ORNL/TM-9864

ornl

OAK RIDGE NATIONAL LABORATORY

MARTIN MARIETTA

On a Central Limit Theorem for Variable Size Simple Random Sampling from a Finite Population

Tommy Wright

OPERATED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A03; Microfiche A01

ORNL/TM-9864

Engineering Physics and Mathematics Division

Mathematical Sciences Section

ON A CENTRAL LIMIT THEOREM FOR VARIABLE SIZE SIMPLE RANDOM SAMPLING FROM A FINITE POPULATION

Tommy Wright

Date Published - February 1986

Research supported by
Applied Mathematics Sciences Research Program,
U.S. Department of Energy
Office of Energy Research

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
operated by
Martin Marietta Energy Systems, Inc.
for the
U.S. Department of Energy
under Contract DE-AC05-84OR21400

CONTENTS

Abstract 1

1. Introduction 2

2. Variable Size Simple Random Sampling 3

3. Estimation of the Population Mean μ_Y Under Variable Size Simple Random Sampling 5

4. A Central Limit Theorem Under VSSRS 12

Acknowledgments 18

References 19

ON A CENTRAL LIMIT THEOREM FOR VARIABLE SIZE SIMPLE RANDOM SAMPLING FROM A FINITE POPULATION

Tommy Wright

ABSTRACT

This paper introduces a sampling plan for finite populations herein called "variable size simple random sampling" and compares properties of estimators based on it with results from the usual fixed size simple random sampling without replacement. Necessary and sufficient conditions (in the spirit of Hájek (1960)) for the limiting distribution of the sample total (or sample mean) to be normal are given.

KEY WORDS AND PHRASES: Finite population; Limiting distribution, Lindeberg-Hájek condition; Poisson sampling; Simple random sampling without replacement; Truncated binomial; Variable sample size.

1. INTRODUCTION

By design, the vast majority of statistical theory, methods, and practice assumes that an investigator is going to make inferences based on data derived from *fixed size sampling procedures*. A *fixed size sampling procedure* is a sampling procedure in which the size of the sample n is fixed before any selection begins, and in theory n is not permitted to vary. For example, this is indeed the usual case when sampling from a finite population of N units for estimation purposes. (See Cochran (1977); Hansen, Hurwitz, and Madow (1953); Kish (1965); Sukhatme and Sukhatme (1970); Hájek (1981); and Brewer and Hanif (1983).) Even though the intent may be to have a fixed size sampling procedure by design, the actual application may yield a sample whose size is different from that planned; this occurs, for example, in survey sampling when there is unit nonresponse (Madow, Nisselson, Olkin, and Rubin (1983)). Removing outliers from an observed sample without replacement also leads to a sample size that is different from that planned (Barnett, 1983). While unit nonresponse and removal of outliers lead to examples of unintentional variable size samples, some variable size samples are the result of careful planning that intentionally leads to variable size sampling procedures such as (1) sequential sampling procedures (Wald, 1947) and (2) Poisson sampling (sometimes called Bernoulli sampling) as discussed by Hájek (1981) and Strand (1979). A *variable size sampling procedure* is a sampling procedure in which the size of the sample n is by design and with intent permitted to vary during the selection process. While Cassel, Sarndal, and Wretman (1977) do not use this exact expression, they do consider such designs. Also, domain estimation (Cochran, 1977, pp. 34-39) in sampling from a finite population is based on samples where the size of the observed sample from a domain of interest is a random variable.

Although it is true in practice that with variable size sampling procedures there is the fear that the realized sample size might be too large to manage or too small to support needed analysis, it is important that properties of these procedures and estimators be examined for theoretical interest and as a first step toward variable sample size procedures where the size of the sample is permitted to vary over a specified subset of $\{1, 2, \ldots, N\}$ containing preferred values for n. In an earlier paper, we considered *variable size simple random sampling* (VSSRS) from a finite population (Wright, 1985), and in Section 2 of this paper, we define and summarize some of the properties of VSSRS. Section 3 compares sampling variances under VSSRS with the sampling variance under the usual (fixed size) simple random sampling (FSSRS). In Section 4, a central limit theorem under VSSRS is given which yields results similar to those of Hájek (1960).

2. VARIABLE SIZE SIMPLE RANDOM SAMPLING

Let $U = \{1, 2, 3, \ldots, N\}$ denote a finite population of N units and assume that the i^{th} unit has associated with it the real number Y_i for $i = 1, \ldots, N$. The value of the parameter of interest, $\mu_Y = \sum_{i=1}^{N} Y_i / N$, is assumed to be unknown. We assume that a sample will be selected to yield an estimate $\hat{\mu}_Y$ for μ_Y. Perhaps the most basic sampling plan, which is the basis for numerous other sampling designs, is simple random sampling without replacement. Under *simple random sampling without replacement*, the size of the sample n is assumed to be fixed and is determined before any sampling begins; sampling is performed so that each of the $\binom{N}{n}$ samples has probability $\binom{N}{n}^{-1}$ of being selected. For the remainder of this paper we will refer to this sampling plan as *fixed size simple random sampling* (FSSRS).

In the following definition, we introduce a variable size sampling plan for a finite population where the size n of the observed sample is not fixed before sampling begins, but is a random variable that takes on integer values between 1 and N inclusively with maximum probability of n being at $n = N/2$ and $(N+2)/2$ if N is even and at $n = (N+1)/2$ if N is odd.

Definition. If the sampling plan is such that each of the $2^N - 1$ nonempty subsets of U has an equal probability of selection, then the sampling plan is called *variable size simple random sampling* (VSSRS), and the observed sample is called a *variable size simple random sample*.

Under VSSRS, n is a random variable whose distribution is given in Lemma 1.

