IME IN PRIIMEK: ______________________ VPISNA ŠT: | | | | | | | | |

FAKULTETA ZA MATEMATIKO IN FIZIKO

Oddelek za matematiko

Statistika

Pisni izpit

7. julij 2017

Navodila

Pazljivo preberite besedilo naloge, preden se lotite reševanja. Nalog je 6, 5 rešenih nalog pa je že 100%. Na razpolago imate 2 uri.

Naloga	a.	b.	c.	d.	
1.			•	•	
2.					
3.				•	
4.			•	•	
5.				•	
6.			•	•	
Skupaj	•	•	•	•	

1. (20) Naj bodo X_1, X_2, X_3, X_4 take slučajne spremenljivke, da so X_1, $X_2 - aX_1$,$X_3 - aX_2$ in $X_4 - aX_3$ neodvisne za znan a. Predpostavite, da je $E(X_i) = 0$ in $\mathrm{var}(X_i) = 1$ za $i = 1, 2, 3, 4$.

a. (10) Izračunajte $E(X_4|X_1)$.

Namig:

$$X_4 = X_4 - aX_3 + a(X_3 - aX_2) + a^2(X_2 - aX_1) + a^3X_1 .$$

Rešitev: Označimo

$$Z_2 = X_2 - aX_1 , \quad Z_3 = X_3 - aX_2 , \quad Z_4 = X_4 - aX_3 .$$

Sledimo namigu in izračunamo

$$\begin{aligned}
&E(X_4|X_1) = \\
&\qquad = E\bigl(Z_4 + aZ_3 + a^2Z_2 + a^3X_1 \bigm| X_1\bigr) \\
&\qquad = E(Z_4|X_1) + aE(Z_3|X_1) + a^2E(Z_2|X_1) + a^3E(X_1|X_1) \\
&\qquad = E(Z_4) + aE(Z_3) + a^2E(Z_2) + a^3X_1 \\
&\qquad = E(X_4) - aE(X_3) + aE(X_3) - a^2E(X_2) + a^2E(X_2) - a^3E(X_1) + a^3X_1 \\
&\qquad = a^3X_1 .
\end{aligned}$$

Tretja vrstica sledi zaradi linearnosti, četrta zaradi neodvisnosti slučajnih spremenljivk Z_k od X_1. Na koncu upoštevamo še, da za $i = 1, 2, 3, 4$ velja $E(X_i) = 0$.

b. (10) Pokažite, da je $E(X_1X_4|X_2, X_3) = aX_3E(X_1|X_2, X_3)$.

Namig: pogojujte najprej na X_1, X_2, X_3.

Rešitev: V skladu z namigom najprej izračunamo

$$\begin{aligned}
E(X_1X_4|X_1, X_2, X_3) &= X_1E(X_4|X_1, X_2, X_3) \\
&= X_1\bigl[E(Z_4|X_1, X_2, X_3) + aE(X_3|X_1, X_2, X_3)\bigr] .
\end{aligned}$$

Vemo, da je Z_4 neodvisna od trojice (X_1, Z_2, Z_3). Toda ker je $X_2 = Z_2 + aX_1$ in $X_3 = Z_3 + aZ_2 + a^2X_1$, mora biti Z_4 neodvisna tudi od trojice (X_1, X_2, X_3) (trojici (X_1, X_2, X_3) in (X_1, Z_2, Z_3) nudita isto informacijo). Sledi $E(Z_4|X_1, X_2, X_3) = E(Z_4) = 0$ in

$$E(X_1X_4|X_1, X_2, X_3) = aX_1X_3 .$$

Zdaj pa uporabimo lastnost gnezdenja in dobimo

$$\begin{aligned}
E(X_1X_4|X_2, X_3) &= E\bigl(E(X_1X_4|X_1, X_2, X_3) \bigm| X_2, X_3\bigr) \\
&= aE(X_1X_3|X_2, X_3) \\
&= aX_3E(X_1|X_2, X_3) .
\end{aligned}$$

2. (20) Včasih je težko dobiti poštene odgovore na delikatna vprašanja kot npr. 'Ste kdaj uporabljali heroin?' ali 'Ste kdaj goljufali na izpitu?' Da bi se izboljšala pristranskost odgovorov, so uvedli metodo *randomiziranega odgovora.* Anketirancu se naključno dodeli ena izmed izjav:

(1) 'Imam lastnost A.'
(2) 'Nimam lastnosti A.'

na katero potem odgovori z DA ali NE. Anketar *ne ve*, na katero izjavo je odgovarjal anketiranec. Privzamemo:

- da anketiranci tvorijo enostavni slučajni vzorec;
- da se jim vprašanja dodeljujejo neodvisno;
- da je dodeljevanje izjav neodvisno od vzorčenja;
- da anketiranci na izjave odgovarjajo po pravici.

