ZAKŁAD WYDAWNICTW STATYSTYCZNYCH
al. Niepodległości 208, 00-925 Warszawa, tel. 608-31-45.
Informacje w sprawach nabywania czasopism tel. 608-32-10, 608-38-10
REDAKCJA pok. 515, tel. 608-32-25

Indeks 381306

Zam. 212/96 — 900

NR 7 (422)

WIADOMOŚCI STATYSTYCZNE

ORGAN: GŁÓWNEGO URZĘDU STATYSTYCZNEGO
I POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO.

LIPIEC 1996

STUDIA METODOLOGICZNE

Paweł J. SZABŁOWSKI, Jacek WESOŁOWSKI, Robert WIECZORKOWSKI

Estymacja w podpopulacjach

W badaniach prowadzonych metodą reprezentacyjną często pojawiającym się problemem jest estymacja parametrów dotyczących podpopulacji. Gdy w ramach schematu losowania nie ustala się liczby elementów próby z danej podpopulacji, to liczba ta jest wielkością losową. W takiej sytuacji najczęstszym sposobem postępowania przy estymowaniu parametru podpopulacji jest zastosowanie standardowo przyjętej, dla danego schematu losowania, procedury estymacyjnej, ale nie dla cechy wyjściowej, powiedzmy Y, lecz dla jej zawężenia do podpopulacji $Y_A = YI(Y \in A)$, gdzie zbiór A oznacza podpopulację, a I jest indykatorem.

Mówiąc prościej dla wszystkich elementów próby spoza podpopulacji kładziemy wartość cechy równą zero, a wartość cechy dla elementów próby należących do podpopulacji pozostawiamy bez zmian. Jest to standardowe postępowanie, opisane np. w monografii Cochran (1977) [2], rozzdz. 2, prowadzące do uzyskania estymatora nieobciążonego. Jednak w przypadku małych podpopulacji estymator taki ma stosunkowo duże prawdopodobieństwo przyjmowania wartości równej zero i w związku z tym występuje efekt przeszacowania, gdy w próbie są elementy z podpopulacji (mowa o nieujemnych parametrach).

Inny mankament takiego podejścia polega na błędach związanych z nierównomiernym rozkładem próby, tzn. sytuacjach takich, gdy liczba elementów w próbie pochodzących z rozważanej podpopulacji różni się znacznie od wartości oczekiwanej. W takiej sytuacji wydaje się naturalne uzależnienie postaci estymatora od tej liczby, polegające na zastąpieniu mnożnika, który w oryginalnym estymatorze jest stały, wielkością losową.

OGÓLNE UWAGI O ESTYMACJI PRZY LOSOWEJ DŁUGOŚCI PRÓBY

Niech Z_1, Z_2, ... będą zmiennymi losowymi o jednakowych wartościach oczekiwanych $m=E(Z_i)$, wariancji $\sigma^2=\mathrm{var}(Z_i)$ i współczynniku korelacji ρ, $\sigma^2\rho=E(Z_iZ_j)-E(Z_i)E(Z_j)$, $i, j=1, 2, ..., i\neq j$. Niech M będzie nie zdegenerowaną zmienną losową taką, że supp $(M)\subset\{1, 2, ...\}$, niezależną od ciągu $Z=\{Z_i\}_{i=1}^{\infty}$, $E(M^2)<\infty$. Celem jest estymacja średniej m, przy założeniu, że ciąg Z jest obserwowany do momentu M, tzn. obserwujemy $Z_1, ..., Z_M$.

Przykład 1. Problemy tego typu napotykamy przy estymacji średniej cechy X w podpopulacji A (o liczności N_A) na podstawie n-krotnego losowania ze zwracaniem z całej populacji (o liczności N), przy czym wnioskowanie prowadzone jest tylko w przypadku trafienia w zbiór A. Wtedy M ma ucięty w zerze rozkład dwumianowy $b(n, p)$, gdzie $p=N_A/N$, tzn.

$$P(M=k)=\frac{1}{1-q^n}\binom{n}{k}p^kq^{n-k} \qquad k=1, ..., n$$

$q=1-p\in(0, 1)$. Natomiast zmienna losowa Z_1 ma rozkład jednostajny w zbiorze $\{x_1, ..., x_{N_A}\}$ wartości cechy X przyjmowanej na elementach ze zbioru A, a ciąg Z jest ciągiem niezależnych zmiennych losowych o tych samych rozkładach.

