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

FAKULTETA ZA MATEMATIKO IN FIZIKO

ODDELEK ZA MATEMATIKO

STATISTIKA

PISNI IZPIT

28. AVGUST 2020

NAVODILA

Pazljivo preberite besedilo naloge, preden se lotite reševanja. Na razpolago imate 120 minut.

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

1. (25) Denimo, da se cena določenega vrednostnega papirja v enem dnevu z verjetnostjo 30% dvigne za 1% trenutne cene, z verjetnostjo 30% pade za 1% trenutne cene, z verjetnostjo 40% pa ostane nespremenjena. Spremembe cen med zaporednimi dnevi so med seboj neodvisne.

a. (10) Utemeljite, da gre verjetnost, da je cena po n dneh večja od izhodiščne, proti nič, ko gre n proti neskončno.

Namig: cene primerno transformirajte.

Rešitev: Naj bodo $Q_1, Q_2, \ldots$ faktorji, ki predstavljajo spremembe cene v posameznih dneh. Velja:

$$Q_k \sim \begin{pmatrix} 0{,}99 & 1 & 1{,}01 \\ 0{,}3 & 0{,}4 & 0{,}3 \end{pmatrix}.$$

Spremembo cene po n dneh predstavlja faktor $Q_1 Q_2 \cdots Q_n$. Logaritem tega faktorja je vsota logaritmov slučajnih spremenljivk Q_k. Ti logaritmi so neodvisni in imajo porazdelitev:

$$\ln Q_k \sim \begin{pmatrix} -0{,}010050336 & 0 & 0{,}009950331 \\ 0{,}3 & 0{,}4 & 0{,}3 \end{pmatrix}.$$

Njihova pričakovana vrednost je enaka $\mu_1 \doteq -3{,}00015 \cdot 10^{-5}$, njihova varianca pa je enaka $\sigma_1^2 \doteq 6{,}00046 \cdot 10^{-5}$.

Dogodek, da je cena po n dneh višja od izhodiščne, lahko zapišemo tudi kot $\{M_n > 0\}$, kjer je M_n povprečje slučajnih spremenljivk $\log Q_1, \log Q_2, \ldots, \log Q_n$. Lahko pa ga zapišemo tudi kot $\{M_n - \mu_1 > -\mu_1\}$. Ker je $\mu_1 < 0$, gre ta verjetnost po šibkem zakonu velikih števil proti nič, ko gre n proti neskončno.

Lahko pa sklepamo tudi iz centralnega limitnega izreka, če poznamo ustrezno enakomerno omejenost napak. Verjetnost, da je cena po n dneh višja, za velike n enaka približno:

$$1 - \Phi\left(-\frac{\mu_1}{\sigma_1}\sqrt{n}\right)$$

(kjer je σ_1 seveda pozitiven). Ker je μ_1 negativen, gre slednji izraz pa gre proti nič, ko gre n proti neskončno.

b. (10) Izračunajte število dni, po katerih je verjetnost, da bo cena višja od izhodiščne, približno enaka 5%.

Rešitev: Če iskano število dni označimo z n, mora po centralnem limitnem izreku približno veljati:

$$1 - \Phi\left(-\frac{\mu_1}{\sigma_1}\sqrt{n}\right) = 0{,}05$$

oziroma

$$-\frac{\mu_1}{\sigma_1}\sqrt{n} = 1{,}645$$

oziroma

$$n = 1{,}645^2\,\frac{\sigma_1^2}{\mu_1^2} \doteq 180365\,.$$

Če bi se trgovalo vsak dan, bi to zneslo približno 494 let.

c. (5) Izračunajte približno *mediano* padca po številu dni iz prejšnje točke.

