COMPARISON OF THE EFFICIENCY OF DRAWING SAMPLES WITH AND WITHOUT REPLACEMENT WHEN THE VARIANCE OF THE GENERAL POPULATION IS UNKNOWN

BY

I. KOŹNIEWSKA (WARSZAWA)

Suppose we are given a general population which contains N elements: $x_1, x_2, \ldots, x_N$. From this population we draw a random sample of n elements. Let μ denote a parameter of the general population while m is the corresponding parameter of the sample.

The parameter m is a discrete random variable with a finite number of saltus; hence there exists its mathematical expectation $E(m)$ and its variance $D(m)$.

The scope of this paper[1] is to compare the efficiency of different estimates of the general population variance. We say that m_1 *is a more efficient estimate of the parameter* μ *than the estimate* m_2 when $D(m_1) < D(m_2)$.

According to the sampling technique we may classify the estimates of the general population parameter into two classes: the first contains those which are parameters of random samples drawn with replacement (denoted by italic letters without asterisks) and the second class contains those which are parameters of samples drawn without replacement (denoted by italic letters with asterisks).

For instance, the arithmetic mean μ of the general population may be estimated by the sample mean $\bar{x}^*$ of elements drawn without replacement and by the sample mean $\bar{x}$ of elements drawn with replacement. It is well known that $\bar{x}^*$ is a more efficient estimate of μ than $\bar{x}$, since

$$D(\bar{x}) = \frac{\sigma^2}{n} \quad \text{and} \quad D(\bar{x}^*) = \frac{\sigma^2}{n}\,\frac{N-n}{N-1}$$

where σ^2 denotes the variance of the general population, N the number of elements in the population, n the respective number of elements in the sample.

It is easily seen that $D(\bar{x}^*) < D(\bar{x})$ when $n > 1$.

[1] Published also in Polish, [2].

One could suppose that estimates deriving from samples without replacement are always more efficient than those deriving from samples with replacement[2]. However, this assumption would be false, as can be seen from the following example, in which the variance of the general population is estimated by the variances of the samples.

Suppose then that we have a population consisting of 9 elements ($N = 9$), namely: $x_i = 1$ for $i = 1, 2, \ldots, 8$, $x_9 = -8$. The mean $\mu = 0$, the variance $\sigma^2 = 8$. Let us now draw random samples of three elements. If we are drawing elements with replacement we may obtain four different samples, 1,1,1; 1,1,−8; 1,−8,−8; −8, −8, −8 with the respective probabilities 512/729, 192/729, 24/729, 1/729. The random variable

$$m_2 = \frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x})^2$$

will have the mathematical expectation $E(m_2) = 5\frac{1}{3}$ and the variance $D(m_2) = 67\frac{5}{9}$.

If, on the contrary, we are drawing elements without replacement, we can obtain only two different samples, 1,1,1 and 1, 1, −8, with the respective probabilities 2/3 and 1/3. The mathematical expectation of the random variable m_2^* will be $E(m_2^*) = 6$ and the variance $D(m_2^*) = 72$. Obviously in this case $D(m_2^*) > D(m_2)$.

This example seems to contradict our intuition, which suggests that drawing without replacement should give us better (in a rather indeterminate sense) results than drawing with replacement.

It may be conjectured that the source of the apparent paradox lies in the bias of the sample variance, since the efficiency of the estimate is not the only quality required from the estimate. For instance, any arbitrary constant could be taken as estimate and its variance would be zero. Such an estimate, however, would be of no use, as its bias is indefinite. Tor unbiased estimates there is no paradox and theorem 1, given further on, holds.

Let N and n denote, as before, the number of elements of the general population and of elements in the sample respectively. It is evident that N and n must fulfil the conditions $2 \leqslant n \leqslant N-1$ and $N > 2$.

Further let $\bar{x}$ and $\bar{x}^*$ denote respectively the arithmetic sample means in drawing with and without replacement, m_2 and m_2^* the respective biased sample variances, M_2 and M_2^* the respective unbiased sample variances, a_2 and a_2^* the respective unbiased second moments taken about the

[2] The problem of comparing the efficiency of the variances as estimates of the population variance when the sampling is with and without replacement has been put forward by S. Szulc.

population mean μ. All these parameters are bound by the following relations:

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad M_2 = \frac{n}{n-1}m_2, \qquad a_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2,$$

(1)

$$m_2^* = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}^*)^2, \qquad M_2^* = \frac{n}{n-1}\cdot\frac{N-1}{N}m_2^*, \qquad a_2^* = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The following theorem is proved:

THEOREM 1. *For the unbiased estimates M_2, M_2^*, a_2, a_2^* of the population variance σ^2 the following relations hold:*

$$D(M_2^*) < D(M_2), \tag{2}$$

$$D(a_2^*) < D(a_2). \tag{3}$$

This theorem states that the estimates M_2^* and a_2^*, which are sample parameters obtained by drawing without replacement, are more efficient estimates of population variance σ^2 than the respective parameters M_2 and a_2, obtained by sampling with replacement.

