Ime in priimek: ____________________ Vpisna št: | | | | | | | | |

FAKULTETA ZA MATEMATIKO IN FIZIKO

Oddelek za matematiko

Statistika

Pisni izpit

31. avgust 2018

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.	
1.			•	•	
2.					
3.					
4.					
Skupaj	•	•	•	•	

1. (25) Ljubiteljski kolesar Bernard vsak dan od ponedeljka do sobote z verjetnostjo 1/2 prevozi 20, z verjetnostjo 1/2 pa 25 km. V nedeljo pa z verjetnostjo 3/4 počiva, z verjetnostjo 1/4 pa naredi izlet, na katerem prevozi 100 km. Privzamemo, da so vse Bernardove odločitve neodvisne.

a. (15) Čim natančneje izračunajte verjetnost, da v 12 tednih prevozi več kot 2000 km.

Rešitev: Označimo z X_{kl} *število kilometrov, ki jih Bernard prevozi* l*-ti dan* k*-tega tedna od ponedeljka do sobote (torej* 20 *ali* 25*), z* N_{12} *pa število nedelj v prvih 12 tednih, ko Bernard naredi izlet. V prvih 12 tednih Bernard prevozi* $S_{12} := T_{12} + 100\,N$ *kilometrov, kjer je* $T_{12} := \sum_{k=1}^{12}\sum_{l=1}^{6} X_{kl}$. *Glede na to, da nedeljski kilometri izstopajo, se splača pogojevati na* N, *ki je porazdeljena binomsko* $\mathrm{Bin}(12, 1/4)$. *Velja torej*

$$
\begin{aligned}
P(S_{12} > 2000) &= \sum_{n=0}^{12} P(N = n)\, P(S_{12} > 2000 \mid N = n) = \\
&= \sum_{n=0}^{12} \binom{12}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{12-n} P(T_{12} > 2000 - 100n)\,.
\end{aligned}
$$

slučajna spremenljivka T_{12} *pa je vsota velikega števila neodvisnih in enako porazdeljenih slučajnih spremenljivk, torej je po centralnem limitnem izreku porazdeljena približno normalno. Ker je*

$$
\begin{aligned}
E(X_{kl}) &= \frac{20+25}{2} = 22{,}5\,, & \mathrm{var}(X_{kl}) &= \left(\frac{5}{2}\right)^2 = 6{,}25\,,\\
E(T_{12}) &= 12 \cdot 6 \cdot 22{,}5 = 1620\,, & \mathrm{var}(T_{12}) &= 12 \cdot 6 \cdot 6{,}25 = 450\,,
\end{aligned}
$$

je približno

$$
\begin{aligned}
P(S_{12} > 2000) &= P(S_{12} > 2002{,}5) \\
&\approx \sum_{n=0}^{12} \binom{12}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{12-n} \left[1 - \Phi\left(\frac{382{,}5 - 100n}{\sqrt{450}}\right)\right].
\end{aligned}
$$

Tabeliramo ustrezne normalne verjetnosti, zaokrožene na 4 decimalke natančno:

n	0	1	2	3	4	≥ 5
$1 - \Phi\left(\frac{382{,}5-100n}{\sqrt{450}}\right)$	0,0000	0,0000	0,0000	0,0001	0,7953	1,0000

in vidimo, da je edina netrivialna normalna verjetnost pri $n = 4$, *poleg tega pa še, da se bolj splača računati nasprotno verjetnost:*

$$
\begin{aligned}
P(S_{12} \le 2000) &\approx \sum_{n=0}^{3} \binom{12}{n} \left(\frac{1}{4}\right)^n \left(\frac{3}{4}\right)^{12-n} + \binom{12}{4} \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right)^8 \cdot 0{,}2047 \\
&\doteq 0{,}6886\,.
\end{aligned}
$$

Verjetnost, da Bernard v prvih 12 tednih prevozi več kot 2000 km, je torej približno 0,3114.

