# Zestaw 2

## Statystyczna Analiza Danych

Emilia Wiśnios

24 marca 2021

1. Niech $X$ i $Y$ zmienne losowe, $E[X] = 1$, $E[Y] = 3$, $Var[X] = Var[Y] = \sigma^2$. Dla jakiej stałej $c$ statystyka $cX^2 + (1-c)Y^2$ jest nieobciążonym estymatorem parametru $\sigma^2$?

   **Rozwiązanie:**

   $T(X_1, ..., X_n)$ jest **estymatorem nieobciążonym** $\theta \Leftrightarrow ET = \theta$

$$E[cX^2 + (1-c)Y^2] = cEX^2 + (1-c)EY^2 = \sigma^2 + 9 - 8c$$

$$EX^2 = VarX + (EX)^2 = \sigma^2 + 1$$

$$EY^2 = \sigma^2 + 9$$

   Zatem

$$9 - 8c = 0 \Leftrightarrow c = \frac{9}{8}$$

2. Liczba wypadków samochodowych zgłoszonych do towarzystwa ubezpieczeniowego w $k$-tym miesiącu jest zmienną losową $W_k$ o rozkładzie Poissona z parametrem $\lambda_{z_k}$, gdzie $z_k$ jest liczbą samochodów zgłoszonych do ubezpieczenia w tym miesiącu, zaś $\lambda$ jest nieznanym parametrem. Zmienne losowe $W_k$ są niezależne. Wyznaczyć estymator największej wiarogodności parametru $\lambda$ na podstawie próby $W_1, ..., W_{12}$.

   **Rozwiązanie:**

   **Estymator największej wiarygodności**
   Wiarygodność:

$$X \sim p_\theta \quad \text{gęstość/ funkcja prawdopodobieństwa}$$

   gdzie $X$ - obserwacje.

$$L(\theta) = p_\theta(X)$$

$$\hat{\theta}_{MLE} = \arg\max L(\theta)$$

$$L(\lambda) = P_\lambda(W_1, ..., W_{12}) = \prod_{i=1}^{12} P_\lambda(W_i) = \prod \frac{(\lambda z_i)^{W_i}}{W_i!} e^{-\lambda z_i} = \lambda^{\sum w_i} e^{-\lambda \sum z_i} \prod \frac{z_i^{W_i}}{W_i!}$$

$$l(\lambda) = \log L(\lambda) = \sum W_i \log \lambda - \lambda \sum z_i + \log \left( \prod \frac{z_i^{W_i}}{W_i!} \right)$$

$$l'(\lambda) = \frac{\sum W_i}{\lambda} - \sum z_i = 0 \quad \Leftrightarrow \quad \lambda = \frac{\sum W_i}{\sum z_i}$$

Pochodna jest funkcją malejącą, zatem to jest maksimum.

3. **Estymacja przedziałowa**

Inwestor chce oszacować ryzyko przedsięwzięcia, które przynosi losowy zysk o rozkładzie normalnym o nieznanych parametrach. Ryzyko: odchylenie standardowe zysku. Po obliczeniu średniej i wariancji próby prostej złożonej z $n = 17$ zysków z przeszłości, otrzymano:

$$\overline{X}_n = 1500EUR$$

$$\hat{S}_n^2 = 6456EUR^2$$

podać przedział ufności dla a) oczekiwanego zysku i b) ryzyka, na poziomie ufności 0.99.

*Wskazówka:* $t(0.995, 16) = 2.921, \chi^2(0.995, 16), \chi^2(0.005, 16) = 5.142.$

**Rozwiązanie:**

$X_1, ..., X_n \sim N(\mu, \sigma^2), \quad \mu, \sigma^2$ nieznane

(a)

$$\mu \in \left[ \overline{X} \pm \frac{t_{(1-\frac{\alpha}{2}, n-1)} S_n}{\sqrt{n-1}} \right]$$

$$S_n^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$$

Zatem

$$\mu \in \left[ 1500 \pm \frac{2.92 \cdot 254}{4} \right]$$

(b)

$$\sigma^2 \in \left[ \frac{nS_n^2}{\chi^2(1-\frac{\alpha}{2}, n-1)}, \frac{nS_n^2}{\chi^2(\frac{\alpha}{2}, n-1)} \right]$$

$$\sigma \in \left[ \frac{\sqrt{n-1} S_n}{\sqrt{\chi^2(1-\frac{\alpha}{2}, n-1)}}, \frac{\sqrt{n-1} S_n}{\sqrt{\chi^2(\frac{\alpha}{2}, n-1)}} \right]$$

$$\sigma \in \left[ \frac{4 \cdot 254}{\sqrt{34.267}}, \frac{4 \cdot 254}{\sqrt{5.142}} \right]$$

4. Wyznaczyć metodą największej wiarogodności estymator parametru $p$ w rozkładzie geometrycznym

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

na podstawie $n$-elementowej próby prostej $X_1, ..., X_n$, pochodzącej z tego rozkładu.