Lemma 1. If n is the size of the observed sample under VSSRS, then the probability function of n is

$$P(n = j) = \binom{N}{j} / (2^N - 1) \qquad \text{for} \qquad j = 1, 2, \ldots, N \ . \tag{1}$$

Proof. The proof is immediate because the probability of each possible sample is $1/(2^N - 1)$ and there are $\binom{N}{j}$ ways of selecting j units from N different units when order is unimportant. ∎

Note that $P(n = j) = P(n = N - j)$ for $j = 1, 2, \ldots, N - 1$ and hence the distribution of n is symmetric except for $n = N$. It is easy to see from Lemma 1 that n is a truncated binomial random variable at zero with parameters N and 1/2 (Johnson and Kotz, 1969, pp. 73-74). The following properties of n and the sampling plan all follow from Lemma 1.

Property 1. The characteristic function of n is

$$E(\exp(itn)) = \left\{(\exp(it) + 1)^N - 1\right\} / (2^N - 1) \qquad \text{for} \qquad t \in \mathbb{R} \ . \tag{2}$$

Property 2. If $\pi = 2^{N-1}/(2^N - 1)$, then

$$E(n) = N\pi \quad , \quad \text{and}$$
$$Var(n) = N\pi(1 - \pi)(2^N + 1 - N) / (2^{N+1} - 2 - 2^N) \ . \tag{3}$$

For large N, $E(n) \approx N/2$ and $Var(n) \approx N/4$. Hence VSSRS is of limited practical use because the average size of the sample is approximately one-half the size of the population, which would be large in many cases. But as was noted earlier, our interest in this paper is theory; and we view VSSRS as a first step towards finding more practical variable size sampling plans.

Property 3. Let i and j be two different but arbitrary fixed units in U. Then

$$\pi_i = P(i \text{ is included in the sample}) = \pi \text{ , and}$$
$$\pi_{ij} = P(i \text{ and } j \text{ are included in the sample}) = \pi/2 \text{ .} \tag{4}$$

Property 3 implies that each unit has an equal probability of sample inclusion and so does each pair. This comment is also true for FSSRS; however in that case, $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$.

It is immediately clear that a variable size simple random sample can be realized by applying one of the following three sampling methods. One method, sometimes referred to as the "mass draw" technique, calls for listing all possible $2^N - 1$ subsets of the population of size N and picking one of them with probability $1/(2^N - 1)$. This method is directly from the definition and seems practical for relatively small values of N. An alternative method is to select the sample in two stages. On the first stage, select the size of the sample $n = j$ with probability $P(n = j) = \binom{N}{j}/(2^N - 1)$ for $j = 1, 2, \ldots, N$. On the second stage given $n = j$, select from among the $\binom{N}{j}$ subsets of j units, one with probability $1/\binom{N}{j}$. That is, on the second stage one selects a fixed size simple random sample of j units. A third method, which gives what is called a sampling scheme in which we select the units for the sample one-by-one, can be obtained via the method presented by Rao (1962).

3. ESTIMATION OF THE POPULATION MEAN μ_Y UNDER VARIABLE SIZE SIMPLE RANDOM SAMPLING

In this section, we consider some results about the statistic $\overline{y}$, the sample mean, under VSSRS and make some comparisons with the statistic $\overline{y}$ under FSSRS.

Lemma 2. Under VSSRS, $\overline{y}$ is an unbiased estimator of μ_Y.

Proof. Let A = {all nonempty samples of U} and A_j = {all samples of size $n = j$, where $1 \leqslant j \leqslant N$} and note that $\sum_{j=1}^{N} \binom{N}{j} = 2^N - 1$. Also for $\alpha \in A$ or $\alpha \in A_j$, let $\overline{y}_\alpha$ be the sample mean of the units in α. Then

$$E(\bar{y}) = \sum_{\alpha \in A} \bar{y}_\alpha / (2^N - 1) = \sum_{j=1}^{N} \sum_{\alpha \in A_j} \bar{y}_\alpha / (2^N - 1)$$

$$= \sum_{j=1}^{N} \binom{N}{j} \mu_Y / (2^N - 1)$$

$$= \mu_Y \ ,$$

as was to be shown. ■

Lemma 3. Under VSSRS, $Var(\bar{y}) = \sigma_Y^2(NE(1/n) - 1)/(N - 1)$, where $\sigma_Y^2 = \sum_{i=1}^{N} (Y_i - \mu_Y)^2/N$.

Proof. $$Var(\bar{y}) = \sum_{\alpha \in A} (\bar{y}_\alpha - \mu_Y)^2 / (2^N - 1)$$

$$= \sum_{j=1}^{N} \sum_{\alpha \in A_j} (\bar{y}_\alpha - \mu_Y)^2 / (2^N - 1)$$

$$= \left[\sum_{j=1}^{N} \binom{N}{j} (N - j)\, \sigma_Y^2 / (N - 1) j \right] / (2^N - 1)$$

$$= \left[\sum_{j=1}^{N} \binom{N}{j} N \sigma_Y^2 / (N - 1) j - \sum_{j=1}^{N} \binom{N}{j} j \sigma_Y^2 / (N - 1) j \right] / (2^N - 1)$$

$$= \sigma_Y^2 \left[N \sum_{j=1}^{N} \binom{N}{j} / j - (2^N - 1) \right] / (N - 1)(2^N - 1)$$

$$= \sigma_Y^2 (NE(1/n) - 1)/(N - 1) \ . \ ■$$

Lemma 4. Let $Var(\bar{y})$ be the variance of $\bar{y}$ under VSSRS and $Var(\bar{y} \mid j) = (N - n)\sigma_Y^2 / (N - 1)n$ be the usual variance of $\bar{y}$ under FSSRS for fixed sample size $n = j$. Then

$$Var(\bar{y}) = E(Var(\bar{y} \mid n)) \ .$$

Proof. The result follows immediately from Lemma 3. ■

It is clear that in general $Var(\bar{y}|n)$ is a nondegenerate random variable in n for $n = 1, 2, \ldots, N$. From Lemma 4, the expected value of $Var(\bar{y}|n)$ under FSSRS is the same as $Var(\bar{y})$ under VSSRS. Thus there are values of n for which $Var(\bar{y}) \leqslant Var(\bar{y}|n)$ and other values of n for which $Var(\bar{y}) \geqslant Var(\bar{y}|n)$. In the remainder of this section, we generalize the direction of the inequality for the various values of n in Theorem 1. Generally, for values of $n > N/2$, $Var(\bar{y}) > Var(\bar{y}|n)$; and the direction is reversed for values of $n < N/2$. The next two lemmas will be used in the proof of Theorem 1.

