Statistika 2 - 2022/2023, Tine Markočič
Domača naloga, 8. 8. 2023

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

$$f(x;a,b^2) = \frac{1}{x\sqrt{2\pi b^2}} e^{-(\ln x - a)^2/(2b^2)} \cdot \mathbf{1}_{(0,\infty)}(x),$$

za parameter $\vartheta = (a, b^2) \in \mathbb{R} \times (0,\infty)$.

a. Poiščite kompletno zadostno statistiko.

b. Obravnavajte problem cenilke največjega verjetja.

c. Izračunajte cenilko $\hat{\vartheta}_{MM}$ za ϑ 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

$$f(x;\vartheta) = e^{-\vartheta} \cdot \frac{\vartheta^x}{x!} \cdot \mathbf{1}_{\{0,1,2,3,\dots\}}(x).$$

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

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 $\vartheta \leqslant 5$ 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

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

Vaš diagram prikazuje grafa funkcij $C_1(p_0)$ in $C_2(p_0)$ za $n = 25$. S pomočjo inverzije konstruirajte (v konkretnih številkah) interval zaupanja za Bernoullijev parameter za vaš n.

b. Za $p_0 = 34/100$ izračunajte pripadajoči konstanti γ_1 in γ_2 preizkusa ϕ_{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 $p_0 = 34/100$.

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

$\boxed{4}$ Obravnavamo diskretno porazdelitev na fiksnih $m+1$ točkah $\xi_0, \dots, \xi_m$ z verjetnostmi $(p_0, p_1, \dots, p_m)$. Preizkušamo domnevo čisto določene porazdelitve $H_0 : (p_0, p_1, p_2, p_3, p_4) = (\frac{1}{12}, \frac{1}{12}, \frac{1}{3}, \frac{1}{4}, \frac{1}{4})$ z asimptotičnimi preizkusi pri nominalni stopnji značilnosti 0.05. Za vzorce velikosti $n = 40, 60, 80, 100$ 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 β_1, β_2, β_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 β_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 ε_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 β_1, β_2 in β_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 15 primerkov tujega in 85 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 `foreign`$_i$ neodvisne in porazdeljene po zakonu $\mathrm{B}(1, p_i)$.

a. Zapišite funkcijo verjetja.
b. Ocenite parametre β_0, β_1 in β_2 po metodi največjega verjetja.
c. Izračunajte Fisherjevo informacijsko matriko.
d. Izračunajte standardne napake za parametre β_0, β_1 in β_2.
e. Na standardni način preizkusite domnevo $H_0 : \beta_1 = \beta_2 = 0$.

$\boxed{7}$ Predpostavimo linearni regresijski model (za neskončni vzorec) oblike

$$\mathbb{X} = Z \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kjer je $Z \in \mathbb{R}^{\infty \times d}$ fiksna matrika z neskončno mnogo vrsticami in so komponente ε_i vektorja $\boldsymbol{\varepsilon}$ neodvisne in enako porazdeljene slučajne spremenljivke s končno (neznano) disperzijo σ^2 in pričakovano vrednostjo 0 in je $\boldsymbol{\beta} = [\beta_1 \quad \ldots \quad \beta_d]^T$ stolpec (preostalih) parameterov. Naj Z_i označuje i-to vrstico matrike Z in naj bo $Z^{(n)} \in \mathbb{R}^{n \times d}$ matrika, sestavljena iz prvih n vrstic matrike Z. Analogno naj bo $\mathbb{X}^{(n)}$ stolpec, sestavljen iz prvih n komponent vektorja $\mathbb{X}$. Privzemite, da imajo matrike $Z^{(n)T} Z^{(n)}$ poln rang za vsa dovolj velika števila n in naj bo

$$\hat{\boldsymbol{\beta}}^{(n)} = (Z^{(n)T} Z^{(n)})^{-1} Z^{(n)T} \mathbb{X}^{(n)}$$

cenilka za β (po metodi najmanjših kvadratov) za restrikcijo na vzorec velikosti n. Naj bo

$$\boldsymbol{\zeta}^{(n)} = \frac{1}{\sigma} (Z^{(n)T} Z^{(n)})^{-1/2} Z^{(n)T} \boldsymbol{\varepsilon}^{(n)}$$

standardizacija (prepričajte se, da to drži) cenilke $\hat{\boldsymbol{\beta}}^{(n)}$.

a. Dokažite, da pri pogoju

$$\lim_{n\to\infty} \max_{1 \leqslant j \leqslant n} \left\langle (Z^{(n)T} Z^{(n)})^{-1} Z_j^T, Z_j^T \right\rangle = 0$$

vektorji $\boldsymbol{\zeta}^{(n)}$ konvergirajo k d-razsežni standardni normalni porazdelitvi.

b. Dokažite, da je pogoj iz prejšnje točke izpolnjen, če zaporedje $d \times d$-matrik $Z^{(n)T} Z^{(n)}/n$ konvergira k neki pozitivo definitni matriki P in za neko število $M > 0$ velja $\|Z_j\| \leqslant M$ za vsa naravna števila j.

c. Kaj pomeni pogoj iz točke **b.** v primeru enostavne linearne regresije ($X_i = \beta_0 + \beta_1 \cdot z_i + \varepsilon_i$)?

d. Ali je pogoj iz točke **b.** v praksi razumen? (Pri razmišljanju o tem bodite inventivni; lahko se odločite za bolj praktično ali bolj teoretično utemeljevanje.)

e. (*) Pokažite, da je pogoj omejenosti $\|Z_j\| \leqslant M$ iz točke **b.** v resnici odveč. Splošni primer dovolj preprosto sledi iz posebnega primera enostavne linearne regresije.

Za točko **a.** si lahko pomagate z naslednjim izrekom: naj bodo za vsako naravno število n slučajne spremenljivke $Y_1^{(n)}, \ldots, Y_n^{(n)}$ neodvisne in enako porazdeljene s pričakovano vrednostjo 0 in disperzijo 1. Dalje naj bodo $\mathbf{c}^{(n)} \in \mathbb{R}^n$ enotski vektorji z lastnostjo

$$\lim_{n\to\infty} \max_{1 \leqslant j \leqslant n} \left(\mathbf{c}_j^{(n)}\right)^2 = 0.$$

Tedaj zaporedje slučajnih spremenljivk $\sum_{j=1}^n \mathbf{c}_j^{(n)} Y_j^{(n)}$ konvergira k standardni normalni porazdelitvi.