# Statistika 2 - 2022/2023

**Domača naloga, 8. 2. 2023, Anže Mramor**

$\boxed{1}$ Predvidevamo, da je vzorec realnih števil realizacija neodvisnih ponovitev slučajne spremenljivke z Lebesguovo gostoto

$$f(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \cdot \mathbf{1}_{(0,\infty)}(x),$$

za parameter $\vartheta = (a,b) \in (0,\infty) \times (0,\infty)$.

**a.** Poiščite kompletno zadostno statistiko.

**b.** Obravnavajte problem cenilke največjega verjetja.

**c.** Izračunajte cenilko $\hat{\vartheta}_{MM}$ za $\vartheta$ po metodi momentov. Komentirajte jo.

**d.** S pomočjo metode delta obravnavajte (dvorazsežno) normalno aproksimacijo $\hat{\vartheta}_{MM}$: formulirajte primeren limitni izrek.

**e.** S pomočjo normalne aproksimacije cenilke $\hat{\vartheta}_{MM}$ konstruirajte aproksimativno območje zaupanja stopnje zaupanja $0.95$: problem prevedite na večrazsežno normalno porazdelitev z znano variančno-kovariančno matriko. Dobljeno območje čim natančneje opišite.

$\boxed{2}$ Predvidevamo, da je naš vzorec realizacija neodvisnih ponovitev diskretne slučajne spremenljivke z verjetnostno funkcijo

$$f(x;\vartheta) = \binom{100}{x} \cdot \vartheta^x (1-\vartheta)^{100-x} \cdot \mathbf{1}_{\{0,1,2,\ldots,99,100\}}(x).$$

za realnoštevilski parameter $\vartheta \in (0,1) \subset \mathbb{R}$.

**a.** Poiščite kompletno zadostno statistiko.

**b.** Kakšni preizkusi domnev $\vartheta \leqslant \vartheta_0$ proti alternativam $\vartheta > \vartheta_0$ so na voljo v tem modelu?

**c.** Za domnevo $\vartheta \leqslant 0.2$ vzemite po vašem prepričanju najboljši možen preizkus stopnje značilnosti $0.05$ za vzorec vaše velikosti in skicirajte graf funkcije moči.

**d.** Na vašem vzorcu realizirajte interval zaupanja stopnje zaupanja $0.95$ po pripadajočem izreku s predavanj.

$\boxed{3}$ **a.** Privzemimo model z $n$ zaporednimi neodvisnimi ponovitvami Bernoullijeve slučajne spremenljivke s parametrom $p$. Za $p_0 \in (0,1)$ naj bo $\phi_{p_0} : \mathbb{R}^n \to [0,1]$ preizkus za $H_0 : p = p_0$ proti $A : p \neq p_0$ velikosti $0.05$, ki je enakomerno najmočnejši med vsemi nepristranskimi preizkusi za $H_0$ proti $A$ stopnje značilnosti $0.05$. Vemo, da obstajajo taki preizkusi oblike

$$\phi_{p_0}(x_1,x_2,\ldots,x_n) = \begin{cases} 1, & \sum_{i=1}^n x_i < C_1(p_0) \text{ ali } C_2(p_0) < \sum_{i=1}^n x_i, \\ \gamma_j(p_0), & \sum_{i=1}^n x_i = C_j(p_0), \\ 0, & \text{sicer.} \end{cases}$$

Vaš diagram prikazuje grafa funkcij $C_1(p_0)$ in $C_2(p_0)$ za $n = 23$. S pomočjo inverzije konstruirajte (v konkretnih številkah) interval zaupanja za Bernoullijev parameter za vaš $n$.

**b.** Za $p_0 = 6/10$ izračunajte pripadajoči konstanti $\gamma_1$ in $\gamma_2$ preizkusa $\phi_{p_0}$ iz **a.**

**c.** Naj bo $C$ območje zaupanja za parameter $\vartheta \in \Theta$. Pokritost pri $\vartheta_0 \in \Theta$ je verjetnost $P_{\vartheta_0}\big(\{\mathbf{X} \mid \vartheta_0 \in C(\mathbf{X})\}\big)$. Koeficient zaupanja območja $C$ je seveda natančna spodnja meja pokritosti.
Za interval zaupanja iz **a.** napravite graf pokritosti (potrudite se!) in ocenite koeficient zaupanja. Koeficient zaupanja lahko tudi natančno izračunate.

**d.** Izračunajte pripadajoče konstante za različico preizkusa iz **a.** za velikost vzorca $n + 10$ (kjer je $n$ iz točke **a.**) in $p_0 = 6/10$.

**e.** (*) Za velikost vzorca $n + 10$ iz **d.** napravite diagram kot v **a.**

$\boxed{4}$ Obravnavamo diskretno porazdelitev na fiksnih $m+1$ točkah $\xi_0,\ldots,\xi_m$ z verjetnostmi $(p_0,p_1,\ldots,p_m)$. Preizkušamo domnevo čisto določene porazdelitve $H_0 : (p_0,p_1,\ldots,p_4) = (\frac{1}{10},\frac{1}{10},\frac{2}{5},\frac{1}{5},\frac{1}{5})$ z asimptotičnimi preizkusi pri nominalni stopnji značilnosti $0.05$. Za vzorce velikosti $n = 30, 50, 70, 90$ izračunajte:

**a.** eksaktno velikost preizkusa domneve $H_0$ na podlagi razmerja verjetij,

**b.** eksaktno velikost preizkusa domneve $H_0$ na podlagi Pearsonove statistike $\sum_j \frac{(T_j - n\pi_j)^2}{n\pi_j}$ (za vaše podatke).

