The Annals of Statistics
1975, Vol. 3, No. 2, 500-503

A NOTE ON SAMPLING WITH REPLACEMENT[1]

BY E. BENTON COBB

University of Kansas

Suppose a finite population is sampled with replacement until the sample contains a fixed number n of distinct units. Let v denote the total number of draws. It is known that $\bar{y}_n$, the mean for the n distinct units, and $\bar{y}_v$, the total sample mean, are both unbiased estimators of the population mean and that $V(\bar{y}_n) \leqq V(\bar{y}_v)$. In this paper the relative difference $\delta = [V(\bar{y}_v) - V(\bar{y}_n)]/V(\bar{y}_n)$ is approximated by a quantity δ_1 which is easy to compute. Upper and lower bounds for $\delta - \delta_1$ are given and it is shown that $\delta < (\lambda + \varepsilon_n)f$ for $n \geqq 3$ and $f \leqq \frac{3}{4}$, where $f = n/N$, N is the population size, $\lambda = [(1-f)^{-\frac{1}{2}} - 1]/f$, and $\varepsilon_n = (1-f)^{-1}/(n-1)$.

1. Introduction and statement of results. Let $Y_1, Y_2, \cdots, Y_N$ denote the values of some characteristic Y for N units in a finite population. Let $\mu = \sum Y_j/N$ and $\sigma^2 = \sum (Y_j - u)^2/(N-1)$. Suppose units are drawn at random with replacement until the sample contains n distinct units. Let $y_1, y_2, \cdots, y_n$ denote the values for the set $s = \{u_1, u_2, \cdots, u_n\}$ of n distinct units in the sample (subscripts do not indicate the order in which values or units are obtained). Define $\bar{y}_n = \sum y_j/n$ and $\bar{y}_v = \sum k_j y_j/v$, where k_j is the frequency with which unit u_j occurs in the total sample and $v = \sum k_j$ is the total number of draws. It is known (cf. Basu (1958)) that $\bar{y}_n$ and $\bar{y}_v$ are both unbiased for μ and that

$$V(\bar{y}_n) \leqq V(\bar{y}_v) .$$

Chikkagoudar (1966) derived an expression for the variance of $\bar{y}_v$, although there is a minor error in his final step. The corrected expression (using the preceding notation) is

$$(1)\qquad V(\bar{y}_v) = \sigma^2\binom{N-1}{n-1}\Delta^{n-1}[t^{-1}(t/N)^{n-1} + Nt^{-2}(t-3)\textstyle\sum_{k\geqq n} k^{-1}(t/N)^k + 2t^{-1}\sum_{k\geqq n} k^{-2}(t/N)^{k-1}]_{t=0}$$

for $n \geqq 2$, where

$$\Delta^k F(t) = \Delta^{k-1}F(t+1) - \Delta^{k-1}F(t) , \qquad k = 1, 2, \cdots ,$$
$$\Delta^0 F(t) = F(t) ,$$

for a function $F(t)$.

The purpose of this note is to obtain a simpler expression for $V(\bar{y}_v)$, involving the first two negative moments of v, from which bounds on the relative difference

$$(2)\qquad \delta = [V(\bar{y}_v) - V(\bar{y}_n)]/V(\bar{y}_n)$$

Received March 1973; revised April 1974.

[1] This research was supported in part by a grant from the General Research Fund, University of Kansas.

AMS 1970 *subject classifications.* Primary 62D05; Secondary 62F10.

Key words and phrases. Sampling with replacement until the sample contains n units, estimation of the mean of a finite population.

can be obtained. Proofs of the following results are given in the next section.

THEOREM. *Let* $f = n/N$ *and* $x = v/n$. *Then* $V(\bar{y}_v) = \sigma^2 n^{-1}\{1 - f + (n-1)^{-1} \times [1 + (n-4)Ex^{-1} - (n-3)Ex^{-2}]\}$ *for* $n \geqq 2$.

COROLLARY 1. *Let* δ *be defined as in* (2) *and let*

$$\delta_1 = (n-1)^{-1}(n-3)(1-f)^{-1}[(Ex)^{-1} - (Ex)^{-2}]\,. \tag{3}$$

If $n \geqq 3$, *then*

$$-2(n-3)(n-1)^{-1}(f/n)(1-f)^{-2} < \delta - \delta_1 < (n-1)^{-1}f(1-f)^{-1}[1 + (n-3)n^{-1}(1-f)^{-1}/256]\,.$$

REMARK. For computation of $Ex = Ev/n$, see (4) in the next section.

COROLLARY 2. *If* $n \geqq 3$ *and* $f \leqq \frac{3}{4}$, *then*

$$\delta < \delta_1 + (n-1)^{-1}f(1-f)^{-1} < (\lambda + \varepsilon_n)f\,, \qquad \textit{where}$$

$$\lambda = [(1-f)^{-\frac{1}{2}} - 1]/f\,, \qquad \varepsilon_n = (1-f)^{-1}/(n-1)\,,$$

and δ_1 *is given by* (3).

An interesting application of Corollary 2 is that if $f \leqq \frac{1}{2}$ and $n \geqq 13$, then $\delta < f$.

2. Proofs. The proof of the Theorem appears to be most easily done by a direct derivation of $V(\bar{y})$ which avoids the use of (1). As in the previous section, s will denote the set of n distinct sample units, v the total number of draws, and k_j the frequency of unit u_j. The derivation depends on the following lemmas:

LEMMA 1. *Let* v *and* s *be fixed. Then* $E(k_j \mid v, s) = v/n$ *and* $V(k_j \mid v, s) = n^{-2}(n-1)^{-1}[n(v-n)(n-2) + (v-n)^2]$.

