I. KOŹNIEWSKA (Warszawa)

PORÓWNANIE EFEKTYWNOŚCI LOSOWANIA ZE ZWRACANIEM I BEZ ZWRACANIA PRZY NIEZNANEJ WARIANCJI POPULACJI GENERALNEJ

Rozpatrzmy populację generalną składającą się z N elementów $x_1, x_2, \ldots, x_N$. Z populacji tej losujemy n elementów do próbki. Oznaczmy przez μ parametr populacji generalnej, a przez m odpowiedni parametr próbki; parametr m jako zmienna losowa dyskretna o skończonej liczbie skoków ma wartość oczekiwaną $E(m)$ oraz wariancję $D(m)$.

Celem niniejszej pracy jest porównanie efektywności estymatorów wariancji populacji generalnej. Będziemy uważali, że m_1 jest efektywniejszym od m_2 estymatorem parametru μ, jeśli $D(m_1)<D(m_2)$.

Stosownie do techniki pobierania próbki podzielmy estymatory parametrów populacji generalnej na dwie klasy: na estymatory otrzymane z próbek, których elementy losowano ze zwracaniem (będziemy je oznaczali literami łacińskimi bez gwiazdki) i na estymatory otrzymane z próbek, których elementy losowano bez zwracania (będziemy je oznaczali odpowiednimi literami łacińskimi z gwiazdką).

Na przykład estymatorami średniej arytmetycznej populacji generalnej są średnie arytmetyczne próbek: $\bar{x}$ — próbki wylosowanej ze zwracaniem, $\bar{x}^*$ — próbki wylosowanej bez zwracania. O estymatorach $\bar{x}$ i $\bar{x}^*$ wiadomo, że efektywniejszy z nich jest $\bar{x}^*$, ponieważ

$$D(\bar{x}) = \frac{\sigma^2}{n},$$

zaś

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

gdzie σ^2 oznacza wariancję populacji generalnej, N — liczebność populacji generalnej, n — liczebność próbki.

Jak widać, dla $n>1$ mamy $D(\bar{x}^*)<D(\bar{x})$.

Można by przypuszczać, że estymatory otrzymane z próbek wylosowanych bez zwracania są zawsze efektywniejsze od odpowiednich

estymatorów z próbek wylosowanych ze zwracaniem. Jednakże takie przypuszczenie byłoby niesłuszne. Można to wykazać na następującym przykładzie, w którym oszacowuje się wariancję populacji generalnej za pomocą wariancji próbek[1]):

Niech populacja generalna składa się z 9 elementów ($N=9$): $x_i=1$ dla $i=1,2,\ldots,8$ i $x_9=-8$, średnia populacji $\mu=0$, wariancja $\sigma^2=8$.

Weźmy próbki składające się z 3 elementów. Jeżeli losujemy elementy ze zwracaniem, to możemy otrzymać następujące składy próbek: 1,1,1; 1,1,−8; 1,−8,−8; −8,−8,−8 z odpowiednimi prawdopodobieństwami 512/729, 192/729, 24/729, 1/729. Zmienna losowa

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

będzie miała wartość oczekiwaną $E(m_2)=5\frac{1}{3}$ oraz wariancję $D(m_2)=67\frac{5}{9}$.

Jeżeli natomiast losujemy elementy do próbki bez zwracania, to możemy otrzymać tylko następujące składy próbek: 1,1,1; 1,1,−8 z odpowiednimi prawdopodobieństwami 2/3 i 1/3; wobec tego zmienna losowa m_2^* będzie miała wartość oczekiwaną $E(m_2^*)=6$ i wariancję $D(m_2^*)=72$. Jak widać, w tym przypadku mamy $D(m_2^*)>D(m_2)$.

Przykład ten wydaje się sprzeczny z intuicją, która mówi, że losowanie bez zwracania powinno dawać wyniki lepsze (w sposób zresztą niezbyt określony) niż losowanie ze zwracaniem. Nasuwa się przypuszczenie, że źródłem tego pozornego paradoksu jest fakt, iż wariancja próbki jest obciążonym estymatorem wariancji populacji generalnej, a o wyższej jakości estymatora decyduje nie tylko efektywność. Na przykład każda dowolna stała wzięta jako estymator daje wariancję równą zeru, ale taki estymator — jako obciążony w nieokreślony sposób — jest bezużyteczny. Dla estymatorów nieobciążonych paradoksu już nie ma; mianowicie zachodzi podane dalej twierdzenie 1.

