A. PLUCIŃSKA (Warszawa)

WARIANCJA ŚREDNIEJ W PEWNYM SCHEMACIE LOSOWANIA

W badaniach metodą reprezentacyjną mamy często do czynienia z następującą sytuacją: Populację generalną stanowi s arkuszy, a i-ty arkusz ($i = 1, 2, \ldots, s$) zawiera u_i pozycji. Celem badania reprezentacyjnego jest oszacowanie wyrażenia

$$\bar{u} = \frac{1}{s}\sum_{i=1}^{s} u_i, \tag{1}$$

gdzie s jest znane, ale u_i są nieznane.

Za przykład może służyć populacja generalna złożona z arkuszy spisowych spisu powszechnego ludności, przy czym każdy arkusz zawiera kilka osób, mianowicie osoby prowadzące wspólne gospodarstwo domowe. Metodą reprezentacyjną chcemy ustalić średnią ilość osób w gospodarstwie.

W rozważanej sytuacji dogodnie jest nieraz stosować następujący schemat losowania: Traktujemy arkusz jako jednostkę losowania i losujemy dla każdego arkusza, czy ma wejść do próby czy nie, przy czym losowanie jest niezależne, a każdy arkusz ma jednakowe prawdopodobieństwo p ($0 < p < 1$) dostania się do próbki. W rozważanym schemacie mamy s jednostek losowania.

Oznaczamy przez X_i ($i = 1, 2, \ldots, s$) liczbę elementów indywidualnych wylosowanych w i-tym losowaniu. Mamy

$$\mathrm{P}(X_i = u_i) = p, \quad \mathrm{P}(X_i = 0) = 1-p = q \quad (i = 1, 2, \ldots, s). \tag{2}$$

Oznaczamy przez X liczbę indywidualnych elementów w próbie. Mamy

$$X = \sum_{i=1}^{s} X_i. \tag{3}$$

Niech N oznacza liczbę jednostek losowania, które weszły do próbki. N jest zmienną losową o rozkładzie dwumianowym

$$\mathrm{P}(N = n) = \binom{s}{n} p^n q^{s-n}. \tag{4}$$

Za estymator nieznanej wielkości $\bar{u}$ określonej wzorem (1) przyjmujemy

$$Y = X/N. \tag{5}$$

Celem niniejszej noty jest znalezienie wariancji $D^2(Y)$ zmiennej losowej Y i porównanie jej z wariancją średniej arytmetycznej z próbki o stałej liczbie ps elementów wybranych z populacji zgodnie ze schematem Bernoulliego. Ta ostatnia wariancja równa jest, jak wiadomo, $D^2(u)/ps$, gdzie

$$D^2(u) = \frac{1}{s}\sum_{i=1}^{s}(u_i-\bar{u})^2.$$

W dalszych rozważaniach będziemy jednak pomijali ewentualność, że $N = 0$, gdyż prawdopodobieństwo tego zdarzenia wynoszące q^s jest przy tych wartościach s, z którymi mamy do czynienia w praktyce, bardzo małe. Ściśle mówiąc znajdziemy wariancję $D^2(Y|N > 0)$, tj. wariancję zmiennej losowej Y pod warunkiem, że $N > 0$. Udowodnimy

TWIERDZENIE. *Niech zmienne losowe X i N mają rozkłady określone wzorami* (2)-(4) *i niech Y będzie dane wzorem* (5). *Wówczas zachodzą relacje*

$$\text{(a)} \qquad E(Y|N > 0) = \bar{u},$$

$$\text{(b)} \qquad D^2(Y|N > 0) = \frac{D^2(u)}{1-q^s}\sum_{n=1}^{s}\frac{q^{s-n}-q^s}{n},$$

przy czym dla $s \to \infty$ zachodzi równość asymptotyczna

$$\text{(c)} \qquad D^2(Y|N > 0) \cong D^2(u)/ps.$$

Dowód. Mamy

$$E(Y|N > 0) = \frac{\sum_{n=1}^{s}E(Y|N = n)P(N = n)}{1-q^s} = \frac{\bar{u}(1-q^s)}{1-q^s} = \bar{u}. \tag{6}$$

Teza (a) została więc udowodniona. Widzimy, że zmienna losowa $(Y|N > 0)$ jest nieobciążonym estymatorem parametru $\bar{u}$.

