Ime in priimek: _____ Vpisna št: | | | | | | | | |

FAKULTETA ZA MATEMATIKO IN FIZIKO

Oddelek za matematiko

Statistika

Pisni izpit

13. junij 2019

Navodila

Pazljivo preberite besedilo naloge, preden se lotite reševanja. Za pozitiven rezultat morate zbrati vsaj 45 točk od 100 možnih. Veliko uspeha!

Naloga	a.	b.	c.	d.	e.	Skupaj
1.			•	•	•	
2.						
3.			•	•	•	
4.			•	•	•	
Skupaj	•	•	•	•	•	

1. (25) Srečko igra igro, v kateri z verjetnostjo 50% izgubi 1 evro, z verjetnostjo 5% dobi 9 evrov, z verjetnostjo 45% pa je na ničli. Igro odigra 500-krat, pri čemer so vse igre med seboj neodvisne.

a. (15) Čim natančneje izračunajte verjetnost, da ima Srečko po seriji 500 iger v denarnici vsaj 50 evrov več kot na začetku.

Rešitev: Če je X_i Srečkov dobiček v i-ti igri, je razlika med Srečkovim stanjem v denarnici na koncu in na začetku enaka S_{500}, kjer je $S_n = \sum_{i=1}^{n} X_i$. To je vsota dovolj neodvisnih in enako porazdeljenih slučajnih spremenljivk, da lahko pri predpisani natančnosti uporabimo centralni limitni izrek. Velja:

$$\begin{aligned} E(X_i) &= -0{,}05\,, & E(S_{500}) &= -500 \cdot 0{,}05 = -25\,, \\ \operatorname{var}(X_i) &= 4{,}5475\,, & \operatorname{var}(S_{500}) &= 500 \cdot 4{,}5475 = 2273{,}75\,, \end{aligned}$$

torej je

$$P(S_{500} \geq 50) = P(S_{500} \geq 49{,}5) \approx 1 - \Phi\left(\frac{49{,}5 + 25}{\sqrt{2273{,}75}}\right) \doteq 0{,}059\,.$$

Točen rezultat: 0,06354382.

b. (10) Recimo, da ima Srečko po seriji 500 iger v denarnici res vsaj 50 evrov več kot na začetku. Čim natančneje izračunajte pogojno verjetnost, da je v prvi igri dobil 9 evrov.

Rešitev: Iskana verjetnost je enaka

$$\begin{aligned} P(X_1 = 9 \mid S_{500} \geq 50) &= \frac{P(X_1 = 9, S_{500} \geq 50)}{P(S_{500} \geq 50)} \\ &= \frac{P(X_1 = 9)\, P(S_{500} \geq 50 \mid X_1 = 9)}{P(S_{500} > 50)} \\ &= \frac{P(X_1 = 9)\, P(S_{499} \geq 41)}{P(S_{500} > 50)}\,. \end{aligned}$$

Podobno kot prej izračunamo:

$$E(S_{499}) = -499 \cdot 0{,}05 = -24{,}95\,, \quad \operatorname{var}(S_{499}) = 499 \cdot 4{,}5475 = 2269{,}2025$$

in

$$P(S_{499} \geq 41) = P(S_{499} \geq 40{,}5) \approx 1 - \Phi\left(\frac{49{,}5 + 24{,}95}{\sqrt{2269{,}2025}}\right) \doteq 0{,}085$$

ter končno

$$P(X_1 = 9 \mid S_{500} \geq 50) \approx 0{,}072\,.$$

Točen rezultat: 0,06951523.

2. (25) V populaciji je N oseb in vsaka je bodisi tipa A bodisi tipa B. Označimo z a delež oseb v populaciji, ki so tipa A.

Iz populacije vzamemo enostavni slučajni vzorec velikosti n. Vsako osebo, izbrano v vzorec, vprašamo, katerega tipa je. Osebe pa ne odgovarjajo nujno po resnici: oseba tipa A bo z verjetnostjo p_A po pravici odgovorila, da je tipa A, sicer pa bo odgovorila, da je tipa B. Oseba tipa B pa bo z verjetnostjo p_B po pravici odgovorila, da je tipa B, sicer pa bo odgovorila, da je tipa A. Verjetnosti p_A in p_B sta znani. Privzamemo, da so odgovori oseb neodvisni tako med seboj kot tudi od vzorčenja.

a. (5) Naj bo S_A število oseb v vzorcu, ki so tipa A, R_A število oseb v vzorcu, ki *odgovorijo*, da so tipa A. Izračunajte $E(R_A \mid S_A)$.

