J. ODERFELD i E. PLESZCZYŃSKA (Warszawa)

*LINIOWY ESTYMATOR ODCHYLENIA ŚREDNIEGO
W POPULACJI NORMALNEJ*

1. Uwagi wstępne. Znaczenie praktyczne estymacji odchylenia średniego w populacji normalnej jest dostatecznie znane. Wiadomo również, że od kilkudziesięciu lat proponuje się coraz inne estymatory, oceniając ich walory nie tylko pod względem poprawności teoretycznej, ale również dogodności.

W niniejszej nocie chcemy przedstawić nowy estymator liniowy oparty na statystykach pozycyjnych.

Został on wyznaczony i zbadany metodą Monte-Carlo na maszynie cyfrowej XYZ zbudowanej przez Zakład Aparatów Matematycznych Polskiej Akademii Nauk.

2. Estymator L . Niech zmienna losowa X o rozkładzie normalnym $N(\mu, \sigma)$ będzie cechą populacji generalnej. Niech $X_1, \dots, X_n$ będą zmiennymi losowymi, z których każda ma taki sam rozkład jak X .

Definiujemy zmienne losowe

$$(1) \quad Y_k = \left| X_k - \frac{1}{n} \sum_{i=1}^n X_i \right| \quad (k = 1, 2, \dots, n).$$

Niech Z_i będzie i -tą statystyką pozycyjną w próbie $(Y_1, \dots, Y_n)$. Definiujemy wreszcie liniowy estymator odchylenia średniego σ w postaci

$$(2) \quad L_n = \sum_{i=1}^n w_i Z_i,$$

gdzie liczby w_i ($i = 1, 2, \dots, n$) są rzeczywiste i stałe; nazywamy je *wagami*. Stawiamy następujące zadanie: znaleźć taki zespół wag, żeby estymator L_n był nieobciążony i żeby jego wariancja była równa minimum ze względu na wszelkie możliwe wartości wag.

Zadanie to wystarczy rozwiązać dla populacji $N(0, 1)$, gdyż wagi nie ulegną zmianie, gdy zamiast zmiennej losowej X wprowadzimy $(X - \mu)/\sigma$.

Przyjęte warunki wyrażają się w postaci

$$(3) \quad EL_n = 1,$$

$$(4) \quad D^2 L_n = \min.$$

Ponieważ $D^2 L_n = EL_n^2 - 1$, można zastąpić (4) przez

$$(5) \quad EL_n^2 = \min.$$

W przyjętej metodzie rozwiązywania maszyna cyfrowa generowała kolejno N próbek $(x_1, \dots, x_n)$ z populacji $N(0, 1)$, obliczała wartości y_k i porządkowała je w ciągi z_k (realizacje zmiennych losowych Z_k) spełniające relacje $z_1 \leq z_2 \leq \dots \leq z_n$. Te czynności maszyny i kilka następnych opiszemy za chwilę. Na razie zwrócimy uwagę, że przy obliczaniu maszynowym odpowiednikami warunków (3) i (5) są

$$(6) \quad \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^n w_i z_{ij} = 1$$

i

$$(7) \quad \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^n w_i z_{ij} \right)^2 = \min.$$

We wzorach (6) i (7) z_{ij} oznacza realizację i -tej statystyki pozycyjnej w j -tej próbce.

Wprowadźmy oznaczenia

$$(8) \quad \frac{1}{N} \sum_{j=1}^N z_{ij} = \bar{z}_i, \quad \frac{1}{N} \sum_{j=1}^N z_{ij} z_{lj} = \overline{z_i z_l}.$$

Stosując do obliczenia wag metodę mnożnika Lagrange'a, otrzymujemy z (6) i (7) przy oznaczeniach (8) układ $(n+1)$ równań liniowych

$$(9) \quad \sum_{i=1}^n w_i \bar{z_i z_l} + \lambda z_l = 0 \quad (l = 1, 2, \dots, n),$$

$$\sum_{i=1}^n w_i \bar{z_i} - 1 = 0.$$

Wynik efektywny, który przedstawiamy w niniejszej nocie, dotyczył przypadku $n = 5$. Wynik ten opiera się na $N = 500$ próbkach.

Program maszyny cyfrowej składał się z dwóch części: generowania liczb losowych z populacji $N(0, 1)$ i programu specjalnego związanego z układem (9).

Działanie generatora opierało się na centralnym twierdzeniu granicznym: maszyna generowała 16 liczb losowych o rozkładzie praktycznie równomiernym (odpowiedni algorytm można znaleźć np. w [1]), a następnie obliczała i normowała średnią arytmetyczną. Normalność otrzymanych liczb sprawdzono w znany sposób testem χ^2 .