Lemma 5. Let n be any discrete random variable with probability function defined for $n = 1, 2, \ldots, N$, where $N < \infty$. Then

$$E(n)E(1/n) \geqslant 1 . \tag{5}$$

Proof. By the Cauchy-Schwarz Inequality, for every $a_1, \ldots, a_N, b_1, \ldots, b_N \in \mathbb{R}$, we have $\left(\sum_{n=1}^{N} a_n b_n\right)^2 \leqslant \sum_{n=1}^{N} a_n^2 \sum_{n=1}^{N} b_n^2$. If we take $a_n = \sqrt{nP(n)}$ and $b_n = \sqrt{P(n)/n}$, then we observe

$$\left(\sum_{n=1}^{N} \sqrt{nP(n)}\sqrt{P(n)/n}\right)^2 \leqslant \sum_{n=1}^{N} \left(\sqrt{nP(n)}\right)^2 \sum_{n=1}^{N} \left(\sqrt{P(n)/n}\right)^2 .$$

Because $P(n)$ and n are nonnegative $\forall n$, we have

$$\left(\sum_{n=1}^{N} P(n)\right)^2 \leqslant \left(\sum_{n=1}^{N} nP(n)\right)\left(\sum_{n=1}^{N} P(n)/n\right),$$

which is the desired result. ■

Lemma 6. If n has the probability function given in (1), then for all N,

$$E(n-1)E(1/n) \leqslant 1 . \tag{6}$$

Proof. From equation (11) of Stephan (1945), we note that for $n > 0$, $1/n$ can be expressed as a series of inverse factorials by

$$\frac{1}{n} = \frac{0!}{n+1} + \frac{1!}{(n+1)(n+2)} + \cdots + \frac{(i-1)!n!}{(n+i)!} + \cdots + \frac{(t-1)!n!}{(n+t)!} + R_t(n) , \tag{7}$$

where $R_t(n) = t!(n-1)!/(n+t)!$ is the remainder after the first t terms. For $t = 3$, we have

$$\frac{1}{n} = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{2!}{(n+1)(n+2)(n+3)} + \frac{3!}{n(n+1)(n+2)(n+3)} \,. \quad (8)$$

For $n \geqslant 1$, it is easy to see that

$$\begin{aligned} \frac{1}{n} &\leqslant \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{2}{(n+1)(n+2)(n+3)} + \frac{3!}{(n+1)(n+2)(n+3)} \\ &= \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{8}{(n+1)(n+2)(n+3)} \,. \end{aligned} \quad (9)$$

Next, we show that

$$\begin{aligned} E(1/n) \leqslant (&12N^2 2^N + 84N 2^N + 528 \cdot 2^N - 17N^3 \\ &- 138N^2 - 421N - 528)/\, 6(N+1)(N+2)(N+3)(2^N - 1) \,. \end{aligned} \quad (10)$$

By (9) and the fact that n is a truncated binomial random variable at zero, we have

$$\begin{aligned} E\left(\frac{1}{n}\right) &\leqslant E\left(\frac{1}{n+1}\right) + E\left(\frac{1}{(n+1)(n+2)}\right) + 8E\left(\frac{1}{(n+1)(n+2)(n+3)}\right) \\ &= \frac{1}{\frac{(N+1)}{2}(1 - \frac{1}{2^N})} \left\{ \sum_{n=1}^{N} \binom{N+1}{n+1} \left(\frac{1}{2}\right)^{N+1} + \frac{2}{(N+2)} \sum_{n=1}^{N} \binom{N+2}{n+2} \left(\frac{1}{2}\right)^{N+2} \right. \\ &\quad \left. + \frac{32}{(N+3)(N+2)} \sum_{n=1}^{N} \binom{N+3}{n+3} \left(\frac{1}{2}\right)^{N+3} \right\} . \end{aligned} \quad (11)$$

The first sum in the brackets is $P(X \geqslant 2)$, where X is a binomial random variable with parameters $N+1$ and 1/2. Similarly, the second and third sums in the brackets are respectively $P(Y \geqslant 3)$ and $P(W \geqslant 4)$, where Y and W are binomial random variables with respective parameters $N + 2$, 1/2 and $N + 3$, 1/2. Thus

$$\sum_{n=1}^{N} \binom{N+1}{n+1} \left(\frac{1}{2}\right)^{N+1} = 1 - \sum_{X=0}^{1} P(X) = 1 - (N+2)/\,(2^{N+1})\,,$$

$$\sum_{n=1}^{N} \binom{N+2}{n+2} \left(\frac{1}{2}\right)^{N+2} = 1 - \sum_{Y=0}^{2} P(Y) = 1 - (N^2 + 5N + 8)/\,2^{N+3}, \quad \text{and}$$

$$\sum_{n=1}^{N} \binom{N+3}{n+3} \left(\frac{1}{2}\right)^{N+3} = 1 - \sum_{W=0}^{3} P(W) = 1 - (N^3 + 9N^2 + 32N + 48)/\,3{\cdot}2^{N+4}\,.$$

Thus substituting these three equalities into (11) yields $E\left(\frac{1}{n}\right) \leqslant (12N^2 2^N + 84N2^N + 528{\cdot}2^N - 17N^3 - 138N^2 - 421N - 528)/6(N+1)(N+2)(N+3)(2^N-1)$, which establishes (10).

