SOME REMARKS ON SAMPLING WITH REPLACEMENT

By Des Raj and Salem H. Khamis

American University of Beirut

1. Introduction and summary. In order to estimate the mean of a finite population from a random sample, the sample units may be selected in two ways. A pre-determined number m of units may be selected with replacement or sampling with replacement may be continued till a desired number n of distinct units is obtained. The first procedure is called sampling with replacement while the second may be called sampling without replacement. A comparison is generally made between the two procedures with $m = n$. This comparison, however, is not fair since costs are usually proportional to the number of distinct units and in the first procedure this number would be less than or equal to n.

In the first procedure the population mean is usually estimated by the sample mean based on all the units in the sample including repetitions while in the second procedure the estimate is generally made to depend on the distinct units only. The object of this paper is to show that the estimate making use of only the distinct units is superior in either procedure. The following results are proved in this paper:

(i) In sampling with replacement the estimate of the mean based on distinct units in the sample is superior to the estimate based on the total sample size when (a) the total sample size is fixed in advance, while the number of distinct units in the sample is a random variable, and, (b) when the total sample size is a random variable while the number of distinct units is fixed in advance.

(ii) The same is true of ratio estimates. It is also shown that the bias is numerically less if the ratio estimate is based on distinct units regardless of whether these are fixed in advance or considered as random variables.

(iii) Expressions for the estimation of the variances of the various estimates considered in this paper are given.

(iv) The above results are extended to multistage sampling.

2. Statement of the problem. Let us consider a finite population consisting of N sampling units. Suppose we are interested in estimating the population mean $\bar{Y}$ for a character y, from a sample selected with replacement with equal probabilities. We consider the following two sampling schemes.

Scheme A. We select with replacement a total sample of size m fixed in advance. We denote the number of distinct sample units selected by the random variable u.

Scheme B. We select with replacement a sample of n distinct units fixed in advance. We denote the total sample size, including repetitions, by the random variable v.

Received February 15, 1957; revised November 18, 1957.

We consider Scheme A first. Let $a_1, a_2, \cdots, a_u$ be the u distinct units selected in the sample and let k_r be the number of times the rth unit a_r occurs in the sample, with $\sum k_r = m$. We compare the bias and variances of the usual estimate

$$\bar{y}_m = \frac{1}{m} \sum_{r=1}^{u} k_r y_r , \tag{1}$$

and the estimate

$$\bar{y}_u = \frac{1}{u} \sum_{r=1}^{u} y_r \tag{2}$$

based on the distinct units only, where y_r is the value of the character y for the rth distinct unit in the sample.

The estimate (1) is well known to be unbiased and that its variance is given by

$$V(\bar{y}_m) = \frac{1}{m}\left(1 - \frac{1}{N}\right)\sigma^2 = \left(Q - \frac{1}{N}\right)\sigma^2 \tag{3}$$

where

$$Q = \frac{1}{m}\left(1 + \frac{m-1}{N}\right) \tag{4}$$

and

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2. \tag{5}$$

For a given u the expected value of $\bar{y}_u$ is $\bar{Y}$ so that $\bar{y}_u$ is an unbiased estimate of $\bar{Y}$. With regard to the variance of $\bar{y}_u$ we have

$$V(\bar{y}_u) = \left[E\left(\frac{1}{u}\right) - \frac{1}{N}\right]\sigma^2. \tag{6}$$

The estimate $\bar{y}_u$ is superior to $\bar{y}_m$ if

$$E\left(\frac{1}{u}\right) < Q. \tag{7}$$

The probability distribution of the random variable u is given by (cf. Feller, [1])

$$P(u) = N^{-m} \binom{N}{u} \Delta^u 0^m, \tag{8}$$

where the sth difference of 0^t is defined by

$$\Delta^s 0^t = \sum_{r=0}^{s} (-1)^{s-r} \binom{s}{r} r^t.$$

Hence

$$E\left(\frac{1}{u}\right) = N^{-m} \sum_{u=1}^{m} \frac{1}{u} \binom{N}{u} \Delta^u 0^m \tag{9}$$

and the expected sample size is

$$E(u) = N^{-m} \sum_{u=1}^{m} u \binom{N}{u} \Delta^u 0^m. \tag{10}$$

We consider now Scheme B. Let $b_1, b_2, \cdots, b_n$ be the n distinct units selected in the sample and let k_r be the number of times the rth unit b_r occurs in the sample with $\sum k_r = v$. We compare the bias and variances of the estimate

