VJEROJATNOST I STATISTIKA

Dodatni zadaci za 11. tjedan (22. predavanje)

10 Matematička statistika

Zadatak 1. Iz intervala $[\alpha, 1]$, gdje je α nepoznat odabrano je na sreću n brojeva: $x_1, x_2, \ldots, x_n$. Da bismo procijenili $1 - \alpha$ duljinu tog intervala odaberimo statistiku

$$Z = 1 - \min\left(X_1, X_2, \ldots, X_n\right).$$

(a) Dokažite da statistika Z nije nepristrana.

(b) S kojim faktorom treba pomnožiti Z kako bismo dobili nepristranu statistiku?

Rješenje. Označimo

$$Z_m = \min\left(X_1, X_2, \ldots, X_n\right).$$

Trebamo izračunati očekivanje slučajne varijable

$$Z = 1 - Z_m.$$

U tu svrhu izračunajmo funkciju razdiobe slučajne varijable Z_m, pomoću koje ćemo dobiti funkciju gustoće a zatim i traženo očekivanje.

$$\begin{aligned} F_{Z_m}(z) &= \mathbf{P}\left(Z_m < z\right) = \mathbf{P}\Big(\min\left(X_1, X_2, \ldots, X_n\right) < z\Big) = 1 - \mathbf{P}\Big(\min\left(X_1, X_2 \ldots X_n\right) \geq z\Big) \\ &= 1 - \mathbf{P}\left(X_1 \geq z, \ldots, X_n \geq z\right). \end{aligned}$$

Budući da brojeve biramo na sreću neovisno jedan o drugome, radi se o nezavisnim slučajnim varijablama te vrijedi

$$F_{Z_m}(z) = 1 - \mathbf{P}\left(X_1 \geq z\right)\mathbf{P}\left(X_2 \geq z\right)\cdots\mathbf{P}\left(X_n \geq z\right) = 1 - \left(1 - \frac{z-a}{b-a}\right)^n = 1 - \left(\frac{1-z}{1-\alpha}\right)^n, \quad z \in [\alpha, 1].$$

Deriviranjem dobivamo funkciju gustoće slučajne varijable Z_m

$$f_{Z_m}(z) = \frac{n\left(1-z\right)^{n-1}}{\left(1-\alpha\right)^n}, \quad z \in [\alpha, 1],$$

a nakon toga i njezino očekivanje

$$\mathbf{E}(Z_m) = \int_\alpha^1 z \cdot \frac{n\left(1-z\right)^{n-1}}{\left(1-\alpha\right)^n}\,\mathrm{d}z = 1 - \frac{n}{n+1}\left(1-\alpha\right).$$

Sada lako možemo izračunati očekivanje slučajne varijable Z

$$\mathbf{E}\left(Z\right) = 1 - \mathbf{E}\left(Z_m\right) = \frac{n}{n+1}\left(1-\alpha\right).$$

Da bi statistika bila nepristrana mora vrijediti $\mathbf{E}\left(Z\right) = 1 - \alpha$. Stoga zaključujemo kako statistika Z nije nepristrana, ali također vidimo da vrlo lako možemo učiniti da ona postane nepristrana pomnožimo li je s $\frac{n+1}{n}$. Tada vrijedi

$$Z' = \frac{n+1}{n}Z \implies \mathbf{E}(Z') = \frac{n+1}{n}\mathbf{E}\left(Z\right) = 1 - \alpha.$$

Zadatak 2. Iz intervala $[\alpha, 1]$, gdje je $\alpha < 1$ nepoznat odabrano je na sreću n brojeva: $x_1, x_2, \ldots, x_n$. Da bismo procijenili α odaberimo statistiku

$$Z = \min\left(X_1, X_2, \ldots, X_n\right) - \frac{1}{n+1}.$$

(a) Dokažite da statistika Z nije nepristrana.

(b) S kojim faktorom treba pomnožiti Z kako bismo dobili nepristranu statistiku?