**Rozwiązanie:**

$X_1, ..., X_n, \quad \theta = p$

$$L(p) = P(X_1, ..., X_n) = \prod_{i=1}^n P(X_i) = \prod_{i=1}^n p(1-p)^{X_i - 1} = p^n (1-p)^{\sum X_i - n}$$

$$l(p) = \log L(p) = n \log(p) + (\sum X_i - n) \log(1-p)$$

$$l'(p) = \frac{n}{p} - \frac{\sum X_i - n}{1-p} = 0$$

$$\frac{n}{p} = \frac{\sum X_i - n}{1-p} \quad \Leftrightarrow \quad p = \frac{n}{\sum X_i}$$

5. **MLE dla rozkładu wielomianowego**

Podaj estymatory największej wiarogodności dla rozkładu wielomianowego Multinomial($n, p_1, ..., p_k$. Zmienna $X$ jest $k$-wymiarową zmienną, zliczającą ile wypadło wyników $i$-tego rodzaju w $n$ próbach, każdy z prawdopodobieństwem $p_1, ..., p_k$, gdzie $\sum_i p_i = 1$. Czyli $X = [n_1, ..., n_k]^T$ oznacza, że w $n$ próbach było $n_i$ wyników typu $i, i \in 1, ..., k$, przy czym $\sum_i^k n_i = n$ oraz $n_i = 0, 1, ..., n$. Mamy

$$P_X([n_1, ..., n_k]^T) = \binom{n}{n_1 ... n_k} p_1^{n_1} p_2^{n_2} ... p_r^{n_k},$$

gdzie $\binom{n}{n_1 ... n_k} = n!/(n_1! n_2! ... n_k!)$.

*Wskazówka:* Aby wyprowadzić warunek, że $\sum_i p_i = 1$, użyj mnożnika Lagrange'a, dodając do log wiarygodności wyrażenie $\lambda(\sum_{i=1}^k p_i - 1)$. Następnie zmaksymalizuj log wiarygodności ze względu na $p_i$, dla $i = 1, ..., k$, a także na $\lambda$.

**Rozwiązanie:**

$$L(p_1, ..., p_k) = \binom{n}{n_1 \, ... \, n_k} \prod_{i=1}^{k} p_i^{n_i}$$

$$L(p) = \log L(p_1, ..., p_k) = \log \binom{n}{n_1 \, ... \, n_k} + \sum n_i \log p_i$$

$$\max_{\sum p_i = 1} \; l(p)$$

$$\mathcal{L}(p, \lambda) = l(p) + \lambda(\sum p_i = 1)$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = 0 \quad \Rightarrow \quad \frac{n_i}{p_i} + \lambda = 0$$

$$p_i = -\frac{n_i}{\lambda}$$

Ale mamy warunek, że $\sum p_i = 1$ zatem

$$-\sum \frac{n_i}{\lambda} = 1 \quad \Rightarrow \quad \lambda = -\sum n_i$$

$$p_i = \frac{n_i}{\sum n_i}$$

6. **MLE dla rozkładu jednostajnego**

Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu jednostajnego na odcinku $(0, 1)$.

- Wyznacz MLE dla parametru $a$.
- Oblicz wartość oczekiwaną i wariancję estymatora z poprzedniego punktu.

**Rozwiązanie:**

$$f(x) = \frac{1}{a} 1(x \in [0, a])$$

$$L(a) = f(x_1, ..., x_n) = \prod_i f(x_i) = \prod_i \frac{1}{a} 1(x_i \in [0, a]) = \frac{1}{a^n} 1(\max x_i \leq a, \min x_i \geq 0)$$

$$\hat{a}_{MLE} = \max x_i$$

Druga kropka:

$$P(\max x_i \leq t) = P(x_1 \leq t, ..., x_n \leq t) = \frac{t^n}{a^n}, \quad t \in [0, a]$$

$$f(t) = F'(t) = n \frac{t^{n-1}}{a^n}, \quad t \in [0, a]$$

$$E(\max x_i) = \int_0^a n \frac{x^{n-1}}{a^n} x dx = \frac{n}{a^n} \frac{1}{n+1} x^{n+1} \Big|_0^a = \frac{n}{n+1} a$$

$$E(\max x_i^2) = \int_0^a n \frac{x^{n-1}}{a^n} x^2 dx = \frac{n}{a^n} \frac{1}{n+2} x^{n+2} \Big|_0^a = \frac{n}{n+2} a^2$$

$$Var(\max x_i) = \frac{n}{n+2} a^2 - \frac{n^2}{(n+1)^2} a^2$$

7. **MLE dla rozkładu normalnego**

   Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu $N(\mu, \sigma^2)$, wyznacz estymatory największej wiarogodności nieznanych parametrów $\mu, \sigma^2$.