$$\bar{y}_v = \frac{1}{v} \sum_{r=1}^{n} k_r y_r \tag{11}$$

and the estimate

$$\bar{y}_n = \frac{1}{n} \sum_{r=1}^{n} y_r \tag{12}$$

where y_r is the same as in Scheme A with obvious modifications. It is easy to see that the estimates $\bar{y}_v$ and $\bar{y}_n$ are unbiased for estimating $\bar{Y}$. Also we have

$$V(\bar{y}_v) = E\left(\frac{1}{v}\right)\frac{N-1}{N}\sigma^2 \tag{13}$$

and

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2. \tag{14}$$

The estimate $\bar{y}_n$ is superior if

$$E\left(\frac{1}{v}\right) > \frac{1}{n}\frac{N-n}{N-1}. \tag{15}$$

The probability distribution of the random variable v can be shown to be

$$P(v) = \binom{N-1}{n-1} N^{1-v} \Delta^{n-1} 0^{v-1}. \tag{16}$$

Hence

$$E\left(\frac{1}{v}\right) = \binom{N-1}{n-1} \sum_{v=n}^{\infty} \frac{1}{v} N^{1-v} \Delta^{n-1} 0^{v-1}. \tag{17}$$

We give in Table 1 a numerical table for selected sample sizes which illustrates the numerical magnitudes of the differences discussed above. The theoretical proofs are given in Section 3 below.

3. Proofs of the inequalities. In order to establish inequality (7), we shall first prove the following

LEMMA. *Let*

$$S_{t,N} = \sum_{u=1}^{m-t} \frac{1}{u+t} \binom{N}{u} \Delta^u 0^{m-t}. \tag{18}$$

TABLE 1

Values of $E(u)$, $E\left(\frac{1}{u}\right)$, Q *and* $1/E(u)$

N	m	$E(u)$	$E\left(\frac{1}{u}\right)$	Q	$\frac{1}{m}$	$\frac{1}{E(u)}$
$N = 10$	1	1.00	1.000	1.000	1.000	1.000
	2	1.90	.550	.550	.500	.526
	3	2.71	.385	.400	.333	.369
	4	3.44	.303	.325	.250	.291
	5	4.10	.253	.280	.200	.244
	6	4.69	.221	.250	.167	.213
$N = 50$	1	1.00	1.000	1.000	1.000	1.000
	2	1.98	.510	.510	.500	.505
	3	2.94	.343	.347	.333	.340
	4	3.88	.260	.265	.250	.258
	5	4.80	.210	.216	.200	.208
	6	5.71	.177	.183	.167	.175
$N = 100$	1	1.00	1.000	1.000	1.000	1.000
	2	1.99	.505	.505	.500	.503
	3	2.97	.338	.340	.333	.337
	4	3.94	.255	.258	.250	.254
	5	4.90	.205	.208	.200	.204
	6	5.85	.172	.175	.167	.171

Then

$$S_{t,N} \leqq (N+1)S_{t+1,N} \text{ for } t = 0, 1, \cdots, m-1; \qquad m > 2, \tag{19}$$

the sign of equality holds only for $t = 0$.

PROOF.

$$\begin{aligned} S_{t,N} &= \sum_{u=1}^{m-t} \frac{u}{u+t}\binom{N}{u}(\Delta^{u-1}0^{m-t-1} + \Delta^u 0^{m-t-1}) \\ &= \sum_{u=2}^{m-t} \frac{N-u+1}{u+t}\binom{N}{u-1}\Delta^{u-1}0^{m-t-1} + \sum_{u=1}^{m-t-1} \frac{u}{u+t}\binom{N}{u}\Delta^u 0^{m-t-1} \\ &= \sum_{u=1}^{m-t-1}\left(\frac{N-u}{u+t+1} + \frac{u}{u+t}\right)\binom{N}{u}\Delta^u 0^{m-t-1} \\ &= \sum_{u=1}^{m-t-1} \frac{(N+1)u + tN}{(u+t)(u+t+1)}\binom{N}{u}\Delta^u 0^{m-t-1}. \end{aligned}$$

Thus $S_{0,N} = (N+1)S_{1,N}$, and $S_{t,N} < (N+1)S_{t+1,N}$ for $1 \leqq t \leqq m-1$. This proves the lemma.