Preprostejši način, ki tipično da manj natančen približek, pa bi bil, da bi zapisali $S_{12} = Y_1 + Y_2 + \cdots + Y_{12}$, kjer je Y_k število kilometrov, ki jih Bernard prevozi v k-tem tednu. Slučajne spremenljivke Y_k so neodvisne in enako porazdeljene, jih je pa le 12, kar je malo. Velja:

$$E(Y_k) = 6 \cdot 22{,}5 + 100 \cdot \frac{1}{4} = 160\,, \qquad \mathrm{var}(Y_k) = 6 \cdot 6{,}25 + 10000 \cdot \frac{1}{4} \cdot \frac{3}{4} = 1912{,}5\,,$$
$$E(S_{12}) = 12 \cdot 160 = 1920\,, \qquad \mathrm{var}(S_{12}) = 12 \cdot 1912{,}5 = 22950\,,$$

kar nam da približek

$$P(S_{12} > 2000) = P(S_{12} > 2002{,}5) \approx 1 - \Phi\left(\frac{2002{,}5 - 1920}{\sqrt{22950}}\right) \doteq 0{,}2930\,.$$

Točen rezultat: 0,3115884.

b. (10) Približno po koliko tednih postane verjetnost, da prevozi več kot 20.000 km, večja od 95%?

Rešitev: Tu uporabimo manj natančni približek iz prve točke. Upoštevajoč

$$E(S_n) = 160\,n \quad \textit{in} \quad \mathrm{var}(S_n) = 1912{,}5\,n\,,$$

nastavimo

$$1 - \Phi\left(\frac{20002{,}5 - 160\,n}{\sqrt{1912{,}5\,n}}\right) = 0{,}95$$

oziroma

$$160\,n - 20002{,}5 = \Phi^{-1}(0{,}95)\sqrt{1912{,}5\,n}\,.$$

Če uvedemo $u = \sqrt{n}$, dobimo rešitvi $u_1 \doteq -10{,}959$ in $u_2 \doteq 11{,}408$. Prava je rešitev s pozitivnim korenom, ker je $\sqrt{n}$ po dogovoru nenegativno število. Iz $u_2^2 \doteq 130{,}15$ dobimo, da postane verjetnost večja od 95% po približno 131 tednih.

V resnici je 131 tednov potrebnih in zadostnih: po 130 tednih verjetnost pride 0,9480694*, po 131 tednih pa* 0,9750782.

2. (25) Populacija velikosti N je za namene vzorčenja razdeljena na K podskupin velikosti M, tako da je $N = KM$. Vzorec velikosti n izberemo tako, da najprej izberemo enostavni slučajni vzorec k izmed K skupin, v vsaki izbrani skupini pa izberemo enostavni slučajni vzorec m enot neodvisno do izbir na prvem koraku vzorčenja, tako da je $n = km$. Populacijsko povprečje ocenimo z vzorčnim povprečjem. Naj bo μ_i populacijsko povprečje, σ_i^2 pa populacijska varianca v i-ti skupini za $i = 1, 2, \ldots, K$; μ naj bo populacijsko povprečje. Označite

$$\sigma_b^2 = \frac{1}{K} \sum_{i=1}^{K} (\mu_i - \mu)^2 .$$

Označite z $\bar{X}$ vzorčno povprečje.

a. (5) Utemeljite, da je vzorčno povprečje nepristranska cenilka populacijskega povprečja μ.

   Rešitev: Če je x_{ij} vrednost spremenljivke na j-ti enoti v i-ti skupini na populaciji, je populacijsko povprečje enako

   $$\mu = \frac{1}{KM} \sum_{i=1}^{K} \sum_{j=1}^{M} x_{ij} .$$

   Nadalje, če z I_i označimo indikator dogodka, da je i-ta skupina izbrana v vzorec, z I_{ij} pa indikator dogodka, da je j-ta enota v i-ti skupini izbrana v vzorec, je vzorčno povprečje enako

   $$\bar{X} = \frac{1}{km} \sum_{i=1}^{K} \sum_{j=1}^{M} x_{ij} I_i I_{ij} .$$

   Iz vzorčnega načrta sledi, da je $E(I_i I_{ij}) = \frac{k}{K} \frac{m}{M}$, od koder sledi, da je $E(\bar{X}) = \mu$.

b. (5) Definirajte

   $$I_i = \begin{cases} 1 & \text{če je } i\text{-ta podskupina izbrana} \\ 0 & \text{sicer.} \end{cases}$$

   Izračunajte

   $$E\left(\bar{X} | I_1, I_2, \ldots, I_K\right) \quad \text{in} \quad \operatorname{var}\left(\bar{X} | I_1, I_2, \ldots, I_K\right) .$$