Naj bo:

- p verjetnost, da je bila posameznemu anketirancu dodeljena izjava (1); ta verjetnost je znana iz načrta poskusa;
- q delež oseb v populaciji, ki imajo lastnost A;
- r verjetnost, da posamezen anketiranec odgovori z DA;
- R delež anketirancev, ki so odgovorili z DA.

Zvezo 'posamezen anketiranec' je potrebno razumeti tako, da anketirance v vzorcu oštevilčimo z $1, 2, \ldots, n$, nakar za fiksen i govorimo o i-tem anketirancu.

a. (5) Izrazite r s p in q in pokažite, da je R nepristranska cenilka za r.

Rešitev: Ker gre za enostavno slučajno vzorčenje, ima vsak fiksen anketiranec lastnost A z verjetnostjo q. Ker je dodeljevanje neodvisno od vzorčenja, je dogodek, da se anketirancu dodeli izjava (1), neodvisen od dogodka, da ima anketiranec lastnost A. Zato in ker anketiranci odgovarjajo po pravici, je pogojna *verjetnost, da anketiranec odgovori z DA, če vemo, da se mu je dodelila izjava (1), enaka q; podobno je pogojna verjetnost, da anketiranec odgovori z DA, če vemo, da se mu je dodelila izjava (2), enaka* $1 - q$. *Izrek o polni verjetnosti nam tako da izražavo* $r = pq + (1-p)(1-q) = 1 - p - q + 2pq$.

Da je R nepristranska cenilka za r, vidimo tako, da izrazimo $R = \frac{1}{n}\sum_{i=1}^{n} I_i$, *kjer je* I_i *indikator dogodka, da i-ti anketiranec odgovori DA. Prej smo pokazali, da je* $E(I_i) = r$ *za vse i. Sledi* $E(R) = r$.

b. (5) Predlagajte nepristransko cenilko za q. Kdaj je to sploh možno? Izrazite varianco predlagane cenilke z $\mathrm{var}(R)$.

Rešitev: Velja

$$q = \frac{p + r - 1}{2p - 1}\,,$$

od koder dobimo cenilko

$$Q = \frac{p + R - 1}{2p - 1}\,.$$

Dokazali smo, da je R nepristranska cenilka za r. Ker se q linearno izraža z r in Q na isti način z R, je tudi Q nepristranska cenilka za q. Velja še

$$\operatorname{var}(Q) = \frac{\operatorname{var}(R)}{(2p-1)^2}\,.$$

Zgornji postopek je seveda možen le za $p \neq 1/2$. Za $p = 1/2$ pa pogojno na izbiro anketirancev vsak odgovori DA z verjetnostjo $1/2$ in anketiranci so med seboj neodvisni. To potem velja tudi brezpogojno. Torej je porazdelitev odgovorov anketirancev neodvisna od q. Kakršno koli cenilko bi konstruirali, bi bila torej njena pričakovana vrednost enaka pri vseh q, torej ne bi mogla biti enaka q za vse q. Pri $p = 1/2$ torej nepristranska cenilka ne obstaja.

c. (5) Naj bo N_A število vseh anketirancev z lastnostjo A, zajetih v vzorec, N_D pa naj bo število vseh anketirancev, zajetih v vzorec, ki na zastavljeno vprašanje odgovorijo DA. Izračunajte $E(N_D|N_A)$ in $\operatorname{var}(N_D|N_A)$.

Rešitev: Za vsako podmnožico M cele populacije označimo z N_M število elementov te množice, izbranih v vzorec. Tedaj je $N_D = N_{A\cap D} + N_{A^c\cap D}$. Pogojno na N_A sta $N_{A\cap D} \sim \operatorname{Bin}(N_A, p)$ in $N_{A^c\cap D} \sim \operatorname{Bin}(N_{A^c}, 1-p)$ neodvisni slučajni spremenljivki. Sledi

$$\begin{aligned} E(N_D|N_A) &= N_A\, p + N_{A^c}\,(1-p) = (2p-1)N_A + n(1-p)\,, \\ \operatorname{var}(N_D|N_A) &= N_A\, p(1-p) + N_{A^c}\, p(1-p) = np(1-p)\,. \end{aligned}$$

d. (5) Izračunajte $\operatorname{var}(R)$.

Rešitev: Dovolj je izračunati $\operatorname{var}(N_D)$, saj je $R = N_D/n$, torej $\operatorname{var}(R) = \operatorname{var}(N_D)/n^2$. Pišimo

$$\operatorname{var}(N_D) = E\big(\operatorname{var}(N_D|N_A)\big) + \operatorname{var}\big(E(N_D|N_A)\big)\,.$$

Iz prejšnje točke takoj dobimo

$$E\big(\operatorname{var}(N_D|N_A)\big) = np(1-p)\,.$$

Nadalje iz teorije vzorčenja vemo, da je

$$\operatorname{var}(N_A) = \frac{N-n}{N-1}\, nq(1-q)\,,$$

od koder sledi

$$\operatorname{var}\big(E(N_D|N_A)\big) = \frac{N-n}{N-1}\, n(2p-1)^2 q(1-q)\,.$$