Niech N i n oznaczają odpowiednio liczebność populacji generalnej i próbki. Z natury rzeczy N i n spełniają nierówności $2\leqslant n\leqslant N-1$ oraz $N>2$.

Niech dalej $\bar{x}$ i $\bar{x}^*$ oznaczają odpowiednio średnie próbki wylosowanej ze zwracaniem i bez zwracania, m_2 i m_2^* — wariancje obciążone próbki, M_2 i M_2^* — wariancje nieobciążone próbki oraz a_2 i a_2^* — wariancje nieobciążone próbki obliczone od średniej μ populacji generalnej. Parametry powyższe są związane następującymi wzorami (zob. [1]):

---

[1]) Zagadnienie porównania efektywności estymatorów wariancji przy losowaniu ze zwracaniem i bez zwracania postawił S. Szulc.

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

(1)

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

Można wypowiedzieć następujące

TWIERDZENIE 1. *Dla estymatorów nieobciążonych* M_2, M_2^*, a_2, a_2^* *wariancji* σ^2 *populacji generalnej zachodzą następujące nierówności:*

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

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

Twierdzenie to orzeka, że estymatory M_2^* i a_2^*, będące parametrami próbek wylosowanych bez zwracania, są efektywniejszymi estymatorami wariancji σ^2 populacji generalnej niż odpowiednie estymatory M_2 i a_2 otrzymane z próbek wylosowanych bez zwracania.

Do dowodu tego twierdzenia, jak również następnych, będzie użyteczny następujący

LEMAT. *Niech*

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

oznacza czwarty moment centralny populacji generalnej i niech A *i* B *będą dowolnymi liczbami spełniającymi nierówności* $A<0$ *i* $B>A$. *Wówczas*

$$A\mu_4-B\sigma^4<0. \tag{a}$$

Dowód lematu. Wiadomo, że współczynnik ekscesu $\gamma_2=\mu_4/\sigma^4$ spełnia nierówność $\gamma_2\geqslant 1$. Jeżeli A i B spełniają warunki założenia, to $B/A<1$, wobec czego $\gamma_2>B/A$, co jest równoważne z nierównością (a).

Dowód twierdzenia 1. Udowodnimy najpierw słuszność nierówności (2). Łatwo sprawdzić, że jest ona spełniona dla $N=3$ i $n=2$. W celu udowodnienia wzoru (2) dla $N>3$ i $2\leqslant n\leqslant N-1$, skorzystamy ze wzorów na $D(m_2^*)$ i $D(m_2)$ podanych przez Hagstroema (zob. [1]):

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

$$+\frac{N(N-n)(n-1)\big(-n(N^2-3)+3(N-1)^2\big)}{n^3(N-1)^2(N-2)(N-3)}\sigma^4,$$

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

(We wzorach tych przyjęto $\mu=0$, co nie zmienia ogólności rozważań.) Z nich oraz z oczywistych zależności

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

otrzymujemy

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

gdzie

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

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

Można łatwo udowodnić, że dla $N>3$ jest $A<0$ i dla $2\leqslant n\leqslant N-1$ jest $B>A$, a więc są spełnione założenia lematu. Teza lematu dowodzi słuszności wzoru (2).

Prawdziwość wzoru (3) wynika bezpośrednio z następujących wzorów podanych również przez Hagstroema [1]:

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

Istotnie, dla $n>1$ mamy $D(a_2^*)<D(a_2)$.

