The variance of the variance of samples from a finite population

Eungchun Cho, Kentucky State University*
Moon Jung Cho, Bureau of Labor Statistics
John Eltinge, Bureau of Labor Statistics

Key Words: Sample Variance; Randomization Variance; Polykays; Moments of Finite Population

Abstract

A direct derivation of the randomization variance of the sample variance $\widehat{V}(\overline{x})$ and related formulae are presented. Examples of the special cases of uniformly distributed population are given.

1 Introduction

Introductory courses in the randomization approach to survey inference generally begin with a relatively parsimonious development based on without-replacement selection of simple random samples. For an arbitrary finite population one establishes the unbiasedness of the sample mean $\overline{x}$ for the corresponding finite population mean; evaluates the randomization variance of the sample mean; and develops as unbiased estimator, $\widehat{V}(\overline{x})$ for this randomization variance. One subsequently uses similar developments for related randomized designs, e.g., stratified random sampling and some forms of cluster sampling. In applications of this material to practical problems, it is often important to evaluate $V(\widehat{V}(\overline{x}))$, the variance of the variance estimator. For example, some cluster sample designs may be considered problematic if the resulting $\widehat{V}(\overline{x})$ is unstable, i.e., has an unreasonably large variance.

This note presents a relatively simple, direct derivation of the randomization variance of $\widehat{V}(\overline{x})$ and related quantities. This derivation is pedagogically appealing because it builds directly on the standard whole sample approach used in introductory texts like Cochran [1], and does not require students to work with the more elaborate "polykay" approach used by Tucky [5], and Wishart [8],

*eccho@gwmail.kysu.edu

2 Functions on Simple Random Samples

Consider a finite population of N numbers $A = [a_1, a_2, \ldots, a_N]$. Let $L_{n,A}$ be the list of all possible samples of n elements selected without replacement from A.

$$L_{n,A} = [S_1, S_2, \ldots, S_\alpha], \quad \alpha = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{1}$$

One selects a *without-replacement simple random sample of size n from A* by selecting one element from $L_{n,A}$ in such a way that each sample S_j has probability $1/\alpha$ of being selected. Consider a function f on $L_{n,A}$, that is, f assigns a each sample $S \in L_{n,A}$ a value $f(S)$. Two prominent examples of f are the sample mean and the sample variance:

$$\overline{a}(S) = \frac{1}{n} \sum_{a_i \in S} a_i \tag{2}$$

$$v(S) = \frac{1}{n-1} \sum_{a_i \in S} \{a_i - \overline{a}(S)\}^2 \tag{3}$$

Evaluation of the randomization properties of $f(S)$ for $S \in L_{n,A}$ is conceptually straightforward. For example, $E\{f(S)\}$, the expectated value of $f(S)$, is obtained by computing its arithmetic average taken over the $\binom{N}{n}$ equally likely samples in $L_{n,A}$:

$$E\{f(S)\} = \frac{1}{\binom{N}{n}} \sum_{S \in L_{n,A}} f(S) \tag{4}$$

$V\{f(S)\}$, the variance of $f(S)$, is defined to be the expectation of the squared deviations $[f(S) - E\{f(S)\}]^2$,

$$V\{f(S)\} = \frac{1}{\binom{N}{n}} \sum_{S \in L_{n,A}} [f(S) - E\{f(S)\}]^2 \tag{5}$$