Upoštevamo, da je $R = N_B/n$, sestavimo skupaj in dobimo

$$\operatorname{var}(R) = \frac{p(1-p)}{n} + \frac{N-n}{N-1}\,\frac{(2p-1)^2 q(1-q)}{n}\,.$$

3. (20) Predpostavite, da so opazovane vrednosti $\mathbf{x}_1, \ldots, \mathbf{x}_n$ nastale kot med sabo neodvisni, enako porazdeljeni d-razsežni vektorji $\mathbf{X}_1, \ldots, \mathbf{X}_n$ z $\mathbf{X}_k \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ za $k = 1, \ldots, n$. Parametra $\boldsymbol{\mu}$ in σ^2 sta neznana.

a. (10) Najdite cenilki parametrov $\boldsymbol{\mu}$ in σ^2 po metodi največjega verjetja.

Rešitev: Logaritemska funkcija verjetja je

$$\ell\left(\boldsymbol{\mu}, \sigma | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) = -\frac{nd}{2}\log(2\pi) - nd\log\sigma - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^T(\mathbf{x}_k - \boldsymbol{\mu})\,.$$

Če vektorje zapišemo po komponentah:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)\,, \quad \mathbf{x}_k = (x_{k1}, \ldots, x_{kd})\,,$$

dobi logaritemska funkcija verjetja obliko

$$\ell\left(\boldsymbol{\mu}, \sigma | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) = -\frac{nd}{2}\log(2\pi) - nd\log\sigma - \frac{1}{2\sigma^2}\sum_{k=1}^{n}\sum_{l=1}^{d}(x_{kl} - \mu_l)^2\,.$$

Parcialno odvajamo po $\mu_1, \ldots, \mu_d$ in σ in izenačimo z nič. Dobimo enačbe

$$-\frac{1}{\hat\sigma^2}\sum_{k=1}^{n}(x_{kl} - \hat\mu_l) = 0\,; \quad l = 1, 2, \ldots, d$$

in

$$-\frac{nd}{\hat\sigma} + \frac{1}{\hat\sigma^3}\sum_{k=1}^{n}\sum_{l=1}^{d}(x_{kl} - \hat\mu_l)^2 = 0\,.$$

Iz prve enačbe sledi $\hat\mu_l = \frac{1}{n}\sum_{k=1}^{n} x_{kl} =: \bar{x}_l$ oziroma $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k =: \bar{\mathbf{x}}$, iz druge pa

$$\hat\sigma^2 = \frac{1}{nd}\sum_{k=1}^{n}\sum_{l=1}^{d}(x_{kl} - \bar{x}_l)^2 = \frac{1}{nd}\sum_{k=1}^{n}\left(\mathbf{x}_k - \bar{\mathbf{x}}\right)^T\left(\mathbf{x}_k - \bar{\mathbf{x}}\right)\,.$$

b. (5) Popravite cenilko parametra σ^2 po metodi največjega verjetja tako, da bo nepristranska, in izračunajte varianco te nepristranske cenilke.

Namig: če so $Y_1, \ldots, Y_n$ neodvisne in enako normalno porazdeljene slučajne spremenljivke z varianco σ^2 in je $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k$, je

$$\sum_{k=1}^{n}\left(Y_k - \bar{Y}\right)^2 \sim \sigma^2\chi^2(n-1)\,.$$

Rešitev: Iz lastnosti večrazsežne normalne porazdelitve sledi, da so vse slučajne spremenljivke X_{kl}, $k = 1, \ldots, n$, $l = 1, \ldots, d$, neodvisne. Torej je vsota:

$$\sum_{k=1}^{n} \sum_{l=1}^{d} (X_{kl} - \bar{X}_l)^2 = \sum_{l=1}^{d} \sum_{k=1}^{n} (X_{kl} - \bar{X}_l)^2$$

vsota d neodvisnih slučajnih spremenljivk s porazdelitvijo $\sigma^2\chi^2(n-1)$, ta pa ima porazdelitev $\sigma^2\chi^2\big(d(n-1)\big)$. To pomeni, da je pričakovana vrednost vsote enaka $\sigma^2 d(n-1)$. Iskana nepristranska cenilka je torej

$$\tilde{\sigma}^2 = \frac{1}{d(n-1)} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})^T (\mathbf{x}_k - \bar{\mathbf{x}}) = \frac{n}{n-1}\, \hat{\sigma}^2 \,.$$

Isti argument uporabimo za izračun variance: varianca vsote je $2\sigma^4 d(n-1)$, varianca iskane cenilke pa je enaka

$$\mathrm{var}(\tilde{\sigma}^2) = \frac{2\sigma^4}{d(n-1)} \,.$$

c. (5) Navedite eksakten interval zaupanja za parameter μ_1 pri stopnji tveganja $\alpha \in (0, 1)$.