The following lemma will be useful for the proofs of this and other theorems:

LEMMA. *Let*

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4$$

denote the fourth central moment of the general population and let A and B be arbitrary numbers fulfilling the inequalities $A < 0$ and $B > A$. Then the inequality $A\mu_4 - B\sigma^4 < 0$ holds.

Proof of the lemma. It is known that the coefficient of excess $\gamma_2 = \mu_4/\sigma^4$ is always not less than 1, $\gamma_2 \geqslant 1$. If A and B have the properties stated in the lemma, $B/A < 1$ and therefore $\gamma_2 > B/A$, which is equivalent to $A\mu_4 - B\sigma^4 < 0$.

Proof of theorem 1. First we shall prove the inequality (2). It is valid for $N = 3$ and $n = 2$, as may be easily calculated. In order to prove (2) for $N > 3$ and $2 \leqslant n \leqslant N-1$ we shall use the formulae, given by Hagstroem, defining the variances of m_2^* and m_2. They are [1]:

$$D(m_2^*) = \frac{(n-1)^2 N\big(N-(n+1)/(n-1)\big)(N-n)}{n^3(N-1)(N-2)(N-3)}\mu_4 + \\ + \frac{N(N-n)(n-1)\big(-n(N^2-3)+3(N-1)^2\big)}{n^3(N-1)^2(N-2)(N-3)}\sigma^4.$$

$$D(m_2) = \frac{(n-1)^2}{n^3}\mu_4 - \frac{(n-1)(n-3)}{n^3}\sigma^4.$$

(Without loss of generality it has been assumed in these relations that $\mu = 0$).

Taking into consideration the obvious formulae

$$D(M_2^*) = \left[\frac{n}{n-1}\cdot\frac{N-1}{N}\right]^2 D(m_2^*), \qquad D(M_2) = \left[\frac{n}{n-1}\right]^2 D(m_2),$$

we obtain $D(M_2^*) - D(M_2) = A\mu_4 - B\sigma^4$, where

$$A = \frac{-n^2(N-1)^2+n(4N^2-5N-1)-N(5N-7)}{nN^2(n-1)(N-2)(N-3)},$$

$$B = \frac{-n^2(N^2-3)+n(8N^2-15N+3)+3N(-3N+5)}{n(n-1)N(N-2)(N-3)}.$$

It may easily be proved that $A < 0$ for $N > 3$ and $B > A$ for $2 \leqslant n \leqslant N-1$, whence the assumptions of the lemma are realized. The lemma proves the theorem.

The validity of formula (3) results from the following relations, given also by Hagstroem [1]:

$$D(a_2^*) = \frac{N-n}{n(N-1)}(\mu_4-\sigma^4), \qquad D(a_2) = \frac{1}{n}(\mu_4-\sigma^4).$$

Indeed, for $n > 1$ we have $D(a_2^*) < D(a_2)$.

THEOREM 2. *a_2 is a more efficient estimate of the population variance σ^2 than M_2, i. e. $D(a_2) < D(M_2)$.*

Proof. The proof results directly from the comparison of the expressions

$$D(a_2) = \frac{1}{n}(\mu_4-\sigma^4), \qquad D(M_2) = \frac{1}{n}\left[\mu_4 - \frac{n-3}{n-1}\sigma^4\right].$$

THEOREM 3. *a_2^* is a more efficient estimate of the population variance σ^2 than M_2^*, i. e. $D(a_2^*) < D(M_2^*)$.*

Proof. If we calculate the difference between the respective variances, we obtain $D(a_2^*) - D(M_2^*) = A_1\mu_4 - B_1\sigma^4$, where

$$A_1 = \frac{N-n}{n(n-1)}\cdot\frac{N^2(4-n)+nN(3-N)+(n-N)-6N+1}{N(N-1)(N-2)(N-3)},$$

$$B_1 = \frac{N-n}{n}\cdot\frac{n(-4N^2+9N-3)+2N^3-4N^2+3N-3}{N(N-1)(N-2)(N-3)}.$$

It is easily seen that here the assumptions of the lemma are fulfilled: $A_1 < 0$ and $B_1 > A_1$. Thus the lemma proves the theorem.