By Property 2 and (10),

$$E(n-1)E\left(\frac{1}{n}\right) \leqslant 1 \quad \text{for all } N$$

if

$$\left[\frac{N}{2^N - 1}\, 2^{N-1} - 1\right] \left[\frac{12N^2 2^N + 84N2^N + 528{\cdot}2^N - 17N^3 - 138N^2 - 421N - 528}{6(N+1)(N+2)(N+3)(2^N-1)}\right] \leqslant 1 \quad \text{for all } N\,,$$

if and only if

$$\begin{aligned} 114N2^{2N} + \frac{23}{2}N^2 2^N + 373N2^N + 1128{\cdot}2^N \leqslant{}& 6N^2 2^{2N} + \frac{17}{2}N^4 2^N + 40N^3 2^N \\ &+ 564{\cdot}2^{2N} + 17N^3 + 138N^2 \\ &+ 421N + 528 \qquad \text{for all } N\,. \end{aligned} \tag{12}$$

Now the left-hand side of (12) is strictly less than the right-hand side for $N > 19$ because

$$114N2^{2N} < 6N^2 2^{2N} \quad \text{if} \quad N > 19\,,$$

$$\frac{23}{2}N^2 2^N < 40N^3 2^N \quad \text{if} \quad N > 1\,,$$

$$373N2^N < \frac{17}{2}N^4 2^N \quad \text{if} \quad N > 4\,,$$

$$\text{and} \quad 1128 \cdot 2^{N} < 564 \cdot 2^{2N} \quad \text{if} \quad N > 1 .$$

Table 1 gives the values for the left-hand and right-hand sides for $N = 1, 2, \ldots, 20$. Hence, Lemma 6 follows. ■

TABLE 1. Values of (12) for $1 \leqslant N \leqslant 10$.

N	Left-hand side	Right-hand side
1	3,481	3,481
2	11,328	13,290
3	40,692	57,192
4	161,600	250,244
5	688,656	1,069,344
6	3,043,584	4,464,558
7	13,625,152	18,441,988
8	61,009,920	76,305,336
9	271,732,992	318,762,480
10	1,201,528,832	1,348,578,002
11	5,273,220,096	5,774,602,692
12	22,980,968,448	24,962,979,372
13	99,520,221,184	108,606,526,576
14	428,563,955,712	474,226,047,222
15	1,836,403,605,504	2,073,666,106,404
16	7,834,678,329,344	9,066,676,074,080
17	33,296,001,073,152	39,598,149,533,712
18	141,015,398,744,064	172,643,510,800,794
19	595,392,030,048,256	751,141,298,547,556
20	2,506,900,339,949,568	3,260,714,072,608,212

Theorem 1. (1) For $n \geqslant E(n)$, $Var(\bar{y}) \geqslant Var(\bar{y} \mid n)$.

(2) For $n < E(n)$, $Var(\bar{y}) \leqslant Var(\bar{y} \mid n)$.

Proof of (1). To prove part (1), first let $n = E(n)$. By Lemma 3, $Var(\bar{y}) = \sigma_Y^2(NE(1/n) - 1) / (N - 1)$. From Property 2, we have

$$Var(\bar{y} \mid E(n)) = (N - E(n))\sigma_Y^2 / (N - 1)E(n)$$
$$= \sigma_Y^2(1 - 1/2^{N-1}) / (N - 1) .$$

By Lemma 5, $E(n) \geqslant 1/E(1/n)$ if and only if

$$NE\left(\frac{1}{n}\right) - 1 \geqslant 1 - \frac{1}{2^{N-1}}$$

or equivalently

$$\sigma_Y^2\left(NE\left(\frac{1}{n}\right) - 1\right)/(N-1) \geqslant \sigma_Y^2(1 - 1/2^{N-1})/(N-1)\,,$$

which is the same as

$$Var(\bar{y}) \geqslant Var(\bar{y} \mid E(n))\,.$$

The rest of part (1) follows because $Var(\bar{y} \mid n) = \dfrac{\sigma_Y^2}{(N-1)}\left(\dfrac{N}{n} - 1\right)$ is a strictly decreasing function in n.

Proof of (2). To prove part (2), let $n = E(n) - 1$ or equivalently $E(n-1)$. Then

$$Var(\bar{y}) \leqslant Var(\bar{y} \mid E(n-1))$$

if and only if

$$\sigma_Y^2(NE(1/n) - 1)/(N-1) \leqslant \sigma_Y^2(N/E(n-1) - 1)/(N-1)\,,$$

if and only if

$$E(n-1)E(1/n) \leqslant 1\,,$$

which is so by Lemma 6. Because $Var(\bar{y} \mid n)$ is a strictly decreasing function in n, it follows that $Var(\bar{y}) \leqslant Var(\bar{y} \mid n)$ when $n < E(n)$. This completes the proof of Theorem 1. ■

From Theorem 1, if $n \geqslant E(n)$, then FSSRS makes $\bar{y}$ as precise an estimator of μ_Y as VSSRS, and VSSRS makes $\bar{y}$ as precise an estimator of μ_Y as FSSRS when $n < E(n)$.

Remark. Note that the inequality in Lemma 5 is a general result, while the inequality in Lemma 6 is not. At first glance, one might think that the inequality in Lemma 6 should be true for any random variable defined over the first N positive integers because it makes one think about the inequality $(n-1)/n < 1$, which is always true for any value of $n \geqslant 1$. The following example illustrates that the inequality in Lemma 6 does not hold in general.

