Metody probabilistyczne
Rozwiązania zadań

13. Statystyka

16.01.2020

Zadanie 1. *Uzasadnij, że zachodzi wzór skróconego mnożenia:*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2$$

Odpowiedź: Ten wzór jest tak naprawdę szczególnym przypadkiem wzoru skróconego mnożenia dla wariancji:

$$E\left((X - EX)^2\right) = E(X^2) - (EX)^2,$$

gdzie za rozkład prawdopodobieństwa weźmiemy rozkład empiryczny na próbie, tzn. każdemu elementowi próby przypiszemy tę samą wartość prawdopodobieństwa $\frac{1}{n}$. Udowodnimy jednak te wzór bezpośrednio. Mamy:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;-\; 2\frac{1}{n}\sum_{i=1}^{n} X_i\overline{X}_n \;+\; \frac{1}{n}\sum_{i=1}^{n}\overline{X}_n^2.$$

Ponieważ $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ jest stałą niezależną od i, można ją wyjąć przed sumę, co daje:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \;&=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;-\; 2\overline{X}_n \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{=\overline{X}_n} \;+\; \overline{X}_n^2 \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1}_{=1} \\ &=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;-\; 2\overline{X}_n^2 \;+\; \overline{X}_n^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 \;-\; (\overline{X}_n)^2.\end{aligned}$$

Zadanie 2. *Pokaż, że estymator $\widehat{\sigma}^2$ wariancji $\sigma^2 = D^2(X)$ zdefiniowany jako:*

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

jest silnie zgodny.

Odpowiedź: Zgodnie z prawem wielkich liczb,

$$\overline{X}_n \;\overset{\text{z pr. 1}}{\to}\; EX$$

Rozważmy zmienną losowe $Y = X^2$. Zdefiniujmy:

$$\overline{Y}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Zgodnie z prawem wielkich liczb:

$$\overline{Y}_n \;\overset{\text{z pr. 1}}{\to}\; EY \;=\; E(X^2)$$

Czyli:

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2 \;=\; \overline{Y}_n - (\overline{X}_n)^2 \;\overset{\text{z pr. 1}}{\to}\; E(X^2) - (EX)^2 \;=\; \sigma^2,$$

gdzie w ostatniej równości użyliśmy wzoru skróconego mnożenia dla wariancji.

Zadanie 3. *Niech* $\widehat{\mu} = \sum_{i=1}^{n} c_i X_i$ *będzie estymatorem wartości oczekiwanej* $\mu = EX$ *dla pewnych współczynników* $c_1, \ldots, c_n$ *niezależnych od danych. Jaki warunek muszą spełniać te stałe, aby estymator był* nieobciążony*?*

Odpowiedź: Stałe muszą się sumować do jedynki: $\sum_{i=1}^{n} c_i = 1$. Wynika to z tego, że:

$$E\widehat{\mu} = E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i \underbrace{EX_i}_{=\mu} = \mu \sum_{i=1}^{n} c_i.$$

Jeśli $\widehat{\mu}$ ma być nieobciążony, to $E\widehat{\mu} = \mu$, a więc $\sum_{i=1}^{n} c_i = 1$.

Zadanie 4. *Rozważ estymator wartości oczekiwanej* μ *postaci:*

$$\widehat{\mu} = \sum_{i=1}^{n} c_i X_i,$$

dla pewnych współczynników $c_1, \ldots, c_n$ *(niezależnych od danych). Załóżmy, że* $\widehat{\mu}$ *jest* nieobciążony *(patrz zadanie 3). Dla jakich wartości* $c_1, \ldots, c_n$ *ma on* najmniejszą wariancję*?*

Odpowiedź: Z poprzedniego zadania wiemy, że jeśli $\widehat{\mu}$ ma być nieobciążonym estymatorem wartości oczekiwanej μ, to musi zachodzić $\sum_{i=1}^{n} c_i = 1$. Policzmy teraz wariancję estymatora. Ponieważ $X_1, \ldots, X_n$ są niezależne, tym samym $c_1 X_1, \ldots, c_n X_n$ są niezależne, a więc:

$$D^2(\widehat{\mu}) = \sum_{i=1}^{n} D^2(c_i X_i) = \sum_{i=1}^{n} c_i^2 \underbrace{D^2(X_i)}_{=\sigma^2} = \sigma^2 \sum_{i=1}^{n} c_i^2,$$

gdzie użyliśmy również prawa skalowania wariancji $D^2(aX) = a^2 D^2(X)$. A więc aby zminimalizować wariancję estymatora, należy zminimalizować $\sum_{i=1}^{n} c_i^2$ przy założeniu, że $\sum_{i=1}^{n} c_i = 1$. Zrobimy to w następujący sposób: za c_n podstawimy $1 - \sum_{i=1}^{n-1} c_i$ i w ten sposób pozbędziemy się jednej zmiennej oraz ograniczenia. Czyli musimy rozwiązać problem minimalizacji funkcji:

$$f(c_1, \ldots, c_{n-1}) = \sum_{i=1}^{n-1} c_i^2 + \left(1 - \sum_{i=1}^{n-1} c_i\right)^2.$$

