Statistička analiza podataka – međuispit

UNIZG FER, ak. god. 2021./2022.

25.11.2021.

Ispit traje 120 minuta i nosi 30 bodova. Svaki zadatak rješavajte na zasebnoj stranici. Pišite uredno i čitko – rješenja koja ispravljači ne mogu pročitati neće se bodovati.

1. (6 bodova) Podatci prikazuju dužinu listova (u milimetrima) prikupljenih od jednog botaničara:

$$53, 39, 39, 33, 69, 30, 25, 67, 130, 40.$$

   a) (4 boda) Iz danih podataka skicirajte kvadratni dijagram (engl. *box plot*) – označite na y-osi sve elemente dijagrama jasno. Neka je područje nestršećih vrijednosti definirano izdancima (engl. *whiskers*) veličine 1.5 IQR. Možemo li iz kvadratnog dijagrama zaključiti koji postotak prikupljenih listova je manji od 47 mm?

   b) (2 boda) Odredite srednju vrijednost i medijan dužine listova. Kakav je odnos te dvije statistike i što one govore o obliku distribucije?

2. (6 bodova) Razmatra se unaprjeđenje pogona za proizvodnju procesora kako bi se povećao broj ispravno proizvedenih jedinica (engl. *yield*). *Yield* postojeće opreme iznosio je 44% na uzorku od 500 procesora, a uzorak jednake veličine odabran je i za novu metodu uz 49% ispravno proizvedenih jedinica. Prema procijenjenom trošku implementacije novog procesa prelazak na napredniju tehnologiju isplati se u slučaju povećanja *yielda* na barem 47%.

   a) (1 bod) Definirajte pogrešku druge vrste. Koja je njena interpretacija u primjeru iz zadatka?

   b) (3 boda) Izračunajte vjerojatnosti pogreške prve i druge vrste.

   c) (2 boda) Izračunajte snagu testa. Obrazložite i računski pokažite jednu metodu kojom bi u ovom primjeru mogli povećati snagu testa.

3. (6 bodova) Provedena je studija Biološkog odsjeka PMF-a kako bi se utvrdile gustoće organizama na dvije različite lokacije. Mjerne stanice smještene su na ušću rijeke Cetine u Jadransko more, te na mjestu izljeva otpadnih voda nuklearne elektrane Krško u rijeku Savu. U tablici se nalaze mjerenja gustoće, u broju organizama po kvadratnom metru, na dvije mjerne stanice. Na razini značajnosti $\alpha = 0.05$ zanima nas je li gustoća organizama na mjernoj stanici Sava i mjernoj stanici Cetina jednaka.

   a) (1 bod) Postavite hipoteze H_0 i H_1 potrebne za statističko testiranje.

   b) (4 boda) Provedite adekvatni statistički test. Koje pretpostavke ste pritom koristili?

   c) (1 bod) Interpretirajte rezultate testa u kontekstu zadatka.

(Okrenuti stranicu!)

Mjerna stanica Sava		Mjerna stanica Cetina	
280	281	503	498
467	133	1370	1191
689	332	1073	813
772	123	1140	2685
		86	1766
		220	2280

4. (6 bodova) U jednoj je školi provedeno istraživanje o sportskoj aktivnosti i fizičkom zdravlju učenika. Svakom je učeniku izračunat indeks tjelesne mase te su po njemu kategorizirani u sljedeće klase: pothranjenost (17 učenika), normalna težina (233 učenika), pretilost (105 učenika). Uz to, za svakog se učenika provjerilo sudjeluje li u izvannastavnim sportskim aktivnostima što se pokazalo potvrdnim za 10 učenika iz kategorije pothranjenih, 148 učenika normalne težine i 48 učenika iz kategorije pretilih.

   a) (1 bod) Odredite kontingencijsku tablicu. Uz pretpostavku nezavisnosti promatranih obilježja, izračunajte vjerojatnost da je učenik normalne tjelesne mase i da se bavi sportom.

   b) (4 boda) Postoji li veza između sportske aktivnosti i tjelesne mase? Provedite test na razini značajnosti 5% i obrazložite zašto ga smijete koristiti.

   c) (1 bod) Skicirajte funkciju gustoće testne statistike iz prethodnog podzadatka te jasno naznačite gdje se nalazi kritično područje. Dodatno, prikažite na istoj skici gdje se nalazi izračunata statistika.

5. (6 bodova) Investicijski fond je u prvoj polovici 2020. godine ostvario sljedeće povrate R. Kao

mjesec	1	2	3	4	5	6
R	4%	-1%	-3%	7%	6%	-1%

   mjeru performansi fonda koristimo tzv. *Sharpeov omjer* koji u obzir uzima i povrat i rizik: $S = \dfrac{\mathbb{E}[R - R_f]}{\sqrt{\mathrm{Var}[R]}}$, gdje je R_f konstanta koja označava bezrizičnu kamatnu stopu.