Rješenje. Postupamo na identičan način kao u prethodnom zadatku. Označimo

$$Z_m = \min\left(X_1, X_2, \ldots, X_n\right).$$

Sada trebamo izračunati očekivanje slučajne varijable

$$Z = Z_m - \frac{1}{n+1}.$$

Funkcija razdiobe slučajne varijable Z_m, njezina funkcija gustoće i očekivanje isti su kao u prethodnom zadatku.

$$F_{Z_m}(z) = 1 - \left(\frac{1-z}{1-\alpha}\right)^n, \quad z \in [\alpha, 1],$$

$$f_{Z_m}(z) = \frac{n\,(1-z)^{n-1}}{(1-\alpha)^n}, \quad z \in [\alpha, 1],$$

$$\mathbf{E}(Z_m) = 1 - \frac{n}{n+1}\,(1-\alpha) = \frac{1}{n+1} + \frac{n}{n+1}\alpha.$$

Sada lako možemo izračunati očekivanje statistike Z

$$\mathbf{E}\left(Z\right) = \mathbf{E}\left(Z_m\right) - \frac{1}{n+1} = \frac{n}{n+1}\alpha.$$

Kao u prethodnom zadatku vidimo kako statistika nije nepristrana. Kako bismo je učinili nepristranom trebamo je pomnožiti s $\frac{n+1}{n}$ jer će tada vrijediti

$$\mathbf{E}(Z') = \mathbf{E}\left(\frac{n+1}{n} \cdot Z\right) = \frac{n+1}{n}\mathbf{E}(Z) = \alpha.$$

Zadatak 3. Uzorak $x_1, x_2, \ldots, x_n$ izvučen je iz populacije koja ima gustoću razdiobe

$$f\left(x\right) = \lambda x^{\lambda-1}, \quad x \in (0, 1).$$

Koristeći kriterij najveće izglednosti, odredite procjenu za parametar λ.

Rješenje. Ovdje se radi o procjeni parametra pomoću kriterija najveće izglednosti pri čemu je nepoznati parametar λ. Funkcija izglednosti je definirana s

$$L\left(\lambda, x_1, x_2, \ldots, x_n\right) = f\left(\lambda, x_1\right) \cdot f\left(\lambda, x_2\right) \cdots f\left(\lambda, x_n\right) = \prod_{i=1}^{n} \lambda x_i^{\lambda-1} = \lambda^n \Big(\prod_{i=1}^{n} x_i\Big)^{\lambda-1}.$$

Trebamo pronaći vrijednost parametra λ za koji ta funkcija poprima globalni maksimum. Označimo

$$k = \prod_{i=1}^{n} x_i.$$

Tada je

$$L(\lambda, x_1, x_2, \ldots, x_n) = \lambda^n k^{\lambda-1}.$$

U svrhu pronalaska maksimuma deriviramo funkciju izglednosti. Međutim, često se više isplati derivirati funkciju:

$$l = \ln L.$$

To možemo uraditi zato što je funkcija izglednosti uvijek pozitivna pa je funkcija l dobro definirana. Također je važno primijetiti da funkcija l poprima maksimum u istim točkama kao i L. Sada računamo

$$l(\lambda, x_1, x_2, \ldots, x_n) = \ln L(\lambda, x_1, x_2, \ldots, x_n) = \ln\left(\lambda^n k^{\lambda-1}\right) = n \ln \lambda + (\lambda - 1)\ln k,$$

$$\frac{\partial}{\partial \lambda} \ln L = \frac{n}{\lambda} + \ln k.$$

Kako bismo pronašli maksimum tražimo stacionarne točke, tj. izjednačavamo prvu derivaciju s 0:

$$\frac{n}{\lambda} + \ln k = 0 \implies \widehat{\lambda} = \frac{-n}{\ln k} = \frac{-n}{\ln\left(\prod_{i=1}^n x_i\right)} = \frac{-n}{\sum_{i=1}^n \ln x_i}.$$

Formalno bismo trebali još jednom derivirati $\frac{\partial}{\partial\lambda}\ln L$ kako bismo odredili predznak i vidjeli da je zaista riječ o maksimumu. To u ovom primjeru možemo lako napraviti

$$\frac{\partial^2}{\partial \lambda^2} \ln L = -\frac{n}{\lambda^2}.$$

Vidimo da je dobivena druga derivacija uvijek negativna, a posebno i za $\widehat{\lambda}$, te da se zaista radi o maksimumu.