Przykład 2. Kolejny interesujący model dotyczy zagadnienia podobnego do problemu przedstawionego w przykładzie 1, lecz w przypadku gdy n-krotne losowanie próby jest bez zwracania. W tej sytuacji zmienne Z_i, $i=1, 2, ...$, mają również jednakowe rozkłady jednostajne na zbiorze $\{x_1, ..., x_{N_A}\}$, lecz są skorelowane ze współczynnikiem korelacji:

$$\rho=\frac{1}{(N_A-1)}$$

Natomiast M, przy naturalnym założeniu $n\leq\min(N_A, N-N_A)$, ma ucięty w zerze rozkład hipergeometryczny $hg(N, N_A, n)$:

$$P(M=k)=\frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}-\binom{N-N_A}{n}} \qquad k=1, 2, ..., n \tag{1}$$

Przykład 3. Inny model dotyczyć może losowania ze zwracaniem prowadzonego do momentu wylosowania elementu spoza zbioru A, przy czym ponownie interesujące są jedynie przypadki, gdy w pierwszym losowaniu trafiamy w zbiór A. Wtedy M ma ucięty w zerze rozkład geometryczny, tzn.

$$P(M=k)=(1-p)p^{k-1} \qquad k=1, 2, ...$$

Rozważymy dwa estymatory wartości oczekiwanej m:

$$T_1=\frac{1}{M}\sum_{i=1}^{M}Z_i \qquad T_2=\frac{1}{E(M)}\sum_{i=1}^{M}Z_i$$

Uwaga 1. Odpowiadające im estymatory wartości globalnych w podpopulacji A o liczności N_A mają postać:

$$\tilde{T}_1=N_AT_1 \text{ i } \tilde{T}_2=N_AT_2$$

W dalszym ciągu będziemy zajmować się jedynie estymatorami średniej, ale otrzymane wyniki w sposób oczywisty można będzie przenieść na estymatory wartości globalnej.

Oto szereg ogólnych własności wprowadzonych estymatorów:

Twierdzenie 1. Estymatory T_1 i T_2 są nieobciążone.

Twierdzenie 2.

$$\mathrm{var}(T_2)\geq\mathrm{var}(T_1)$$

wtedy i tylko wtedy, gdy

$$\left(\frac{\mathrm{var}(M)}{E(M)}\right)\eta^2\geq(E(M)E(M^{-1})-1)(1-\rho)-\rho\frac{\mathrm{var}(M)}{E(M)} \tag{2}$$

gdzie $\eta=\frac{m}{\sigma}$ jest odwrotnością współczynnika zmienności zmiennych losowych z ciągu Z (tj. $1/\eta=\gamma=\sigma/m$).

W szczególności, gdy zmienne losowe Z_i, $i=1, 2,...$, są nie skorelowane powyższa nierówność ma postać:

$$\left(\frac{\mathrm{var}(M)}{E(M)}\right)\eta^2\geq E(M)E(M^{-1})-1 \tag{3}$$

Uwaga 2. Zauważmy, że $E(M)E(M^{-1})\geq 1$. Zatem, gdy $m=0$ (czyli gdy $\eta=0$) i $\rho\leq 0$, to dla każdego rozkładu zmiennej M **lepszy jest estymator** T_2.

W kolejnych dwu rozdziałach niniejszej pracy zastosujemy otrzymane wyniki ogólne do uzasadnienia teoretycznego następującej praktycznej reguły wyboru estymatorów.

Jeżeli $n\eta^2\geq\frac{N}{N_A}$, to należy stosować estymator T_1, w przeciwnym razie należy stosować T_2.

Rozważmy teraz bardziej szczegółowo przykłady 1 i 2.

LOSOWANIE ZE ZWRACANIEM

Obecnie przeprowadzimy analizę otrzymanych zależności w przypadku losowania ze zwracaniem opisanego w przykładzie 1. Zmienne losowe Z_i, $i=1, 2,...$, są niezależne, więc podstawą rozważań będzie wzór (3).

