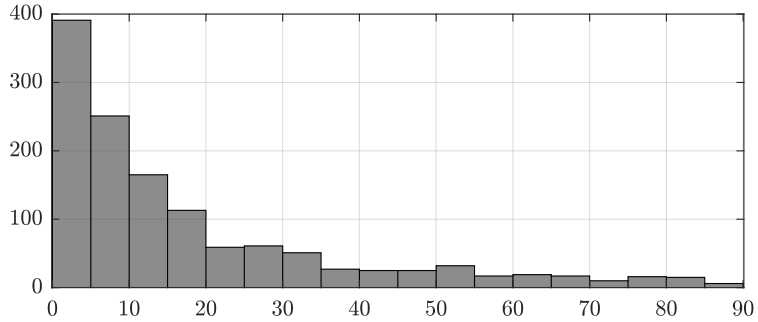
Statistička analiza podataka – međuispit

UNIZG FER, ak. god. 2020./2021.

16.11.2020.

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) Zadan je sljedeći histogram:



   a) (3 boda) Poredajte sljedeće mjere po vrijednostima koje biste očekivali iz zadanog histograma (od najveće prema najmanjoj): medijan, mod, aritmetička sredina, 10% podrezana aritmetička sredina.

   b) (3 boda) Neka su 1., 2., i 3. kvartil podataka iz histograma redom: 4, 10, i 24. Skicirajte pripadni oblik kvadratnog dijagrama (engl. *box plot*). Pritom označite na y-osi sve elemente dijagrama, uz pretpostavku da su rubovi stupaca histograma ujedno i najmanji/najveći podatci u njima.

2. (6 bodova) Provedeno je istraživanje u dva grada u Virginiji kako bi se utvrdio sentiment glasača za dva kandidata za guvernera na predstojećim izborima. Nasumično je izabrano 500 glasača iz svakog grada i zabilježeni su sljedeći podatci: u gradu Richmond 204 glasača favorizira kandidata A, 211 favorizira kandidata B, a ostatak je neodlučan. Zatim, u gradu Norfolk zabilježeno je sljedeće: 225 glasača favorizira kandidata A, 198 kandidata B, a ostatak je neodlučan.

   Zanima nas jesu li proporcije glasača koji favoriziraju kandidata A, kandidata B ili su neodlučni isti za oba grada.

   a) (1 bod) Obrazložite koji test ćete koristiti (koja je testna statistika te kako se test zove, odnosno što ispituje).

   b) (1 bod) Jasno postavite hipoteze H_0 i H_1.

   c) (3 boda) Za razinu značajnosti $\alpha = 5\%$, provedite test i interpretirajte rezultate testa u kontekstu zadatka.

   d) (1 bod) Skicirajte distribuciju testne statistike iz prethodnih podzadataka te jasno naznačite gdje se nalazi kritično područje. Dodatno, prikažite na istoj skici gdje se nalazi izračunata statistika iz c) dijela zadatka.

3. (6 bodova) Neka je nul-hipoteza za vjerojatnost dobijanja glave bacanjem novčića $H_0 : p_g = 1/2$, te je varijabla X broj pojavljivanja glave u 5 bacanja novčića, čija je razdioba uz uvjet da je H_0 istinita: $P(X = 0) = 0.031, P(X = 1) = 0.156, P(X = 2) = 0.313, P(X = 3) = 0.313, P(X = 4) = 0.156, P(X = 5) = 0.031$.

   (a) (2 boda) Izračunajte očekivanje i varijancu slučajne varijable X.

   (b) (2 boda) Želite testirati hipotezu H_0 uz alternativu $H_1 : p_g > 1/2$. Bacili ste novčić 5 puta i dobili $X = 4$ glave. Kolika je p-vrijednost?

   (c) (2 boda) Neka je kritična vrijednost navedenog testa $X_c = 3.5$, a alternativna hipoteza $H_1 : p_g = 0.7$, te je za nju razdioba broja glava X u 5 bacanja novčića: $P(X = 0) = 0.003, P(X = 1) = 0.028, P(X = 2) = 0.132, P(X = 3) = 0.309, P(X = 4) = 0.360, P(X = 5) = 0.168$. Kolika je snaga testa?

4. (6 bodova) Znanstvenik želi ispitati pristranost svoga uređaja koji mjeri pH vrijednosti. Odlučio je testirati pH neutralne supstance (pH=7.00) te je dobio sljedeće rezultate:

   7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

   (a) (4 boda) Na razini značajnosti od 5% testirajte pristranost uređaja, odnosno ispitajte hipotezu da je vrijednost pH neutralnih supstanci za uređaj jednaka 7.00. Pritom jasno napišite koji test koristite, hipoteze testa, koje pretpostavke trebaju biti zadovoljene te zaključak testa u kontekstu zadatka.

   (b) (2 boda) Je li 95%-tni interval pouzdanosti općenito podskup/nadskup/ništa od navedenog 99%-tnog intervala pouzdanosti za dani uzorak? Obrazložite svoju tvrdnju.
   Je li 95%-tni interval pouzdanosti manjeg uzorka općenito podskup/nadskup/ništa od navedenog 95%-tnog interval pouzdanosti većeg uzorka? Obrazložite svoju tvrdnju.

5. (6 bodova) Zadani su sljedeći podatci:

x	2	5	15	30	10	20	45
y	7	9	50	100	40	70	120

   (a) (2 boda) Odredite koeficijente modela linearne regresije $y = \beta_1 x + \beta_0$.