Zadatak 4. Uzorak $x_1, x_2, \ldots, x_n$ izvučen je iz populacije koja ima Poissonovu razdiobu. Koristeći kriterij najveće izglednosti, odredite procjenu parametra λ te provjerite je li dobivena procjena nepristrana.

Rješenje. Za slučajnu varijablu X sa Poissonovom razdiobom vrijedi

$$f(\lambda, x) = \mathbf{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots$$

Funkcija izglednosti definana je s

$$\begin{aligned} L(\lambda, x_1, x_2, \ldots, x_n) &= f(\lambda, x_1) \cdot f(\lambda, x_2) \cdots f(\lambda, x_n) \\ &= \mathbf{P}(X = x_1)\, \mathbf{P}(X = x_2) \cdots \mathbf{P}(X = x_n) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} e^{-n\lambda}. \end{aligned}$$

Sada imamo

$$\ln L = \ln\left(\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} e^{-n\lambda}\right) = -n\lambda + \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!),$$

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0 \implies \widehat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

Vidimo da je dobivena procjena nepristrana.

Zadatak 5. Neka je X diskretna slučajna varijabla zadana sljedećom razdiobom koja ovisi o parametru ϑ:

$$X \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{2\vartheta}{3} & \frac{\vartheta}{3} & \frac{2(1-\vartheta)}{3} & \frac{1-\vartheta}{3} \end{pmatrix}.$$

Uzet je uzorak od 10 nezavisnih mjerenja slučajne varijable X i dobivene su vrijednosti $3, 0, 2, 1, 3, 2, 1, 0, 2, 1$. Koristeći kriterij najveće izglednosti, odredite procjenu za parametar ϑ na temelju dobivenog uzorka.

Rješenje. Ovaj zadatak rješavamo na isti način kao i prethodne zadatke, samo sada radimo s konkretnim brojevima x_i. Računamo:

$$\begin{aligned} L(\lambda, x_1, x_2, \ldots, x_{10}) &= \mathbf{P}(X = x_1)\,\mathbf{P}(X = x_2) \cdots \mathbf{P}(X = x_{10}) \\ &= \left(\frac{2\vartheta}{3}\right)^2 \left(\frac{\vartheta}{3}\right)^3 \left(\frac{2(1-\vartheta)}{3}\right)^3 \left(\frac{1-\vartheta}{3}\right)^2 \\ &= \frac{2^5}{3^{10}} \vartheta^5 (1-\vartheta)^5, \\ \ln L &= \ln\left(\frac{2^5}{3^{10}}\right) + 5\ln\vartheta + 5\ln(1-\vartheta), \\ \frac{\partial \ln L}{\partial \lambda} &= \frac{5}{\vartheta} - \frac{5}{1-\vartheta} = 0 \implies \widehat{\vartheta} = \frac{1}{2}. \end{aligned}$$

Zadatak 6. Iz intervala $[a, b]$, gdje su rubovi a i b nepoznati, odabrano je na sreću n brojeva: $x_1, x_2, \ldots, x_n$. Kako bismo procijenili sredinu tog intervala, odabiremo statistiku

$$\Theta = \frac{X_m + X_M}{2},$$

gdje je $X_m = \min\{X_1, X_2, \ldots, X_n\}$ i $X_M = \max\{X_1, X_2, \ldots, X_n\}$.

(a) Dokažite da je statistika Θ nepristrana procjena sredine intervala $[a, b]$.

(b) Dokažite da je ta procjena valjana.