**Opozorilo:** Če vaši rezultati močno odstopajo od nominalne značilnosti in ste prepričani, da vaš algoritem deluje, preverite, ali je njegova implementacija numerično zanesljiva.

**c.** (*) Obravnavajte moč na alternativah oblike $p_j = \pi_j + \delta$, $p_k = \pi_k - \delta$ (za vaše podatke).

$\boxed{5}$ Med različnimi holesteroli je holesterol LDL (*low-density lipoprotein* - lipoprotein nizke gostote) relativno težko oziroma drago meriti. Zato je še marsikje v uporabi cenilka oblike LDL$= \beta_1 \cdot$TCH$+\beta_2\cdot$HDL$+\beta_3\cdot$TRI (za konkretne vrednosti parametrov $\beta_1$, $\beta_2$, $\beta_3$; ocena je znana pod imenom Friedewaldova formula), kjer so TCH (*total cholesterol* - skupni holesterol), HDL (*high-density lipoprotein* - lipoprotein visoke gostote) in TRI (trigliceridi) količine, ki jih je relativno lahko meriti. Ocenjevanje parametrov $\beta_i$ je torej regresijski problem.

**a.** Predpostavite linearni model $\mathrm{LDL_i} = \beta_0 + \beta_1 \cdot \mathrm{TCH_i} + \beta_2 \cdot \mathrm{HDL_i} + \beta_3 \cdot \mathrm{TRI_i} + \varepsilon_i$, kjer so $\varepsilon_i$ neodvisne slučajne spremenljivke, vse porazdeljene po zakonu $N(0,\sigma^2)$.

(i) Ocenite parametre $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$ po metodi najmanjših kvadratov.

(ii) Preizkusite domnevo $\beta_1 = \beta_2 = \beta_3 = 0$ na standardni način pri stopnji značilnosti $0.05$.

(iii) Preizkusite domnevo $\beta_0 = 0$ na standardni način pri stopnji značilnosti $0.05$.

**b.** Predpostavite model brez prostega člena $\mathrm{LDL_i} = \beta_1 \cdot \mathrm{TCH_i} + \beta_2 \cdot \mathrm{HDL_i} + \beta_3 \cdot \mathrm{TRI_i} + \varepsilon_\mathrm{i}$.

(i) Ocenite parametre $\beta_1$, $\beta_2$, $\beta_3$ po metodi najmanjših kvadratov.

(ii) Preizkusite domnevo $(\beta_1,\beta_2,\beta_3) = (1,-1,-0.45)$.

(iii) Realizirajte območje zaupanja za $(\beta_1,\beta_2,\beta_3)$ stopnje zaupanja $0.95$ in ga čim natančneje opišite.

(iv) Realizirajte hkratne intervale zaupanja za $\beta_1$, $\beta_2$ in $\beta_3$ z Bonferronijevim popravkom in primerjajte dobljeni kvader zaupanja z območjem iz prejšnje točke.

$\boxed{6}$ Za dani avtomobil želimo na podlagi podatka o učinkovitosti `mpg` („miles per gallon") in mase `weight` pojasniti, ali gre za avtomobil „tujega" izvora (`foreign=1`) ali za avtomobil „domačega" izvora (`foreign=0`). Za dani vzorec, v katerem je 26 primerkov tujega in 74 primerkov domačega izvora predpostavimo, da je realizacija slučajnega vektorja iz modela

$$p_i = P(\texttt{foreign}_i = 1) = 1 - \exp(-\exp(\beta_0 + \beta_1 * \texttt{weight}_i + \beta_2 * \texttt{mpg}_i)),$$

kjer so komponente $\texttt{foreign}_i$ neodvisne in porazdeljene po zakonu $\mathrm{B}(1, p_i)$.

**a.** Zapišite funkcijo verjetja.

**b.** Ocenite parametre $\beta_0$, $\beta_1$ in $\beta_2$ po metodi največjega verjetja.

**c.** Izračunajte Fisherjevo informacijsko matriko.

**d.** Izračunajte standardne napake za parametre $\beta_0$, $\beta_1$ in $\beta_2$.

**e.** Na standardni način preizkusite domnevo $H_0 : \beta_1 = \beta_2 = 0$.

$\boxed{7}$ **a.** Za (ne nujno zvezno) slučajno spremenljivko $X$ s kumulativno porazdelitveno funkcijo $F$ čim bolj natančno obravnavajte kumulativno porazdelitveno funkcijo slučajne spremenljivke $Y = F(X)$.

**b.** Dokažite točko (i) splošnega izreka s predavanj o konstrukciji intervala zaupanja v primeru enoparametričnega modela.