   Rešitev: Predstavljamo si lahko, da najprej vzamemo vzorce iz vseh skupin, nakar upoštevamo le izbrane skupine. Naj bo $\bar{X}_i$ vzorčno povprečje v i-ti skupini. Tedaj je

   $$\bar{X} = \frac{1}{k} \sum_{i=1}^{K} \bar{X}_i I_i .$$

   Iz vzorčnega načrta sledi, da je $\bar{X}_i$ neodvisen od $I_1, \ldots, I_K$. Iz standardne teorije vzorčenja dobimo

   $$\begin{aligned} E(\bar{X}_i | I_1, \ldots, I_K) &= E(\bar{X}_i) = \mu_i , \\ \operatorname{var}(\bar{X}_i | I_1, \ldots, I_K) &= \operatorname{var}(\bar{X}_i) = \frac{M - m}{(M-1)m}\, \sigma_i^2 . \end{aligned}$$

Torej je tudi

$$E(\bar{X}_i I_i | I_1, \ldots, I_K) = \mu_i I_i \quad \textit{in} \quad \operatorname{var}(\bar{X}_i I_i | I_1, \ldots, I_K) = \frac{M-m}{(M-1)m}\, \sigma_i^2 I_i \,.$$

Spet iz vzorčnega načrta sledi, da so slučajne spremenljivke $X_i I_i$ pogojno na $I_1, \ldots, I_K$ neodvisne. Sledi

$$E(\bar{X} | I_1, \ldots, I_K) = \frac{1}{k} \sum_{i=1}^{K} \mu_i I_i \quad \textit{in} \quad \operatorname{var}(\bar{X} | I_1, \ldots, I_K) = \frac{M-m}{(M-1)m} \frac{1}{k^2} \sum_{i=1}^{K} \sigma_i^2 I_i \,.$$

c. (10) Naj bo $\bar{X}_i$ vzorčno povprečje v i-ti skupini. Izračunajte

$$E\big(\bar{X}_i^2 I_i\big) \quad \text{in} \quad E\big(\bar{X}^2\big)\,.$$

Rešitev: Iz vmesnih rezultatov prejšnje točke dobimo

$$\begin{aligned} E\big(\bar{X}_i^2 I_i \mid I_1, \ldots, I_K\big) &= E\big(\bar{X}_i^2 \mid I_1, \ldots, I_K\big) I_i = \\ &= \Big[\operatorname{var}\big(\bar{X}_i \mid I_1, \ldots, I_K\big) + \big(E(\bar{X}_i | I_1, \ldots, I_K)\big)^2\Big] I_i \\ &= \left(\frac{M-m}{(M-1)m}\, \sigma_i^2 + \mu_i^2\right) I_i \,, \end{aligned}$$

torej je

$$E\big(\bar{X}_i^2 I_i\big) = E\Big(E\big(\bar{X}_i^2 I_i \mid I_1, \ldots, I_K\big)\Big) = \frac{k}{K} \left(\frac{M-m}{(M-1)m}\, \sigma_i^2 + \mu_i^2\right) .$$

Podobno iz rezultatov prejšnje točke dobimo

$$\begin{aligned} E\big(\bar{X}^2 \mid I_1, \ldots, I_K\big) &= \operatorname{var}\big(\bar{X} \mid I_1, \ldots, I_K\big) + \big(E(\bar{X} | I_1, \ldots, I_K)\big)^2 \\ &= \frac{1}{k^2} \frac{M-m}{(M-1)} \sum_{i=1}^{K} \sigma_i^2 I_i + \left(\frac{1}{k} \sum_{i=1}^{K} \mu_i I_i\right)^2 . \end{aligned}$$

Izračunajmo

$$\begin{aligned} E\left[\left(\frac{1}{k} \sum_{i=1}^{K} \mu_i I_i\right)^2\right] &= \frac{1}{k^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \mu_i \mu_j \, E(I_i I_j) \\ &= \frac{1}{Kk} \sum_{i=1}^{K} \mu_i^2 + \frac{k-1}{K(K-1)k} \sum_{\substack{1 \le i,j \le K \\ i \ne j}} \mu_i \mu_j \\ &= \frac{k-1}{K(K-1)k} \left(\sum_{i=1}^{K} \mu_i\right)^2 + \frac{K-k}{K(K-1)k} \sum_{i=1}^{K} \mu_i^2 \\ &= \frac{K(k-1)}{k(K-1)}\, \mu^2 + \frac{K-k}{K(K-1)k} \sum_{i=1}^{K} \mu_i^2 \,. \end{aligned}$$