   **Rozwiązanie:**

$$L(\mu, \sigma^2) = f_{\mu, \sigma^2}(X_1, ..., X_n) = \prod_{i=1}^n f_{\mu, \sigma^2}(X_i) =$$

$$= \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (X_i - \mu)^2\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum (X_i - \mu)^2\right)$$

$$l(\mu, \sigma^2) = \log L(\mu, \sigma^2) = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - n \log \sigma - \frac{1}{2\sigma^2} \sum (X_i - \mu)^2$$

$$\frac{\partial l}{\partial \mu} = \frac{\sum (X_i - \mu)}{\sigma^2}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum (X_i - \mu)^2$$

$$\frac{\partial l}{\partial \mu} = 0 \quad \Rightarrow \quad \mu = \frac{\sum X_i}{n} = \overline{X}$$

$$\frac{\partial l}{\partial \sigma} = 0 \quad \Rightarrow \quad \sigma^2 = \frac{\sum (X_i - \mu)^2}{n}$$

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma^2} = \frac{\sum (X_i - \overline{X})^2}{n}$$

8. Jaka powinna być długość próbki pochodzącej z rozkładu normalnego $N(\mu, \sigma^2)$, ze znanyn $\sigma > 0$ aby dwustronny przedział ufności dla $\mu$ na poziomie istotności 0.95 miał długość mniejszą niż 0.01?

   *Wskazówka:* Kwantyl rozkładu normalnego na poziomie 0.975 jest równy $z_{0.975} = 1.96$.

   **Rozwiązanie:**

$$\overline{X} \pm \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}$$

   Długość:

$$\frac{2x_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} < \frac{1}{100}$$

$$z_{0.975} = 1.96 \approx 2$$

$$\sqrt{n} > 400\sigma$$

$$n > 160000\sigma^2$$

9. Niech $X_1, ..., X_n$ będzie próbą prostą z rozkładu o skończonej wariancji. Oblicz obciążenie estymatora wariancji:

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

   gdzie $\overline{X} = \sum_{i=1}^{n} X_i/n$. Znajdź $c \in \mathbb{R}$, takie że $cS_n^2$ jest nieobciążonym estymatorem wariancji.

   **Rozwiązanie:**

   Wprowadźmy oznaczenie

$$VarX_i = \sigma^2, EX_i = \mu$$

$$ES_n^2 = E\left(\frac{1}{n}\sum(X_i - \overline{X})^2\right) = E\left(\frac{1}{n}(\sum X_i^2 - 2\sum X_i\overline{X} + n\overline{X}^2)\right) =$$

$$= E\left(\frac{1}{n}\sum X_i^2 - \frac{2}{n}\sum X_i \sum X_j + \frac{1}{n^2}\sum X_i \sum X_j\right) = E\left(\frac{1}{n}\sum X_i^2 - \frac{1}{n^2}\sum X_i \sum X_j\right) =$$

$$= E\left(\left(\frac{1}{n} - \frac{1}{n^2}\right)\sum X_i^2 - \frac{1}{n^2}\sum_{i \neq j} X_i X_j\right) = \left(\frac{1}{n} - \frac{1}{n^2}\right)\sum EX_i^2 - \frac{1}{n^2}\sum_{i \neq j} EX_i EX_j =$$

$$= \left(1 - \frac{1}{n}\right)(\sigma^2 + \mu^2) - \left(1 - \frac{1}{n}\right)\mu^2 = \left(1 - \frac{1}{n}\right)\sigma^2 = \frac{n-1}{n}\sigma^2$$

$$c = \frac{n}{n-1}$$

   Ogólnie $c = \frac{1}{n-1}\sum(X_i - \overline{X})^2$

10. Pokaż, że częstość empiryczna $p_n = S_n/n$, gdzie $S_n$ jest liczbą sukcesów i $n$ próbach w schemacie Bernoulliego jest estymatorem zgodnym prawdopodobieństwa sukcesu.

    **Rozwiązanie:**

    Estymator zgodny

$$\lim_{n\to\infty} P(|\hat{\theta_n} - \theta| > \varepsilon) = 0$$

    Zbieżność według prawdopodobieństwa

    Jak pokazać, że $\frac{S_n}{n}$ dąży do $p$? Skorzystamy z prawa wielkich liczb.

$$S_n = \sum X_i, \quad X_i - iid$$

$$P(X_i = 0) = 1 - p = 1 - P(X_i = 1)$$

Zatem z prawa wielkich liczb:

$$\frac{S_n}{n} = \frac{\sum X_i}{n} \to EX_i = p \text{ zbieżność prawie na pewno}$$

Zbieżność prawie na pewno jest silniejsza niż zbieżność według prawdopodobieństwa, więc teza zachodzi.

11. Dla próby prostej $U_1, ..., U_n$ z rozkładu jednostajnego na odcinku $[a, 1]$ skonstruuj przedział ufności na poziomie ufności $1 - \alpha = 0.95$.

    **Rozwiązanie:**

    *To zadanie nie jest fajne, więc go nie robiliśmy*