The Variance of Sample Variance for a Finite Population

Eungchun Cho *

November 11, 2004

Key Words: variance of variance, variance estimator, sampling variance, randomization variance, moments

Abstract

The variance of variance of sample from a finite population is given in terms of the second and the fourth moments of the population. Examples of the special case when the population is of uniform distribution is given.

1 Introduction

If the population variance is estimated by a sample variance, it is important to know a relationship between the variance of the population and the variance of the variance of all the samples of a certain fixed size. Though the sample variance is an unbiased estimator of the population variance, it is problematic if the variance of the variance estimator is too large.

For an arbitrary finite population one establishes the unbiasedness of the sample mean $\overline{x}$ for the population mean; evaluates the randomization variance of the sample mean. 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 the variance of the variance estimator. For example, some cluster sample designs may be considered problematic if the resulting variance estimator is unstable, i.e., has an unreasonably large variance.

Historically, introductory sampling textbooks have addressed this issue in a relatively limited form, either through a brief reference to an elaborate algebra of "polykays" developed by Tukey [4, pages 37-54] and Wishart [8, pages 1-13] or through a direct appeal to large-sample approximations based on the normal and chi-square distributions. See, e.g., Cochran [1, pages. 29, 96] for examples of these two approaches. In addition, the statistical literature has developed computational tools to implement the above mentioned "polykay" results.

*Kentucky State University, 400 E. Main Street, Frankfort, KY 40601, USA, ec-cho@gwmail.kysu.edu

We present a relatively simple, direct derivation of the variance of variance, that is, the variance of the variance of samples from a finite population. We provide results for the important special case when the population is of uniformly distribution.

2 The Variance of the Sample Variance

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

$$E\{\overline{a}(S)\} = \overline{A} \quad (1)$$

$$E\{v(S)\} = V(A) \quad (2)$$

$$V\{\overline{a}(S)\} = \frac{1-\frac{n}{N}}{n} V(A) \quad (3)$$

where $\overline{a}(S)$ is the mean of the sample S, $v(S)$ the variance of the sample S, $\overline{A}$ the mean of A, and $V(A)$ the full finite-population analogue of the sample variance $v(S)$. That is,

$$\overline{a}(S) = \frac{1}{n} \sum_{a_i \in S} a_i \quad (4)$$

$$v(S) = \frac{1}{n-1} \sum_{a_i \in S} (a_i - \overline{a}(S))^2 \quad (5)$$

$$\overline{A} = \frac{1}{N} \sum_{i=1}^{N} a_i \quad (6)$$

$$V(A) = \frac{1}{N-1} \sum_{i=1}^{N} \left(a_i - \overline{A}\right)^2 \quad (7)$$

The equation (2) shows that $v(S)$ is an unbiased estimator of $V(A)$. The principal task in this paper is to obtain a relatively simple expression for the variance of the variance estimator $v(S)$,

$$V\{v(S)\} = \frac{1}{\binom{N}{n}} \sum_{S \in L_{n,A}} \{v(S) - V(A)\}^2 \quad (8)$$

in terms of a_i's in the underlying population. The formula will be useful for estimating the variance of the variance estimator $v(S)$ when the straight forward computation by the definition is impractical due the combinatorial explosion when N is not very small.

3 The Main Thheorem and Formua

A relationship between the variance of a given population A of size N and the variance of the variance of the samples (of size n) of A is given. The formula

for the variance of the variance estimator is given in terms of the fourth and the second moments of the population, the size of the population, and the sample size. Recall μ_4 and μ_2, the fourth and the second moments of the population A.

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^4 \tag{9}$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 \tag{10}$$

Since the variance is invariant under a shifting by a constant, we will assume the mean of the population is zeo, which simplifies our formula a lot.

Theorem 1 *Let A be a population of size $N \geq 4$, which is represented as a list of N numbers, $A = [a_1, a_2, \ldots, a_N]$. Assume the mean of A be 0. Consider the list of all possible samples of n ($2 \leq n \leq N-1$) numbers selected without replacement from A, $L_{n,A} = [S_1, S_2, \ldots, S_\alpha]$ where $S_i \subset A$, $|S_i| = n$ for all i, and $\alpha = N!/n!(N-n)!$ (the number of sublists of size n). Let $V_n = [v(S_1), v(S_2), \ldots, v(S_\alpha)]$ be the list of the sample variance $v(S_i)$ for each $S_i \in L_{n,A}$. Then $V(v(S)) = E\left([v(S) - E\{v(S)\}]^2\right)$, the variance of the variance of estimator $v(S)$ taken over all S in V_n, is a linear combination*

$$V(v(S)) = a_1\,\mu_4 + a_2\,{\mu_2}^2 \tag{11}$$

where

$$a_1 = \frac{N(N-n)(Nn-N-n-1)}{(N-3)(N-2)(N-1)(n-1)n} \tag{12}$$

$$a_2 = \frac{N(N-n)\left(N^2n-3n-3N^2+6N-3\right)}{(N-3)(N-2)(N-1)^2(n-1)n} \tag{13}$$

Remark 1 *Following are alternate forms of the formual*

$$V(v(S)) = \frac{N(N-n)}{(N-3)(N-2)(N-1)^2(n-1)n}\left(b_1\,\mu_4 + b_2\,{\mu_2}^2\right) \tag{14}$$

where

$$b_1 = (N-1)(Nn-N-n-1) \tag{15}$$

$$b_2 = N\left(N^2n-3n-3N^2+6N-3\right) \tag{16}$$

Note b_1 and b_2 can be written as

$$b_1 = (N-1)((N-1)(n-1)-2) \tag{17}$$

$$\begin{aligned} b_2 &= N\left(n\left(N^2-3\right)-3(N-1)^2\right) \\ &= N\left(N^2(n-3)+3(N-n)+3(N-1)\right) \end{aligned} \tag{18}$$

$$V\left(v(S)\right) = \frac{N\left(N-n\right)}{\left(N-3\right)\left(N-2\right)\left(N-1\right)^{2}\left(n-1\right)n}\left(\left(N-1\right)\left(Nn-N-n-1\right)\mu_4 - \left(N\left(N^2n-3n-3N^2+6N-3\right)\right)\mu_2{}^2\right) \quad (19)$$

A preliminary formula in terms of all the possible fourth order products of the elements in the population will be derived in the lemma, which will be simplified into a form in terms of the second and the fourth moments of the population.