Rešitev: Če označimo $S_n = \ln Q_1 + \ln Q_2 + \cdots + \ln Q_n$*, iščemo tisti* s*, za katerega približno velja* $P(S_n < s) = \frac{1}{2}$*, torej približno*

$$\Phi\left(\frac{s - n\mu_1}{\sigma_1\sqrt{n}}\right) = \frac{1}{2}\,,$$

to pa je res za $s = m\mu_1 \doteq -5{,}411231$*. To odgovarja faktorju* $e^s \doteq 0{,}004466138$*, torej padcu na približno* $0{,}4466\%$ *prvotne vrednosti oziroma za faktor* $e^{-s} \doteq 223{,}9$*.*

2. (25) Predpostavite, da vsaki enoti populacije velikosti N pripadata dve vrednosti statističnih spremenljivk. Enotam torej pripadajo pari $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Oceniti želimo povprečje vseh vrednosti

$$\lambda = \frac{1}{2N} \sum_{k=1}^{N} (x_k + y_k)\,.$$

Če k-to enoto izberemo v vzorec, bo navedla vrednost x_k z verjetnostjo $\frac{1}{2}$ in vrednost y_k z verjetnostjo $\frac{1}{2}$, neodvisno od postopka vzorčenja in neodvisno med enotami. Anketarji ne vedo ali je enota navedla vrednost x_k ali y_k. Privzemimo, da izberemo enostavni slučajni vzorec n enot, parameter λ pa ocenimo z vzorčnim povprečjem. Opisano cenilko lahko zapišemo kot

$$\hat{\lambda} = \frac{1}{n} \sum_{k=1}^{N} I_k \left(x_k J_k + y_k (1 - J_k) \right)\,,$$

kjer je I_k indikator tega, da k-to enoto izberemo v vzorec, J_k pa indikator dogodka, da je k-ta enota navedla vrednost x_k.

a. (5) Pokažite, da je vzorčno povprečje nepristranska cenilka λ.

Rešitev: Iz predpostavk sledi, da je vektor $(I_1, \ldots, I_N)$ neodvisen od vektorja $(J_1, \ldots, J_N)$ in so indikatorji $J_1, \ldots, J_N$ neodvisni med sabo. Vemo, da je $E(I_k) = \frac{n}{N}$, po predpostavki pa je $P(J_k = 1) = \frac{1}{2}$. Sledi

$$E(\hat{\lambda}) = \frac{1}{n} \sum_{k=1}^{N} E(I_k) \left(x_k E(J_k) + y_k E(1 - J_k) \right) = \lambda\,.$$

b. (10) Izračunajte

$$E(\hat{\lambda} | I_1, I_2, \ldots, I_n) \qquad \text{in} \qquad \mathrm{var}(\hat{\lambda} | I_1, \ldots, I_N)\,.$$

Rešitev: Iz predpostavk o neodvisnosti sledi

$$E(\hat{\lambda} | I_1, I_2, \ldots, I_n) = \frac{1}{n} \sum_{k=1}^{N} I_k \left(\frac{x_k + y_k}{2} \right)\,.$$

in

$$\mathrm{var}(\hat{\lambda} | I_1, \ldots, I_N) = \frac{1}{n^2} \sum_{k=1}^{N} I_k \left(\frac{(x_k - y_k)^2}{4} \right)\,.$$

c. (5) Označite $z_k = (x_k + y_k)/2$ in $w_k = (x_k - y_k)^2/4$. Izrazite standardno napako cenilke $\hat{\lambda}$ s populacijsko varianco σ_z^2 vrednosti $z_1, z_2, \ldots, z_N$ in povprečjem θ vrednosti $w_1, w_2, \ldots, w_N$.

Rešitev: Iz poglavja o enostavnem slučajnem vzorčenju sledi, da je

$$\operatorname{var}(E(\hat{\lambda}|I_1, I_2, \ldots, I_n)) = \frac{\sigma_z^2}{n} \cdot \frac{N-n}{N-1}\,.$$

Po drugi strani je

$$E\left(\operatorname{var}(\hat{\lambda}|I_1, \ldots, I_N)\right) = \frac{\theta}{n}\,.$$

Po formuli za razcep variance je

$$\operatorname{var}(\hat{\lambda}) = \frac{\sigma_z^2}{n} \cdot \frac{N-n}{N-1} + \frac{\theta}{n}\,.$$

d. (5) Ali v praksi lahko nepristransko ocenimo standardno napako cenilke $\hat{\lambda}$? Podajte samo mnenje.