Example. Let the probability function of the random variable n be defined by

$$P(n) = \begin{cases} 1/2 & \text{for } n = 1, N \\ 0 & \text{otherwise}. \end{cases}$$

Then it is easy to show that $E(n-1)E\left(\frac{1}{n}\right) = (N-1)(N+1)/4N > 1$ if $N \geqslant 5$.

4. A CENTRAL LIMIT THEOREM UNDER VSSRS

In this section, we justify necessary and sufficient conditions for the limiting distribution of $\xi = n\overline{y}$ under VSSRS to be normally distributed. The approach and result are quite similar to that taken initially by Hájek (1960) and used by others, including Scott and Wu (1981). The main theorem follows from Lemma 7 and the proof of Theorem 3.1 of Hájek (1960).

Recall the definition of VSSRS and the probability function of n given in (1). Another example of variable size sampling from a finite population that is a special case of what Hájek (1981) refers to as Poisson sampling of mean size n is given by

$$P(s_k \mid n) = \left(\frac{n}{N}\right)^k \left(1 - \frac{n}{N}\right)^{N-k}, \tag{13}$$

where s_k is a subset of U consisting of k units and $P(s_k \mid n)$ is the probability of s_k given n. It is important to note that we think of Poisson sampling here as being conditional on n.

If $\xi = n\overline{y} = \sum_{i=1}^{n} Y_i$, where $\{Y_1, \ldots, Y_n\}$ is the set of observed values of the variable size simple random sample s_n, then it follows that, given n, ξ has finite conditional mean value

$$E(\xi | n) = \frac{n}{N} \sum_{i=1}^{N} Y_i \tag{14}$$

and conditional variance

$$Var(\xi | n) = \frac{n}{N} \frac{N-n}{N-1} \sum_{i=1}^{N} (Y_i - \mu_Y)^2 . \tag{15}$$

Next consider an infinite sequence of VSSRS experiments $\{(Y_{\nu 1}, Y_{\nu 2}, \ldots, Y_{\nu N_\nu}), n_\nu, k_\nu, N_\nu\}$, where N_ν is the size of the ν^{th} population with population values $Y_{\nu 1}, Y_{\nu 2}, \ldots, Y_{\nu N_\nu}$; n_ν is the size of the variable size simple random sample from the ν^{th} population; and k_ν is the size of the conditional Poisson sample s_{k_ν} from the ν^{th} population for $\nu = 1, 2, \ldots$ In what follows, we consider the ν^{th} experiment and for simplicity will omit the subscripts ν until needed again.

Recall that under VSSRS, n is a truncated binomial random variable at zero with parameters N and 1/2. It is easy to see that under Poisson sampling (conditional on n), k is a binomial random variable with parameters N and n/N. Hence

$$E(k | n) = n \quad \text{and} \quad Var(k | n) = E((k - n)^2 | n) = n \left(1 - \frac{n}{N} \right). \tag{16}$$

Next we define an experiment that is an extension of the one proposed by Hájek (1960) and yields simultaneously a variable size simple random sample s_n and a Poisson sample s_k, where $s_n \subset s_k$ if $n \leqslant k$ and $s_k \subset s_n$ if $k \leqslant n$. The experiment consists of the following three steps:

1. First realize a value n according to

$$P(n) = \binom{N}{n} / (2^N - 1) \qquad \text{for} \quad 1 \leqslant n \leqslant N .$$

2. Next realize a value k given n according to

$$P(k | n) = \binom{N}{k} \left(\frac{n}{N} \right)^k \left(1 - \frac{n}{N} \right)^{N-k} \qquad \text{for} \quad 0 \leqslant k \leqslant N .$$

3. There are three cases to consider for step 3.

   (a) When $k = n$, select a fixed size simple random sample $s_n = s_k$ that is a simultaneous realization of VSSRS and Poisson sampling.

   (b) When $k > n$, select a fixed size simple random sample s_k and (given s_k) select a fixed size simple random sample s_n from s_k. The observed s_k is our realization of Poisson sampling, and s_n is our realization of VSSRS.

   (c) When $k < n$, select a fixed size simple random sample s_n and (given s_n) select a fixed size simple random sample s_k from s_n. The observed s_k is our realization of Poisson sampling, and s_n is our realization of VSSRS.

Now for the overlapping samples s_n and s_k just described, let

$$\phi = \sum_{i \in s_n} (Y_i - \mu_Y) = \xi - n\mu_Y \tag{17}$$

and

$$\phi^* = \sum_{i \in s_k} (Y_i - \mu_Y) . \tag{18}$$

Note that

$$\phi - \phi^* = \begin{cases} 0 & \text{if } k = n \\ \sum_{i \in s_n - s_k} (Y_i - \mu_Y) & \text{if } k < n \\ \sum_{i \in s_k - s_n} (Y_i - \mu_Y) & \text{if } k > n. \end{cases} \tag{19}$$

Lemma 7.. If $N_\nu \to \infty$ as $\nu \to \infty$, then $\phi - \phi^*$ converges in probability to zero as $\nu \to \infty$.