Zmienna losowa M ma ucięty w zerze rozkład dwumianowy $b(n, p)$, więc

$$E(M)=\frac{np}{1-q^n}$$

$$\operatorname{var}(M)=\frac{npq+n^2p^2}{1-q^n}-\frac{n^2p^2}{(1-q^n)^2}$$

$$E(M^{-1})=(1-q^n)^{-1}h_n(q)$$

gdzie

$$h_n(q)=\sum_{k=1}^{n}\frac{1}{k}\binom{n}{k}p^kq^{n-k}$$

Lemat 1. Mamy

$$h_n(q)=\sum_{k=1}^{n}q^{n-k}/k-q^n\sum_{i=1}^{n}\frac{1}{k}$$

Wobec tego z (3) wynika, że wybór estymatora zależy od położenia pierwiastków wielomianu:

$$P(q)=q(1-q^n)^2-n(1-q)(1-q^n)q^n-\gamma^2[n(1-q)h_n(q)-(1-q^n)^2]$$

względem $\frac{N-N_A}{N}$, $\gamma=\eta^{-1}$. Dla małych wartości n przykłady przedziałów dla q, w których należy wybierać estymator T_1 podano w tablicy 1. W tablicach 2, 3, 4 przedstawiono wyniki symulacji potwierdzające dokładne reguły wyboru z tablicy 1. Ogólnie, gdy wartość η może być wstępnie oszacowana (np. z poprzednich badań) i przy znanych N i N_A, wystarczy sprawdzić znak wyrażenia $P(q)$. Gdy $P(q)>0$ wybieramy T_1, w przeciwnym przypadku T_2.

Obecnie przeprowadzimy analizę asymptotyczną (duże wartości n) nierówności (3). Zacznijmy od podstawowego wyniku o charakterze technicznym dotyczącego asymptotyki ciągu funkcji h_n.

Lemat 2.

$$\lim_{n\to\infty}n\left(nh_n(q)-\frac{1}{p}\right)=\frac{1-p}{p^2} \tag{4}$$

W konsekwencji, jeżeli M_n ma ucięty w zerze rozkład dwumianowy $b(n,p)$, $n=1, 2, \ldots$, to mamy:

Lemat 3.

$$0\le E(M_n)E(M_n^{-1})-1=\frac{1-p}{np}+o\left(\frac{1}{n}\right) \tag{5}$$

Ponieważ $(q=1-p)$

$$\frac{\operatorname{var}(M_n)}{E(M_n)}=\frac{(1-q^n)q-n(1-q)q^n}{1-q^n}=q+o\left(\frac{1}{n}\right) \tag{6}$$

więc, zgodnie z (6), nierówność (3) przyjmuje postać

$$q\eta^2\ge\frac{q}{np}+o\left(\frac{1}{n}\right)$$

Ze względu na to, że $p=\frac{N_A}{N}$, więc pomijając małe wyrazy rzędu $o\left(\frac{1}{n}\right)$ otrzymamy nierówność

$$n\eta^2\ge\frac{N}{N_A}$$

skąd wynika reguła podana w zakończeniu rozdziału poprzedniego.

Różnice w wartościach n pochodzące ze wzoru dokładnego i reguły asymptotycznej pokazuje tablica 5. Trafność podanej reguły przybliżającej potwierdzają symulacje, których wyniki przedstawione są w tablicy 6.

LOSOWANIE BEZ ZWRACANIA

W przypadku losowania bez zwracania zmienne losowe Z_i, $i=1, 2, \ldots$, nie są niezależne, więc podstawą rozważań jest wzór (2) ze współczynnikiem korelacji $\rho=-\frac{1}{N_A-1}$ (zob. przykład 2). Jak zauważono w rozdziale 2, jeżeli

$$n\le\min(N_A, N-N_A) \tag{7}$$

(a tylko takim przypadkiem będziemy się tu zajmowali), to zmienna losowa M_n ma ucięty w zerze rozkład hipergeometryczny $hg(N, N_A, n)$. Własności tego rozkładu zsumowane są w poniższym stwierdzeniu

Stwierdzenie 1. Niech $p=N_A/N$. Wówczas

i)

$$EM=\frac{pn}{\alpha} \tag{8}$$

ii)

$$\operatorname{var}(M)=\frac{np(1-p)(N-n)}{\alpha(N-1)}-\frac{(1-\alpha)}{\alpha^2}n^2p^2 \tag{9}$$

w szczególności

$$\frac{\operatorname{var}(M)}{E(M)}=\frac{(1-p)(N-n)}{(N-1)}-\frac{(1-\alpha)}{\alpha}np$$

iii)

$$EM^{-1}=\frac{1}{\alpha\binom{N}{n}}\sum_{k=1}^{n}\frac{1}{k}\binom{N_A}{k}\binom{N-N_A}{n-k}$$

gdzie oznaczyliśmy $\alpha=1-\dfrac{\binom{N-N_A}{n}}{\binom{N}{n}}$

Wobec tego nierówność (2) przybiera postać:

$$n\eta^2\geq\frac{nN_A}{N_A-1}\,\frac{npE(M^{-1})-\alpha}{\dfrac{\alpha(1-p)(N-n)}{N-1}-np(1-\alpha)}+\frac{n}{N_A-1} \tag{10}$$