Program specjalny obejmował generowanie próbek 5-elementowych, obliczanie wartości y_k i porządkowanie ich w ciągi z_k , obliczanie $\bar{z}_i$ oraz

$z_i z_i$ i rozwiązanie układu (9), przy czym ostatnia operacja odbywała się według standardowego podprogramu rozwiązywania algebraicznych równań liniowych.

Ostatecznie dla $n = 5$ otrzymano następujący estymator

$$(10) \quad L_5 = -0,065Z_1 + 0,150Z_2 + 0,175Z_3 + 0,312Z_4 + 0,405Z_5.$$

Praca maszyny trwała około 45 minut.

3. Własności estymatora L_5 . Własności estymatora zbadaliśmy również za pomocą maszyny cyfrowej. W tym celu utworzono 1000 próbek 5-elementowych z populacji $N(0, 1)$; oczywiście były to próbki inne niż te, których użyto do obliczenia wag.

Dla każdej próbki maszyna obliczała wartość l_5 zmiennej losowej L_5 , stopniowo budowała histogram i obliczała EL_5 oraz D^2L_5 . Jednocześnie dla porównania prowadzono analogiczne obliczenia dla estymatora

$$S_5 = \sqrt{\frac{5}{2}} \cdot \frac{\Gamma_2}{\Gamma_2^5} \sqrt{\frac{1}{5} \sum_{i=1}^5 Z_i^2}$$

i obliczano przeciętną wartość $\overline{l_5 s_5}$ potrzebną do wyznaczenia korelacji. Praca maszyny trwała około 30 minut.

Estymator S_5 wybraliśmy do porównania dlatego, że stanowi nieobciążony wariant (por. np. [2]) estymatora klasycznego $\sqrt{\frac{1}{5} \sum_{i=1}^5 (X_i - \bar{X})^2}$ i że ze wszystkich znanych estymatorów nieobciążonych odchylenia średniego, opartych na wartościach z próbki z populacji $N(0, 1)$, ma najmniejszą wariancję 0,1318. Przy sposobności przypominamy, że najmniejsza możliwa wartość regularnego i nieobciążonego (zresztą niezanego) estymatora parametru σ jest $1/2n$, czyli w naszym przypadku $1/10$.

Ostatecznie, otrzymaliśmy wyniki spisane w tabelicy 1.

TABLICA 1

Wynik badania

Estymator	L_5		S_5	
	empir.	teoret.	empir.	teoret.
Wartość oczekiwana	1,024	1	1,006	1
Wariancja	0,145	?	0,140	0,1318

Jak widzimy, zgodność empirycznej wartości EL_5 z wartością teoretyczną (ściślej — założoną) jest bardzo dobra. Równie dobra jest zgodność empirycznych wartości EL_5 i ES_5 .

Teoretycznej wariancji estymatora L_5 nie znamy, przypuszczamy jednak, że nie wiele różni się od teoretycznej wariancji estymatora S_5 . Przemawia za tym porównanie empirycznych wariancji: są one praktycz-

Rys. 1. Histogramy L_5 i S_5 .

nie równe, z drobną różnicą na niekorzyść estymatora liniowego. Dla porządku notujemy, że empiryczna wariancja estymatora S_5 okazała się nieco większa od teoretycznej, co przypisujemy przypadkowi w generowaniu liczb losowych.

O podobnych rozkładach estymatorów L_5 i S_5 świadczą podobne histogramy pokazane na rysunku 1. Zresztą empiryczny współczynnik korelacji obu estymatorów jest bardzo wysoki: wynosi 0,88.

4. Zastosowania. Oszacowanie odchylenia średniego przy użyciu wzoru (10) jest bardzo proste.

Przypuśćmy, że w próbie 5-elementowej, o której zakłada się, że pochodzi z populacji normalnej, otrzymano wartości x_i podane w pierwszej kolumnie tablicy 2.

TABLICA 2

Schemat stosowania wzoru (10)

(1)	(2)	(3)	
x_i	y_i	z_i	wagi
44	7	$1 \times -0,065 =$	$-0,065$
38	1	$5 \times 0,150 =$	$0,750$
32	5	$5 \times 0,175 =$	$0,875$
42	5	$7 \times 0,312 =$	$2,184$
29	8	$8 \times 0,405 =$	$3,240$
185			$\sigma = 6,984$
$\bar{x} = 37$			

Po obliczeniu $\bar{x} = 37$ znajdujemy w kolumnie drugiej wartości y_i jako $|x_i - 37|$. Po ustawieniu tych liczb w kolejności nie malejącej znajdujemy wartości z_i spisane w kolumnie trzeciej. Reszta obliczeń nie wymaga wyjaśnień. Znajdujemy oszacowanie 6,984 odchylenia średniego w populacji generalnej. Estymator S_5 daje w tym przypadku wartość 6,814.