COROLLARY. Applying the lemma $m-2$ times beginning with $t = 1$, we have, for $m > 2$

$$S_{1,N} < (N+1)^{m-2}S_{m-1,N} = (N+1)^{m-2}\frac{N}{m}$$

so that

$$S_{1,N-1} < N^{m-2}\frac{N-1}{m}. \tag{20}$$

Now, inequality (7) is proved in the following

THEOREM.

$$E\left(\frac{1}{u}\right) \leqq \frac{1}{N} + \frac{N-1}{N}\frac{1}{m}; \tag{21}$$

the sign of equality holds only for $m = 2$.

PROOF.

$$\begin{aligned} E\left(\frac{1}{u}\right) &= N^{-m}\sum_{u=1}^{m}\binom{N}{u}(\Delta^u 0^{m-1} + \Delta^{u-1}0^{m-1}) \\ &= \frac{1}{N} + N^{1-m}\sum_{2}^{m}\frac{1}{u}\binom{N-1}{u-1}\Delta^{u-1}0^{m-1} \\ &= \frac{1}{N} + N^{1-m}\sum_{1}^{n-1}\frac{1}{u+1}\binom{N-1}{u}\Delta^u 0^{m-1} \\ &= \frac{1}{N} + N^{1-m}S_{1,N-1}. \end{aligned}$$

On using (20) we have for $m > 2$,

$$E\left(\frac{1}{u}\right) < \frac{1}{N} + \frac{N-1}{N}\frac{1}{m}.$$

For $m = 2$, $S_{1,N-1} = \frac{1}{2}(N-1)$, so that the last inequality reduces to an equality in this case.

To prove inequality (15), we make use of the following

LEMMA. *If* v *is a positive random variable, then*

$$E\left(\frac{1}{v}\right) \geqq \frac{1}{E(v)}. \tag{22}$$

The proof of this lemma follows from Cauchy's inequality (cf. Hardy and others, [3]),

$$(\sum a^2)(\sum b^2) > (\sum ab)^2, \tag{23}$$

by substituting $a = \sqrt{vP(v)}$, $b = \sqrt{v^{-1}P(v)}$, and noting that the two sides of the inequality are convergent because $E(v)$ and $E(1/v)$ are finite.

Now (cf. Feller, [2])

$$E(v) = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{N-n+1}\right),$$

$$\therefore E\left(\frac{1}{v}\right) > \frac{1}{N\dfrac{n}{N-n+1}} = \frac{N-n+1}{nN} > \frac{N-n}{n(N-1)} \quad \text{for} \quad n > 1.$$

It is easy to see that for $n = 1$ the two procedures are equivalent and the inequality (15) reduces to an equality.

4. Estimation of variances. We shall consider the problem of estimating from the sample the variances of $\bar{y}_u$ and $\bar{y}_v$. Under Scheme A, it is easy to see that for a given $u \geqq 2$, an unbiased estimate of σ^2 is provided by

$$s_u^2 = \frac{1}{u-1} \sum_{i=1}^{u} (y_i - \bar{y}_u)^2. \tag{24}$$

Thus considering

$$G_u = \left[\left(\frac{1}{u} - \frac{1}{N}\right) + N^{1-m}\left(1 - \frac{1}{u}\right)\right] s_u^2, \tag{25}$$

we have

$$E[G_u \mid u \geqq 2] = V(\bar{y}_u); \tag{26}$$

so that G_u provides an unbiased estimate of the variance of $\bar{y}_u$. It is unbiased in the conditional sense, namely when the number of distinct units in the sample exceeds unity. An alternative unbiased estimate is provided by G_u' where

$$\begin{aligned} G_u' &= \left[\left(\frac{1}{u} - \frac{1}{n}\right) + \frac{N-1}{N^m - N}\right] s^2, \\ s^2 &= \frac{1}{u-1} \sum_{i=1}^{u} (y_i - \bar{y}_u)^2, \qquad \text{for } u \geqq 2, \\ s^2 &= 0, \qquad \text{for } u = 1. \end{aligned} \tag{27}$$

and

Under Scheme B, it is easy to see that

$$E\left[\frac{1}{v} s_v^2 \mid v\right] = \frac{1}{v} \frac{N-1}{N} \sigma^2 \tag{28}$$

where

$$s_v^2 = \frac{1}{v-1} \sum_{i=1}^{v} (y_i - \bar{y}_v)^2. \tag{29}$$

Thus

$$E\left[\frac{1}{v} s_v^2\right] = E\left(\frac{1}{v}\right) \frac{N-1}{N} \sigma^2, \tag{30}$$

so that $1/v\, s_v^2$ is an unbiased estimate of $V(\bar{y}_v)$.