Obliczymy teraz drugi moment. Mamy

$$E(Y^2|N > 0) = \frac{\sum_{n=1}^{s}E(Y^2|N = n)P(N = n)}{1-q^s}. \tag{7}$$

Dla każdego n

$$\mathrm{E}(Y^2|N=n) = \mathrm{D}^2(Y|N=n)+[\mathrm{E}(Y|N=n)]^2,$$

a więc korzystając z wzorów (5) i (6) otrzymujemy

$$\mathrm{E}(Y^2|N=n) = \mathrm{D}^2(u)/n+\bar{u}^2.$$

Stąd i z wzoru (7) wynika

$$\mathrm{E}(Y^2|N>0) = \frac{\sum_{n=1}^{s}\left[\frac{\mathrm{D}^2(u)}{n}+\bar{u}^2\right]\mathrm{P}(N=n)}{1-q^s} = \mathrm{D}^2(u)\frac{\sum_{n=1}^{s}\frac{1}{n}\mathrm{P}(N=n)}{1-q^s}+\bar{u}^2.$$

Ostatecznie więc

$$\mathrm{D}^2(Y|N>0) = \frac{\mathrm{D}^2(u)\sum_{n=1}^{s}\frac{1}{n}\mathrm{P}(N=n)}{1-q^s}.$$

Oznaczmy

$$\varphi(P) = \sum_{n=1}^{s}\frac{1}{n}\mathrm{P}(N=n) = \sum_{n=1}^{s}\frac{1}{n}\binom{s}{n}p^n(1-p)^{s-n}. \tag{8}$$

W celu obliczenia tej sumy różniczkujemy ją najpierw względem p:

$$\varphi'(p) = \sum_{n=1}^{s}\binom{s}{n}p^{n-1}(1-p)^{s-n} - \sum_{n=1}^{s}\frac{s-n}{n}\binom{s}{n}p^n(1-p)^{s-n-1} =$$

$$= \frac{1}{p(1-p)}[1-(1-p)^s] - \frac{s}{1-p}\varphi(p).$$

Otrzymaliśmy równanie różniczkowe liniowe rzędu pierwszego. Rozwiązaniem równania jednorodnego

$$\varphi'(p) = -\frac{s}{1-p}\varphi(p)$$

jest

$$\varphi(p) = c(1-p)^s.$$

Po uzmiennieniu stałej i dokonaniu drobnych przekształceń otrzymujemy

$$c'(p) = \frac{1}{p(1-p)^{s+1}}[1-(1-p)^s].$$

Aby obliczyć $\int \frac{dp}{p(1-p)^{s+1}}$ korzystamy z wzoru

$$\frac{1}{p(1-p)^{s+1}} = \frac{1}{p} + \frac{1}{(1-p)^{s+1}} + \ldots + \frac{1}{1-p},$$

którego słuszność łatwo wykazać indukcyjnie, i znajdujemy

$$c = \sum_{n=1}^{s} \frac{(1-p)^{n-1-s}}{s+1-n} + A,$$

gdzie A jest stałą całkowania.

Wstawiając wartość c otrzymujemy

$$\varphi(p) = \sum_{n=1}^{s} \frac{(1-p)^{n-1}}{s+1-n} + A(1-p)^{s}. \tag{9}$$

Stałą A wyznaczamy przyjmując w wyrażeniu (8) $p = 0$. Stąd otrzymujemy $\varphi(0) = 0$, a następnie porównujemy z (9).

Ostatecznie

$$A = \sum_{n=1}^{s} \frac{1}{s+1-n},$$

$$\varphi(p) = \sum_{n=1}^{s} \frac{(1-p)^{n-1} - (1-p)^{s}}{s+1-n} = \sum_{n=1}^{s} \frac{q^{s-n} - q^{s}}{n}.$$

Tym samym wykazany został wzór (b).

Przechodzimy do dowodu wzoru (c).

Oszacujemy sumę

$$\sum_{n=1}^{s} \frac{q^{s-n}}{n} = \sum_{n=0}^{s-1} \frac{q^{n}}{s-n}.$$

W tym celu skorzystamy z rozwinięcia

$$\frac{1}{s-n} = \frac{1}{s} + \frac{n}{s^2} + \frac{n^2}{s^3} + \ldots,$$

które daje

$$\sum_{n=0}^{s-1} \frac{q^{n}}{s-n} = \frac{1}{s} \sum_{n=0}^{s-1} q^{n} + \frac{1}{s^2} \sum_{n=0}^{s-1} nq^{n} + \ldots = \frac{1}{s} \sum_{n=0}^{s-1} q^{n} + O\left(\frac{1}{s}\right) \cong \frac{1}{sp}.$$

Zauważmy dalej, że

$$\lim_{s=\infty} q^s \sum_{n=1}^{s} \frac{1}{s+1-n} = O\left(\frac{1}{s}\right),$$

mamy bowiem

$$\lim_{s=\infty} sq^s \sum_{n=1}^{s} \frac{1}{s+1-n} < \lim_{s=\infty} s^2 q^s = 0.$$

Jeżeli ponadto uwzględnimy, że $\lim_{s=\infty}(1-q^s) = 1$, to będziemy mieli

$$\mathrm{D}^2(Y|N > 0) \cong \mathrm{D}^2(u)/ps.$$

Wzór (c) został więc udowodniony.