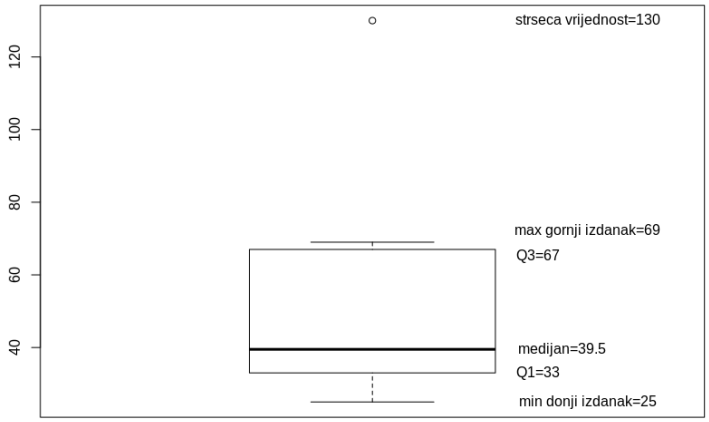
   a) (1 bod) Jasno napišite izraz za procjenitelj Sharpeovog omjera iz podataka.

   b) (2 boda) Obrazložite prednosti i nedostatke korištenja jackknife metode u odnosu na parametarske metode za ovaj procjenitelj.

   c) (3 boda) Izračunajte jackknife procjenu Sharpeovog omjera i 95% intervala pouzdanosti za fond koji razmatramo, uz $R_f = 1\%$.

Rješenja zadataka

1. a) Kvadratni dijagram:



Za maksimalni gornji izdanak priznavala se i vrijednost $118 = Q3 + 1.5 \times IQR$, dok su se za minimalni donji izdanak prihvaćale i vrijednosti $-18 = Q1 - 1.5 \times IQR$ i 0 (jer u kontekstu zadatka, dužina lista ne može biti negativna).
Ne možemo zaključiti iz kvadratnog dijagrama koji postotak prikupljenih listova je manji od 47mm. Možemo zaključiti da ih je više od 50%, ali manje od 75% njih.

b) $\mu = 52.5$, med. $= 39.5$ – distribucija je zakrivljena u desno (s težim lijevim repom).

2. a) Pogreška druge vrste je kad ne odbacimo nul-hipotezu H_0 u slučaju kad je H_1 istinita. U ovom slučaju bi pogreška druge vrste značila zadržavanje postojećeg proizvodnog procesa iako nova metoda nudi superioran *yield*.

b) Označimo *yield* s p i postavljamo hipoteze kao

$$H_0 : p = 0.44$$
$$H_1 : p = 0.49.$$

Kritična vrijednost iznosi 234.5 (*napomena: priznaje se i 235*). Izračunavamo α korištenjem aproksimacije normalnom distribucijom uz parametre

$$\mu = np = 500 \cdot 0.44 = 220,$$
$$\sigma = \sqrt{npq} = \sqrt{500 \cdot 0.44 \cdot 0.56} \approx 11.0995.$$

Odgovarajuća z-vrijednost jednaka je

$$z = \frac{234.5 - 220}{11.0995} \approx 1.31. \quad (\approx 1.35 \text{ uz k.v. } 235)$$

Iščitavanjem odgovarajuće vrijednosti iz tablica normalne distribucije dobivamo vjerojatnost pogreške prve vrste

$$\begin{aligned}\alpha &= P(X \geq 235 \text{ uz } p = 0.44) \\ &\approx P(Z > 1.31) = P(Z < -1.31) = 0.0951. \quad (\approx 0.0885 \text{ uz k.v. } 235)\end{aligned}$$

Sličan postupak ponavljamo za β uz parametre

$$\mu = np = 500 \cdot 0.49 = 245,$$
$$\sigma = \sqrt{npq} = \sqrt{500 \cdot 0.49 \cdot 0.51} \approx 11.1781.$$

Sada je z-vrijednost $(234.5 - 245)/11.1781 \approx -0.94$ (odnosno -0.89 uz k.v. 235) pa je vjerojatnost pogreške druge vrste jednaka

$$\begin{aligned}\beta &= P(X < 235 \text{ uz } p = 0.49) \\ &\approx P(Z < -0.94) = 0.1736. \quad (\approx 0.1867 \text{ uz k.v. } 235)\end{aligned}$$

c) Snaga testa jednaka je $1 - \beta = 0.8264$. Na snagu testa moguće je utjecati povećanjem uzorka (obrazloženje na predavanjima); primjerice za $n = 1000$ imamo

$$z = \frac{469.5 - 490}{\sqrt{1000 \cdot 0.49 \cdot 0.51}} \approx -1.30.$$

Vjerojatnost pogreške druge vrste je $\beta \approx P(Z < -1.30) = 0.0968$ pa nova snaga testa iznosi $1 - \beta = 0.9032$.