Poberemo skupaj in dobimo

$$\begin{aligned} E(\bar{X}^2) &= E\Big(E\big(\bar{X}^2 \mid I_1, \ldots, I_K\big)\Big) \\ &= \frac{1}{Kk}\,\frac{M-m}{(M-1)m}\sum_{i=1}^K \sigma_i^2 + \frac{K(k-1)}{k(K-1)}\,\mu^2 + \frac{K-k}{K(K-1)k}\sum_{i=1}^K \mu_i^2\,. \end{aligned}$$

d. (5) Predlagajte nepristransko cenilko količine σ_b^2.

Rešitev: Pišimo

$$\sigma_b^2 = \frac{1}{K}\sum_{i=1}^K \mu_i^2 - \mu^2\,.$$

To količino bomo znali nepristransko oceniti, če bomo znali nepristransko oceniti μ_i^2 *in* μ^2. *Kot osnova cenilke za* μ_i^2 *se ponuja opazljiva količina* $\bar{X}_i^2 I_i$, *a njena pričakovana vrednost se izraža še s* σ_i^2, *kar pa znamo nepristransko oceniti. Iz teorije vzorčenja namreč vemo, da bi bila količina*

$$\frac{M-1}{M(m-1)}\sum_{j=1}^M (x_{ij} - \bar{X}_i)^2 I_{ij}$$

nepristranska cenilka za σ_i^2, *če bi bila opazljiva; tu si spet mislimo, da iz vsake skupine vzamemo podvzorec. Opazljiva pa je količina*

$$\frac{M-1}{M(m-1)}\sum_{j=1}^M (x_{ij} - \bar{X}_i)^2 I_i I_{ij}$$

in iz vzorčnega načrta sledi, da ima le-ta pričakovano vrednost $\frac{k}{K}\,\sigma_i^2$. *Tako dobimo, da je*

$$\widehat{\sigma_i^2} := \frac{K}{k}\,\frac{M-1}{M(m-1)}\sum_{j=1}^M (x_{ij} - \bar{X}_i)^2 I_i I_{ij}$$

nepristranska cenilka za σ_i^2. *Sledi, da je*

$$\widehat{\mu_i^2} := \frac{K}{k}\bar{X}_i^2 I_i - \frac{M-m}{(M-1)m}\,\widehat{\sigma_i^2}$$

nepristranska cenilka za μ_i^2. *Nadalje je*

$$\begin{aligned} \widehat{\mu^2} &:= \frac{k(K-1)}{K(k-1)}\left(\bar{X}^2 - \frac{1}{Kk}\,\frac{M-m}{(M-1)m}\sum_{i=1}^K \widehat{\sigma_i^2} - \frac{K-k}{K(K-1)k}\sum_{i=1}^K \widehat{\mu_i^2}\right) \\ &= \frac{k(K-1)}{K(k-1)}\Bigg(\bar{X}^2 - \frac{1}{Kk}\,\frac{M-m}{(M-1)m}\sum_{i=1}^K \widehat{\sigma_i^2} - \frac{K-k}{k^2(K-1)}\sum_{i=1}^K \bar{X}_i^2 I_i \\ &\qquad + \frac{K-k}{K(K-1)k}\,\frac{M-m}{(M-1)m}\sum_{i=1}^K \widehat{\sigma_i^2}\Bigg) \\ &= \frac{k(K-1)}{K(k-1)}\,\bar{X}^2 - \frac{K-k}{Kk(k-1)}\sum_{i=1}^K \bar{X}_i^2 I_i - \frac{1}{K^2}\,\frac{M-m}{(M-1)m}\sum_{i=1}^K \widehat{\sigma_i^2} \end{aligned}$$

nepristranska cenilka za μ^2*. Iskana nepristranska cenilka za* σ_b^2 *pa je*

$$\begin{aligned}
\widehat{\sigma_b^2} &:= \frac{1}{K}\sum_{i=1}^{K}\widehat{\mu_i^2} - \widehat{\mu^2} \\
&= \frac{K-1}{K(k-1)}\sum_{i=1}^{K}\bar{X}_i^2 I_i - \frac{k(K-1)}{K(k-1)}\,\bar{X}^2 - \frac{K-1}{K^2}\,\frac{M-m}{(M-1)m}\sum_{i=1}^{K}\widehat{\sigma_i^2} \\
&= \frac{K-1}{K}\sum_{i=1}^{K}\left(\frac{1}{k-1}\,(\bar{X}_i-\bar{X})^2 - \frac{1}{k}\,\frac{M-m}{Mm(m-1)}\sum_{j=1}^{M}(x_{ij}-\bar{X}_i)^2 I_{ij}\right) I_i\,.
\end{aligned}$$