Rešitev: Morali bi nepristransko oceniti σ_z^2 in θ. Tega ne moremo, saj kot opazovalec ne vidimo nobenega z_k in nobenega w_k.

3. (25) Opazovane vrednosti naj bodo $x_1, x_2, \ldots, x_m$ in $y_1, y_2, \ldots, y_n$. Predpostavljamo, da so opazovane vrednosti nastale kot med sabo neodvisne slučajne spremenljivke $X_1, X_2, \ldots, X_m$ in $Y_1, Y_2, \ldots, Y_n$ z $X_k \sim \exp(\mu)$ za $k = 1, 2, \ldots, m$ in $Y_k \sim \exp(\nu)$ za $k = 1, 2, \ldots, n$. Preizkusiti želimo domnevo

$$H_0 \colon \mu = \nu \qquad \text{proti} \qquad H_1 \colon \mu \neq \nu\,.$$

Pri tem sta $\mu, \nu > 0$.

a. (15) Najdite testno statistiko po metodi kvocienta verjetij in navedite njeno aproksimativno porazdelitev v primeru, ko H_0 drži.

Rešitev: Logaritemska funkcija verjetja je

$$\ell(\mu, \nu | \mathbf{x}, \mathbf{y}) = m \log \mu - \mu \sum_{k=1}^{m} x_k + n \log \nu - \nu \sum_{k=1}^{n} y_k\,.$$

Če sta μ in ν prosta, je maksimum dosežen pri

$$\hat{\mu} = \frac{m}{\sum_{k=1}^{m} x_k} = \frac{1}{\bar{x}} \quad \mathit{in} \quad \hat{\nu} = \frac{n}{\sum_{k=1}^{n} y_k} = \frac{1}{\bar{y}}\,.$$

Vstavljanje v logaritemsko funkcijo verjetja nam da

$$\ell(\hat{\nu}, \hat{\mu} | \mathbf{x}, \mathbf{y}) = m \log \hat{\mu} - m + n \log \hat{\nu} - n\,.$$

Pri pogoju $\nu = \mu$ je račun podoben in dobimo

$$\tilde{\mu} = \tilde{\nu} = \frac{\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} y_k}{m + n}$$

in

$$l(\tilde{\mu}, \tilde{\nu} | \mathbf{x}, \mathbf{y}) = (m + n) \log \tilde{\mu} - m - n\,.$$

Sledi

$$\lambda = 2m \log \hat{\mu} + 2n \log \hat{\nu} - 2(m + n) \log \tilde{\mu}\,.$$

Aproksimativna porazdelitev testne statistike λ je $\chi^2(1)$.

b. (5) Utemeljite, da v primeru, ko H_0 drži, velja

$$U := \frac{\bar{Y}}{\bar{X}} \sim F_{2n,2m}\,.$$

Rešitev: Če ničelna domneva drži, torej če je $\mu = \nu$, velja

$$\bar{X} \sim \Gamma(m, m\mu) \quad \mathit{in} \quad \bar{Y} \sim \Gamma(n, n\mu)\,.$$

Iz tega sledi, da je

$$2m\mu\bar{X} \sim \Gamma(m, \tfrac{1}{2}) = \chi^2(2m) \quad \mathit{in} \quad 2n\mu\bar{Y} \sim \Gamma(n, \tfrac{1}{2}) = \chi^2(2n)$$

in posledično

$$U = \frac{\bar{Y}}{\bar{X}} \sim F_{2n,2m}\,.$$

c. (5) Pokažite, da je eksaktno kritično območje po metodi kvocienta verjetij za zgornjo domnevo pri dani stopnji tveganja $\alpha \in (0,1)$ oblike $C_\alpha = \{u\colon f(u) > c\} = (0,a) \cup (b,\infty)$ za $a < 1 < b$ za tako konstanto c, da je $P_{H_0}(f(U) > c) = \alpha$, pri čemer je

$$f(u) = 2(m+n)\log\left(\frac{m+nu}{m+n}\right) - 2n\log u$$

Kot znano privzemite, da je funkcija f padajoča na $(0,1)$, naraščajoča na $(1,\infty)$ in je $\lim_{u\to\infty} f(u) = \infty$.