Liniowa postać wzoru (10) nastęrcza wiele możliwości częściowego lub nawet całkowitego zautomatyzowania obliczeń.

Jedną z nich jest przygotowanie tablicy iloczynów $w_i z_i$, na przykład w układzie następującym:

TABLICA 3

Iloczyny $w_i z_i$

z_i	$-0,065 z_1$	$0,150 z_2$	$0,175 z_3$	$0,312 z_4$	$0,405 z_5$
0	0	0	0	0	0
1	-0,065	0,150	0,175	0,312	0,405
2	-0,130	0,300	0,350	0,624	0,810
3	-0,195	0,450	0,525	0,936	1,215
$\vdots$					
99	-6,435	14,850	17,325	30,888	40,095
100	-6,500	15,000	17,500	31,200	40,500

Zastosowanie tablicy jest oczywiste.

Można również zbudować bardzo prosty nomogram zastępujący wiernie wzór (10). Nomogram zyska jeszcze na prostocie, gdy się uwzględni, że z_1 jest nie większe od z_2, z_3, z_4, z_5 i że waga w_1 jest bardzo mała.

Pomijając z_1 wprowadzamy błąd systematyczny nie większy od 1,3%; przeciętnie wynosi on kilka dziesiątych procenta, to znaczy mniej

niż przypadkowy błąd średni spowodowany wykreśleniem i odczytaniem nomogramu. Można więc zadowolić się bardzo prostym nomogramem pokazanym na rysunku 2. Sposób posługiwania się wynika z narysowanego przykładu.

Rys. 2. Schemat nomogramu. Przykład: $z_2 = 5$, $z_3 = 5$, $z_4 = 7$, $z_5 = 8$, $\sigma = 7,0$

Dalszym krokiem w kierunku automatyzacji byłaby realizacja wzoru (10) za pomocą sumatora dźwigienkowego lub zębatego, a nawet opartego na zasadzie ważenia. Projekty takich urządzeń wykraczają poza zakres niniejszej noty.

5. Zakończenie. Celem niniejszej noty było przedstawienie nowego estymatora L_5 . Oczywiście podaną metodą można wyznaczyć podobne estymatory dla innych licznosci niż 5; w każdym razie warto to zrobić w zakresie $3 \leq n \leq 10$.

Ponadto chcieliśmy na prostym przykładzie pokazać możliwości maszyn cyfrowych w zakresie probabilistyki.

Koncepcja estymatora powstała w Instytucie Matematycznym PAN, a obliczenia przeprowadziliśmy na maszynie XYZ w Biurze Obliczeń i Programów Zakładu Produkcji Doświadczalnej Maszyn Matematycz-

nych w Warszawie. W archiwum Biura zachowano jako program pokazowy generowanie liczb normalnych i budowę histogramów L_5 i S_5 , których powstawanie można obserwować w toku obliczania na ekranie oscyloskopu.

Prace cytowane

- [1] H. A. Meyer, *Symposium on Monte-Carlo method*, New York, London 1954, str. 17-19.
 [2] M. G. Kendall, *The advanced theory of statistics*, London 1948.

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK
 BIURO OBLICZEŃ I PROGRAMÓW ZPDMM

Praca wpłynęła 25. 6. 1960

ЯН ОДЕРФЕЛЬД и Е. ПЛЕШИНСКАЯ (Варшава)

ЛИНЕЙНАЯ ОЦЕНКА СРЕДНЕГО ОТКЛОНЕНИЯ В НОРМАЛЬНОЙ СОВОКУПНОСТИ

РЕЗЮМЕ

На цифровой машине определена методом Монте Карло линейная оценка (10), основанная на позиционных статистиках, среднего отклонения нормальной совокупности. Оценка эта несмещена и удовлетворяет условию наименьшей дисперсии в своем классе.

Тоже на машине определены распределение оценки и первый и второй центральные моменты.

Оценка удобна для приложений, особенно с применением номограммы (черт. 2).

J. ODERFELD and E. PLESZCZYŃSKA (Warszawa)

A LINEAR ESTIMATE OF THE MEAN DEVIATION IN NORMAL POPULATION

SUMMARY

The linear estimate (10) of mean deviation from the normal population has been determined on an electronic digital computer by using the Monte-Carlo method. The estimate obtained is unbiased and satisfies the condition of minimum variance in its class.

The distribution and the first and second central moments of the estimate also have been determined by using the electronic digital computer.

The estimate is convenient for usage, particularly when applying nomogram (Fig. 2).