3. (25) Prepostavljajte, da so podatki $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$ nastali kot med sabo neodvisne slučajne spremenljivke $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ z $X_k \sim \mathrm{N}(\mu, \sigma^2)$ in $Y_k \sim \mathrm{N}(\mu, \tau^2)$ za $k = 1, 2, \ldots, n$. Za cenilko parametra μ predlagamo

$$\hat{\mu} = \frac{\bar{X} \sum_{k=1}^n (Y_k - \bar{Y})^2 + \bar{Y} \sum_{k=1}^n (X_k - \bar{X})^2}{\sum_{k=1}^n (X_k - \bar{X})^2 + \sum_{k=1}^n (Y_k - \bar{Y})^2} .$$

Označite

$$U = \frac{\sum_{k=1}^n (Y_k - \bar{Y})^2}{\sum_{k=1}^n (X_k - \bar{X})^2 + \sum_{k=1}^n (Y_k - \bar{Y})^2}$$

in

$$V = \frac{\sum_{k=1}^n (X_k - \bar{X})^2}{\sum_{k=1}^n (X_k - \bar{X})^2 + \sum_{k=1}^n (Y_k - \bar{Y})^2} .$$

a. (5) Izračunajte $E(\bar{X}|U,V)$ in $E(\bar{Y}|U,V)$. Je cenilka $\hat{\mu}$ nepristranska?

Rešitev: Vemo, da so slučajne spremenljivke $\bar{X}$, $\bar{Y}$, U in V med sabo neodvisne, torej je $E(\bar{X}|U,V) = E(\bar{X}) = \mu$ in $E(\bar{Y}|U,V) = E(\bar{Y}) = \mu$. Nadalje velja $\hat{\mu} = \bar{X}\, U + \bar{Y}\, V$, torej je

$$E(\hat{\mu}) = E(\bar{X})\, E(U) + E(\bar{Y})\, E(V) = \mu\big(E(U) + E(V)\big) .$$

Ker je $U + V = 1$, je tudi $E(U) + E(V) = 1$, zato je $\hat{\mu}$ nepristranska cenilka za μ.

b. (5) Izračunajte $\mathrm{var}(\hat{\mu}|U,V)$.

Rešitev: Zaradi neodvisnosti je

$$\begin{aligned}
\mathrm{var}(\hat{\mu}|U,V) &= \mathrm{var}(\bar{X}U + \bar{Y}V|U,V) \\
&= U^2\,\mathrm{var}(\bar{X}|U,V) + 2UV\,\mathrm{cov}(\bar{X},\bar{Y}|U,V) + V^2\,\mathrm{var}(\bar{Y}|U,V) \\
&= U^2\,\mathrm{var}(\bar{X}) + 2UV\,\mathrm{cov}(\bar{X},\bar{Y}) + V^2\,\mathrm{var}(\bar{Y}) \\
&= \frac{\sigma^2 U^2}{n} + \frac{\tau^2 V^2}{n} .
\end{aligned}$$

c. (5) Pokažite, da je

$$\mathrm{var}(\hat{\mu}) = \frac{\sigma^2}{n} E(U^2) + \frac{\tau^2}{n} E(V^2)$$

in sklepajte, da je vedno

$$\mathrm{var}(\hat{\mu}) \le \frac{\sigma^2 + \tau^2}{n} .$$

Rešitev: Velja

$$\mathrm{var}(\hat{\mu}) = E\big[\mathrm{var}(\hat{\mu}|U,V)\big] + \mathrm{var}\big[E(\hat{\mu}|U,V)\big] .$$

Iz prejšnje točke sledi

$$E\big[\mathrm{var}(\hat{\mu}|U,V)\big] = \frac{\sigma^2}{n}\, E(U^2) + \frac{\tau^2}{n}\, E(V^2) ,$$

iz neodvisnosti pa dobimo

$$E(\hat{\mu}|U,V) = U\,E(\bar{X}) + V\,E(\bar{Y}) = (U+V)\mu = \mu\,,$$

torej je $\operatorname{var}\big[E(\hat{\mu}|U,V)\big] = 0$. *Prva trditev je tako dokazana. Druga trditev pa sledi iz dejstva, da je* $U^2 \le 1$ *in posledično* $E(U^2) \le 1$ *ter podobno za* $E(V^2)$.

d. (10) Naj bo $\tilde{\mu}$ cenilka μ po metodi največjega verjetja. Cenilke vam ni treba izračunati, navedite pa aproksimacijo standardne napake z uporabo Fisherjeve matrike.