Rešitev: Zaradi neodvisnosti odgovarjanja od vzorčenja verjetnosti odgovorov za posamezni tip veljajo tudi pogojno na S_A. Sledi

$$E(R_A \mid S_A) = S_A p_A + (n - S_A)(1 - p_B) = S_A(p_A + p_B - 1) + n(1 - p_B)\,.$$

b. (5) Predlagajte nepristransko cenilko deleža a.

Rešitev: Cenilko je smiselno zastaviti kot funkcijo števila R_A (in konstant). Število R_A je namreč od doslej omenjenih količin edina, ki je opazljiva, slučajna in katere porazdelitev je tipično odvisna od a. Izračunajmo kar

$$E(R_A) = E\big(E(R_A \mid S_A)\big) = na(p_A + p_B - 1) + n(1 - p_B)\,.$$

Od tod sledi, da je

$$\hat{a} := \frac{1}{p_A + p_B - 1}\left(\frac{R_A}{n} - 1 + p_B\right)$$

iskana nepristranska cenilka. Le-ta obstaja, brž ko je $p_A + p_B \neq 1$.

Opomba. *Količina S_A/n, ki ima prav tako pričakovano vrednost a, ni cenilka za a, ker ni opazljiva.*

c. (5) Utemeljite, kdaj je delež a sploh možno nepristransko oceniti.

Rešitev: Iz prejšnje točke vemo, da je delež možno nepristransko oceniti, brž ko je $p_A + p_B \neq 1$. Če pa je $p_A + p_B = 1$, pogojno na S_A vsaka oseba z enako verjetnostjo p_A odgovori, da je tipa A, in osebe so pri tem neodvisne. Skupna pogojna porazdelitev odgovorov oseb, izbranih v vzorec, je torej enaka ne glede na S_A, kar pomeni, da je to tudi skupna brezpogojna porazdelitev odgovorov. Ta pa je neodvisna tudi od a. Odgovori izprašanih oseb pa so poleg konstant edina informacija, ki jo imamo pri ocenjevanju na voljo: vsaka cenilka mora biti funkcija odgovorov izprašanih oseb. Ker je skupna porazdelitev neodvisna od a, mora to veljati tudi za pričakovano vrednost katere koli cenilke. Pričakovana vrednost torej ne bo vedno *enaka a, kar je pogoj za nepristransko cenilko.*

Sklep: nepristranska cenilka za a obstaja natanko tedaj, ko je $p_A + p_B = 1$.

d. (5) Izračunajte $\operatorname{var}(R_A \mid S_A)$.

Rešitev: Pogojno na S_A je R_A vsota dveh neodvisnih slučajnih spremenljivk s porazdelitvama $\operatorname{Bin}(S_A, p_A)$ *in* $\operatorname{Bin}(n - S_A, 1 - p_B)$. *Sledi*

$$\operatorname{var}(R_A \mid S_A) = S_A p_A (1 - p_A) + (n - S_A) p_B (1 - p_B) \,.$$

e. (5) Izračunajte standardno napako vaše cenilke deleža a.

Namig: $S_A \sim \operatorname{HiperGeom}(n, aN, N)$.

Rešitev: Najprej s pomočjo dekompozicije variance izračunamo

$$\begin{aligned} \operatorname{var}(R_A) &= E\big(\operatorname{var}(R_A \mid S_A)\big) + \operatorname{var}\big(E(R_A \mid S_A)\big) \\ &= p_A(1 - p_A)\, E(S_A) + p_B(1 - p_B)\, \big(n - E(S_A)\big) \\ &\qquad + (p_A + p_B - 1)^2 \operatorname{var}(S_A) \,. \end{aligned}$$

Iz hipergeometrijske porazdelitve dobimo

$$E(S_A) = na \quad \textit{in} \quad \operatorname{var}(S_A) = \frac{N - n}{N - 1}\, na(1 - a) \,,$$

torej

$$\begin{aligned} &\operatorname{var}(R_A) \\ &\quad = n \left[a p_A (1 - p_A) + (1 - a) p_B (1 - p_B) + \frac{N - n}{N - 1}\, a(1 - a)(p_A + p_B - 1)^2 \right] . \end{aligned}$$

Iskana standardna napaka pa je

$$\begin{aligned} \operatorname{se} &= \sqrt{\operatorname{var}(\hat{a})} = \frac{\sqrt{\operatorname{var}(R_A)}}{n |p_A + p_B - 1|} \\ &= \frac{1}{\sqrt{n}} \left[\frac{a p_A (1 - p_A) + (1 - a) p_B (1 - p_B)}{(p_A + p_B - 1)^2} + \frac{N - n}{N - 1}\, a(1 - a) \right]^{1/2} . \end{aligned}$$

