International Journal of Pure and Applied Mathematics

Volume 21 No. 3 2005, 387-394

THE VARIANCE OF SAMPLE VARIANCE FROM A FINITE POPULATION

Eungchun Cho[1] §, Moon Jung Cho[2], John Eltinge[3]

[1]Department of Mathematics and Sciences
Kentucky State University
400, E. Main Street, Frankfort, KY 40601, USA
e-mail: eccho@gwmail.kysu.edu

[2,3]Office of Survey Methods Research
Bureau of Labor Statistics
Washington, DC 20212, USA
[2]e-mail: Cho.Moon@bls.gov
[3]e-mail: Eltinge.John@bls.gov

Abstract: A formula for the variance of variance of all the samples taken from a finite population is given in terms of the second and the fourth moments of the population. Formulae for the variance of variance of the samples from a finite population of the discrete uniform and of the binomial distribution, and numerical values of those for the populations of uniform, binomial, normal, Poisson and hypergeometric distributions are given.

AMS Subject Classification: 62D05, 62J10
Key Words: variance of sample variance, variance estimator, sampling variance, randomization variance, moments

1. Introduction

For an arbitrary finite population, one can establish the sample mean as an unbiased estimator for the population mean and evaluate the randomization variance of the sample mean. One can subsequently use similar developments for related randomized designs, for example, stratified random sampling and

Received: May 10, 2005

§Correspondence author

some forms of cluster sampling. In applications of this approach to practical problems, it is often important to evaluate the variance of the variance estimator. In the situation when the variance of a finite population is estimated from the variance of a sample it is important to know the relationship between the variances of the population and of the sample, the sizes of the population and of the sample. Though the sample variance is an unbiased estimator of the population variance, it is problematic if the variance of the sample variance is too large. Some cluster sample designs, for example, may be considered problematic if the resulting variance estimator is unstable, that is, the variance estimator 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, pp. 37-54] and Wishart [8, pp. 1-13] or through a direct appeal to large-sample approximations based on the normal and chi-square distributions. See Cochran [1, pp. 29, 96] for examples of these two approaches. In addition, the statistical literature has developed computational tools to implement the above mentioned "polykay" results. However, a simple formula that can be directly coded in programming languages or in computer algebra systems (such as *Maple* or *Mathematica*) does not exist in the literature. We present a relatively simple formula of the variance of sample variance. The formula is derived directly by considering all possible samples (without replacement) from a finite population. We show the application of the formula for the special cases when the population is of uniform, binomial, normal, Poisson, hypergeometric distribution.

2. Functions of Random Samples

Consider a finite population A of N numbers and the list $L_{n,A}$ of all possible $\binom{N}{n}$ samples of size n selected without replacement from A. One can select 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 the same probability $1/\alpha$ of being selected.

Two prominent examples of functions f defined on $L_{n,A}$ are the sample mean and the sample variance. Evaluating the randomization properties of $f(S)$ for $S \in L_{n,A}$ is conceptually straightforward. For example, $E\left(f(S)\right)$, the expected value of $f(S)$, is obtained by computing its arithmetic average taken over the α equally likely samples in $L_{n,A}$, and $V\left(f(S)\right)$, the variance of $f(S)$, is

defined to be the expectation of the squared deviations $E\left(\left(f(S)-E\left(f(S)\right)\right)^2\right)$:

$$E\left(f(S)\right) = \frac{1}{\alpha}\sum_{S\in L_{n,A}} f(S), \quad V\left(f(S)\right) = \frac{1}{\alpha}\sum_{S\in L_{n,A}} \left(f(S)-E\left(f(S)\right)\right)^2 .$$

3. The Variance of Sample Variance

Let S be a sample S of size n from A, $m(S)$ and $v(S)$ represent the mean and the variance of S. Let μ and $v(A)$ represent the mean and the full finite-population analogue of the sample variance of A

$$m(S) = \frac{1}{n}\sum_{a_i\in S} a_i, \quad v(S) = \frac{1}{n-1}\sum_{a_i\in S}\left(a_i - m(S)\right)^2 , \tag{1}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} a_i, \quad v(A) = \frac{1}{N-1}\sum_{i=1}^{N}\left(a_i-\mu\right)^2 . \tag{2}$$

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

$$E\{m(S)\} = \mu, \quad E\{v(S)\} = v(A), \quad v\{m(S)\} = \frac{1-\frac{n}{N}}{n}\, v(A)\,. \tag{3}$$

Obviously, $v(S)$ is an unbiased estimator of $v(A)$. We present a simple expression for the variance of $v(S)$,

$$v\left(v(S)\right) = \frac{1}{\alpha}\sum_{S\in L_{n,A}} \left\{v(S)-v(A)\right\}^2 \tag{4}$$

in terms of n, N, and the moments of A. We note any direct computation of the variance of $v(S)$ using Definition 4 is impractical due to the large value of α unless the size N of the population is small.