Proof. First we will show that

$$E((\phi - \phi^*)^2) \leqslant \sigma_Y^2 \sqrt{E(n(1 - \frac{n}{N}))} . \tag{20}$$

If n and k are fixed, then $s_n - s_k$ (or $s_k - s_n$) represents a fixed size simple random sample from U of size $|k - n|$ by Results 0 and 1 of Wright and Tsao (1985). Thus from (15) we have

$$E((\phi-\phi^*)^2 | n,k) = Var((\phi-\phi^*) | n,k)$$

$$= \frac{|k-n|}{N}\ \frac{N-|k-n|}{N-1} \sum_{i=1}^{N} (Y_i-\mu_Y)^2 \tag{21}$$

$$\leqslant |k-n| \sigma_Y^2 .$$

Thus from (21), an application of the Cauchy-Schwarz Inequality, and (16), we have

$$E((\phi-\phi^*)^2 | n) = E(E((\phi-\phi^*)^2 | n,k) | n)$$

$$\leqslant \sigma_Y^2\, E(|k-n|\ | n)$$

$$= \sigma_Y^2\, E(\sqrt{(k-n)^2}\, | n) \tag{22}$$

$$= \sigma_Y^2 \sum_k \sqrt{(k-n)^2}\, P(k | n)$$

$$\leqslant \sigma_Y^2 \sqrt{\sum_k (k-n)^2 P(k | n)}$$

$$= \sigma_Y^2 \sqrt{n(1-\frac{n}{N})} .$$

Also by (22) and another application of the Cauchy-Schwarz Inequality,

$$E((\phi-\phi^*)^2) = E(E((\phi-\phi^*)^2 | n))$$

$$\leqslant E\left|\sigma_Y^2 \sqrt{n(1-\frac{n}{N})}\right| \tag{23}$$

$$= \sigma_Y^2 \left|\sum_n \sqrt{n(1-\frac{n}{N})}\, P(n)\right|$$

$$\leqslant \sigma_Y^2 \sqrt{E(n(1-\frac{n}{N}))} .$$

Thus (20) has been shown. Next we show that

$$Var(\phi^*) = \sigma_Y^2 E(n(1-\frac{n}{N})) . \tag{24}$$

To show (24), first note that Poisson sampling as described in (13) can be achieved as follows: "For each unit of the population of size N, perform one Bernoulli trial. If a success occurs, the trial unit is accepted as part of the (Poisson) sample; otherwise the unit is passed up. The probability of success $\frac{n}{N}$ is assumed to be the same for all trials, and the trials are mutually independent" (Strand, 1979). This implies that ϕ^* based on s_k can be presented as a sum of N independent random variables,

$$\phi^* = \sum_{i=1}^{N} \zeta_i \, . \tag{25}$$

where

$$\zeta_i = \begin{cases} Y_i - \mu_Y & \text{with probability } \frac{n}{N} \quad (\text{if } i \in s_k) \\ \\ 0 & \text{with probability } 1 - \frac{n}{N} \quad (\text{if } i \in U - s_k). \end{cases} \tag{26}$$

Thus $E(\zeta_i | n) = (Y_i - \mu_Y)n / N$, $Var(\zeta_i | n) = (Y_i - \mu_Y)^2 n(N - n)/ N^2$, $E(\phi^* | n) = 0$, and $Var(\phi^* | n) = \sum_{i=1}^{N} Var(\zeta_i | n) = \sigma_Y^2 n(N - n)/ N$. Hence

$$\begin{aligned} Var(\phi^*) &= Var(E(\phi^* | n)) + E(Var(\phi^* | n)) \\ &= E(Var(\phi^* | n)) \\ &= \sigma_Y^2 E(n(1 - \frac{n}{N})) \, , \end{aligned} \tag{27}$$

which demonstrates (24).

From (20) and (24), it follows that

$$\begin{aligned} E((\phi - \phi^*)^2)/ Var(\phi^*) &\leq \sigma_Y^2 \sqrt{E(n(1 - \frac{n}{N}))} / \sigma_Y^2 E(n(1 - \frac{n}{N})) \, , \\ &= \sqrt{2N/ (N - 1)E(n)} \, , \end{aligned} \tag{28}$$

because from Properties 1 and 2 one can show that $E(n^2) = (N + 1)E(n)/ 2$. Using a generalization of Chebyshev's Inequality given on pp. 54-55 of Hogg and Craig (1972) with the random variable $(\phi - \phi^*)^2$ and constant $c = k^2 Var(\phi^*)$, where $k > 0$, it follows that

$$P\left\{(\phi - \phi^*)^2 \geq k^2 Var(\phi^*)\right\} \leq E((\phi - \phi^*)^2)/\, k^2 Var(\phi^*) \,,$$

or equivalently

$$P\left\{|\phi - \phi^*| \geq k\sqrt{Var(\phi^*)}\right\} \leq E((\phi - \phi^*)^2)/\, k^2 Var(\phi^*) \,. \tag{29}$$

Thus $\forall \epsilon > 0$ we can have $P(|\phi - \phi^*| \geq \epsilon) = P\left\{|\phi - \phi^*| \geq k\sqrt{Var(\phi^*)}\right\}$, where $k = \epsilon/\sqrt{Var(\phi^*)}$.

Now reconsidering the sequence of experiments $\{(Y_{\nu 1}, \ldots, Y_{\nu N_\nu}), n_\nu, k_\nu, N_\nu\}$, let ϕ_ν and ϕ_ν^* be the random variables given in (17) and (18) corresponding to the ν^{th} experiment. Thus $\forall \epsilon > 0$

$$\begin{aligned}
\lim_{\nu \to \infty} P\left\{|\phi_\nu - \phi_\nu^*| \geq \epsilon\right\} &= \lim_{N_\nu \to \infty} P\left\{|\phi_\nu - \phi_\nu^*| \geq k\sqrt{Var(\phi_\nu^*)}\right\} \\
&\leq \lim_{N_\nu \to \infty} \left\{E((\phi_\nu - \phi_\nu^*)^2)/\, k^2 Var(\phi_\nu^*)\right\} \qquad \text{by (29)} \\
&\leq \lim_{N_\nu \to \infty} \sqrt{2N_\nu/\,(N_\nu - 1)E(n_\nu)}/\, k^2 \qquad \text{by (28)} \\
&= 0 \qquad \text{since } E(n_\nu) = N_\nu 2^{N_\nu - 1}/\,(2^{N_\nu} - 1)\,.
\end{aligned}$$

Hence the proof of Lemma 7 is complete. ■

From Lemma 7, it is clear that the limiting distributions of ϕ_ν and ϕ_ν^* are the same provided that limiting distributions exist. Hence in talking about the limiting distribution of ϕ_ν(or ξ_ν), it is enough to consider the limiting distribution of ϕ_ν^*.

