kpt. dr inż. Mariusz BODJAŃSKI
Wojskowy Instytut Techniczny Uzbrojenia

ESTYMATORY WARIANCJI – DOWÓD NA OBCIĄŻENIE

Streszczenie: W artykule zdefiniowano podstawowe pojęcia z zakresu analizy statystycznej. Wyjaśniono pojęcia populacji, próby n-elementowej, realizacji, zmiennej losowej, rozkładu zmiennej losowej, parametrów rozkładu, statystyki z próby, estymacji punktowej oraz estymatora. Opisano własności wartości oczekiwanej i wariancji – podstawowych parametrów rozkładu zmiennych losowych. Przedstawiono wzory estymatorów wartości oczekiwanej i wariancji. Przeprowadzono dowód na obciążenie estymatora wariancji w przypadku braku wiedzy o wartości oczekiwanej. Wyprowadzono wzór na nieobciążony estymator wariancji.

1. Wstęp

W statystyce matematycznej rozważa się pewien zbiór nazywany populacją oraz pewną cechę X tej populacji. W praktyce przebadanie wszystkich elementów populacji w celu określenia badanej cechy nie zawsze jest możliwe. W takim przypadku o wartości cechy wnioskujemy na podstawie określonej części populacji zwanej próbą n-elementową. Realizację próby $(X_1,X_2,...,X_n)$ można potraktować jak ciąg zmiennych losowych niezaleznych o rozkładzie takim samym jak rozkład cechy X. Ponieważ poszczególne wartości elementów populacji nie są jednolite, modelowanie cechy X wygodnie jest przeprowadzać za pomocą zmiennej losowej X o odpowiednim rozkładzie i parametrach tego rozkładu oraz reguł rachunku prawdopodobieństwa. W tym celu należy zdefiniować zmienną losową $U_n=f(X_1,X_2,...,X_n)$ czyli funkcję próby zwaną statystyką z próby lub estymatorem. Oszacowanie wartości parametru Q rozkładu zmiennej losowej X (badanej cechy) sprowadza się do zdefiniowania odpowiedniej statystyki U_n i obliczenia jej wartości [1]. W artykule przeprowadzono dowód na obciążenie niewłaściwie zdefiniowanego estymatora wariancji oraz wyprowadzono wzór na nieobciążony estymator wariancji.

2. Parametry rozkładu zmiennych losowych

Wartość oczekiwaną zmiennej losowej X oznaczamy

$$E(X)=m \tag{1}$$

Własności wartości oczekiwanej [2]:

$$E(aX)=aE(X) \tag{2}$$

$$E(X+Y)=E(X)+E(Y)^{*} \tag{3}$$

Wariancję zmiennej losowej X oznaczamy

$$D^2(X)=E\left((X-m)^2\right)=\sigma^2 \tag{4}$$

Własności wariancji [2]:

$$D^2(aX)=a^2D^2(X) \tag{5}$$

[*] jeśli istnieją wartości oczekiwane zmiennych losowych X i Y

$$E(X+Y)=E(X)+E(Y)^{**} \tag{6}$$

3. Estymatory parametrów rozkładu

Estymatorem parametru Q nazywamy statystykę U_n z próby n-elementowej $(X_1,X_2,...,X_n)$. Estymator U_n nazywamy [1]:

- obciążonym, gdy jego wartość oczekiwana jest różna od prawdziwej wartości parametru Q

$$E(U_n)\neq Q \tag{7}$$

- nieobciążonym, gdy jego wartość oczekiwana jest prawdziwą wartością parametru Q

$$E(U_n)=Q \tag{8}$$

Estymator wartości oczekiwanej

$$\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k \tag{9}$$

Nieobciążony estymator wariancji

$$S_n^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2 \tag{10}$$

4. Obciążony estymator wariancji (dowód na obciążenie)

Dla przejrzystości wywodu w dalszej części artykułu pominięto indeks n przy oznaczeniach estymatorów.