4. The Main Formula

Notation.

$$A = [a_1, a_2, \ldots, a_N], \quad L_{n,A} = [S_1, S_2, \ldots, S_\alpha] ,$$
$$V_n = [v(S_1), v(S_2), \ldots, v(S_\alpha)], \quad v(S) = \frac{1}{n-1}\sum_{a_i\in S}(a - \frac{1}{n}\sum_{a_i\in S} a_i)^2 ,$$

$$v\left(v(S)\right) = E\left(\left[v(S) - E\{v(S)\}\right]^2\right), \quad \mu_4 = \frac{1}{N}\sum_{i=1}^{N}(a_i-\mu)^4,$$

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{N}(a_i-\mu)^2, \quad \alpha = \binom{N}{n} = \frac{N!}{n!(N-n)!}.$$

The following lemma expresses the variance of sample variance $v(v(S))$ in terms of all the fourth order products of the elements in A, i.e., ${a_i}^4$, ${a_i}^2{a_j}^2$, ${a_i}^2a_ja_k$, ${a_i}^3a_j$, and $a_ia_ja_ka_l$. The terms in the formula are combined so that they appear only once. For example, $a_ia_ja_ka_l$ appears only for the indices arranged in increasing order $i<j<k<l$.

Lemma 1.

$$v\left(v(S)\right) \;=\; C\left(C_1s_4 + C_2s_{31} + C_3s_{22} + C_4s_{211} + C_5s_{1111}\right),$$

where

$$\begin{aligned}
C &= \frac{N-n}{(N-1)^2N^2n}, \quad C_1 = (N-1)^2, \\
C_2 &= -4(N-1), \quad C_3 = \frac{-2(Nn-3N-3n+3)}{(n-1)}, \\
C_4 &= \frac{8(2Nn-3N-3n+3)}{(N-2)(n-1)}, \quad C_5 = \frac{-48(2Nn-3N-3n+3)}{(N-2)(N-3)(n-1)},
\end{aligned}$$

and

$$\begin{aligned}
s_d &= \sum_{i}^{N}{a_i}^d, \quad d=1,2,3,4, \\
s_{31} &= \sum_{i\neq j}^{N}a_i^3a_j, \quad s_{22} = \sum_{i<j}^{N}a_i^2a_j^2, \\
s_{211} &= \sum_{\substack{i\neq j, i\neq k\\ j<k}}^{N}a_i^2a_ja_k, \quad s_{1111} = \sum_{i<j<k<l}^{N}a_ia_ja_ka_l.
\end{aligned}$$

Proof. Expand the equation (4), then determine the coefficients C_i of the terms ${a_i}^4$, ${a_i}^2{a_j}^2$, ${a_i}^2a_ja_k$, ${a_i}^3a_j$, and $a_ia_ja_ka_l$ that appears in the summation.

We assume, to avoid trivial cases, $N \geq 4$ and $3 \leq n \leq N-1$. Since the variance is invariant under the shifting by a constant, we will assume μ is zero by shifting every element of A by μ. That $\mu = 0$ simplifies the formula substantially. □

Theorem 1. *The variance of sample variance $v\,(v(S))$ for $S \in L_{n,A}$ is a linear combination of μ_4 and ${\mu_2}^2$ with coefficients a_1 and a_2, which are rational expressions of N and n,*

$$v\,(v(S)) \;=\; a_1\,\mu_4 - a_2\,{\mu_2}^2\,, \tag{5}$$

where

$$\begin{aligned} a_1 &= \frac{N\,(N-n)\,(Nn-N-n-1)}{(N-3)\,(N-2)\,(N-1)\,(n-1)\,n}\,, \\ a_2 &= \frac{N\,(N-n)\left(N^2n-3n-3N^2+6N-3\right)}{(N-3)\,(N-2)\,(N-1)^2\,(n-1)\,n}\,. \end{aligned}$$

Factoring a common factor from a_1 and a_2,

$$v\,(v(S)) = c\left(b_1\,\mu_4 - b_2\,{\mu_2}^2\right)\,, \tag{6}$$

where

$$\begin{aligned} c &= \frac{N\,(N-n)}{(N-3)\;\;(N-2)\,(N-1)\,(n-1)\,n}\,, \\ b_1 &= Nn-N-n-1, \qquad b_2 = \frac{N^2n-3n-3N^2+6N-3}{N-1}\,. \end{aligned}$$

Corollary 1. *If N is very large compared to n, then*

$$v(v(S)) \approx \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}{\mu_2}^2).$$