3 The Variance of the Sample Variance

Routine arguments (e.g., Cochran [1, Theorems 2.1, 2.2 and 2.4]) show

$$E\{\overline{a}(S)\} = \overline{A} \tag{6}$$

$$E\{v(S)\} = V(A) \tag{7}$$

$$V\{\overline{a}(S)\} = \frac{1 - \frac{n}{N}}{n} V(A) \tag{8}$$

where $\overline{a}(S)$ is the mean of the sample S, $v(S)$ is the variance of the sample S, $\overline{A} = \sum_{i=1}^{N} a_i/N$, the mean of A, and

$$V(A) = \frac{1}{N-1} \sum_{i=1}^{N} \left(a_i - \overline{A}\right)^2 \tag{9}$$

is the full finite-population analogue of the sample variance $v(S)$. The principal task in this paper is to obtain a relatively simple expression (formula) for the variance of $v(S)$,

$$V\{v(S)\} = \frac{1}{\binom{N}{n}} \sum_{S\in L_{n,A}} \{v(S) - V(A)\}^2 \tag{10}$$

in terms of a_i's in the underlying population. The formula will be useful for estimating the variance of the variance when the straight forward computation by the definition is practically impossible due the combinatorial explosion.

4 The Main Result

The following theorem gives a formula for the variance of the sample variance under simple random sampling without replacement.

Theorem. *Let $A = [a_1, a_2, \ldots, a_N]$ be a list of N numbers, ($N \geq 4$). Let $L_{n,A}$ be the list of all possible samples of n numbers selected without replacement from A, ($2 \leq n \leq N-1$).*

$$L_{n,A} = [S_1, S_2, \ldots, S_\alpha]$$

where $\alpha = \binom{N}{n} = N!/n!(N-n)!$, $S_i \subset A$, and $|S_i| = n$ Let V_n be the list of the sample variance $v(S_i)$ for each $S_i \in L_{n,A}$,

$$V_n = [v(S_1),\, v(S_2),\, \ldots,\, v(S_\alpha)]$$

Then $E\left([v(S) - E\{v(S)\}]^2\right)$, the variance of the variances of all the samples of A of size n, is given by

$$\begin{aligned} V\{v(S)\} \quad &= \quad C_1 \sum_{i=1}^{N} a_i^4 + C_2 \sum_{i\neq j} a_i^3 a_j + C_3 \sum_{i<j} a_i^2 a_j^2 + \\ &+ \quad C_4 \sum_{\substack{i\neq j, i\neq k \\ j<k}} a_i^2 a_j a_k + C_5 \sum_{i<j<k<l} a_i a_j a_k a_l \end{aligned}$$

where

$$\begin{aligned} C_1 &= \frac{N-n}{N^2 n} \\ C_2 &= -4\left\{\frac{N-n}{N^2\,(N-1)\,n}\right\} \\ C_3 &= 2\left\{\frac{N(N-1)((n-1)^2+2) - n(n-1)((N-1)^2+2)}{N^2\,(N-1)^2\,n\,(n-1)}\right\} \\ C_4 &= 4\left\{\frac{N(N-1)(n-2)(n-3) - n(n-1)(N-2)(N-3)}{N^2\,(N-1)^2\,(N-2)\,n\,(n-1)}\right\} \\ C_5 &= 24\left\{\frac{N(N-1)(n-2)(n-3) - n(n-1)(N-2)(N-3)}{N^2\,(N-1)\,(N-2)\,(N-3)^2\,n\,(n-1)}\right\} \end{aligned}$$

Sketch of Proof. The proof involves determining the coefficients of all the fourth degree terms $a_i{}^4$, $a_i{}^2 a_j{}^2$, $a_i{}^2 a_j a_k$, $a_i{}^3 a_j$, and $a_i a_j a_k a_l$ that appears in the summation. The summations in the formula are such that all like terms are combined thus appear only once. For example, $a_i a_j a_k a_l$ appears only for the indices arranged in increasing order $i < j < k < l$. Recall that α is the number of *without-replacement* samples S of A of size n and the variance of the variances of the samples S of A of size n is

$$\begin{aligned} & \frac{\sum\{v(S) - V(A)\}^2}{\alpha} \\ = \; & \frac{\sum_S v(S)^2}{\alpha} - V(A)^2 \end{aligned}$$

where the sum is taken over all the samples S in $L_{n,A}$.