Rešitev: Lahko si predstavljamo, da so podatki dobljeni kot neodvisni pari $(X_1, Y_1), \ldots, (X_n, Y_n)$. *Fisherjevo informacijo je zato dovolj izračunati za* $n = 1$. *Za ta primer velja*

$$\ell(\mu, \sigma, \tau \mid x, y) = -\log\sigma - \log\tau - \log(2\pi) - \frac{(x-\mu)^2}{2\sigma^2} - \frac{(y-\mu)^2}{2\tau^2}\,.$$

Prvi parcialni odvodi so

$$\frac{\partial \ell}{\partial \mu} = \frac{x-\mu}{\sigma^2} + \frac{y-\mu}{\tau^2}\,, \qquad \frac{\partial \ell}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}\,, \qquad \frac{\partial \ell}{\partial \tau} = -\frac{1}{\tau} + \frac{y-\mu}{\tau^3}\,,$$

drugi parcialni odvodi pa so

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{1}{\sigma^2} - \frac{1}{\tau^2}\,, \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = -\frac{1}{\sigma^2} + \frac{3(x-\mu)^2}{\sigma^4}\,, \qquad \frac{\partial^2 \ell}{\partial \tau^2} = -\frac{1}{\tau^2} + \frac{3(y-\mu)^2}{\tau^4}\,,$$
$$\frac{\partial^2 \ell}{\partial \mu\,\partial \sigma} = -\frac{2(x-\mu)}{\sigma^3}\,, \qquad \frac{\partial^2 \ell}{\partial \mu\,\partial \tau} = -\frac{2(y-\mu)}{\tau^3}\,, \qquad \frac{\partial^2 \ell}{\partial \sigma\,\partial \tau} = 0\,.$$

Iz $E(X_1) = E(Y_1) = \mu$ *ter* $E\big[(X_1-\mu)^2\big] = \sigma^2$ *in* $E\big[(Y_1-\mu)^2\big] = \tau^2$ *dobimo, da je Fisherjeva matrika za* $n = 1$ *enaka*

$$\mathrm{I}(\mu, \sigma, \tau) = \begin{pmatrix} \frac{1}{\sigma^2} + \frac{1}{\tau^2} & 0 & 0 \\ 0 & \frac{2}{\sigma^2} & 0 \\ 0 & 0 & \frac{2}{\tau^2} \end{pmatrix}.$$

Iz nje dobimo aproksimativno standardno napako

$$\operatorname{se}(\tilde{\mu}) \approx \frac{\sigma\tau}{\sqrt{n(\sigma^2+\tau^2)}}\,.$$

Opomba. *Morda se zdi, kot da je za izračun aproksimativne standardne napake dovolj izračunati pričakovano vrednost drugega parcialnega odvoda po* μ, *saj je aproksimativna standardna napaka enaka* $\left[n\,E\big(-\frac{\partial^2\ell}{\partial\mu^2}(\mu,\sigma,\tau|X_1,Y_1)\big)\right]^{-1/2}$. *To pa ni vedno res. Res je, če je Fisherjeva matrika diagonalna, kar se zgodi tudi v našem primeru. Tega pa ne moremo vedeti vnaprej, zato je treba izračunati celo Fisherjevo matriko ali pa na določen način pokazati, da je diagonalna.*

4. (20) Predpostavimo naslednji regresijski model:

$$\begin{aligned} Y_{i1} &= \beta x_{i1} + \epsilon_i \,, \\ Y_{i2} &= \beta x_{i2} + \eta_i \end{aligned}$$

za $i = 1, 2, \ldots, n$. Predpostavljamo, da so pari $(\epsilon_1, \eta_1), \ldots, (\epsilon_n, \eta_n)$ med sabo neodvisni in enako porazdeljeni z $E(\epsilon_i) = E(\eta_i) = 0$, $\mathrm{var}(\epsilon_i) = \mathrm{var}(\eta_i) = \sigma^2$ in $\mathrm{corr}(\epsilon_i, \eta_i) = \rho$. Privzemite, da je ρ znan in $|\rho| < 1$.

a. (5) Oglejmo si cenilko

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1} x_{i1} + Y_{i2} x_{i2})}{\sum_{i=1}^n (x_{i1}^2 + x_{i2}^2)} \,.$$

Je ta cenilka nepristranska? Izračunajte njeno standardno napako.