Dla małych liczności populacji N (i w konsekwencji małych liczności podpopulacji N_A i próby n) należy kierować się przy wyborze estymatora dokładną nierównością (10).

Dla dużych wartości n (a więc i dużych N i N_A) obliczenie prawej strony (10) jest skomplikowaną procedurą ze względu na postać $E(M^{-1})$ oraz α. Stąd konieczność znalezienia przybliżonej reguły asymptotycznej.

Ze względu na następującą uwagę

Uwaga 3. Jeśli $X_N\sim hg(N, N_A, n)$ i $\lim\limits_{N\to\infty}\dfrac{N_A}{N}=p$, to $\lim\limits_{N\to\infty}P(X_N=k)=\binom{n}{k}p^k(1-p)^{n-k}$, czyli X_N ma w granicy (przy $N\to\infty$) rozkład $b(n, p)$.

można oczekiwać rezultatu podobnego do tego, który uzyskano w poprzednim rozdziale dla rozkładu dwumianowego. Hipotezę tę potwierdzają rozważania przeprowadzone w dalszej części rozdziału.

Uwaga 4. Nietrudno zauważyć, że $1-\exp\left(-\dfrac{N_An}{N}\right)\leq\alpha\leq1-\exp\left(-\dfrac{N_An}{N-n+1}\right)$. Zatem dla dużych n i $\dfrac{N}{N_A}$ niezbyt dużych (rzędu kilku), α praktycznie równa się 1.

Zmienimy teraz nieco oznaczenia dodając do odpowiednich symboli indeks n, aby podkreślić ich zależność od rozmiaru próby n. Dalsze rozważania prowadzone będą przy założeniu, że:

$$\lim_{n\to\infty}p_n=p\qquad\lim_{n\to\infty}\frac{n}{N_{A,n}}=0 \tag{11}$$

Dodatkowo ze względów technicznych przyjmujemy:

$$n(N_{A,n}+3)\geq2(N_n-N_{A,n}+1)$$

Lemat 4. i) Mamy:

$$\lim_{n\to\infty}n\left[nE(M_n^{-1})-\frac{1}{p}\right]=\frac{1-p}{p^2}$$

ii) W szczególności wykorzystując wzór (8) dostaniemy

$$\lim_{n\to\infty}n\left[EM_n^{-1}EM_n-1\right]=\frac{1-p}{p} \tag{12}$$

$$\lim_{n\to\infty}\frac{\operatorname{var}(M_n)}{EM_n}=1-p=q \tag{13}$$

Zatem przy $n\to\infty$ prawa strona wzoru (10) ma granicę równą $\dfrac{1}{p}$. W konsekwencji dla dużych wartości n i znacząco większych $N_{A,n}$ i N otrzymujemy w przybliżeniu nierówność

$$\eta^2n>\frac{1}{p}=\frac{N}{N_A} \tag{14}$$

potwierdzającą regułę praktyczną zaproponowaną wcześniej. Ilustracją podanej reguły są wyniki symulacji zawarte w tablicy 7.

LOSOWANIE DO WYPADNIĘCIA POZA PODPOPULACJĘ

Obecnie zajmujemy się przypadkiem opisanym w przykładzie 3. Ponieważ Z_i, $i=1, 2, \ldots$, są niezależne bierzemy pod uwagę wzór (3).

Gdy M ma ucięty w zerze rozkład geometryczny, to

$$E(M)=\frac{1}{1-p}$$

$$\operatorname{var}(M)=\frac{p}{(1-p)^2}$$

oraz

$$E(M^{-1})=\frac{p-1}{p}\ln(1-p)$$

Więc nierówność (3) przyjmuje postać:

$$p^2\leq\eta^2(p-1)(\ln(1-p)+p)$$

gdzie $p=\dfrac{N_A}{N}$. Jeżeli spełniona jest powyższa nierówność wybieramy estymator T_1, w przeciwnym przypadku T_2.