TWIERDZENIE 2. *Estymator a_2 jest efektywniejszym estymatorem wariancji σ^2 populacji generalnej niż M_2, czyli*

$$D(a_2)\leqslant D(M_2). \tag{4}$$

Dowód. Słuszność tego twierdzenia wynika bezpośrednio z porównania wzorów

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

TWIERDZENIE 3. *Estymator a_2^* jest efektywniejszym estymatorem wariancji σ^2 populacji generalnej niż M_2^**, czyli

$$D(a_2^*)<D(M_2^*). \tag{5}$$

Dowód. Obliczając różnicę między odpowiednimi wariancjami otrzymujemy

$$D(a_2^*)-D(M_2^*)=A_1\mu_4-B_1\sigma^4,$$

gdzie

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

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

Łatwo sprawdzić, że są tu spełnione założenia lematu: $A_1<0$ i $B_1>A_1$. Teza lematu dowodzi twierdzenia.

Pozostaje obecnie do porównania efektywność estymatorów M_2^* i a_2. Można udowodnić następujące

TWIERDZENIE 4. *Warunkiem koniecznym i dostatecznym na to, żeby M_2^* było efektywniejszym estymatorem wariancji σ^2 populacji generalnej niż a_2, jest, by współczynnik ekscesu γ_2 populacji generalnej spełniał nierówność*

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

Dowód. Porównanie odpowiednich wariancji daje

$$D(M_2^*)-D(a_2)=C\mu_4-D\sigma^4,$$

gdzie

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

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

Łatwo zauważyć, że jest tu $C<0$ i $D\leqslant C$, więc nierówność $C\mu_4-D\sigma^4\leqslant 0$ będzie spełniona wtedy i tylko wtedy, gdy $\gamma_2\geqslant B/A$, przy czym

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

co dowodzi twierdzenia.

Wnioski. Pierwsze trzy twierdzenia są ogólne i intuicyjne, natomiast twierdzenie 4 nie ma żadnej z tych zalet. Mówi nam ono, że przy warunku (6) spełnionym przez współczynnik ekscesu γ_2 populacji generalnej, estymator M_2^*, uzyskany bez znajomości średniej μ populacji generalnej z próbki wylosowanej bez zwracania, jest efektywniejszy niż estymator a_2 uzyskany przy znajomości średniej μ z próbki wylosowanej ze zwracaniem. Okazuje się, że losowanie elementów do próbki bez zwracania ma tu większy wpływ na zwiększenie efektywności estymatora niż znajomość średniej populacji generalnej.

Postarajmy się znaleźć konkretne przypadki, w których mamy zapewnione zachodzenie lub niezachodzenie warunku (6).

1. Ponieważ maksimum współczynnika ekscesu γ_2 dla danego N jest określone przez wzór (zob. [2])

$$\max \gamma_2 = \frac{N^2 - 3N + 3}{N - 1}, \tag{7}$$

więc warunek (6) nie będzie zachodził, jeżeli

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

a więc przy $n=2$ dla żadnego N. Wynik ten oznacza, że jeżeli do próbki weźmiemy tylko 2 elementy, to będzie zawsze

$$D(a_2) < D(M_2^*). \tag{8}$$

Jeżeli natomiast do próbki wylosujemy $N-1$ elementów, to zawsze będzie spełniona nierówność przeciwna względem (8).

2. Jeżeli populacja generalna ma duży współczynnik asymetrii $\gamma_1 = \mu_3/\sigma^3$, to ma również duży współczynnik ekscesu, ponieważ parametry te są związane zależnością (zob. [3])

$$\gamma_2 \geqslant \gamma_1^2 + 1, \tag{9}$$

a wtedy nierówność (6) może być spełniona dla małych n. Wynik ten jest zgodny z intuicją, ponieważ przy dużej asymetrii populacji generalnej znajomość średniej niewiele nam mówi o populacji.

W przykładzie rozpatrywanym na stronicy 298 mamy właśnie do czynienia z populacją generalną o dużej asymetrii. Dla próbki o liczebności $n=3$ warunek (6) jest spełniony, tj. $D(M_2^*) < D(a_2)$, natomiast dla próbki o liczebności $n=2$ — zgodnie z wynikami punktu 1 — warunek (6) nie jest spełniony i $D(a_2) < D(M_2^*)$.

3. W praktyce mamy często do czynienia z rozkładami zbliżonymi do normalnego, dla którego $\gamma_2 = 3$. Wtedy warunek (6) przybiera postać

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

lub

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

ta ostatnia nierówność jest spełniona dla $n \geqslant 1 + \sqrt{N}$.