For the second term, $\sum_{S \in L_{n,A}} V(A)^2$. it follows that

$$\begin{aligned} V(A)^2 \;\; = \;\; & \frac{\sum\limits_i a_i^4 + 2 \sum\limits_{i<j} a_i^2 a_j^2}{N^2} - 4 \left\{ \frac{\sum\limits_{i \neq j} a_i^3 a_j + \sum\limits_{\substack{i \neq j, i \neq k \\ j<k}} a_i^2 a_j a_k}{N^2 (N-1)} \right\} + \\ + \;\; & 4 \left\{ \frac{\sum\limits_{i<j} a_i^2 a_j^2 + 2 \sum\limits_{\substack{i \neq j, i \neq k \\ j<k}} a_i^2 a_j a_k + 6 \sum\limits_{i<j<k<l} a_i a_j a_k a_l}{N^2 (N-1)^2} \right\} \end{aligned}$$

Similarly, for the term $\sum_S v(S)^2$ it follows

$$\begin{aligned} v(S)^2 \;\; = \;\; & \frac{\sum\limits_{a_i \in S} a_i^4 + 2 \sum\limits_{\substack{i<j \\ a_i, a_j \in S}} a_i^2 a_j^2}{n^2} - 4 \left\{ \frac{\sum\limits_{\substack{i \neq j \\ a_i, a_j \in S}} a_i^3 a_j + \sum\limits_{\substack{i \neq j, i \neq k, j<k \\ a_i, a_j, a_k \in S}} a_i^2 a_j a_k}{n^2 (n-1)} \right\} + \\ + \;\; & 4 \left\{ \frac{\sum\limits_{\substack{i<j \\ a_i, a_j \in S}} a_i^2 a_j^2 + 2 \sum\limits_{\substack{i \neq j, i \neq k, j<k \\ a_i, a_j, a_k \in S}} a_i^2 a_j a_k + 6 \sum\limits_{\substack{i<j<k<l \\ a_i, a_j, a_k, a_l \in S}} a_i a_j a_k a_l}{n^2 (n-1)^2} \right\} \end{aligned}$$

Each term in $\sum_S v(S)^2$ is transformed to give

$$\sum_{S \in L_{n,A}} v(S)^2 \;\; = \;\; \frac{\binom{N-1}{n-1} \sum\limits_i a_i^4 + 2 \binom{N-2}{n-2} \sum\limits_{i<j} a_i^2 a_j^2}{n^2} +$$

$$
\begin{aligned}
- & \left\{ \frac{4\binom{N-2}{n-2}\sum\limits_{i\neq j} a_i^3 a_j + 4\binom{N-3}{n-3}\sum\limits_{\substack{i\neq j, i\neq k\\ j<k}} a_i^2 a_j a_k}{n^2\,(n-1)} \right\} + \\
+ & \left\{ \frac{4\binom{N-2}{n-2}\sum\limits_{i<j} a_i^2 a_j^2 + 8\binom{N-3}{n-3}\sum\limits_{\substack{i\neq j, i\neq k\\ j<k}} a_i^2 a_j a_k + 24\binom{N-4}{n-4}\sum\limits_{i<j<k<l} a_i a_j a_k a_l}{n^2\,(n-1)^2} \right\}
\end{aligned}
$$

Simplification of the binomial coefficients leads to,

$$
\begin{aligned}
\frac{1}{\alpha}\sum_{S\in L_{n,A}} v(S)^2 \;=\; & \frac{\sum\limits_i a_i^4}{nN} + 2\,(n-1)\,\frac{\sum\limits_{i<j} a_i^2 a_j^2}{n\,(N-1)\,N} - 4\,\frac{\sum\limits_{i\neq j} a_i^3 a_j}{n\,(N-1)\,N} \\
- \; & 4\,(n-2)\,\frac{\sum\limits_{\substack{i\neq j, i\neq k\\ j<k}} a_i^2 a_j a_k}{n\,(N-2)\,(N-1)\,N} + \\
+ \; & 4\,\frac{\sum\limits_{i<j} a_i^2 a_j^2}{n\,(n-1)\,(N-1)\,N} + 8\,(n-2)\,\frac{\sum\limits_{\substack{i\neq j, i\neq k\\ j<k}} a_i^2 a_j a_k}{n\,(n-1)\,(N-2)\,(N-1)\,N} + \\
+ \; & 24\,(n-2)\,(n-3)\,\frac{\sum\limits_{i<j<k<l} a_i a_j a_k a_l}{n\,(n-1)\,(N-3)\,(N-2)\,(N-1)\,N}
\end{aligned}
$$