EKSPERYMENTY NUMERYCZNE

Oddzielnie rozważamy dwa schematy losowania: ze zwracaniem i bez zwracania. ***Losowanie ze zwracaniem.*** W pierwszej zamieszczonej tablicy prezentujemy dla wybranych parametrów γ^2 oraz n przedziały, dla których wielkość proporcji $x=\frac{N-N_A}{N}$ zapewnia nieujemność wielomianu $P(x)$, co jak wykazaliśmy jest równoważne mniejszej wariancji estymatora T_1.

TABL. 1. PRZEDZIAŁY ZMIENNEJ x, DLA KTÓRYCH ZACHODZI $P(x)>0$

n \ γ^2	0,01	0,1	1,0	4,0	16	100
5	(0;1)	(0;1)	(0;1)	0	0	0
10	(0;1)	(0;1)	(0;1)	(0;0,437)	0	0
15	(0;1)	(0;1)	(0;1)	(0;0,613)	0	0
20	(0;1)	(0;1)	(0;1)	(0;0,706)	(0;0,142)	0
25	(0;1)	(0;1)	(0;1)	(0;0,763)	(0;0,306)	0
30	(0;1)	(0;1)	(0;1)	(0;0,802)	(0;0,417)	0
50	(0;1)	(0;1)	(0;1)	(0;0,880)	(0;0,645)	0
100	(0;1)	(0;1)	(0;1)	(0;0,940)	(0;0,821)	(0;0,645)

Tablice 2, 3, 4 zawierają wyniki eksperymentów symulacyjnych dotyczących porównywania odchyleń standardowych $S(T_1)$ i $S(T_2)$ dla rozważanych estymatorów T_1 i T_2. Jako populacje generowano tablice $pop[1..N]$ liczb losowych z rozkładu normalnego $N\left(\frac{1}{\gamma}, 1\right)$ (przy takim wyborze współczynnik zmienności rozkładu wynosi γ). Podpopulację stanowiła początkowa część tablicy $pop[1..N_A]$. Następnie wielokrotnie (10000 iteracji) dokonywano losowania ze zwracaniem prób n-elementowych, za każdym razem obliczając wartości obu estymatorów. Po zakończeniu pętli iteracyjnej obliczano odchylenia standardowe z próby $S(T_1)$ i $S(T_2)$ dla badanych estymatorów. W poniższych przykładach przyjęto $N=1000$, zmiany parametru $p=1-x=\frac{N_A}{N}$ dokonywano przez odpowiedni wybór N_A.

Obliczenia przeprowadzono na komputerze typu PC486 z systemem operacyjnym *Linux* z użyciem języka C. Program napisany w języku C używał generatora rozkładu równomiernego z pracy [3].

Wyniki symulacyjne potwierdzają rezultaty teoretyczne z tablicy 1.

TABL. 2

$p=1-x$	γ^2	$n=10$ $S(T_1)$	$n=10$ $S(T_2)$	$n=20$ $S(T_1)$	$n=20$ $S(T_2)$	$n=50$ $S(T_1)$	$n=50$ $S(T_2)$
0,1	0,01	0,877	4,87	0,764	5,28	0,502	4,23
0,2	0,01	0,740	5,16	0,540	4,35	0,322	2,86
0,3	0,01	0,660	4,48	0,449	3,45	0,266	2,18
0,4	0,01	0,550	3,80	0,375	2,77	0,227	1,76
0,5	0,01	0,483	3,18	0,325	2,23	0,203	1,42
0,6	0,01	0,423	2,62	0,294	1,83	0,184	1,17
0,7	0,01	0,392	2,12	0,275	1,48	0,170	0,947
0,8	0,01	0,360	1,60	0,253	1,15	0,157	0,722
0,9	0,01	0,332	1,09	0,236	0,783	0,148	0,497