Liczmy pochodne cząstkowe po c_i:

$$\frac{\partial f}{\partial c_i} = 2c_i - 2\left(1 - \sum_{i=1}^{n-1} c_i\right),$$

i przyrównujemy je do zera:

$$\frac{\partial f}{\partial c_i} = 0 \quad \iff \quad c_i = 1 - \sum_{i=1}^{n-1} c_i.$$

Ponieważ prawa strona powyższego wyrażenia jest *taka sama* dla dowolnego i, wnioskujemy, że w optimum wszystkie wartości c_i są sobie równe, tzn. $c_1 = c_2 = \ldots = c_{n-1}$, co po podstawieniu do powyższego wyrażenia daje:

$$c_i = 1 - (n-1)c_i \quad \Longrightarrow \quad c_i = \frac{1}{n}.$$

Wtedy również $c_n = \frac{1}{n}$, a więc wszystkie współczynniki są równe $\frac{1}{n}$. Tym samym pokazaliśmy, że estymator nieobciążony ma najmniejszą wariancję, gdy jest zwykłą średnią arytmetyczną.

Uwaga: powinniśmy formalnie argumentować, że znalezione optimum to minimum, a nie maksimum. Ponieważ jest to jedyny punkt zerowania się pochodnej, funkcja nie ma więcej minimów ani maksimów. Aby przekonać się, że jest to minimum, wystarczy zauważyć, że f rośnie do nieskończoności gdy któryś ze współczynników rośnie do nieskończoności, więc znaleziony punkt nie może być maksimum.

Zadanie 5. *Pokaż, że funkcja informacji Fishera $I(p)$ dla rozkładu dwupunktowego $B(p)$ ma postać:*

$$I(p) \;=\; \frac{1}{p(1-p)}$$

Następnie pokaż, że estymator $\overline{X}_n$ wartości oczekiwanej p jest efektywny dla tego rozkładu

Odpowiedź: W rozkładzie dwupunktowym $X \in \{0,1\}$, zapisując $q(x) = P(X = x)$ (zmieniamy oznaczenie rozkładu prawdopodobieństwa, żeby nie myliło się z parametrem p) mamy $q(1) = p$ i $q(0) = 1 - p$. Tym samym:

$$\frac{\partial \ln q(x)}{\partial p} \;=\; \begin{cases} \frac{\partial \ln p}{\partial p} \;=\; \frac{1}{p} & \text{dla } x = 1, \\ \frac{\partial \ln(1-p)}{\partial p} \;=\; -\frac{1}{1-p} & \text{dla } x = 0. \end{cases}$$

Funkcja informacji Fishera ma więc postać:

$$\begin{aligned} I(p) \;&=\; E\left(\left(\frac{\partial \ln q(X)}{\partial p}\right)^2\right) \;=\; q(1)\left(\frac{\partial \ln q(1)}{\partial p}\right)^2 + q(0)\left(\frac{\partial \ln q(0)}{\partial p}\right)^2 \\ &=\; p\frac{1}{p^2} + (1-p)\frac{1}{(1-p)^2} \;=\; \frac{1}{p} + \frac{1}{1-p} \;=\; \frac{1-p+p}{p(1-p)} \;=\; \frac{1}{p(1-p)} \end{aligned}$$

Estymator $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ ma wartość oczekiwaną i wariancję równą:

$$E\overline{X}_n \;=\; EX \;=\; p, \qquad D^2(\overline{X}_n) \;=\; \frac{D^2(X)}{n} \;=\; \frac{p(1-p)}{n}.$$

A więc $\overline{X}_n$ jest estymatorem nieobciążonym parametru p. Z kolei z nierówności Craméra-Rao wynika, że dla dowolnego estymatora nieobciążonego $\widehat{\mu}$ parametru p mamy:

$$D^2(\widehat{\mu}) \;\geqslant\; \frac{1}{nI(\theta)} \;=\; \frac{p(1-p)}{n}.$$

A więc $\overline{X}_n$ jest efektywnym estymatorem parametru p dla rozkładu dwupunktowego.