Rešitev: Iz predpostavk sledi, da je $\tilde{\sigma}^2$ neodvisna od $\bar{\mathbf{X}}$. Po definiciji Studentove t-porazdelitve so komponente kvocienta

$$\frac{\sqrt{n}\, \bar{\mathbf{X}}}{\tilde{\sigma}}$$

porazdeljene po Studentovem t-zakonu z $m = d(n-1)$ prostostnimi stopnjami. Eksakten interval zaupanja bo torej določen s krajiščema

$$\hat{\mu}_1 \pm t_\alpha \cdot \frac{\tilde{\sigma}}{\sqrt{n}} \,,$$

kjer je t_α tak, da je

$$P\left(-t_\alpha \le T \le t_\alpha\right) = \alpha \,,$$

T pa je slučajna spremenljivka, ki ima Studentovo t-porazdelitev z m prostostnimi stopnjami.

4. (20) Predpostavite, da so opazovane vrednosti $\mathbf{x}_1, \ldots, \mathbf{x}_n$ nastale kot med sabo neodvisni, enako porazdeljeni d-razsežni vektorji $\mathbf{X}_1, \ldots, \mathbf{X}_n$ z $\mathbf{X}_k \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ za $k = 1, \ldots, n$. Parametra $\boldsymbol{\mu}$ in σ^2 sta neznana. Preizkusiti želimo domnevo

$$H_0\colon \boldsymbol{\mu}^T \boldsymbol{\mu} = 1 \qquad \text{proti} \qquad H_1\colon \boldsymbol{\mu}^T \boldsymbol{\mu} \neq 1\,.$$

a. (10) Najdite testno statistiko po metodi kvocienta verjetij in navedite njeno aproksimativno porazdelitev.

Rešitev: Logaritemska funkcija verjetja je

$$\ell\left(\boldsymbol{\mu}, \sigma | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) = -\frac{nd}{2} \log(2\pi) - nd \log \sigma - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})^T (\mathbf{x}_k - \boldsymbol{\mu})\,.$$

Če vektorje zapišemo po komponentah:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)\,, \quad \mathbf{x}_k = (x_{k1}, \ldots, x_{kd})\,,$$

dobi logaritemska funkcija verjetja obliko

$$\ell\left(\boldsymbol{\mu}, \sigma | \mathbf{x}_1, \ldots, \mathbf{x}_n\right) = -\frac{nd}{2} \log(2\pi) - nd \log \sigma - \frac{1}{2\sigma^2} \sum_{k=1}^{n} \sum_{l=1}^{d} (x_{kl} - \mu_l)^2\,.$$

V širšem modelu funkcijo verjetja maksimiziramo tako, da parcialno odvajamo po $\mu_1, \ldots, \mu_d$ in σ ter izenačimo z nič. Dobimo enačbe

$$-\frac{1}{\hat{\sigma}^2} \sum_{k=1}^{n} (x_{kl} - \hat{\mu}_l) = 0\,; \quad l = 1, 2, \ldots, d$$

in

$$-\frac{nd}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3} \sum_{k=1}^{n} \sum_{l=1}^{d} (x_{kl} - \hat{\mu}_l)^2 = 0\,.$$

Iz prve enačbe sledi $\hat{\mu}_l = \frac{1}{n} \sum_{k=1}^{n} x_{kl} =: \bar{x}_l$ oziroma $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k =: \bar{\mathbf{x}}$, iz druge pa

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{k=1}^{n} \sum_{l=1}^{d} (x_{kl} - \bar{x}_l)^2 = \frac{1}{nd} \sum_{k=1}^{n} \left(\mathbf{x}_k - \bar{\mathbf{x}}\right)^T \left(\mathbf{x}_k - \bar{\mathbf{x}}\right)\,.$$

Ocena je v skladu z našim modelom, brž ko je $\hat{\sigma}^2 > 0$, to pa je takrat, ko sta vsaj dve opaženi vrednosti $\mathbf{x}_1, \ldots, \mathbf{x}_n$ različni.

V ožjem modelu, ko imamo omejitev $\boldsymbol{\mu}^T \boldsymbol{\mu} = 1$ oziroma $\sum_{l=1}^{d} \mu_l^2 = 1$, gre na vsaj dva načina.

Prvi način. *Nastavimo Lagrangeovo funkcijo z multiplikatorjem a:*

$$F = -\frac{nd}{2} \log(2\pi) - nd \log \sigma - \frac{1}{2\sigma^2} \sum_{k=1}^{n} \sum_{l=1}^{d} (x_{kl} - \mu_l)^2 - a \sum_{l=1}^{d} \mu_l^2\,.$$

Če cenilki v tem primeru označimo s $\tilde{\sigma}$ in $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$, po odvajanju dobimo enačbe

$$-\frac{1}{\tilde{\sigma}^2}\sum_{k=1}^n (x_{kl} - \tilde{\mu}_l) - a\tilde{\mu}_l = 0\,; \quad l = 1, 2, \ldots, d$$

in

$$-\frac{nd}{\tilde{\sigma}} + \frac{1}{\tilde{\sigma}^3}\sum_{k=1}^n \sum_{l=1}^d (x_{kl} - \tilde{\mu}_l)^2 = 0\,.$$

Prvo serijo enačb prepišemo v vektorski obliki

$$(n - a\tilde{\sigma}^2)\tilde{\boldsymbol{\mu}} = n\bar{\mathbf{x}}\,.$$

Ta enačba pa je v našem kontekstu ekvivalentna zahtevi, da sta $\tilde{\boldsymbol{\mu}}$ in $\bar{\mathbf{x}}$ kolinearna. Brž ko namreč obstaja ustrezni $\tilde{\sigma} > 0$, obstaja tudi ustrezni Lagrangeov multiplikator a in z njim faktor v enačbi.