3. (25) Opazovane vrednosti naj bodo pari $(x_1, y_1), \ldots, (x_n, y_n)$, $n \geq 2$, za katere privzamemo, da so vzorec neodvisnih realizacij neizrojene dvorazsežne normalne porazdelitve $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$, kjer je

$$\mathbf{\Sigma} = \begin{pmatrix} a & b \\ b & \frac{1+b^2}{a} \end{pmatrix},$$

$a > 0$ in $b \in \mathbb{R}$ pa sta neznana parametra.

a. (15) Poiščite cenilki za oba parametra po metodi največjega verjetja.

Rešitev: Najprej izračunamo

$$\det(\mathbf{\Sigma}) = 1\,, \qquad \mathbf{\Sigma}^{-1} = \begin{pmatrix} \frac{1+b^2}{a} & -b \\ -b & a \end{pmatrix}.$$

Označimo še $\mathbf{x} = (x_1, \ldots, x_n)$ *in* $\mathbf{y} = (y_1, \ldots, y_n)$. *Funkcijo verjetja tako lahko zapišemo kot*

$$L(a, b|\mathbf{x}, \mathbf{y}) = \left(\frac{1}{2\pi}\right)^n \exp\left(-\frac{1+b^2}{2a}\sum_{k=1}^n x_k^2 + b\sum_{k=1}^n x_k y_k - \frac{a}{2}\sum_{k=1}^n y_k^2\right).$$

Če označimo

$$m_{xx} = \frac{1}{n}\sum_{k=1}^n x_k^2\,, \qquad m_{xy} = \frac{1}{n}\sum_{k=1}^n x_k y_k\,, \qquad m_{yy} = \frac{1}{n}\sum_{k=1}^n y_k^2\,,$$

lahko logaritem verjetja zapišemo v obliki

$$\ell(a, b|\mathbf{x}, \mathbf{y}) = n\left(-\log(2\pi) - \frac{(1+b^2)\,m_{xx}}{2a} + b\,m_{xy} - \frac{a\,m_{yy}}{2}\right).$$

Odvajamo:

$$\frac{\partial \ell}{\partial a} = \frac{n}{2}\left(\frac{(1+b^2)\,m_{xx}}{a^2} - m_{yy}\right), \qquad \frac{\partial \ell}{\partial b} = n\left(-\frac{b\,m_{xx}}{a} + m_{xy}\right)$$

in ko izenačimo z nič, po nekaj računanja dobimo ustrezni cenilki:

$$\hat{a} = \frac{m_{xx}}{\sqrt{m_{xx}m_{yy} - m_{xy}^2}}\,, \qquad \hat{b} = \frac{m_{xy}}{\sqrt{m_{xx}m_{yy} - m_{xy}^2}}\,.$$

b. (10) Izračunajte aproksimativni standardni napaki obeh cenilk pri velikem vzorcu.

Rešitev: Če se omejimo na en sam opažen par (x, y), *so drugi parcialni odvodi enaki*

$$\frac{\partial^2 \ell}{\partial a^2} = -\frac{(1+b^2)x^2}{a^3}\,, \qquad \frac{\partial^2 \ell}{\partial a\,\partial b} = \frac{bx^2}{a^2}\,, \qquad \frac{\partial^2 \ell}{\partial b^2} = -\frac{x^2}{a}\,.$$

Če opaženi par (x, y) *nadomestimo s slučajno spremenljivko* (X, Y) *in upoštevamo, da je* $E(X^2) = a$, *dobimo Fisherjevo matriko*

$$I(a, b) = \begin{pmatrix} \frac{1+b^2}{a^2} & -\frac{b}{a} \\ -\frac{b}{a} & 1 \end{pmatrix},$$

ki ima inverz

$$I^{-1}(a,b) = \begin{pmatrix} a^2 & ab \\ ab & 1+b^2 \end{pmatrix} .$$

Aproksimativni standardni napaki sta torej:

$$\mathrm{se}(\hat{a}) = \frac{a}{\sqrt{n}} \quad \textit{in} \quad \mathrm{se}(\hat{b}) = \frac{\sqrt{1+b^2}}{\sqrt{n}} .$$

4. (25) Privzemite regresijski model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

kjer je $\mathbf{Z}$ matrika velikosti $r \times n$ ranga r in velja $n \geq r$, za $\mathbf{u}$ pa velja $E(\mathbf{u}) = 0$ in $\mathrm{var}(\mathbf{u}) = \sigma^2\mathbf{I}$.

a. (15) Pokažite, da je

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1}\mathbf{Y}$$

najboljša nepristranska linearna cenilka za $\boldsymbol{\beta}$.