(c) Dokažite da se pomoću kriterija najveće izglednosti dobiva ista procjena za sredinu intervala.

Rješenje. (a) Nezavisne slučajne varijable $X_1, \ldots, X_n$ imaju uniformnu razdiobu s funkcijom razdiobe, odnosno gustoće

$$F(x) = \frac{x-a}{b-a},\ x \in [a, b], \quad f(x) = \frac{1}{b-a},\ x \in [a, b].$$

Na isti način kao u prvom zadatku određujemo funkciju razdiobe slučajne varijable $X_m = \min\{X_1, X_2, \ldots, X_n\}$:

$$F_{X_m}(x) = 1 - (1 - F(x))^n, \quad x \in [a, b],$$

a zatim i funkciju gustoće

$$f_{X_m}(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{X_m}(x) = n(1 - F(x))^{n-1} f(x) = \frac{n(b-x)^{n-1}}{(b-a)^n}, \quad x \in [a, b].$$

Očekivanje slučajne varijable X_m je

$$\mathbf{E}(X_m) = \int_{-\infty}^{\infty} x f_{X_m}(x)\,\mathrm{d}x = a + \frac{b-a}{n+1}.$$

Slično radimo i za maksimum $X_M = \max\{X_1, \ldots, X_n\}$:

$$\begin{aligned} F_{X_M}(x) &= \mathbf{P}(X_M < x) = \mathbf{P}(\max\{X_1, \ldots, X_n\} < x) = \mathbf{P}(X_1 < x, \ldots, X_n < x) \\ &= \mathbf{P}(X_1 < x) \cdots \mathbf{P}(X_n < x) = F(x)^n, \\ f_{X_M}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_{X_M}(x) = nF(x)^{n-1} f(x) = \frac{n(x-a)^{n-1}}{(b-a)^n}, \quad x \in [a, b]. \end{aligned}$$

Očekivanje slučajne varijable X_M je

$$\mathbf{E}(X_M) = \int_{-\infty}^{\infty} x f_{X_M}(x)\,\mathrm{d}x = b - \frac{b-a}{n+1}.$$

Sada možemo izračunati i očekivanje statistike Θ

$$\mathbf{E}(\Theta) = \frac{\mathbf{E}(X_m) + \mathbf{E}(X_M)}{2} = \frac{a+b}{2},$$

te zaključujemo da je Θ nepristrana procjena sredine $\frac{a+b}{2}$ intervala $[a, b]$.

(b) Da bi statistika bila valjana dovoljno je da njena disperzija teži u nulu. Disperziju $\mathbf{D}(\Theta)$ računamo na način:

$$\mathbf{D}(\Theta) = \mathbf{D}\left(\frac{X_m + X_M}{2}\right) = \frac{1}{4}\mathbf{D}(X_m + X_M) = \frac{1}{4}\left(\mathbf{E}\left((X_m + X_M)^2\right) - \big(\mathbf{E}(X_m) + \mathbf{E}(X_M)\big)^2\right).$$

Odredimo razdiobu slučajnog vektora (X_m, X_M):

$$\begin{aligned} F_{X_m, X_M}(x, y) &= \mathbf{P}(X_m < x, X_M < y) \\ &= \mathbf{P}(X_M < y) - \mathbf{P}(X_m \geq x, X_M < y) \\ &= F_{X_M}(y) - \mathbf{P}(x \leq X_1 < y, \ldots, x \leq X_n < y) \\ &= F(y)^n - \big(F(y) - F(x)\big)^n, \qquad x, y \in [a, b],\ x \leq y. \end{aligned}$$

Deriviranjem dolazimo do funkcije gustoće

$$\begin{aligned} f_{X_m, X_M}(x, y) &= \frac{\partial^2}{\partial x \partial y} F_{X_m, X_M}(x, y) \\ &= n(n-1) f(x) f(y) (F(y) - F(x))^{n-2} \\ &= \frac{n(n-1)}{(b-a)^n} (y-x)^{n-2}, \qquad x, y \in [a, b],\ x \leq y. \end{aligned}$$