Če je $\bar{\mathbf{x}} \neq 0$, sta $\tilde{\boldsymbol{\mu}}$ in $\bar{\mathbf{x}}$ kolinearna, če je

$$\tilde{\boldsymbol{\mu}} = \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}} \quad \text{ali} \quad \tilde{\boldsymbol{\mu}} = -\frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\,.$$

Prava cenilka pa je le prva, saj za vsak $\sigma > 0$ velja:

$$\begin{aligned}
&\ell\left(\frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\,,\, \sigma \;\middle|\; \mathbf{x}_1, \ldots, \mathbf{x}_n\right) - \ell\left(-\frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\,,\, \sigma \;\middle|\; \mathbf{x}_1, \ldots, \mathbf{x}_n\right) \\
&\quad = \frac{1}{2\sigma^2}\sum_{k=1}^n \left[\left(\mathbf{x}_k + \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right)^T \left(\mathbf{x}_k + \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right) - \left(\mathbf{x}_k - \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right)^T \left(\mathbf{x}_k - \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right)\right] \\
&\quad = \frac{1}{\sigma^2\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\sum_{k=1}^n \left(\mathbf{x}_k^T\bar{\mathbf{x}} + \bar{\mathbf{x}}^T\mathbf{x}_k\right) \\
&\quad = \frac{2n\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}{\sigma^2} \\
&\quad > 0\,.
\end{aligned}$$

Iz enačbe z odvodom po σ dobimo še cenilko za ta parameter:

$$\tilde{\sigma}^2 = \frac{1}{nd}\sum_{k=1}^n (\mathbf{x}_k - \tilde{\boldsymbol{\mu}})^T(\mathbf{x}_k - \tilde{\boldsymbol{\mu}})$$

in ta je v skladu z našim modelom (torej strogo pozitivna), brž ko sta vsaj dve opaženi vrednosti različni ali pa so vse enake, a ne enotske.

Če pa je $\bar{\mathbf{x}} = 0$, sta $\tilde{\boldsymbol{\mu}}$ in $\bar{\mathbf{x}}$ kolinearna pri poljubnem $\tilde{\boldsymbol{\mu}}$. Vrednost cenilke $\tilde{\sigma}$ pa je neodvisna od $\tilde{\boldsymbol{\mu}}$, saj je v tem primeru

$$\tilde{\sigma}^2 = \frac{1}{nd}\left[\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k - \sum_{k=1}^n \mathbf{x}_k^T\tilde{\boldsymbol{\mu}} - \sum_{k=1}^n \tilde{\boldsymbol{\mu}}^T\mathbf{x}_k + n\tilde{\boldsymbol{\mu}}^T\tilde{\boldsymbol{\mu}}\right] = \frac{1}{nd}\sum_{k=1}^n \mathbf{x}_k^T\mathbf{x}_k + \frac{1}{d}\,.$$

Drugi način. *Opazimo, da je, če je verjetje maksimalno, izraz*

$$\sum_{k=1}^{n} (\mathbf{x}_k - \boldsymbol{\mu})^T (\mathbf{x}_k - \boldsymbol{\mu})$$

minimalen. Ta izraz se ob upoštevanju omejitve poenostavi v

$$\sum_{k=1}^{n} \mathbf{x}_k^T \mathbf{x}_k - 2n\boldsymbol{\mu}^T \bar{\mathbf{x}} + n\,.$$

To pomeni, da moramo maksimizirati $\boldsymbol{\mu}^T \bar{\mathbf{x}}$, seveda pri danem pogoju. Če je $\bar{\mathbf{x}} \neq 0$, maksimum nastopi pri

$$\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} := \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T \bar{\mathbf{x}}}}\,,$$

če pa je $\bar{\mathbf{x}} = 0$, je izraz neodvisen od $\boldsymbol{\mu}$. Dobili smo torej isto kot pri prvem načinu. Vstavimo v verjetje, maksimiziramo še po σ in spet dobimo isto cenilko kot pri prvem načinu.

Wilksova testna statistika λ se izraža z maksimumoma logaritemske funkcije verjetja v širšem in ožjem modelu. Opazimo, da je ta maksimum tako v širšem kot tudi v ožjem modelu enak

$$\frac{nd}{2}\log(2\pi) - nd\log\sigma - \frac{nd}{2}\,,$$

kjer σ nadomestimo z ustrezno cenilko. Iz tega lahko izrazimo Wilksovo testno statistiko λ kot

$$\lambda = nd\log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)\,.$$

Omejitev odvzame eno dimenzijo, zato bo aproksimativna porazdelitev testne statistike $\chi^2(1)$.