TABL. 3

$p=1-x$	γ^2	$n=10$ $S(T_1)$	$n=10$ $S(T_2)$	$n=20$ $S(T_1)$	$n=20$ $S(T_2)$	$n=50$ $S(T_1)$	$n=50$ $S(T_2)$
0,1	4,0	0,879	0,848	0,756	0,719	0,503	0,497
0,2	4,0	0,739	0,712	0,550	0,546	0,324	0,350
0,3	4,0	0,652	0,619	0,450	0,453	0,265	0,282
0,4	4,0	0,556	0,538	0,377	0,380	0,226	0,240
0,5	4,0	0,479	0,470	0,322	0,327	0,202	0,211
0,6	4,0	0,417	0,414	0,292	0,298	0,183	0,190
0,7	4,0	0,391	0,394	0,268	0,274	0,171	0,176
0,8	4,0	0,356	0,360	0,252	0,256	0,156	0,160
0,9	4,0	0,333	0,335	0,238	0,240	0,147	0,148

TABL. 4

$p=1-x$	γ^2	$n=10$ $S(T_1)$	$n=10$ $S(T_2)$	$n=20$ $S(T_1)$	$n=20$ $S(T_2)$	$n=50$ $S(T_1)$	$n=50$ $S(T_2)$
0,1	100	0,880	0,813	0,763	0,666	0,504	0,446
0,2	100	0,743	0,654	0,555	0,494	0,324	0,312
0,3	100	0,651	0,583	0,443	0,412	0,270	0,264
0,4	100	0,559	0,502	0,376	0,361	0,229	0,228
0,5	100	0,476	0,441	0,327	0,318	0,200	0,198
0,6	100	0,426	0,404	0,293	0,288	0,184	0,183
0,7	100	0,396	0,384	0,272	0,269	0,170	0,169
0,8	100	0,357	0,352	0,251	0,249	0,156	0,156
0,9	100	0,330	0,328	0,235	0,234	0,149	0,148

W tablicy 5 przedstawiono wyniki obliczeń liczności próby koniecznej dla zapewnienia lepszej precyzji estymatora T_1, z użyciem nierówności asymptotycznej

$$n>\frac{N}{N_A}\gamma^2$$

oraz za pomocą badania znaku wielomianu $P(x)$.

W obliczeniach przyjęto parametr $\gamma^2=100$.

TABL. 5. MINIMALNE LICZNOŚCI, DLA KTÓRYCH MNIEJSZĄ WARIANCJĘ MA ESTYMATOR T_1

$x=1-p$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
n asymptotyczne	111	125	142	166	200	250	333	500	1000
n według $P(x)$	113	127	145	170	204	255	340	510	1020

Symulacje ilustrujące użyteczność formuły asymptotycznej zawarto również w tablicy 6.

W obliczeniach przyjęto $N=10000$.

TABL. 6. ODCHYLENIA STANDARDOWE Z PRÓBY DLA BADANYCH ESTYMATORÓW

N_A	$\frac{N}{N_A}$	γ^2	n	$S(T_1)$	$S(T_2)$
1000	10	100	900	0,105	0,105
1000	10	100	1000	0,100	0,100
1000	10	100	1500	0,080	0,080
5000	2	100	100	0,141	0,141
5000	2	100	200	0,101	0,101
5000	2	100	300	0,082	0,082
100	10	100	1000	0,334	0,317
100	10	100	5000	0,142	0,142
1000	10	4	40	0,559	0,541
1000	10	4	100	0,333	0,345
5000	2	4	10	0,481	0,467
5000	2	4	50	0,200	0,208
100	100	4	300	0,664	0,639
100	100	4	400	0,571	0,558
100	100	4	500	0,504	0,508

Losowanie bez zwracania. W przypadku losowania bez zwracania wykonano również doświadczenia symulacyjne dla wybranych parametrów. Użyto analogicznego schematu jak opisany dla przypadku losowania ze zwracaniem. Różnica polegała na użyciu procedury losowania bez zwracania napisanej w języku C według procedury RANKSB zamieszczonej w książce [1].

Wykonane symulacje pozwalają na sprawdzenie użyteczności wyprowadzonej wcześniej formuły asymptotycznej dla liczności próby zapewniającej mniejszą wariancję estymatora T_1. Formuła ta miała postać $n > \frac{N}{N_A}\gamma^2$ (przy spełnieniu dodatkowych założeń $n < N - N_A$ oraz $n > \frac{2(N-N_A+1)}{N_A+3}$). W obliczeniach do tablicy 7 przyjęto $N = 10000$.