Lemma 8. Let $U_\nu = \{1, 2, \ldots, N_\nu\}$ and let $U_{\nu\tau} = \{i \mid i \in U_\nu$ and $|Y_{\nu i} - \mu_{Y\nu}| > \tau\sqrt{Var(\phi_\nu^*)}\}$ where $\tau > 0$. Let $N_\nu \to \infty$ as $\nu \to \infty$. Then the random variable ϕ_ν^* is asymptotically normal with mean 0 and variance $Var(\phi_\nu^*)$ if and only if

$$\lim_{\nu \to \infty} \frac{\sum_{i \in U_{\nu\tau}} (Y_{\nu i} - \mu_{Y\nu})^2}{\sum_{i \in U_\nu} (Y_{\nu i} - \mu_{Y\nu})^2} = 0 \text{ for } \tau > 0\,. \tag{30}$$

Proof. The proof is essentially equivalent to that of Theorem 3.1 in Hájek (1960). ■

A CENTRAL LIMIT THEOREM UNDER VSSRS. Let $N_\nu \to \infty$ as $\nu \to \infty$. Then under VSSRS,

$$\lim_{\nu \to \infty} P\{\xi_\nu - E(\xi_\nu) < x\sqrt{Var(\xi_\nu)}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt$$

if and only if (30) holds for $\{(Y_{\nu 1}, \ldots, Y_{\nu N_\nu}), n_\nu, N_\nu\}$.

Proof. Follows from Lemmas 7 and 8. ■

The condition (30) has been referred to as the Lindeberg - Hájek Condition (see, e.g., Scott and Wu, 1981) because it occurs in Hájek's Theorem for finite populations. Hájek's Theorem is proved using the Lindeberg Condition of the Central Limit Theorem for independently distributed random variables.

Acknowledgments

I am grateful to my colleagues, G. Ostrouchov and V. R. R. Uppuluri, whose discussions were helpful in the preparation of parts of this paper.

REFERENCES

BARNETT, V. (1983). Principles and Methods for Handling Outliers in Data Sets. (In *Statistical Methods and The Improvement of Data Quality*, T. Wright, ed.) 131-166. Academic Press, New York.

BREWER, K. R. W. and HANIF, M. (1983). *Sampling with Unequal Probabilities*. Springer - Verlag, New York.

COCHRAN, W. G. (1977). *Sampling Techniques* (3rd Edition). John Wiley and Sons, New York.

HÁJEK, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population. *Pub. Math. Inst. Hungarian Acad. Sci.*, 5, 361-374.

HÁJEK, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

HALDANE, J. B. S. (1945). On a Method of Estimating Frequencies. *Biometrika*, 33, 222-225.

HANSEN, M. H., HURWITZ, W. N., and MADOW, W. G. (1953). *Sample Survey Methods and Theory, Vols. I and II.* John Wiley and Sons, New York.

HOGG, R. V. and CRAIG, A. T. (1970). *Introduction to Mathematical Statistics* (3rd Edition). Macmillan, New York.

JOHNSON, N. I. and KOTZ, S. (1969). *Distributions in Statistics: Discrete Distributions*. John Wiley and Sons, New York.

KISH, L. (1965). *Survey Sampling*. John Wiley and Sons, New York.

MADOW, W. G., NISSELSON, H., OLKIN, I., and RUBIN, D. (eds.) (1983). *Incomplete Data in Sample Surveys, Vols. 1-3*, Academic Press, New York.

RAO, T. V. H. (1962). An Existence Theorem in Sampling Theory. *Sankhyá*, 24, 327-330.

SCOTT, A. and WU, C. -F. (1981). On the Asymptotic Distribution of Ratio and Regression Estimators. *Journal of the American Statistical Association*, 76, #373, 98-102.

STEPHAN, F. F. (1945). The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate. *Annals of Mathematical Statistics*, 16, 50-61.

STRAND, M. M. (1979). Estimation of a Population Total Under a 'Bernoulli Sampling' Procedure. *The American Statistician*, 33, 81-84.

SUKHATME, P. V. and SUKHATME, B. V. (1970). *Sampling Theory of Surveys with Applications* (2nd Revised Ed.). Iowa State University Press, Ames, Iowa.

WALD, A. (1947). *Sequential Analysis*. John Wiley and Sons, New York.

WRIGHT, T. and TSAO, H. J. (1985). Some Useful Notes on Simple Random Sampling. *Journal of Quality Technology*, 17, #2, 67-73.

WRIGHT, T. (1985). On Variable Size Simple Random Sampling From a Finite Population. College of Bus. Adm./Statistics Dept. Working Paper Series #213, University of Tennessee, Knoxville.