Opomba. *Wilksova testna statistika je nedefinirana, če so vse opažene vrednosti $\mathbf{x}_1, \ldots, \mathbf{x}_n$ enake. Toda ta dogodek ima po modelu vselej verjetnost nič.*

b. (10) Privzemite, da je σ^2 znana in da je $\sigma^2 = 1$. Pokažite, da je v tem primeru

$$\lambda = n\left(1 - \sqrt{\bar{\mathbf{x}}^T \bar{\mathbf{x}}}\right)^2\,.$$

Kot znano privzemite, da je porazdelitvena funkcija slučajne spremenljivke $\sqrt{\bar{\mathbf{x}}^T \bar{\mathbf{x}}}$, če H_0 velja, enaka $F(x)$, kjer je F zvezna funkcija z $F(0) = 0$, ki je strogo monotono naraščajoča za $x \geq 0$. Obrazložite, kako bi s pomočjo $F(x)$ našli tak λ_α, da bi pri dani stopnji tveganja $\alpha \in (0, 1)$ veljalo

$$P_{H_0}(\lambda > \lambda_\alpha) = \alpha\,.$$

Rešitev: Iz računov prvega dela sledi, da sta oceni za $\boldsymbol{\mu}$ *tudi v spremenjenem širšem in ožjem modelu enaki* $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ *in* $\tilde{\boldsymbol{\mu}} = \bar{\mathbf{x}}/\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}$*. Vstavimo v funkcijo verjetja in dobimo*

$$\begin{aligned}
\lambda &= 2\Big(\ell(\hat{\boldsymbol{\mu}}, 1|\mathbf{x}_1, \ldots, \mathbf{x}_n) - \ell(\tilde{\boldsymbol{\mu}}, 1|\mathbf{x}_1, \ldots, \mathbf{x}_n)\Big) \\
&= -\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \sum_{k=1}^{n}\left(\mathbf{x}_k - \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right)^T\left(\mathbf{x}_k - \frac{\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\right) \\
&= \sum_{k=1}^{n}\mathbf{x}_k^T\bar{\mathbf{x}} + \sum_{k=1}^{n}\bar{\mathbf{x}}^T\mathbf{x}_k - n\,\bar{\mathbf{x}}^T\bar{\mathbf{x}} - \frac{1}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\sum_{k=1}^{n}\mathbf{x}_k^T\bar{\mathbf{x}} - \frac{1}{\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}}\sum_{k=1}^{n}\bar{\mathbf{x}}^T\mathbf{x}_k + \frac{n}{\bar{\mathbf{x}}^T\bar{\mathbf{x}}}\,\bar{\mathbf{x}}^T\bar{\mathbf{x}} \\
&= n\left(\bar{\mathbf{x}}^T\bar{\mathbf{x}} - 2\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}} + 1\right) \\
&= n\left(\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}} - 1\right)^2
\end{aligned}$$

(tokrat je ta statistika definirana za vsak nabor opaženih vrednosti $\mathbf{x}_1, \ldots, \mathbf{x}_n$*). Ničelno domnevo zavrnemo, če je* $\lambda > \lambda_\alpha$*, kar je natanko tedaj, ko je*

$$\sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}} < 1 - \sqrt{\frac{\lambda_\alpha}{n}} \quad \textit{ali} \quad \sqrt{\bar{\mathbf{x}}^T\bar{\mathbf{x}}} > 1 + \sqrt{\frac{\lambda_\alpha}{n}}\,.$$

Izbrati moramo tak λ_α*, da bo zgornja neenakost veljala z verjetnostjo* α*, če* H_0 *drži. Verjetnost zgornjega dogodka je enaka*

$$F\left(1 - \sqrt{\frac{\lambda_\alpha}{n}}\right) + 1 - F\left(1 + \sqrt{\frac{\lambda_\alpha}{n}}\right)\,.$$

Zgornji izraz je zvezna in strogo padajoča funkcija spremenljivke $\lambda_\alpha \geq 0$ *(čeprav sama* F *ni povsod strogo naraščajoča). Nadalje je zgornji izraz pri* $\lambda_\alpha = 0$ *enak* 1 *in gre proti nič, ko gre* λ_α *proti neskončno. Zato obstaja natanko en* λ_α*, za katerega je ta izraz enak* α*, in to je zahtevani prag za eksakten test.*