Now, it remains to compare the efficiency of the estimates M_2^* and a_2. We shall prove the following

THEOREM 4. *In order that M_2^* should be a more efficient estimate of population variance σ^2 than a_2, it is necessary and sufficient that the population coefficient of excess γ should fulfil the inequality*

$$\gamma_2 \geqslant 1+\frac{2(N-n)(N-n-1)(N-2)}{n^2(N-1)^2-n(4N^2-5N-1)+N(5N-7)}. \tag{6}$$

Proof. If we compare the respective variances, we get $D(M_2^*)-D(a_2) = S\mu_4 - T\sigma^4$, where

$$S = \frac{-n^2(N-1)^2+n(4N^2-5N-1)-N(5N-7)}{nN(N-1)(N-2)(N-3)},$$

$$T = \frac{-n^2(N^2-3)+n(8N^2-15N+3)+N(-2N^2+N+3)}{n(n-1)N(N-2)(N-3)}.$$

It may be verified that $S < 0$ and $T \leqslant S$. Thus $S\mu_4 - T\sigma^4 \leqslant 0$ will hold if and only if $\gamma_2 \geqslant T/S$ and

$$\frac{T}{S} = 1+\frac{2(N-n)(N-n-1)(N-2)}{n^2(N-1)^2-n(4N^2-5N-1)+N(5N-7)},$$

and this proves the theorem.

Conclusions. The first three theorems are general and intuitive, whereas the fourth theorem possesses none of these advantages. It asserts that, provided condition (6) is fulfilled by the population coefficient of excess, the estimate M_2^* found in random sampling without replacement, the population mean μ being unknown, is more efficient than the estimate a_2, obtained in sampling with replacement and with population mean μ known. It appears that here random sampling without replacement has more influence on the efficiency of the estimate than the knowledge of the population mean.

Let us try to find cases in which condition (6) definitely holds and those for which it definitely does not hold.

1. The maximum value of the coefficient of excess γ_2 for given N is defined by the formula [3]:

$$\max\gamma_2 = \frac{N^2-3N+3}{N-1};$$

thus (6) will not hold if

$$1+\frac{2(N-n)(N-n-1)(N-2)}{n^2(N-1)-n(4N^2-5N-1)+N(5N-7)} > \frac{N^2-3N+3}{N-1};$$

this is the case when $n = 2$ with any N. This result indicates that if we draw only two elements to the sample, we shall always have $D(a_2) < D(M_2^*)$.

On the contrary, if we draw $N-1$ elements to the sample, for any N we shall always have $D(a_2) > D(M_2^*)$.

2. If the general population coefficient of skewness $\gamma_1 = \mu_3/\sigma^3$ is large, the population coefficient of excess is also large, since these parameters are bound by the relation [4]: $\gamma_2 \geqslant \gamma_1^2+1$.

In this case the inequality (6) may hold even for small n. This fact agrees with our intuition, which suggests that the knowledge of the mean of a very asymmetric population does not extend our knowledge about the population.

The example on p. 233 concerns such an asymmetric population. If we draw samples of three elements, condition (6) holds and therefore $D(M_2^*) < D(a_2)$. If, on the contrary, we draw samples of two elements, condition (6) does not hold, according to p. 232, and $D(a_2) < D(M_2^*)$.

3. In practice we often deal with distributions approximately normal, for which $\gamma_2 = 3$. In these cases condition (6) reduces to

$$\frac{(N-n)(N-n-1)(N-2)}{n^2(N-1)^2-n(4N^2-5N-1)+N(5N-7)} < 1,$$

or $n^2(N^2-3N+3)+n(-2N^2+3)-N(N^2-8N+9) > 0$.

It is worth noting that the last inequality holds for $n \geqslant 1+\sqrt{N}$. Thus, if a general population has the coefficient of excess $\gamma_2 = 3$, then for $n \geqslant 1+\sqrt{N}$ the relation $D(M_2^*) < D(a_2)$ holds. It means that M_2^* is a more efficient estimate of the population variance σ^2 than the estimate a_2.

This result indicates that for a general population with the coefficient of excess $\gamma_2 = 3$, if we sample without replacement at least $1+\sqrt{N}$ elements, the estimate

$$M_2^* = \frac{N-1}{(n-1)N}\sum_{i=1}^{n}(x_i-\bar{x}^*)^2,$$

which is based on the sample mean $\bar{x}^*$, will be a more efficient estimate of the population variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

than the estimate

$$a_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2,$$

calculated with the aid of the population mean μ from the random sample drawn with replacement. This fact should be exploited in practice.

REFERENCES

[1] K. G. Hagstroem, *Alcune formule appartenenti alla statistica rappresentativa,* Giornale dell'Istituto Italiano degli Attuari 3 (1932).

[2] I. Koźniewska, *Porównanie efektywności losowania ze zwracaniem i bez zwracania przy nieznanej wariancji populacji generalnej,* Zastosowania Matematyki 2 (1955), p. 297 - 303.

[3] H. C. Picard, *A note on the maximum value of kurtosis,* The Annals of Mathematical Statistics 22 (1951), p. 480 - 482.

[4] J. E. Wilkins Jr., *A note on skewness and kurtosis,* The Annals of Mathematical Statistics 15 (1944), p. 333 - 335.