Rešitev: Da je cenilka nepristranska, sledi iz definicije in dejstva, da je $E(Y_{ij}) = \beta x_{ij}$. Za izračun variance pa cenilko zapišimo kot poseben primer cenilke oblike

$$\hat{\beta} = \sum_{i=1}^n (a_i Y_{i1} + b_i Y_{i2}) \,.$$

Iz neodvisnosti ter dejstva, da je $\mathrm{var}(Y_{i1}) = \mathrm{var}(Y_{i2}) = \sigma^2$ in $\mathrm{cov}(Y_{i1}, Y_{i2}) = \rho\sigma^2$, dobimo

$$\mathrm{var}(\hat{\beta}) = \sum_{i=1}^n \mathrm{var}(a_i Y_{i1} + b_i Y_{i2}) = \sigma^2 \sum_{i=1}^n (a_i^2 + 2\rho a_i b_i + b_i^2) \,.$$

V našem primeru je $a_i = x_{i1} \big/ \sum_{j=1}^n (x_{j1}^2 + x_{j2}^2)$ in $b_i = x_{i2} \big/ \sum_{j=1}^n (x_{j1}^2 + x_{j2}^2)$. Dobimo

$$\mathrm{var}(\hat{\beta}) = \sigma^2 \left(\frac{1}{\sum_{i=1}^n (x_{i1}^2 + x_{i2}^2)} + 2\rho \, \frac{\sum_{i=1}^n x_i y_i}{\left(\sum_{i=1}^n (x_{i1}^2 + x_{i2}^2)\right)^2} \right) .$$

b. (5) Enačbi lahko seštejemo in dobimo

$$Y_{i1} + Y_{i2} = \beta(x_{i1} + x_{i2}) + \xi_i \,,$$

kjer je $\xi_i = \epsilon_i + \eta_i$. Napake $\xi_1, \ldots, \xi_n$ so med sabo nekorelirane in velja $E(\xi_i) = 0$ in $\mathrm{var}(\xi_i) = \sigma^2(2 + \rho)$. Parameter β lahko ocenimo kot

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1} + Y_{i2})(x_{i1} + x_{i2})}{\sum_{i=1}^n (x_{i1} + x_{i2})^2} \,.$$

Je ta cenilka nepristranska? Izračunajte njeno standardno napako.

Rešitev: Tudi ta cenilka je nepristranska, kar se ponovno vidi neposredno iz definicije. Za izračun variance pa si pomagamo z nastavkom iz točke a.: tokrat je $a_i = b_i = (x_{i1} + x_{i2}) \big/ \sum_{j=1}^n (x_{j1} + x_{j2})^2$. Dobimo

$$\mathrm{var}(\hat{\beta}) = \frac{2(1 + \rho)\sigma^2}{\sum_{i=1}^n (x_{i1} + x_{i2})^2} \,.$$

c. (5) Drugo enačbo lahko nadomestimo z zvezo

$$\frac{Y_{i2}-\rho Y_{i1}}{\sqrt{1-\rho^2}} = \beta\,\frac{x_{i2}-\rho x_{i1}}{\sqrt{1-\rho^2}} + \frac{\eta_i-\rho\epsilon_i}{\sqrt{1-\rho^2}}\,.$$

Označimo

$$\tilde{Y}_{i2} = \frac{Y_{i2}-\rho Y_{i1}}{\sqrt{1-\rho^2}} \qquad \text{in} \qquad \tilde{x}_{i2} = \frac{x_{i2}-\rho x_{i1}}{\sqrt{1-\rho^2}}\,,$$

parameter β pa ocenimo z

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_{i1}x_{i1} + \tilde{Y}_{i2}\tilde{x}_{i2})}{\sum_{i=1}^n (x_{i1}^2 + \tilde{x}_{i2}^2)}\,.$$

Je ta cenilka nepristranska? Izračunajte njeno standardno napako.