Rešitev: Kvocient verjetja prepišemo v

$$\lambda = 2(m+n)\log\left(\frac{m+nU}{m+n}\right) - 2n\log U = f(U)\,.$$

Domnevo zavrnemo, ko je $\lambda > \lambda_\alpha$. To se zgodi natanko takrat, ko je U v opisanem kritičnem območju. Da je kritično območje dane oblike, sledi iz lastnosti funkcije f.

4. (25) Dana imamo dva kompleta regresijskih enačb

$$Y_i = \alpha_1 + \beta x_i + \epsilon_i$$

za $i = 1, 2, \ldots, m$ in

$$Z_j = \alpha_2 + \beta w_j + \eta_j$$

za $j = 1, 2, \ldots, n$. Predpostavljamo, da vsak model zase ustreza standardnim predpostavkam regresijskega modela, torej, da za vse i, j velja $E(\epsilon_i) = E(\eta_j) = 0$, $\mathrm{var}(\epsilon_i) = \mathrm{var}(\eta_i) = \sigma^2$ in so tako ϵ_i kot tudi η_j med sabo nekorelirani. Predpostavljamo tudi, da so vsi ϵ_i in η_j nekorelirani.

a. (5) Utemeljite, da lahko modela združimo v

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \\ Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_m \\ 1 & 1 & w_1 \\ 1 & 1 & w_2 \\ \vdots & \vdots & \vdots \\ 1 & 1 & w_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}.$$

Utemeljite, da ta model ustreza standardnim predpostavkam modela linearne regresije.

Rešitev: Z množenjem matrik ugotovimo, da dobimo povsem iste enačbe. Da ta novi model ustreza vsem predpostavkam standardnega modela, sledi iz navedenih predpostavk.

b. (10) Predpostavite, da je

$$\sum_{i=1}^{m} x_i = 0 \qquad \text{in} \qquad \sum_{j=1}^{n} w_j = 0$$

ter

$$\sum_{i=1}^{m} x_i^2 = 1 \qquad \text{in} \qquad \sum_{j=1}^{n} w_j^2 = 1\,.$$

Podajte najboljšo nepristransko cenilko parametra β. Izračunajte njeno standardno napako.

Rešitev: Ugotovimo, da je

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} m+n & n & 0 \\ n & n & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Matriko obrnemo in dobimo

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} 1/m & -1/m & 0 \\ -1/m & (m+n)/mn & 0 \\ 0 & 0 & 1/2 \end{pmatrix}.$$

Ugotovimo še

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^m Y_i + \sum_{j=1}^n Z_j \\ \sum_{j=1}^n Z_j \\ \sum_{i=1}^m x_i Y_i + \sum_{j=1}^n w_j Z_j \end{pmatrix} .$$

Sledi

$$\hat{\beta} = \frac{1}{2}\left(\sum_{i=1}^m x_i Y_i + \sum_{j=1}^n w_j Z_j\right) .$$

Če pišemo še $\mathbf{C} = [c_{ij}]_{i,j} = (\mathbf{X}^T\mathbf{X})^{-1}$, *za standardno napako po formuli dobimo*

$$\mathrm{se}(\hat{\beta}) = \sigma\sqrt{c_{33}} = \frac{\sigma}{\sqrt{2}} .$$

c. (10) Podajte najboljšo nepristransko cenilko parametra α_2. Podajte njeno standardno napako.

Rešitev: Če imamo najboljšo cenilko $\alpha_2 - \alpha_1$ *in najboljšo cenilko* α_1, *je po izreku Gauss–Markova njuna vsota najboljša linearna cenilka* α_2. *Z množenjem matrik dobimo, da je* $\hat{\alpha}_1 = \bar{Y}$ *in* $\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{Z} - \bar{Y}$. *Torej je* $\hat{\alpha}_2 = \bar{Z}$. *Standardna napaka je* $\sigma/\sqrt{n}$ *(kar lahko dobimo bodisi z množenjem matrik bodisi neposredno iz variance povprečja).*