The Annals of Mathematical Statistics
1970, Vol. 41, No. 6, 2132–2134

ON AVERAGING OVER DISTINCT UNITS IN SAMPLING WITH REPLACEMENT

BY R. M. KORWAR AND R. J. SERFLING

Florida State University

0. Summary. Raj and Khamis [2], and Raj [3], have shown that in sampling with replacement the average over distinct units possesses lower variance than the average over the entire sample including repetitions. But no usable results are available regarding the amount of reduction in variance. The relevant expression given by Raj and Khamis is prohibitively cumbersome even for numerical computations. The present note provides the simplest possible exact expression as well as a convenient approximation. In particular, it follows readily that the relative benefit due to averaging over distinct units only is basically an increasing function of the sampling fraction.

1. Background. Let u be the number of distinct units obtained in a sample of size n drawn with replacement from a population of size N.

To estimate a mean, one may average over all units in the sample including repetitions, or alternatively average over the distinct units only. Each estimate is unbiased but the variances differ. For the first, the variance is proportional to

$$Q - \frac{1}{N} \tag{1.1}$$

where

$$Q = \frac{1}{n} + \frac{n-1}{nN} . \tag{1.2}$$

In the second case, the quantity Q becomes replaced by $E(1/u)$. Also, in comparison of the mean square errors of certain ratio estimates, the quantities Q and $E(1/u)$ again are involved.

It is thus important to determine $E(1/u)$ and to compare it with Q. Raj and Khamis [2] derive an expression for $E(1/u)$ and show that

$$E(1/u) < Q, \qquad n \geqq 3. \tag{1.3}$$

(For $n = 1$ and 2, $E(1/u) = Q$.) However, their expression is too complicated to yield readily any further theoretical insight, nor is it well suited to numerical computations. It is of interest to make a more refined comparison than that given by (1.3). For further background details, see [2] and [3].

2. Results. It will be shown that

$$E(1/u) = N^{-n} \textstyle\sum_{j=1}^{N} j^{n-1}, \qquad n \geqq 1, \tag{2.1}$$

Received December 15, 1969; revised April 2, 1970.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

and that

$$Q^* - \frac{1}{720N} < E(1/u) \leqq Q^*, \qquad 3 \leqq n \leqq N, \tag{2.2}$$

where

$$Q^* = \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}\,. \tag{2.3}$$

(For $n = 3$, $E(1/u) = Q^*$.) Formula (2.1) affords the simplest possible exact representation for $E(1/u)$ and is amenable to numerical computations. In addition, Q^* gives an excellent approximation, suitable not only for computations but also for theoretical considerations. For example, we have from (1.2) and (2.3) the approximation

$$\frac{Q-1/N}{Q^*-1/N} \doteq \frac{1}{1-\frac{1}{2}f}, \tag{2.4}$$

where $f = n/N$, showing that the relative benefit due to averaging over distinct units only is an increasing function of f.

3. Proofs. Formula (2.1) may be obtained simply by appealing to the symmetry and independence inherent in the sampling scheme, without invoking the probability distribution of the random variable u. To this effect, let us denote the population units by $U_1, \cdots, U_N$. Let $i_1, \cdots, i_n$ be the subscripts of units selected in the sample and let $i_{(1)} < i_{(2)} < \cdots < i_{(u)}$ be their ordered distinct values. Then

$$\begin{aligned} P[i_1 = i_{(u)}] &= EP[i_1 = i_{(u)} \mid i_{(1)}, \cdots, i_{(u)}] \\ &= E(1/u), \end{aligned} \tag{3.1}$$

utilizing the symmetry. On the other hand,

$$\begin{aligned} P[i_1 = i_{(u)}] &= P[i_1 \geqq i_2, i_1 \geqq i_3, \cdots, i_1 \geqq i_n] \\ &= \textstyle\sum_{j=1}^{N} P[i_1 = j] P[i_2 \leqq j, \cdots, i_n \leqq j] \\ &= \textstyle\sum_{j=1}^{N} \left(\frac{1}{N}\right)\left(\frac{j}{N}\right)^{n-1}, \end{aligned} \tag{3.2}$$

again appealing to symmetry and also using the independence of $i_1, \cdots, i_n$. Thus (2.1) is proved. It may be of interest that (2.1) was first obtained by a combinatorial argument, the present argument being discovered after knowing the result.

To obtain (2.2), use will be made of the formulas

$$\textstyle\sum_{j=1}^{N} j^{n-1} = \frac{N^n}{n} + \frac{N^{n-1}}{2} + \sum_{j=0}^{L} (-1)^j \frac{1}{2(j+1)} \binom{n-1}{2j+1} B_{j+1} N^{n-2-2j}, \tag{3.3}$$

where $L = (n-4)/2$ if n is even and $= (n-3)/2$ if n is odd, and

$$\frac{2^{2r-1}\pi^{2r}}{(2r)!} B_r = \textstyle\sum_{k=1}^{\infty} k^{-2r}, \tag{3.4}$$

from Adams [1] pages 27 and 140. (The B_r are the Bernoulli numbers.) By (2.1), (2.3) and (3.3), and since $B_1 = \frac{1}{6}$, we have, for $n \geqq 3$,

$$E(1/u) = Q^* + \textstyle\sum_{j=0}^{L}(-1)^j A_j \tag{3.5}$$

where $A_0 = 0$ and

$$A_j = N^{-2(j+1)} \frac{1}{2(j+1)} \tbinom{n-1}{2j+1} B_{j+1}, \qquad 1 \leqq j \leqq L. \tag{3.6}$$

Now, for $j \geqq 1$, (3.6) and (3.4) yield

$$\begin{aligned} -A_{2j-1} + A_{2j} &= N^{-4j} \frac{1}{4j} \tbinom{n-1}{4j-1} \left[-B_{2j} + B_{2j+1} \frac{(2j)(n-4j+1)(n-4j)}{(2j+1)(4j)(4j+1)N^2} \right] \\ &< B_{2j} N^{-4j} \frac{1}{4j} \tbinom{n-1}{4j-1} \left[-1 + \frac{(n-4j+1)(n-4j)}{4\pi^2 N^2} \right] \\ &< -B_{2j} N^{-4j} \frac{1}{4j} \tbinom{n-1}{4j-1} \left(1 - \frac{1}{4\pi^2} \right). \end{aligned} \tag{3.7}$$

Hence $\sum_0^L(-1)^j A_j \leqq 0$, so that Q^* is an upper bound for $E(1/u)$. On the other hand, it may similarly be shown that

$$A_{2j} - A_{2j+1} > B_{2j+1} N^{-4j-2} \frac{1}{4j+2} \tbinom{n-1}{4j+1} \left(1 - \frac{1}{4\pi^2} \right), \tag{3.8}$$

whence $\sum_0^L(-1)^j A_j > -A_1$. Since $B_2 = 1/30$, we have $A_1 = \binom{n-1}{3}/(120N^4) < 1/720N$. Thus (2.2) is proved.

Acknowledgment. The authors are grateful to Dr. R. R. Davidson for encouragement in the investigation and to the referee for constructive and helpful suggestions on the presentation.

REFERENCES

[1] Adams, Edwin P. (1939). *Smithsonian Mathematical Formulae and Tables of Elliptic Functions.* Smithsonian Institution, Washington, D.C.

[2] Raj, Des and Khamis, Salem H. (1958). Some remarks on sampling with replacement. *Ann. Math. Statist.* **29** 550–557.

[3] Raj, Des (1968). *Sampling Theory.* McGraw-Hill, New York.