A zatem dla populacji generalnej o współczynniku ekscesu $\gamma_2=3$ dla $n\geqslant 1+\sqrt{N}$ zachodzi nierówność $D(M_2^*)<D(a_2)$, tzn. estymator M_2^* jest efektywniejszym estymatorem wariancji σ^2 niż estymator a_2.

Wynik ten oznacza, że dla populacji o współczynniku ekscesu $\gamma_2=3$ i przy losowaniu do próbki co najmniej $1+\sqrt{N}$ elementów estymator M_2^*, uzyskany bez znajomości średniej μ populacji generalnej z próbki wylosowanej bez zwracania, jest efektywniejszy, niż estymator a_2 uzyskany przy znajomości średniej μ z próbki wylosowanej ze zwracaniem. Wynik ten można wykorzystać w praktyce do przypadków, gdy nie jest znana średnia μ populacji generalnej.

Prace cytowane

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

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

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

Praca wpłynęła dnia 4. 5. 1954 r.

И. КОЗЬНЕВСКАЯ (Варшава)

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ВЫБОРА С ВОЗВРАЩЕНИЕМ И БЕЗ ВОЗВРАЩЕНИЯ ПРИ НЕИЗВЕСТНОЙ ДИСПЕРСИИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

РЕЗЮМЕ

В работе рассматривается генеральная совокупность состоящая из n элементов, со средней μ и дисперсией σ^2. Оценками дисперсии σ^2 являются параметры определенные формулами (1), причем m_2, M_2, a_2 — параметры выборки состоящей из n элементов выбранных с возвращением, а $m_2^*, M_2^*; a_2^*$ — соответствующие параметры выборки также состоящей из n элементов, но выбранных без возвращения. Для несмещенных оценок M_2^*, M_2, a_2^*, a_2 доказаны неравенства (2), (3), (4), и (5), справедливые для $2\leqslant n\leqslant N-1$ и $N\geqslant 3$. Необходимым и достаточным условием для неравенства $D(M_2^*)\leqslant D(a_2)$ является условие (6) для коэффициента эксцесса γ_2 генеральной совокупности.

Принимая во внимание условие (7) для коэффициента эксцесса получаем следующие результаты:

1. Для выборки с $n=2$ всегда имеет место неравенство (8).

2. Для выборки с $n=N-1$ всегда имеет место неравенство противоположного смысла неравенству (8).

3. Для генеральной совокупности с коэффициентом эксцесса $\gamma_2=3$ и для выборки с $n\geqslant 1+\sqrt{N}$ имеет место неравенство $D(M_2^*)<D(a_2)$.

I. KOŹNIEWSKA (Warszawa)

COMPARISON OF THE EFFICIENCY OF DRAWING LOTS WITH AND WITHOUT RETURNING THEM, WHEN THE VARIANCE OF THE GENERAL POPULATION IS UNKNOWN

SUMMARY

In the paper a general population with N elements, the mean μ and the variance σ^2 is considered. The estimates of variance σ^2 are the parameters defined by formulas (1), m_2, M_2, a_2 being the parameters of a sample consisting of n elements drawn and returned, and m_2^*, M_2^*, a_2^* being the corresponding parameters of a sample consisting of n elements drawn and not returned. For the unbiased estimates M_2^*, M_2, a_2^*, a_2 we prove inequalities (2), (3), (4) and (5), valid for $2\leqslant n\leqslant N-1$ and $N\geqslant 3$. The necessary and sufficient condition of the validity of the inequality $D(M_2^*)\leqslant D(a_2)$ is the validity of condition (6) concerning the coefficient of excess, γ_2, of the general population.

Taking into account condition (7) we reach the following conclusions:

1. For a sample of the size $n=2$ inequality (8) always holds.

2. For a sample of the size $n=N-1$ an inequality inverse to inequality (8) always holds.

3. For a general population with the coefficient of excess $\gamma_2=3$ for a sample of a size not less than $n\geqslant 1+\sqrt{N}$ the inequality $D(M_2^*)<D(a_2)$ holds.