Proof. $\lim_{N\to\infty} a_1 = 1/n$ and $\lim_{N\to\infty} a_2 = (n-3)/(n-1)n$. □

Corollary 2. *If $n = N-1$, as in the jackknife set up, then*

$$v\,(v(S)) = \left(\frac{N}{(N-1)\,(N-2)}\right)^2\left(\mu_4 - {\mu_2}^2\right)\,.$$

Proof. Simple substitution of $N-1$ into n in the formula (5). □

Proof of Theorem 1. A sketch of the proof of the theorem is given. The sums s_{31}, s_{22}, s_{211} and s_{1111} can be expressed in terms of s_2 and s_4 by routine expansion and simplification together with the fact that $\sum_i^N a_i = 0$,

$$\begin{aligned} s_{31} &= -Ns_4, \quad s_{22} = \frac{N}{2}\,(N\,{s_2}^2 - s_4)\,, \\ s_{211} &= -\frac{N}{2}\,(N\,{s_2}^2 - 2\,s_4), \quad s_{1111} = \frac{N}{8}\,(N\,{s_2}^2 - 2\,s_4)\,. \end{aligned}$$

Substituting these relations to the equation of the lemma, we get the formula (5) of Theorem 1. □

Example 1. (Discrete Uniform Distribution) Let A be the population of N numbers distributed uniformly on the interval $[0,1]$.

$$A = \left[\frac{i}{N-1} : \ i = 0, \ldots, N-1 \right] .$$

It follows from the definition that

$$\begin{aligned} \mu &= \frac{1}{2}, \quad \mu_2 = \frac{N+1}{12(N-1)}, \\ \mu_4 &= \frac{50\,N^5 + 15\,N^4 + 3\,N^3 + 90\,N^2 + 67\,N + 15}{240\,N\,(N-1)^4}. \end{aligned}$$

Substitution of μ_2 and μ_4 into the formula (5) of the theorem and taking the limit as N approaches to $+\infty$ gives

$$v\,(v(S)) \approx \frac{7\,n-12}{360\ (n-1)\,n}.$$

Example 2. (Binomial Distribution) Let A be the population of $N = 2^k$ integers between 0 and k such that the integer r appears in A $\binom{k}{r}$ times,

$$A = \left[0, 1, \ldots, \overbrace{r, \ldots, r}^{\binom{k}{r}}, \ldots, k-1, k \right] .$$

It follows from the definition that

$$\mu = \frac{k}{2}, \quad \mu_2 = \frac{k}{4}, \quad \mu_4 = \frac{3}{16}\,k^2 - \frac{1}{8}\,k\,.$$

If N is large, $a_1 \approx 1/n$ and $a_2 \approx (n-3)/n(n-1)$, and

$$v(v(S)) \approx \frac{k}{8n} \left(\frac{n-2}{n-1} k - 1 \right) .$$

Example 3. (Numerical Examples) In this examples, A is a population of size 2048 generated by a random variable of the given distribution (using built-in random number generators in Matlab)and the sample size is 32.

Uniform Distribution. A is a population generated by a random variable uniformly distributed on $[0,1]$.

$$\mu = 0.5059,\ \mu_2 = 0.0843,\ \mu_4 = 0.0126,\ v(v(S)) = 0.0006.$$

Note that $\mu = 0.5$, $\mu_2 \approx 0.083415$, $\mu_4 \approx 0.012555$, and $v(v(S)) \approx .000588$, if A were an ideal population of 2024 numbers uniformly distributed on $[0, 1]$.

Binomial Distribution. A is a population generated by a random variable of the binomial distribution with $k = 11$ and $p = 0.5$.

$$\mu = 5.4985,\ \mu_2 = 2.7783,\ \mu_4 = 22.4375,\ v(v(S)) = 0.9148.$$

Poisson Distribution. A is a population generated by a random variable of Poisson distribution with $\lambda = 3$

$$\mu = 3.0176,\ \mu_2 = 3.0973,\ \mu_4 = 36.2807,\ v(v(S)) = 1.3958.$$

Hypergeometric Distribution. A is a population generated by a random variable of hypergeometric distribution with $M = 10, K = 5$ and $n = 5$.

$$\mu = 2.4829,\ \mu_2 = 0.7048,\ \mu_4 = 1.3838,\ v(v(S)) = 0.0570.$$

Normal Distribution. A is a population generated by a random variable of the standard normal distribution.

$$\mu = 0.0040,\ \mu_2 = 1.0826,\ \mu_4 = 3.6113,\ v(v(S)) = 0.1452.$$

Acknowledgements

The first author was partially supported by Kentucky State University.

References

[1] W.G. Cochran, *Sampling Techniques*, Third Edition, John Wiley (1977), 29, 96.

[2] R.L. Graham, D.E. Knuth, 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.