Nieobciążony estymator wariancji (znamy wartość oczekiwaną m)

$$S^2=\frac{1}{n}\sum_{k=1}^{n}(X_k-m)^2 \tag{11}$$

Obciążony estymator wariancji (nie znamy wartości oczekiwanej m, zastępujemy ją estymatorem $\overline{X}$)

$$S_*^2=\frac{1}{n}\sum_{k=1}^{n}(X_k-\overline{X})^2 \tag{12}$$

Rozwijając równanie (12) otrzymujemy

$$\begin{aligned}
S_*^2 &= \frac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X}+\underbrace{m-m}_{0}\right)^2 = \\
&= \frac{1}{n}\sum_{k=1}^{n}\left((X_k-m)-(\overline{X}-m)\right)^2 = \\
&= \frac{1}{n}\sum_{k=1}^{n}\left((X_k-m)^2+(\overline{X}-m)^2-2(X_k-m)(\overline{X}-m)\right)= \\
&= \underbrace{\frac{1}{n}\sum_{k=1}^{n}(X_k-m)^2}_{S^2}+\frac{1}{n}n(\overline{X}-m)^2-2(\overline{X}-m)\frac{1}{n}\sum_{k=1}^{n}(X_k-m)=
\end{aligned}$$

** jeśli istnieją wartości oczekiwane zmiennych losowych X i Y oraz zmienne te są niezależne

$$
\begin{aligned}
&= S^2 + \left(\overline{X} - m\right)^2 - 2\left(\overline{X} - m\right)\left(\underbrace{\frac{1}{n}\sum_{k=1}^{n} X_k}_{\overline{X}} - \frac{1}{n} n m\right) = \\
&= S^2 + \left(\overline{X} - m\right)^2 - \underbrace{2\left(\overline{X} - m\right)\left(\overline{X} - m\right)}_{2\left(\overline{X}-m\right)^2} = \\
&= S^2 - \left(\overline{X} - m\right)^2
\end{aligned}
\tag{13}
$$

Chcąc dowieść, że estymator (13) jest obciążony należy zbadać jego wartość oczekiwaną

$$E\left(S_*^2\right) = E\left(S^2 - \left(\overline{X} - m\right)^2\right) \tag{14}$$

Na podstawie (3) możemy zapisać

$$E\left(S_*^2\right) = E\left(S^2\right) - E\left(\left(\overline{X} - m\right)^2\right) \tag{15}$$

Pierwszy czynnik po prawej stronie równania (15) to wartość oczekiwana nieobciążonego estymatora wariancji

$$E\left(S^2\right) = \sigma^2 \tag{16}$$

W celu rozwinięcia drugiego czynnika po prawej stronie równania (15) wprowadźmy nową zmienną losową

$$Y = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{17}$$

Zakładamy, że ciąg X_k jest ciągiem zmiennych losowych niezależnych o jednakowym rozkładzie i parametrach m oraz σ^2. Założenie to jest prawdziwe, ponieważ ciąg X_k jest realizacją próby $(X_1, X_2, ..., X_n)$.

Wartość oczekiwana zmiennej losowej Y

$$E\left(Y\right) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \tag{18}$$

Na podstawie własności wartości oczekiwanej (2) i (3) możemy rozwinąć równanie (18) do następującej postaci

$$
\begin{aligned}
E\left(Y\right) &= E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \\
&= \frac{1}{n} E\left(\sum_{k=1}^{n} X_k\right) = \\
&= \frac{1}{n}\sum_{k=1}^{n} E\left(X_k\right) = \\
&= \frac{1}{n} n m = m
\end{aligned}
\tag{19}
$$

Wariancja zmiennej losowej Y

$$D^2\left(Y\right) = D^2\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) \tag{20}$$

Na podstawie własności wariancji (5) i (6) możemy rozwinąć równanie (20) do następującej postaci

$$\begin{aligned} D^2(Y) &= D^2\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \\ &= \frac{1}{n^2} D^2\left(\sum_{k=1}^{n} X_k\right) = \\ &= \frac{1}{n^2}\sum_{k=1}^{n} D^2(X_k) = \\ &= \frac{1}{n^2} n\sigma^2 = \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{21}$$