TABL. 7. ODCHYLENIA STANDARDOWE Z PRÓBY DLA BADANYCH ESTYMATORÓW

N_A	$\frac{N}{N_A}$	γ^2	n	$S(T_1)$	$S(T_2)$
1000	10	100	900	0,100	0,100
1000	10	100	1000	0,094	0,094
1000	10	100	1500	0,074	0,074
5000	2	100	100	0,140	0,139
5000	2	100	200	0,100	0,100
5000	2	100	300	0,080	0,080
100	10	100	1000	0,321	0,302
100	10	100	5000	0,101	0,101
1000	10	4	40	0,557	0,551
1000	10	4	100	0,326	0,341
5000	2	4	10	0,485	0,474
5000	2	4	50	0,200	0,208
100	100	4	300	0,652	0,663
100	100	4	400	0,566	0,560
100	100	4	500	0,490	0,491

MODYFIKACJA ESTYMATORÓW W MIKROSPISIE 1995

Dwustopniowy schemat losowania zastosowany w Mikrospisie'95 (opracowany przez Czesława Brachę, przy współpracy Andrzeja Szarkowskiego) przygotowany został pod kątem optymalizacji efektywności estymatorów dla półwojewództw (miejskich i wiejskich). Jednostki pierwszego stopnia (jps), którymi najczęściej były obwody spisowe, losowano zgodnie ze schematem Suntera (zob. [4]) z prawdopodobieństwem π_j wejścia j-tej jps do próby, przy czym π_j było proporcjonalne do umownej liczby mieszkań w j-tej jps (choć nie zawsze). Następnie, w ramach każdej jps, losowano mieszkania (jds), w taki sposób, aby każde z nich wchodziło do próby z tym samym prawdopodobieństwem f. W konsekwencji otrzymano próbę samowyważoną, dla której estymator wartości globalnej dowolnej cechy y ma postać:

$$\hat{Y} = \frac{1}{f} \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij} \tag{15}$$

gdzie:
n — liczba wylosowanych jps,
m_i — liczba jds wylosowanych z i-tej jps,
y_{ij} — wartość cechy y dla wylosowanej j-tej jds z i-tej jps, $j = 1, ..., m_i$, $i = 1, ..., n$.
Uogólnianie na półwojewództwo ma zatem niezwykle wygodną formę. Wystarczy zsumować wartości cechy wylosowane w całym półwojewództwie, a następnie przemnożyć tak otrzymaną wielkość przez mnożnik $\frac{1}{f}$.

W sposób naturalny przy opracowywaniu wyników Mikrospisu'95 pojawiła się potrzeba uogólnienia otrzymanych wyników, nie tylko na populacje brane pod uwagę przy projektowaniu schematu losowania, ale również na wielkie miasta (Warszawa, Kraków, Łódź, Poznań, Wrocław) liczące ponad 500 tys. mieszkańców. Problem ten jest szczególnym przypadkiem ogólniejszego problemu estymacji cechy w podpopulacji utworzonej po losowaniu, rozważanego w niniejszej pracy.

Stosując podejście standardowe (jak w [2]) otrzymujemy estymator postaci:

$$\hat{Y}_0 = \frac{1}{f} \sum_{i=1}^{n(A)} \sum_{j=1}^{m_i} y_{ij}$$

gdzie $n(A)$ oznacza liczbę elementów próby pochodzących z podpopulacji A. Estymator ten, ze względu na dużą długość próby, uważać można za odpowiednik estymatora T_2.

Natomiast odpowiednik estymatora T_1 ma postać

$$\widehat{Y(A)} = \frac{\Pi(A)}{n(A)f} \sum_{i=1}^{n(A)} y_i \quad y_i = \sum_{j=1}^{m_i} y_{ij} \tag{16}$$

gdzie $\Pi(A) = \sum_{i=1}^{N(A)} \pi_i$, przy czym sumowanie rozciąga się jedynie na elementy z podpopulacji A ($N(A)$ oznacza liczebność podpopulacji). Ponieważ $\frac{N}{N(A)}$ w przypadku dużych miast nie przekracza 1,5, natomiast n było rzędu kilkuset, a η również dla podstawowych cech jest większe od 1, więc reguły wcześniej podane pozwalają sądzić, że drugi z estymatorów jest efektywniejszy.