Substitution of the expressions of $V(A)^2$ and of $v(S)^2$ into the expression of the variance of the samples S of A leads to the result of the theorem. $\|$

The formula for $V\{v(S)\}$ becomes considerably simpler for the population of a discrete uniform distribution on a finite interval. In this case, A is a finite arithmetic sequence. This occurs, for example, in the important special cases of equal-probability systematic sampling (Cochran [1], Chapter 8).

Corollary 1. *Let $A = [1, 2, \ldots, N]$, $N \geq 3$. Let S, $L_{n,A}$ and $v(S)$ be as in the theorem above. The variance of the sample variances*

$$
V\{v(S)\} \;=\; \frac{N(N+1)(N-n)\,(2\,nN + 3n + 3N + 3)}{360\,n(n-1)} \qquad (11)
$$

For a more general arithmetic sequence,

Corollary 2. *Let $A = [a_0, a_0 + d, \ldots, a_0 + (N-1)\,d]$, $N \geq 3$. Then the variance of $v(S)$,*

$$
V\{v(S)\} \;=\; \frac{N(N+1)(N-n)\,(2\,nN + 3n + 3N + 3)}{360\,n(n-1)}\,d^4
$$

Corollary 3. *Let A be a list of numbers uniformly distributed on the interval $[1/N, 1]$, $A = \left[\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, \frac{N}{N}\right]$, $N \geq 3$. The variance of $v(S)$,*

$$
V\{v(S)\} \;=\; \frac{(1 + \frac{1}{N})(1 - \frac{n}{N})\left(2\,n + 3 + \frac{3n}{N} + \frac{3}{N}\right)}{360\,n(n-1)} \qquad (12)
$$

$V\{v(S)\}$ approaches $2\,n+3/360\,n\,(n-1)$ as N approaches ∞. For simplest two special cases when $n=2$ and $n=N-1$, we have

$$V\{v(S)\} = \frac{(1+\frac{1}{N})(1-\frac{2}{N})(7+\frac{9}{N})}{720}, \quad \text{if } n=2 \tag{13}$$

and

$$V\{v(S)\} = \frac{(1+\frac{1}{N})(1+\frac{2}{N})}{180}, \quad \text{if } n=N-1 \tag{14}$$

References

1. W. G. Cochran, *Sampling Techniques (3rd ed.)*, John Wiley, 1977.

2. R. L. Graham, D. E. Knuth and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.

3. J. W. Tukey, Some sampling simplified, *Journal of the American Statistical Association*, 45 (1950), 501-519.

4. J. W. Tukey, Keeping moment-like sampling computation simple, The Annals of Mathematical Statistics, 27 (1956), 37-54.

5. J. W. Tukey, Variances of variance components: I. Balanced designs. *The Annals of Mathematical Statistics*, 27 (1956), 722-736.

6. J. W. Tukey, Variances of variance components: II. Unbalanced single classifications. *The Annals of Mathematical Statistics*, 28 (1957), 43-56.

7. J. W. Tukey, Variance components: III. The third moment in a balanced single classification. *The Annals of Mathematical Statistics*, 28 (1957), 378-384.

8. J. Wishart, Moment Coefficients of the k-Statistics in Samples from a Finite Population. *Biometrika*, 39 (1952), 1-13.