Widzimy więc, że dla dużych wartości s jednakowa jest efektywność omawianego w niniejszej pracy estymatora $(Y|N > 0)$ określonego przez (5) i średniej arytmetycznej z próbki o stałej liczbie ps elementów wybranych z populacji. Asymptotyczną wariancję średniej dla rozważanego typu schematu losowania znalazł Fisz [1].

Dla niezbyt dużych wartości s średnia arytmetyczna z próbki o stałej liczbie ps elementów jest estymatorem efektywniejszym. Łatwo to sprawdzić przez zbadanie różnicy wariancji obu estymatorów, która równa jest

$$R = \frac{\mathrm{D}^2(u)}{ps} - \frac{\mathrm{D}^2(u)}{1-q^s} \sum_{n=1}^{s} \frac{q^{s-n}-q^s}{n} =$$

$$= \frac{\mathrm{D}^2(u)}{sp(1-q^s)}\left[1-q^s-ps\left(\frac{1}{s}\sum_{n=1}^{s} q^{s-n} + \sum_{n=1}^{s} \frac{s-n}{sn}\, q^{s-n} - \sum_{n=1}^{s} \frac{q^s}{n}\right)\right] =$$

$$= C \sum_{n=1}^{s} \frac{-(s-n)q^{1-n}+sq^s}{n} < C \sum_{n=1}^{s} [-(s-n)q^{s-n}+sq^s] =$$

$$= C \sum_{k=0}^{s-1} (-kq^k+sq^s) = C \sum_{k=0}^{s-1} q^k(-k+sq^{s-k}) < C \sum_{k=0}^{s-1} (-k+sq^{s-k}) =$$

$$= C\left(-\frac{s(s-1)}{2} + sq\,\frac{1-q^s}{1-q}\right),$$

gdzie

$$C = \frac{\mathrm{D}^2(u)}{s^2 p(1-q^s)}.$$

Zbadamy dla jakich wartości q różnica $R < 0$, czyli kiedy są spełnione nierówności

$$sq(1-q^s)/(1-q) < s(s-1)/2,$$
$$q-q^{s+1} < (s-1)/2-q(s-1)/2,$$
$$q-2q^{s+1}/(s+1) < (s-1)/(s+1).$$

Zatem $R < 0$, przynajmniej gdy $q < (s-1)/(s+1)$.

Nierówność $q < (s-1)/(s+1)$ obejmuje wszystkie praktycznie ważne przypadki. Za pomocą podobnych oszacowań można wykazać, że estymator (5) jest efektywniejszy, gdy

$$q-2q^{s+1}/(s+1) > (s-1)/(s+1).$$

Nierówność tę spełnia np. wartość $q = (s-1)/s$.

Praca cytowana

[1] M. Fisz, *Efektywność dwóch schematów losowania*, Studia i prace statystyczne 1 (1951).

Praca wpłynęła 12. 2. 57

А. ПЛУЦИНСКАЯ (Варшава)

ДИСПЕРСИЯ СРЕДНЕГО В НЕКОТОРОЙ СХЕМЕ ВЫБОРОВ

РЕЗЮМЕ

Генеральная совокупность состоит из s единиц выбора, каждая i-ая единица ($i = 1, 2, ..., s$) содержит u_i индивидуальных элементов. Целью репрезентативного исследования является оценка выражения (1)

$$\bar{u} = \frac{1}{s}\sum_{i=1}^{s} u_i.$$

В качестве оценки этой величины принимаем Y, определенное формулой (5). N есть случайная величина равная числу единиц выбора, которые вошли в состав выборки и которая по определению больше нуля. Доказывает, что

1) Y является несмещенной оценкой,

2) дисперсия $D^2(Y)$ определена формулой (b),

3) асимптотическая эффективность оценки Y и среднего арифметического выборка с постоянным числом ps элементов одинаковы, но для небольших значений s среднее арифметическое быборки с постоянным числом ps элементов является более эффективной оценкой, чем оценка Y, по крайней мере для $q < (s-1)/(s+1)$.

A. PLUCIŃSKA (Warszawa)

VARIANCE OF THE MEAN IN A CERTAIN SAMPLING SCHEME

SUMMARY

The general population consists of s sampling units, every i-th unit ($i = 1, 2, \ldots, s$) contains u_i individual elements. The object of the survey is the estimation of expression (1)

$$\bar{u} = \frac{1}{s}\sum_{i=1}^{s} u_i .$$

We take Y defined by (5) as the estimator of that quantity. N is a random variable denoting the number of sampling units which have entered into the sample, that number being greater than zero by hypothesis. We prove that

1) Y is an unbiased estimator,

2) the variance $D^2(Y)$ is defined by (b),

3) the asymptotic effectiveness of the estimator Y and that of the arithmetic mean of a sample with a constant number ps of elements are identical, but for not very large s the arithmetic mean of a sample with the constant number ps of elements is a more effective estimator than the estimator Y, at least for $q < (s-1)/(s+1)$.