Zastosowanie wzoru (16) jako estymatora wartości globalnych cech w wielkich miastach prowadzi do formuły

$$\widehat{Y(M)} = \frac{\Pi(M)}{fn(M)} \sum_{i=1}^{n(M)} y_i$$

gdzie y_i oznacza sumę wartości cechy w wylosowanych elementach i-tej jps, a M wielkie miasto. Ostatecznie przyjęto

$$\widehat{Y(M)} = F \sum_{i=1}^{n(M)} y_i \tag{17}$$

gdzie

$$F = \left[\frac{\Pi(M)}{fn(M)} + \frac{1}{2}\right]$$

jest nowym mnożnikiem powstałym poprzez przemnożenie starego mnożnika (f^{-1} z estymatora Y) przez $\sum_{i=1}^{n(M)} \frac{\pi_i}{n(M)}$ i zaokrąglenie do najbliższej liczby całkowitej.

Konkretne obliczenia nowych mnożników dla wielkich miast obrazuje poniższa tablica:

Podpopulacje	$n(M)$	$\Pi(M)$	f^{-1}	F
Warszawa	229	257,5900	108	121
Kraków	284	276,6122	41	40
Łódź	252	253,8479	58	58
Poznań	198	198,6969	44	44
Wrocław	230	241,5514	41	43

Ostatnia kolumna tablicy zawiera wartości mnożników, które wykorzystano do estymacji dla dużych miast. Szczególnie dobre efekty wystąpiły dla cech o dużej korelacji z umowną liczbą mieszkań — parametrem, na którego podstawie obliczane były wielkości π_i.

Jak widać zmianie ulegają jedynie mnożniki dla trzech miast. Podobne rozumowanie należy zastosować do pozostałych części półwojewództw miejskich w odpowiednich trzech województwach. W rezultacie otrzymujemy nowe mnożniki dla reszt odpowiednich półwojewództw miejskich przedstawione w (ostatniej kolumnie) tablicy:

Podpopulacje	$n(M)$	$\Pi(M)$	f^{-1}	F
Warszawa	95	66,4123	108	75
Kraków	26	33,3912	41	53
Wrocław	77	65,4581	41	35

BIBLIOGRAFIA

[1] A. Nijeuhuis and H. S. Wilf: *Combinatorial Algorithms.* J. Wiley and Sons, New York, 1978
[2] W. G. Cochran: *Sampling Techniques.* J. Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1977

[3] G. Marsaglia and A. Zaman: *Towards a universal random number generator.* Statistics and Probability Letters, No 8, p. 35—39, 1990
[4] A. Sunter: *Solutions to the problem of unequal probability sampling without replacement.* International Statistical Review, No 54, p. 33—50, 1986
[5] N. J. Wilenkin: *Kombinatoryka.* PWN, Warszawa 1972

dr hab. Paweł J. Szabłowski, dr inż. Jacek Wesołowski, dr Robert Wieczorkowski — *Instytut Matematyki Politechniki Warszawskiej*

Mariusz SZAŁAŃSKI

Wybrane zagadnienia analizy szeregów czasowych

Celem artykułu jest wykazanie, iż w analizie szeregów czasowych jako funkcję trendu można efektywnie wykorzystywać specjalne funkcje elementarne o jednym stopniu swobody. Funkcje te znane są z matematyki finansowej, a reprezentują charakter wzrostu liniowy lub wykładniczy. Przedstawione zostaną także różne metody liczenia parametrów tych funkcji.

*
* *

Zbiór wartości, traktowanych jako dane ekonomiczne, może być przedstawiony w formie szeregu liczbowego lub czasowego. Szeregiem liczbowym nazwiemy zbiór wartości, zaś szeregiem czasowym może być ten sam szereg liczbowy, lecz liczą się w nim kolejności występowania poszczególnych wartości. Ogólnie przyjmuje się, że wartości w szeregu czasowym są okresowe, czyli występują co pewną jednostkę czasu (np. rok lub miesiąc). Dopuszczalne są też szeregi o nieregularnych przedziałach czasowych. Dla obu typów powyższych szeregów jedną z najważniejszych wielkości je charakteryzujących jest **wartość średnia**. Jeżeli szereg liczbowy przekształcimy w szereg czasowy, a uzyskane wartości skumulujemy, to dla takiego szeregu możemy obliczyć jeszcze jedną wielkość — **stopę wzrostu**.

Istnieje szereg wzorów służących obliczaniu wartości średniej w szeregach liczbowych.

Przytaczamy nazwy najważniejszych z nich:

— średnia arytmetyczna,
— średnia geometryczna,
— średnia harmoniczna.

Należy stwierdzić, że uzyskana z powyższych wzorów wielkość jest wartością średnią dla danego szeregu.