PROOF. The results are derived by conditioning on the last distinct unit selected, say u_i (u_i can be considered as a random selection from s). Then $k_i = 1$, and the joint conditional distribution of the "excess frequencies" $(k_j - 1)$, $j \neq i$, is multinomial with uniform probabilities $(n-1)^{-1}$, and $\sum_{j \neq i}(k_j - 1) = v - n$.

LEMMA 2. *Let* v *and* s *be fixed and let* $k_1, k_2, \cdots, k_n$ *be a random permutation of the components of a fixed vector* $c = (c_1, c_2, \cdots, c_n)$ *with* $\sum c_j = v$. *Then* $E(\bar{y}_v \mid v, s, c) = \bar{y}_n$ *and* $V(\bar{y}_v \mid v, s, c) = v^{-2}n(n-1)^{-1}V(k_1 \mid v, s, c) \sum (y_j - \bar{y}_n)^2$, *where* $V(k_1 \mid v, s, c) = \sum (c_j - \bar{c})^2/n$.

PROOF. The expression for the conditional mean is obvious. The expression for the conditional variance follows from

$$V(\textstyle\sum k_j y_j \mid v, s, c) = \sum \sum y_i y_j \operatorname{Cov}(k_i, k_j \mid v, s, c)\,.$$

PROOF OF THEOREM. Let E_1 and V_1 denote the conditional mean and variance operators, respectively, for v and s fixed. Then $E_1(k_1)$ and $V_1(k_1)$ are given by Lemma 1. Keep v and s fixed and condition on a fixed set of frequencies as in

Lemma 2. Then, since $V_1(\bar{y}_n) = 0$, we have

$$V_1(\bar{y}_v) = v^{-2}n(n-1)^{-1}V_1(k_1)\sum (y_j - \bar{y}_n)^2 .$$

Finally, using the independence of v and s and the fact that $(n-1)^{-1}\sum (y_j - \bar{y}_n)^2$ is unbiased for σ^2, we have

$$\begin{aligned} V(\bar{y}_v) &= EV_1(\bar{y}_v) + V(\bar{y}_n) \\ &= E[v^{-2}V_1(k_1)]n\sigma^2 + (1-f)n^{-1}\sigma^2 , \end{aligned}$$

which reduces to the desired expression.

It is known (cf. Feller (1968, page 225)) that $v - 1$ has a representation as a sum of independent geometric random variables. The mean and variance of v are

$$Ev = N \textstyle\sum_0^{n-1} (N-j)^{-1} , \tag{4}$$

and

$$\sigma_v^2 = N \textstyle\sum_1^{n-1} j(N-j)^{-2} . \tag{5}$$

It can be shown, using integral approximations and some elementary calculus, that

$$-(1/f)\log(1-g) < Ex < -(1/f)\log(1-f) < (1-f)^{-\frac{1}{2}} , \tag{6}$$

and

$$\sigma_x^2 < (1/n)[(1-f)^{-1} + (1/f)\log(1-f)] < (f/n)(1-f)^{-1} , \tag{7}$$

where

$$x = v/n , \qquad f = n/N , \qquad \text{and} \qquad g = n/(N+1) . \tag{8}$$

Proof of Corollary 1. From the Theorem and the fact that $V(\bar{y}_n) = (1-f)n^{-1}\sigma^2$, we have

$$\delta = (n-1)^{-1}(1-f)^{-1}[1 - Ex^{-1} + (n-3)Eh(x)] , \tag{9}$$
$$\text{where } h(x) = x^{-1}(1-x^{-1}),\ x \geqq 1 .$$

Note that $h''(x)$, the second derivative of h, satisfies

$$-4 \leqq h''(x) \leqq 1/128 , \qquad x \geqq 1 . \tag{10}$$

Applying Taylor's formula with remainder and (10), it can be verified that

$$h(Ex) - 2\sigma_x^2 \leqq Eh(x) \leqq h(Ex) + (1/256)\sigma_x^2 . \tag{11}$$

Also, it follows from the convexity of x^{-1} and from (6) that

$$0 < 1 - Ex^{-1} < 1 - (Ex)^{-1} < 1 - (1-f)^{\frac{1}{2}} < f . \tag{12}$$

Corollary 1 can now be verified by applying inequalities (7), (11), and (12) to expression (9).

Proof of Corollary 2. Since $h(x) = x^{-1}(1-x^{-1})$ is both concave and increasing for $1 \leqq x < 2$ and is decreasing for $x > 2$, it is not difficult to verify that

$$Eh(x) < h(Ex) , \qquad \text{if} \quad Ex < 2 . \tag{13}$$

The condition $f \leqq \frac{3}{4}$ implies that $Ex < 2$ (see (6)). Then, inequality (13) gives a sharper upper bound for $Eh(x)$ than the one given in (11). Using the new upper bound and repeating the steps in the proof of Corollary 1, it is easy to verify

$$\delta < \delta_1 + (n-1)^{-1}f(1-f)^{-1}. \tag{14}$$

Since $Ex < \lambda f + 1$ (see (6)) and $h(x)$ is increasing in x for $x < 2$, we have

$$\delta_1 < (1-f)^{-1}h(Ex) < (1-f)^{-1}h(\lambda f + 1) = \lambda f.$$

From this and (14), we conclude $\delta < (\lambda + \varepsilon_n)f$, with $\varepsilon_n = (1-f)^{-1}/(n-1)$.

REFERENCES

[1] BASU, D. (1958). On sampling with and without replacement. *Sankhyā* **20** 287–294.

[2] CHIKKAGOUDAR, M. S. (1966). A note on inverse sampling with equal probabilities. *Sankhyā Ser. A* **28** 93–96.

[3] FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**. Wiley, New York.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044