5. (20) Dan je statistični model

$$Y_k = a + bx_k + \epsilon_k\,; \quad k = 0, 1, \ldots, n\,,$$

kjer so $x_0, \ldots, x_n$ konstante, $Y_0, \ldots, Y_n$ opažene vrednosti, a in b parametra, $\epsilon_0, \ldots, \epsilon_n$ pa šumi. Za slednje privzamemo, da je

$$E(\epsilon_k) = 0 \quad \text{in} \quad \operatorname{cov}(\epsilon_k, \epsilon_l) = \rho^{|k-l|}\sigma^2$$

za vse $k, l \in \{0, 1, \ldots, n\}$; tu je $\rho \in (-1, 1)$ še ena konstanta modela, $\sigma > 0$ pa še en parameter modela.

a. (10) Dokažite, da obstaja taka konstanta $\gamma \in \mathbb{R}$, da

$$Z_k = Y_k + \gamma Y_{k-1}\,; \quad k = 1, 2, \ldots, n$$

ustrezajo standardnemu linearnemu regresijskemu modelu. Zapišite matriko tega modela.

Rešitev: Zapišimo

$$Z_k = (1+\gamma)a + bx_k + \gamma b x_{k-1} + \eta_k\,,$$

kjer je $\eta_k = \epsilon_k + \gamma\epsilon_{k-1}$. *To ustreza standardnemu linearnemu regresijskemu modelu, če imajo* $\eta_1, \ldots, \eta_n$ *enake variance in če so paroma nekorelirane. Za* $k = 1, 2, \ldots, n$ *izračunajmo*

$$E(\eta_k) = E(\epsilon_k) + \gamma E(\epsilon_{k-1}) = 0$$

in

$$\begin{aligned}\operatorname{var}(\eta_k) &= \operatorname{var}(\epsilon_k) + 2\gamma\operatorname{cov}(\epsilon_k, \epsilon_{k-1}) + \gamma^2\operatorname{var}(\epsilon_{k-1})\\ &= \big(1 + 2\gamma\rho + \gamma^2\big)\sigma^2\,,\end{aligned}$$

za $1 \le k < l \le n$ *pa izračunajmo*

$$\begin{aligned}\operatorname{cov}(\eta_k, \eta_l) &= \operatorname{cov}(\epsilon_k, \epsilon_l) + \gamma\operatorname{cov}(\epsilon_k, \epsilon_{l-1}) + \gamma\operatorname{cov}(\epsilon_{k-1}, \epsilon_l) + \gamma^2\operatorname{cov}(\epsilon_{k-1}, \epsilon_{l-1})\\ &= (1+\gamma\rho)(\gamma+\rho)\rho^{l-k-1}\sigma^2\,.\end{aligned}$$

Za $\gamma = -\rho$ *in za* $\gamma = -1/\rho$ *model ustreza vsem predpostavkam in ima matrični zapis*

$$\boldsymbol{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}\,,$$

kjer je:

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1+\gamma & x_1 + \gamma x_0 \\ 1+\gamma & x_2 + \gamma x_1 \\ \vdots & \vdots \\ 1+\gamma & x_n + \gamma x_{n-1} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}.$$

b. (5) Model iz prejšnje točke ima najboljši nepristranski linearni cenilki za a in b. Izračunajte varianci teh dveh cenilk za posebni primer, ko je $n = 2$ in $x_k = k$ za $k = 0, 1, 2$.

Rešitev: Varianci cenilk, ki ju dobimo na podlagi modela iz prejšnje točke, sta diagonalca matrike:

$$(1 + 2\gamma\rho + \gamma^2)\sigma^2 \big(\mathbf{X}^T\mathbf{X}\big)^{-1} .$$

Matrika $\mathbf{X}$ *je obrnljiva, zato lahko izračunamo kar:*

$$\mathbf{X}^{-1} = \frac{1}{(1+\gamma)^2} \begin{bmatrix} 2+\gamma & -1 \\ -1-\gamma & 1+\gamma \end{bmatrix} ,$$

$$\big(\mathbf{X}^T\mathbf{X}\big)^{-1} = \mathbf{X}^{-1}\mathbf{X}^{-T} = \frac{1}{(1+\gamma)^4} \begin{bmatrix} 5+4\gamma+\gamma^2 & -(3+\gamma)(1+\gamma) \\ (3+\gamma)(1+\gamma) & -2(1+\gamma)^2 \end{bmatrix} .$$

Torej je

$$\operatorname{var}(\hat{a}) = \frac{(1+2\gamma\rho+\rho^2)(5+4\gamma+\gamma^2)}{(1+\gamma)^4}\,\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = \frac{2(1+2\gamma\rho+\rho^2)}{(1+\gamma)^2}\,\sigma^2 .$$

Končni rezultat je odvisen od tega, kateri γ *izberemo. Pri* $\gamma = -\rho$ *pride:*

$$\operatorname{var}(\hat{a}) = \frac{(1+\rho)(5-4\rho+\rho^2)}{(1-\rho)^3}\,\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = \frac{2(1+\rho)}{1-\rho}\,\sigma^2 ,$$

pri $\gamma = -1/\rho$ *pa pride:*

$$\operatorname{var}(\hat{a}) = \frac{(1+\rho)(1-4\rho+5\rho^2)}{(1-\rho)^3}\,\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = \frac{2(1+\rho)}{1-\rho}\,\sigma^2 .$$

c. (5) Cenilki iz prejšnje točke sta najboljši linearni nepristranski cenilki na podlagi vrednosti $Z_1, Z_2, \ldots, Z_n$. Sta to nujno najboljši nepristranski linearni cenilki tudi v prvotnem modelu, se pravi na podlagi prvotnih opaženih vrednosti $Y_0, Y_1, \ldots, Y_n$?