Lemma 1 *The variance $V\left(v(S)\right)$ of the variance of the samples can be represented as a linear combination of the sums of fourth degree terms;*

$$V\left(v(S)\right) = C_1\sum_i a_i^4 + C_2\sum_{i\neq j} a_i^3a_j + C_3\sum_{i<j} a_i^2a_j^2 + C_4\sum_{\substack{i\neq j, i\neq k\\ j<k}} a_i^2a_ja_k + C_5\sum_{i<j<k<l} a_ia_ja_ka_l \quad (20)$$

where i,j,k,l runs from 1 to N and

$$C_1 = \frac{N-n}{N^2n} \quad (21)$$

$$C_2 = \frac{-4\left(N-n\right)}{N^2\left(N-1\right)n} \quad (22)$$

$$C_3 = \frac{-2\left(N-n\right)\left(Nn-3\,N-3\,n+3\right)}{N^2\,\left(N-1\right)^2n\,\left(n-1\right)} \quad (23)$$

$$C_4 = \frac{8\left(N-n\right)\left(2\,Nn-3\,N-3\,n+3\right)}{N^2\,\left(N-1\right)^2\left(N-2\right)n\,\left(n-1\right)} \quad (24)$$

$$C_5 = \frac{-48\left(N-n\right)\left(2\,Nn-3\,N-3\,n+3\right)}{N^2\,\left(N-1\right)^2\left(N-2\right)\left(N-3\right)n\,\left(n-1\right)} \quad (25)$$

Remark 2 *Note the coefficient c_i's have the common factor $c_1 = \left(N-n\right)/N^2\,n$. This makes it obvious, as expected, that $V\left(v\left(S\right)\right) = 0$ when $N = n$.*

Sketch of the Proof of the Theorem Let s_d represent the sum $\sum_i^N {a_i}^d$ and $s_{31} = \sum_{i\neq j}^N a_i^3a_j$, $s_{22} = \sum_{i<j}^N a_i^2a_j^2$, $s_{211} = \sum_{\substack{i\neq j, i\neq k\\ j<k}}^N a_i^2a_ja_k$, and $s_{1111} = \sum_{i<j<k<l}^N a_ia_ja_ka_l$. Then

$$s_{31} = N^2s_3s_1 - Ns_4 \quad (26)$$

$$s_{22} = \frac{1}{2}N^2{s_2}^2 - \frac{1}{2}Ns_4 \quad (27)$$

$$s_{211} = \frac{1}{2}N^3s_2{s_1}^2 - N^2s_3s_1 + Ns_4 - \frac{1}{2}N^2{s_2}^2 \quad (28)$$

$$s_{1111} = \frac{1}{24}N^4{s_1}^4 - \frac{1}{4}Ns_4 + \frac{1}{3}N^2s_3s_1 + \frac{1}{8}N^2{s_2}^2 - \frac{1}{4}N^3s_2{s_1}^2 \quad (29)$$

Now the double, triple and quadruple summations in the formula given by fthe lemma are replaced by simpler summations using these relations and our simplifying assumption, $\mu = 0$, hence $s_1 = 0$. Collecting the similar terms after the cancellation of various intermediate terms, we get the formula in terms of μ_4 and ${\mu_2}^2$ only.

The formula in the lemma follows from the definition of the variance of the sample variances. The proof of the formula 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 $i < j < k < l$.

Example. Consider the first, second and fourth moments of the population A of size N that is uniformly distributed on [-1/2,1/2]

$$A = [-\frac{1}{2}, -\frac{1}{2} + h, -\frac{1}{2} + 2h, \ldots, \frac{1}{2}]$$

where $h = 1/(N-1)$. Obviously, the first moment $\mu = 0$. The sencond and fourth moments are also easily computed from the mement generating function.

$$\begin{aligned} \mu_2 &= \frac{1}{12}\,\frac{N^2-1}{N^2} \\ \mu_4 &= \frac{1}{240}\,\frac{3\,N^4+7-10\,N^2}{N^4} \end{aligned}$$

Substituting μ_2 and μ_4 into the equation 11, we get the variance of the variance of the sample of size n

$$V\left(v(S)\right) = \frac{1}{360}\,\frac{\left(-12\,N^4+7\,nN^4+5\,N^3n-9\,N^3-20\,N^2n+21\,N^2-15\,Nn+21\,N+3\,n+3\right)(N-n)}{(n-1)\,n\,(N-3)\,(N-2)\,N^3}$$

When N is large,

$$\begin{aligned} \mu_2 &\approx \frac{1}{12} \\ \mu_4 &\approx \frac{1}{80} \end{aligned}$$

and

$$V\left(v(S)\right) \approx \frac{1}{360}\,\frac{-12+7\,n}{(n-1)\,n}$$

References

[1] W. G. Cochran, Sampling Techniques (3rd ed.), John Wiley, 1977,pages 29,96.

[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.