3. a) Hipoteze postavljamo kao

$$H_0 : \mu_s = \mu_c$$
$$H_1 : \mu_s \neq \mu_c$$

b) Na stanici Sava imamo $n_s = 8$ mjerenja; procjene srednje vrijednosti μ_s i standardne devijacije s_s iznose

$$\mu_s = 384.624 \qquad s_s = 240.745,$$

a na stanici Cetina $n_c = 12$ mjerenja procjene iznose:

$$\mu_c = 1135.417 \qquad s_c = 798.887.$$

Broj stupnjeva slobode procjenjujemo iz

$$v = \frac{(s_s^2/n_s + s_c^2/n_c)^2}{\dfrac{(s_s^2/n_s)^2}{n_s - 1} + \dfrac{(s_c^2/n_c)^2}{n_c - 1}} = 13.79$$

te uzimamo najveće cijelo pa je $v = 13$. Iščitavanjem iz tablica dobivamo kritično područje za $t < -2.160$ i $t > 2.160$. Vrijednost t-statistike određujemo kao

$$t = \frac{\mu_s - \mu_c}{\sqrt{\mu_s^2/s_s + \mu_c^2/s_c}} = 3.054.$$

Pretpostavljamo da su podatci nezavisni i normalno distribuirani.

c) Na razini značajnosti $\alpha = 0.05$ odbacujemo hipotezu H_0 i zaključujemo $\mu_s \neq \mu_c$.

4. a) Kontingencijska tablica (očekivane frekvencije u zagradi):
$P = 206/355 \cdot 233/355 = 0.3809$.

sport/BMI	pothranjenost	normalna težina	pretilost	
DA	10 (9.865)	148 (135.206)	48 (60.93)	206
NE	7 (7.135)	85 (97.794)	57 (44.07)	149
	17	233	105	355

b) Provodimo χ^2 test o nezavisnosti. Smijemo ga koristiti jer su sve očekivane frekvencije veće ili jednake 5.

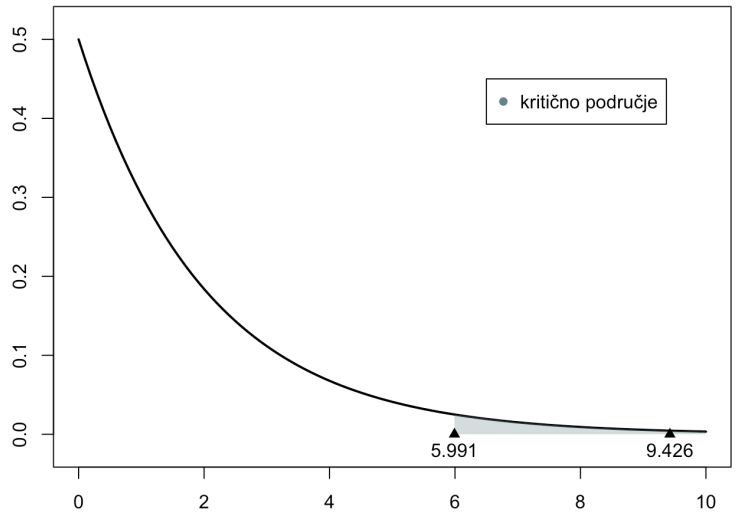
H_0 : Sportska aktivnost i tjelesna masa su nezavisni.
H_1 : Sportska aktivnost i tjelena masa su zavisni.

Iz $\alpha = 0.05$ imamo $\chi^2_\alpha = 5.991$ uz $\nu = 2$. Kritično područje je $\chi^2 > 5.991$. Slijedi

$$\chi^2 = \frac{(10 - 9.865)^2}{9.865} + \ldots + \frac{(57 - 44.07)^2}{44.07} = 9.426,$$

$0.005 < p$-vrijednost < 0.01
$9.426 > 5.991 \Rightarrow$ odbacujemo nul-hipotezu o nezavisnosti obilježja

c) Skica:



5. a) $\hat{S} = \dfrac{\hat{R} - R_f}{\hat{\sigma}_R}$, gdje su $\hat{R} = \dfrac{1}{n}\sum_{i=1}^{n} R_i$ i $\hat{\sigma}_R = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(R_i - \hat{R})^2}$

b) Jackknife nam omogućuje da izračunamo pristranost i varijabilnost procjenitelja bez da znamo njegovu distribuciju uzorkovanja, što u ovom slučaju može biti istina zbog složenosti procjenitelja ali i zbog potencijalne nepoznate distribucije povrata. S druge strane, jackknife postupci imaju manju statističku snagu od parametarskih.

c) Jackknife replikati:

$$\hat{S}_{(1)} = 0.1316$$
$$\hat{S}_{(2)} = 0.3642$$
$$\hat{S}_{(3)} = 0.5252$$
$$\hat{S}_{(4)} = 0$$
$$\hat{S}_{(5)} = 0.0482$$
$$\hat{S}_{(6)} = 0.3642.$$

Pseudovrijednosti:

$$ps_1 = 0.7724$$
$$ps_2 = -0.3908$$
$$ps_3 = -1.1959$$
$$ps_4 = 1.4302$$
$$ps_5 = 1.1891$$
$$ps_6 = -0.3908.$$

Slijedi $\hat{S}_{\text{jack}} = \overline{ps} = 0.236$, $\hat{SE}_{\text{jack}}(\hat{S}) = 0.427$
95% interval pouzdanosti: $\hat{S}_{\text{jack}} \pm t_{\alpha=0.025,v=5} SE_{\text{jack}}(\hat{S}) = 0.236 \pm 2.571 \cdot 0.427 = [-0.861, 1.332]$