Rešitev: Tudi ta cenilka je nepristranska, kar se ponovno vidi neposredno iz definicije. Za izračun variance si lahko tudi tokrat pomagamo z nastavkom iz točke a. – cenilka ustreza izbiri

$$\begin{aligned}
a_i &= \frac{1}{\sum_{j=1}^n (x_{j1}^2 + \tilde{x}_{j2}^2)}\left(x_{i1} - \frac{\rho\,\tilde{x}_{i2}}{\sqrt{1-\rho^2}}\right) = \frac{x_{i1}-\rho x_{i2}}{\sum_{j=1}^n (x_{j1}^2 - 2\rho x_{j1}x_{j2} + x_{j2}^2)}\,,\\
b_i &= \frac{1}{\sum_{j=1}^n (x_{j1}^2 + \tilde{x}_{j2}^2)}\,\frac{\tilde{x}_{i2}}{\sqrt{1-\rho^2}} = \frac{x_{i2}-\rho x_{i1}}{\sum_{j=1}^n (x_{j1}^2 - 2\rho x_{j1}x_{j2} + x_{j2}^2)}\,.
\end{aligned}$$

Po krajšem računu dobimo

$$\operatorname{var}(\hat{\beta}) = \frac{\sigma^2(1-\rho^2)}{\sum_{i=1}^n (x_{i1}^2 - 2\rho x_{i1}x_{i2} + x_{i2}^2)}\,.$$

d. (10) Katera od zgornjih cenilk ima po vašem mnenju najmanjšo standardno napako? Je cenilka z najmanjšo standardno napako tudi najboljša linearna nepristranska cenilka? Utemeljite vaš odgovor.

Rešitev: Pišimo $\tilde{Y}_{i2} = \beta\tilde{x}_{i2} + \tilde{\eta}_i$, kjer je

$$\tilde{\eta}_i = \frac{\eta_i - \rho\epsilon_i}{\sqrt{1-\rho^2}}\,.$$

Krajši račun pokaže, da je ta slučajna spremenljivka nekorelirana z ϵ_i ter da je $E(\tilde{\eta}_i) = 0$ in $\operatorname{var}(\tilde{\eta}_i) = \sigma^2$. V primeru c. so tako izpolnjeni pogoji izreka Gauss–Markova. Če postavimo

$$\tilde{\mathbf{X}} := \begin{bmatrix} x_{11} \\ \cdots \\ x_{n1} \\ \tilde{x}_{12} \\ \cdots \\ \tilde{x}_{n2} \end{bmatrix}, \qquad \tilde{\boldsymbol{Y}} := \begin{bmatrix} Y_{11} \\ \cdots \\ Y_{n1} \\ \tilde{Y}_{12} \\ \cdots \\ \tilde{Y}_{n2} \end{bmatrix},$$

se cenilka iz točke c. ujema z $\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}\tilde{\boldsymbol{Y}}$*, torej je to najboljša linearna nepristranska cenilka na podlagi opaženih vrednosti* $Y_{11}, \ldots, Y_{n1}$ *in* $\tilde{Y}_{12}, \ldots \tilde{Y}_{n2}$*. Ta nabor opaženih vrednosti pa je v bijektivni linearni zvezi z izvirnim naborom* $Y_{11}, \ldots, Y_{n1}$ *in* $Y_{12}, \ldots, Y_{n2}$*, zato gre v tem primeru tudi za najboljšo linearno nepristransko cenilko v izvirnem modelu, torej tudi za najboljšo izmed vseh treh predlaganih cenilk.*

Opomba*. Pogoji izreka Gauss–Markova so izpolnjeni tudi pri točki b., a za nabor opaženih vrednosti* $Y_{11} + Y_{12}, \ldots, Y_{n1} + Y_{n2}$*. Ta nabor pa ni v v bijektivni linearni zvezi z izvirnim naborom* $Y_{11}, \ldots, Y_{n1}$ *in* $Y_{12}, \ldots, Y_{n2}$*. Cenilka iz točke b. je torej najboljša linearna nepristranska cenilka na podlagi opaženih vrednosti* $Y_{11}+Y_{12}, \ldots, Y_{n1}+Y_{n2}$*, ni pa najboljša linearna nepristranska cenilka na podlagi izvirnega nabora, ki nudi več informacije.*