Należy zauważyć, że zdefiniowana równaniem (17) zmienna losowa Y jest estymatorem wartości oczekiwanej $\overline{X}$. Ponadto biorąc pod uwagę definicję wariancji oraz równania (19) i (21), drugi czynnik po prawej stronie równania (15) możemy rozwinąć do następującej postaci

$$\begin{aligned} E\left((\overline{X} - m)^2\right) &= E\left((Y - m)^2\right) = \\ &= D^2(Y) = \frac{\sigma^2}{n} \end{aligned} \tag{22}$$

Jest to wariancja średniej arytmetycznej (zastosowanie prawa wielkich liczb Chinczyna [3]). Podstawiając równania (16) i (22) do równania (15), wartość oczekiwana obciążonego estymatora wariancji przyjmuje postać

$$E\left(S_*^2\right) = \sigma^2 - \underbrace{\left(\frac{\sigma^2}{n}\right)}_{bias} \tag{23}$$

Wartość oczekiwana obciążonego estymatora wariancji jest różna od prawdziwej wartości wariancji

$$E\left(S_*^2\right) \neq \sigma^2 \quad \text{c.n.d.} \tag{24}$$

5. Nieobciążony estymator wariancji (nie znamy wartości oczekiwanej *m*)

Przekształcając równanie (23) otrzymujemy

$$\begin{aligned} E\left(S_*^2\right) &= \sigma^2 - \frac{\sigma^2}{n} = \\ &= \frac{n-1}{n}\sigma^2 \end{aligned} \tag{25}$$

Mnożąc obie strony równania (25) przez czynnik

$$\frac{n}{n-1} \tag{26}$$

oraz uwzględniając własność wartości oczekiwanej (2) otrzymujemy

$$\frac{n}{n-1}E\left(S_*^2\right)=E\left(\frac{n}{n-1}S_*^2\right)= \\ =\frac{n}{n-1}\left(\frac{n-1}{n}\sigma^2\right)=\sigma^2 \tag{27}$$

Z definicji estymatora nieobciążonego (8) i równania (27) wynika, że mnożąc obciążony estymator wariancji (12) przez czynnik (26) otrzymujemy nieobciążony estymator wariancji

$$S^2=\frac{n}{n-1}S_*^2= \\ =\frac{n}{n-1}\left(\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2\right)= \\ =\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2 \tag{28}$$

6. Podsumowanie

Estymacja punktowa parametru Q (parametru rozkładu cechy populacji X) polega na:

- wybraniu odpowiedniej statystyki U_n,
- obliczeniu wartości statystyki U_n na podstawie próby n-elementowej.

Statystykę U_n nazywamy wówczas estymatorem parametru Q.

Nieobciążony estymator wariancji zdefiniowany jest zależnością (10), gdy wartość oczekiwana cechy X nie jest znana. Wartość oczekiwaną m zastępujemy wówczas jej estymatorem (9). W przypadku, gdy znamy wartość oczekiwaną m nieobciążony estymator wariancji definiujemy za pomocą zależności (11).

Dowiedziono, że estymator (12) jest obciążony to znaczy, że jego wartość oczekiwana nie jest równa prawdziwej wartości parametru Q.

Dowiedziono również, że mnożąc obciążony estymator wariancji (12) przez czynnik (26) otrzymujemy nieobciążony estymator wariancji.

7. Literatura

[1] Leitner R., Zacharski J.: *Zarys matematyki wyższej dla studentów część III*, 289-295, Wydawnictwo Naukowo-Technichniczne, Warszawa 1995.

[2] Leitner R., Zacharski J.: *Zarys matematyki wyższej dla studentów część III*, 239-246, Wydawnictwo Naukowo-Technichniczne, Warszawa 1995.

[3] Leitner R., Zacharski J.: *Zarys matematyki wyższej dla studentów część III*, 281-282, Wydawnictwo Naukowo-Technichniczne, Warszawa 1995.

VARIANCE ESTIMATORS – PROOF ON BIAS

There were defined basic concepts of statistical analysis in the article. Concept of population, n-element sample, realization, random variable, distribution of random variable, parameters of distribution, function of sample, point estimation and astimator were explain. Properties of

basic parameters of random variable distribution – expected value and variance were described. Formula of expected value estimator and variance estimator was presented. Proof on biased variance estimator in case of lack of expected value was done. Formula of unbiased variance estimator was conducted.