Rešitev:
Prvi način: *neposredno. Nepristranskost lahko takoj preverimo. Naj bo* $\tilde{\boldsymbol{\beta}} = \mathbf{L}\mathbf{Y}$ *alternativna linearna nepristranska cenilka. Iz*

$$E\left(\tilde{\boldsymbol{\beta}}\right) = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$$

za vse $\boldsymbol{\beta}$ *sledi* $\mathbf{L}\mathbf{X} = \mathbf{I}$. *Računamo*

$$\begin{aligned}\mathrm{var}\left(\tilde{\boldsymbol{\beta}}\right) &= \mathrm{var}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}\right) \\ &= \mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) + \mathrm{var}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right) + 2\,\mathrm{cov}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\right).\end{aligned}$$

Cenilka $\tilde{\boldsymbol{\beta}}$ *bo najboljša, brž ko bo kovarianca na desni enaka* 0. *Upoštevajoč*

$$\mathrm{cov}\left(\mathbf{Y}, \mathbf{Y}\right) = \sigma^2\mathbf{Z}\mathbf{Z}^T,$$

izračunamo, da je to res:

$$\begin{aligned}&\mathrm{cov}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\right) = \\ &\quad= \left(\mathbf{L} - \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\right)\mathrm{cov}(\mathbf{Y}, \mathbf{Y})\cdot \\ &\qquad\qquad\cdot\left(\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\right)^T \\ &\quad= \sigma^2\left(\mathbf{L} - \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\right)\mathbf{Z}\mathbf{Z}^T \\ &\qquad\qquad\cdot(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= \sigma^2\left(\mathbf{L} - \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\right)\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= \sigma^2\left(\mathbf{L}\mathbf{X} - \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= \sigma^2(\mathbf{I} - \mathbf{I})\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= 0\,.\end{aligned}$$

Drugi način: *prevedemo na standardno linearno regresijo. Razlika je namreč le v tem, da šum* $\mathbf{Z}\mathbf{u}$ *nima kovariančne matrike* $\sigma^2\mathbf{I}$, *temveč* $\sigma^2\mathbf{Z}\mathbf{Z}^T$. *Če definiramo*

$\mathbf{Y}' = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Y}$, $\mathbf{X}' = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{X}$ *in* $\boldsymbol{\varepsilon} = (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\mathbf{u}$, *velja* $\mathbf{Y}' = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *kar je standardni linearni regresijski model, saj je:*

$$\begin{aligned}\operatorname{var}(\boldsymbol{\varepsilon}) &= (\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\operatorname{var}(\mathbf{u})\,\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2} \\ &= \sigma^2(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Z}\mathbf{I}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2} \\ &= \sigma^2\mathbf{I}\,.\end{aligned}$$

Seveda se linearne cenilke v podanem nestandardnem modelu ujemajo z linearnimi cenilkami v prirejenem standardnem modelu, nepristranskost in standardna napaka pa sta tako ali tako univerzalna pojma. Zato je iskana cenilka tudi najboljša nepristranska linearna cenilka v prirejenem standardnem modelu, to pa je:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'^T\mathbf{X}'\right)^{-1}\mathbf{X}'^T\mathbf{Y}' \\ &= \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2}(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1/2}(\mathbf{Z}\mathbf{Z}^T)^{-1/2}\mathbf{Y} \\ &= \left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Y}\,.\end{aligned}$$

Opomba. *Množenje z* $\mathbf{Z}^{-1}$ *dani model prav tako prevede na standardnega, a za to mora biti matrika* $\mathbf{Z}$ *obrnljiva, to pa je res le za* $n = r$.

b. (10) Naj bo

$$\hat{\mathbf{u}} = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

cenilka za $\mathbf{u}$. Izračunajte

$$\operatorname{cov}(\hat{\mathbf{u}}, \hat{\boldsymbol{\beta}})\,.$$

Rešitev: Računamo

$$\begin{aligned}&\operatorname{cov}(\hat{\mathbf{u}}, \hat{\boldsymbol{\beta}}) = \\ &\quad= \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\operatorname{cov}(\mathbf{Y}, \mathbf{Y})\cdot \\ &\qquad\qquad\cdot(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= \sigma^2\,\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\Big)\cdot \\ &\qquad\qquad\cdot\mathbf{Z}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\left(\mathbf{X}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{X}\right)^{-1} \\ &\quad= \mathbf{0}\,.\end{aligned}$$