ORNL/TM-9864

INTERNAL DISTRIBUTION

1. L. S. Abbott
2. V. Alexiades
3. J. J. Beauchamp
4. T. S. Darland
5. M. V. Denson
6. D. J. Downing
7. D. M. Flanagan
8. E. L. Frome
9. L. J. Gray
10. R. F. Harbison
11. T. L. Hebble
12. H. Hwang
13. W. F. Lawkins
14. E. Leach
15. W. E. Lever
16. F. C. Maienschein
17. S. A. McGuire
18. T. J. Mitchell
19. M. D. Morris
20. G. Ostrouchov
21. S. Thompson
22. V. R. R. Uppuluri
23. R. C. Ward
24. D. G. Wilson

25-29. T. Wright

30. A. Zucker
31. P. W. Dickson, Jr. (Consultant)
32. G. H. Golub (Consultant)
33. R. M. Haralick (Consultant)
34. D. Steiner (Consultant)
35. Central Research Library
36. K-25 Plant Library
37. ORNL Patent Office
38. Y-12 Technical Library /Document Reference Section
39. Laboratory Records - RC

40-41. Laboratory Records Department

EXTERNAL DISTRIBUTION

42. Donald M. Austin, ER-7, Applied Mathematical Sciences, Scientific Computing Staff, Office of Energy Research, Office G-437 Germantown, Washington, DC 20545
43. Barbara A. Bailar, Statistical Standards and Methodology, Bureau of the Census, Washington, DC 20233
44. Leroy Bailey, Statistical Research Division, U. S. Bureau of the Census, 4413 Weldon Drive, Temple Hills, Maryland 20748
45. Charles B. Bell, Jr., Department of Mathematical Sciences, San Diego State University, College Avenue, San Diego, California 92182
46. Yvonne M. Bishop, Office of Statistical Standards, Energy Information Administration, Department of Energy, 1000 Independence Avenue, Washington, DC 20585
47. Michael R. Chernick, Mission Information Systems Division, The Aerospace Corporation, P. O. Box 92957, MS M4-955, Los Angeles, California 90009
48. Douglas J. DePriest, Mathematical Sciences Division, Office of Naval Research, 800 N. Quincy Street, Code 411, Arlington, Virginia 22217
49. Keith R. Eberhardt, Statistical Engineering Division, National Bureau of Standards, Administration Bldg., Rm. A337, Gaithersburg, Maryland 20899
50. Wayne A. Fuller, Statistics Department, Iowa State University, 221 Snedecor Hall, Ames, Iowa 50011

51. V. P. Godambe, Statistics Department, University of Waterloo, University Avenue, Waterloo, Ontario, Canada N2L 3G1
52. Louis Gordon, Mathematics Department, University of Southern California, DRB 306, Los Angeles, California 90089-1113
53. David Hall, Statistics, Systems Department, Pacific Northwest Laboratory, P.O. Box 999, Richland, Washington 99352
54. Morris H. Hansen, Westat Inc., 1650 Research Blvd., Rockville, Maryland 20850
55. Robert H. Harvey, Research Initiation, National Science Foundation, 1800 G. Street, N.W., Washington, DC 20550
56. Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106
57. Benjamin F. King, Jr., Survey Methods, Research Statistics Group, Educational Testing Service, Princeton, New Jersey 08541
58. Leslie Kish, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106
59. George Legall, Statistics Department, University of Tennessee, Knoxville, Tennessee 37996
60. Roderick J. Little, Biomathematics Department, University of California at Los Angeles, Los Angeles, California 90024
61. William G. Madow, 2293 Greer Road, Palo Alto, California 94303
62. George P. McCabe, Jr., Department of Statistics, Purdue University, West Lafayette, Indiana 47907
63. Maj. C. E. Oliver, ER-7, Applied Mathematical Sciences, Scientific Computing Staff, Office of Energy Research, Office G-437 Germantown, Washington, DC 20545
64. Ronald Peierls, Applied Mathematics Department, Brookhaven National Laboratory, Upton, New York 11973
65. Peter Purdue, Probability and Statistics, National Science Foundation, 1800 G. Street, N.W., Washington, DC 20550
66. B. S. Rajput, Mathematics Department, Ayres Hall, University of Tennessee, Knoxville, Tennessee 37996
67. Gipsie B. Ranney, Statistics Department, Stokely Management Center, University of Tennessee, Knoxville, Tennessee 37996
68. J. N. K. Rao, Mathematics and Statistics Department, Carleton University, Ottawa, Ontario, Canada K1S 5B6
69. Douglas S. Robson, Biometrics Unit, Cornell University, Ithaca, New York 14850
70. Richard M. Royall, Biostatistics Department, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205
71. Donald B. Rubin, Statistics Department, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138
72. Jagdish S. Rustagi, Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210

73. A. J. Scott, Department of Mathematics and Statistics, University of Auckland, Auckland, New Zealand
74. L. R. Shenton, Office of Computing and Information Service, Boyd Graduate Studies Building, University of Georgia, Athens, Georgia 30602
75. Milton Sobel, Department of Mathematics, University of California at Santa Barbara, Santa Barbara, California 93017
76. Daniel L. Solomon, Department of Statistics, North Carolina State University, P.O. Box 5457, Raleigh, North Carolina 27650
77. Louis Sucheston, Mathematics Department, Ohio State University, Columbus, Ohio 43210
78. David L. Sylwester, Statistics Department, Stokely Management Center, University of Tennessee, Knoxville, Tennessee 37996
79. Lawrence A. Thibodeau, Mathematics and Statistics Analysis, Martin Marietta, Mathtech Division, 1401 Wilson Blvd., Suite 930, Arlington, Virginia 22209
80. How J. Tsao, Eastman Kodak Company, MSD - Kodak Park, Bldg. 56, Rochester, New York 14650
81. Carl Wagner, Mathematics Department, Ayres Hall, University of Tennessee, Knoxville, Tennessee 37996
82. Ray A. Waller, S-1, Statistics, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, New Mexico 87545
83. William E. Winkler, Statistical Standards, Energy Information Administration, Department of Energy, 1000 Independence Avenue, Washington, DC 22003
84. Kirk M. Wolter, Statistical Research Division, Bureau of the Census, Washington, DC 20233
85. Chien - Fu Jeff Wu, Statistics Department, University of Wisconsin, 1210 W. Dayton Street, Madison, Wisconsin 53706
86. Office of Assistant Manager for Energy Research and Development, Department of Energy, Oak Ridge Operations Office, Oak Ridge, Tennessee 37830

87-113. Technical Information Center.