Namigi:

- *Se dajo izvirna opažanja* $Y_0, Y_1, \ldots, Y_n$ *še kako drugače transformirati?*
- *Razmislite za posebni primer iz prejšnje točke.*
- $\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}^{-1} = \dfrac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 1+\rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix}.$

Rešitev: Transformacija iz točke a. ne ohrani vse informacije o izvirnih opažanjih $X_0, X_1, \ldots, X_n$*, saj slika v prostor nižje dimenzije. Zato lahko upravičeno sumimo, da cenilki iz prejšnje točke nista najboljši. Transformacija, ki ohrani vso* *informacijo, obenem pa izvirni model prevede na standardni linearni regresijski model, je recimo:*

$$\tilde{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}\,,$$

kjer je $\boldsymbol{\Sigma}$ *kovariančna matrika izvirnega modela:*

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^n \\ \rho & 1 & \rho & \cdots & \rho^{n-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^n & \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}.$$

Varianci cenilk iz tega modela sta diagonalca matrike:

$$\sigma^2\Big(\big(\boldsymbol{\Sigma}^{-1/2}\mathbf{W}\big)^T\big(\boldsymbol{\Sigma}^{-1/2}\mathbf{W}\big)\Big)^{-1} = \sigma^2\big(\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\big)^{-1},$$

kjer je

$$\mathbf{W} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Za predlagani posebni primer po krajšem računu dobimo

$$\big(\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\big)^{-1} = \frac{(1+\rho)\sigma^2}{2(3-\rho)}\begin{bmatrix} 5-4\rho+\rho^2 & -(1-\rho)(3-\rho) \\ -(1-\rho)(3-\rho) & (1-\rho)(3-\rho) \end{bmatrix}.$$

Sum, da cenilki iz prejšnje točke nista vedno najboljši, je dovolj potrditi na določenem ρ*. Pri* $\rho = 1/2$ *varianci cenilk iz te točke prideta*

$$\operatorname{var}(\hat{a}) = \tfrac{39}{40}\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = \tfrac{3}{8}\sigma^2\,.$$

Varianci iz prejšnje točke pa pri $\gamma = -1/2$ *prideta*

$$\operatorname{var}(\hat{a}) = 13\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = 4\sigma^2\,,$$

pri $\gamma = -2$ *pa prideta*

$$\operatorname{var}(\hat{a}) = 9\sigma^2 \quad \textit{in} \quad \operatorname{var}(\hat{b}) = 18\sigma^2\,.$$

Za obe izbiri imata obe cenilki iz prejšnje točke strogo večjo varianco, torej res nista vedno najboljši.

6. (20) Berti odpre stojnico z igro s tremi kockami. V vsaki igri, ki stane 1 euro, se vse tri kocke vržejo. Če ne pade nobena šestica, Berti obdrži vplačani znesek. Če pade natanko ena šestica, Berti igralcu vrne vplačani znesek in še en euro. Če padeta natanko dve šestici, Berti igralcu vrne vplačani znesek in še dva eura. Če pa padejo tri šestice, Berti igralcu vrne vplačani znesek in še 14 eurov. Privzamemo, da so kocke standardne in da so vsi meti neodvisni.

a. (10) Izračunajte pričakovano vrednost in varianco Bertijevega dobička po n igrah.

Rešitev: Naj bo X_i Bertijev dobiček v i-ti igri. Velja:

$$X_i \sim \begin{pmatrix} -14 & -2 & -1 & 1 \\ \frac{1}{216} & \frac{15}{216} & \frac{75}{216} & \frac{125}{216} \end{pmatrix},$$

od koder po krajšem računu dobimo $E(X_i) = 1/36$ in $\mathrm{var}(X_i) = 2735/1296$. Če torej z S_n označimo Bertijev dobiček po n igrah, velja $E(S_n) = n/36$ in $\mathrm{var}(S_n) = 2735n/1296$.

b. (10) Po približno koliko igrah ima Berti s približno 95% verjetnostjo pozitiven dobiček?

Rešitev: Označimo spet število iger z n. Iz centralnega limitnega izreka sledi, da mora približno veljati

$$1 - \Phi\left(\frac{-\frac{1}{36}\,n}{\sqrt{\frac{2735}{1296}\,n}}\right) = \Phi\left(\frac{\sqrt{n}}{\sqrt{2735}}\right) = 0{,}95$$

oziroma

$$\frac{\sqrt{n}}{\sqrt{2735}} \doteq 1{,}645$$

kar je res, če je n približno 7400.