selection without replacement from within strata with probability proportional to size", *J. R. Statist. Soc.*, **B**, **15**, 253–261.