5. Extension to ratio estimation. We shall now extend the above results to ratio estimates. We make use of the notation and approximate results given by Cochran [1]. The object is to estimate the population ratio $R = Y/X$. Under Scheme A we compare the two estimates

$$\bar{R}_m = \sum_{i=1}^{u} k_i y_i \Big/ \sum_{i=1}^{u} k_i x_i \tag{31}$$

and

$$\bar{R}_u = \sum_{i=1}^{u} y_i \Big/ \sum_{i=1}^{u} x_i. \tag{32}$$

We have

$$\begin{aligned} V(\bar{R}_m) &\doteq R^2\left(Q - \frac{1}{N}\right)[C_{yy} + C_{xx} - 2C_{xy}] \\ B(\bar{R}_m) &\doteq R\left(Q - \frac{1}{N}\right)[C_{xx} - C_{xy}] \\ V(\bar{R}_u) &\doteq R^2\left[E\left(\frac{1}{u}\right) - \frac{1}{N}\right][C_{yy} + C_{xx} - 2C_{xy}] \\ B(\bar{R}_u) &\doteq R\left[E\left(\frac{1}{u}\right) - \frac{1}{N}\right][C_{xx} - C_{xy}] \end{aligned} \tag{33}$$

where $B(\bar{R})$ stands for the absolute value of the bias of the estimate $\bar{R}$. Using inequality (7), we have

$$V(\bar{R}_u) < V(\bar{R}_m) \text{ and } B(\bar{R}_u) < B(\bar{R}_m). \tag{34}$$

Under Scheme B, the estimates to be compared are

$$\bar{R}_v = \sum_{i=1}^{n} k_i y_i \Big/ \sum_{i=1}^{n} k_i x_i \tag{35}$$

and

$$\bar{R}_n = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i. \tag{36}$$

It is easy to see that the estimate $\bar{R}_n$ is superior to $\bar{R}_v$ from the point of view of variance and bias.

6. Extension to multistage designs. The result obtained for unistage designs will now be extended to multistage designs. Let a population consist of N first stage sampling units, of which m or n are selected with equal probabilities with replacement according to the Schemes A or B respectively. For the ith first stage unit, let t_i (based on sampling at second and subsequent stages) be an unbiased estimate of y_i, the total value of the character y for the unit. For Schemes A and B the unbiased estimates considered are

$$\bar{y}'_u = \frac{1}{u}\sum_{i=1}^{n} t_i, \tag{37}$$

$$\bar{y}'_m = \frac{1}{m}\sum_{i=1}^{u} k_i t_i, \tag{38}$$

and

$$\bar{y}_n' = \frac{1}{n}\sum_{i=1}^{n} t_i, \tag{39}$$

$$\bar{y}_v' = \frac{1}{v}\sum_{i=1}^{n} k_i t_i. \tag{40}$$

The variances of the estimates are given by

$$V(\bar{y}_u') = \left[\left(1 + \frac{\delta^2}{\sigma^2}\right) E\left(\frac{1}{u}\right) - \frac{1}{N}\right]\sigma^2 \tag{41}$$

$$V(\bar{y}_m') = \left[\frac{\delta^2}{\sigma^2} Q + \frac{N-1}{mN}\right]\sigma^2 \tag{42}$$

$$V(\bar{y}_n') = \frac{\delta^2}{n} + \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2, \tag{43}$$

$$V(\bar{y}_v') = \frac{N-1}{N} E\left(\frac{1}{v}\right)(\delta^2 + \sigma^2) + \frac{\delta^2}{N}, \tag{44}$$

where

$$\delta^2 = \frac{1}{N}\sum_{1}^{N} V(t_i).$$

Using inequalities (7) and (15) it is found that $\bar{y}_u'$ is superior to $\bar{y}_m'$ while $\bar{y}_n'$ is superior to $\bar{y}_v'$.

REFERENCES

[1] W. G. COCHRAN, "*Sampling Techniques*", John Wiley and Sons, New York, 1953, pp. 117, 118.

[2] W. FELLER, "*An introduction to probability theory and its applications*", John Wiley and Sons, New York, 1950, pp. 69, 175.

[3] G. H. HARDY, J. E. LITTLEWOOD, AND G. POLYA, "*Inequalities*", Cambridge University Press, 1934, pp. 16, 124.