Računamo očekivanje kvadrata zbroja minimuma i maksimuma:

$$\begin{aligned}
\mathbf{E}\Big((X_m+X_M)^2\Big) &= \int\int (x+y)^2 f_{X_m,X_M}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{(b-a)^n}\int_a^b \left(\int_x^b n(n-1)(x+y)^2(y-x)^{n-2}\,\mathrm{d}y\right)\mathrm{d}x \\
&= \frac{1}{(b-a)^n}\int_a^b \left(\int_x^b n(n-1)\Big((y-x)^n+4x(y-x)^{n-1}+4x^2(y-x)^{n-2}\Big)\,\mathrm{d}y\right)\mathrm{d}x \\
&= \frac{1}{(b-a)^n}\int_a^b \left(\frac{n(n-1)}{n+1}(y-x)^{n+1}+4x(n-1)(y-x)^n+4x^2n(y-x)^{n-1}\right)\Bigg|_{y=x}^{b}\mathrm{d}x \\
&= \frac{1}{(b-a)^n}\int_a^b \left(\frac{n(n-1)}{n+1}(b-x)^{n+1}+4x(n-1)(b-x)^n+4x^2n(b-x)^{n-1}\right)\mathrm{d}x \\
&= \frac{1}{(b-a)^n}\int_a^b \left(\left(\frac{n(n-1)}{n+1}+4\right)(b-x)^{n+1}-4b(n+1)(b-x)^n+4b^2n(b-x)^{n-1}\right)\mathrm{d}x \\
&= \frac{1}{(b-a)^n}\left(-\frac{n^2+3n+4}{(n+1)(n+2)}(b-x)^{n+2}+4b(b-x)^{n+1}-4b^2(b-x)^n\right)\Bigg|_{x=a}^{b} \\
&= \frac{1}{(b-a)^n}\left(\frac{n^2+3n+4}{(n+1)(n+2)}(b-a)^{n+2}-4b(b-a)^{n+1}+4b^2(b-a)^n\right) \\
&= \frac{n^2+3n+4}{(n+1)(n+2)}(b-a)^2-4b(b-a)+4b^2 \\
&= \frac{n^2+3n+4}{(n+1)(n+2)}(b-a)^2+4ab.
\end{aligned}$$

Sada je

$$\begin{aligned}
\mathbf{D}(\Theta) &= \frac{1}{4}\left(\frac{n^2+3n+4}{(n+1)(n+2)}(b-a)^2+4ab-(a+b)^2\right) \\
&= \frac{1}{4}\left(\frac{n^2+3n+4}{(n+1)(n+2)}(b-a)^2-(b-a)^2\right) \\
&= \frac{(b-a)^2}{2(n+1)(n+2)} \xrightarrow{n\to\infty} 0,
\end{aligned}$$

odakle slijedi da je statistika Θ valjana.

(c) Primijetimo da funkcija izglednosti ovisi i o dva nepoznata parametra a i b:

$$L(a,b,x_1,\ldots,x_n) = f(a,b,x_1)\cdots f(a,b,x_n) = \frac{1}{(b-a)^n}.$$

Vidimo da funkcija L poprima maksimum kada izraz $b-a$ poprima minimum. Budući da je

$$a \le x_i \le b,\ \forall i \in \{1,\ldots,n\},$$

odnosno

$$a \le \min(x_1,\ldots,x_n) \le \max(x_1,\ldots,x_n) \le b,$$

razlika $b-a$ je najmanja moguća za

$$a = \min(x_1,\ldots,x_n) \text{ i } b = \max(x_1,\ldots,x_n).$$

Dakle, za procjenu sredine intervala $[a, b]$ uzimamo vrijednost

$$c = \frac{a+b}{2} = \frac{\min(x_1, \ldots, x_n) + \max(x_1, \ldots, x_n)}{2}$$

što je ista procjena koju smo imali ranije.