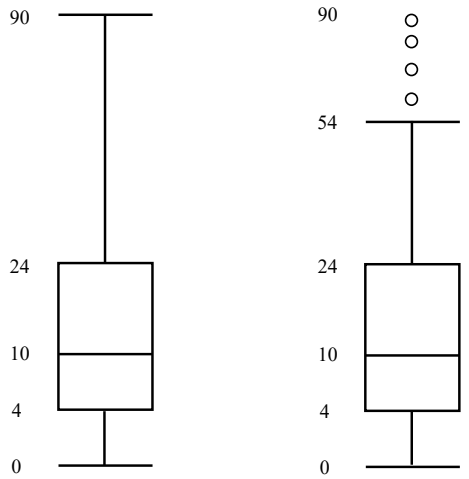
   (b) (2 boda) Izračunajte procjenu za srednju vrijednost $\mu_{Y|x_0}$ i njezin 95% interval pouzdanosti za $x_0 = 35$.

   (c) (2 boda) Izračunajte procjenu Y_0 i njezin 95% interval pouzdanosti za $x_0 = 35$, te objasnite razliku u odnosu na b).

1.

a) Aritmetička sredina > 10% podrezana aritmetička sredina > medijan > mod.

b) Priznaju se oba rješenja (sa i bez stršećih vrijednosti):



2.

a) χ^2 testna statistika, test homogenosti

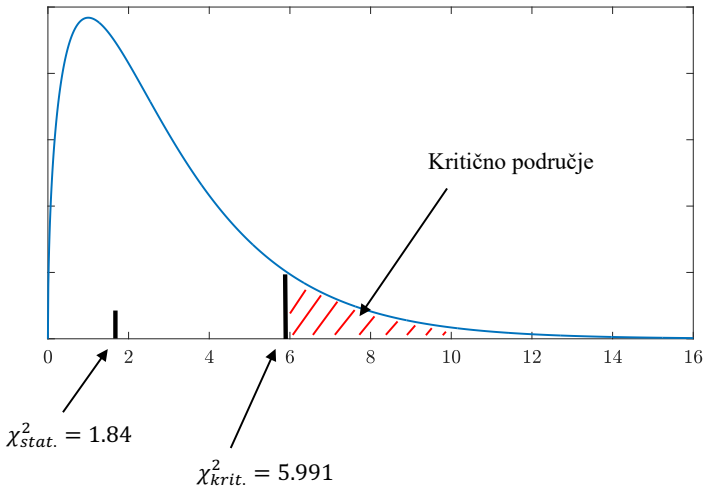
b) H_0 : Proporcije glasača koji favoriziraju A, B ili su neodlučni iste su za oba grada.
H_1 : Proporcije glasača koji favoriziraju A, B ili su neodlučni nisu iste za oba grada.

c) $\alpha = 5\%$
$\chi^2_\alpha = 5.991$ uz $\nu = 2$
Kritično područje: $\chi^2 > 5.991$
Kontingencijska tablica (s ostvarenim i očekivanim frekvencijama u zagradi):

	R	N	
A	204 (214.5)	225 (214.5)	429
B	211 (204.5)	198 (204.5)	409
neodlučni	85 (81)	77 (81)	162
	500	500	1000

$\chi^2 = \frac{(204-214.5)^2}{214.5} + ... + \frac{(77-81)^2}{81} = 1.84$
$1.84 < 5.991 \Rightarrow$ ne odbacujemo H_0 hipotezu o jednakosti proporcija glasača

d) Skica:



3.

a) $\mathrm{E}(X) = \sum [x \cdot P(X = x)] = 2.5$,
$\mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2) = 1.246$,
(priznaje se i izračun preko binomne distribucije, u tom slučaju nema greške zbog zaokruživanja vjerojatnosti i varijanca iznosi 1.25)

b) p-vrijednost $= P(X \geq 4|H_0) = P(X = 4) + P(X = 5) = 0.187$.

c) $\beta = P(X \leq X_c|H_1) = P(X = 3|H_1) + P(X = 2|H_1) + P(X = 1|H_1) + P(X = 0|H_1) = 0.472$,
Snaga testa $= 1 - \beta = 0.528$.

4.

a) Koristimo t-test uz pretpostavke slučajnosti uzorka i normalnosti podataka. Hipoteze testa:

$$H_0 : \mu_{P_H} = 7.00$$
$$H_1 : \mu_{P_H} \neq 7.00$$

Testna statistika glasi $T = \frac{\overline{x} - \mu_{P_H}}{\frac{s}{\sqrt{n}}} \sim t(n-1)$ gdje za:

$$n = 10,\ \overline{x} = 7.0250,\ s = 0.044,\ \frac{s}{\sqrt{n}} = 0.0139,\ \alpha = 0.05$$

dobivamo realizaciju testne statistike $T = 1.80$.

Kritična vrijednost:

$$t_{\frac{\alpha}{2}, n-1} \approx 2.262.$$

Obzirom da $T \notin \langle -\infty, -2.262\rangle \cup \langle 2.262, \infty\rangle$, zaključujemo da ne odbacujemo H_0 u korist H_1, odnosno možemo reći da uređaj nije pristran i radi kako treba.

b) (95% interval pouzdanosti) $\subset$ (99% interval pouzdanosti), odnosno iz veće pouzdanosti sljedi veći interval pouzdanosti (što možemo vidjeti iz površina ispod krivulje distribucije ili ekvivalentno iz formula intervala pouzdanosti za tada različite kritične vrijednosti).

Većim skupom podataka (većim n-om) imamo uži interval pouzdanosti (što možemo primjetiti iz same formule intervala pouzdanost gdje povećanjem n-a, smanjujemo izraz $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, odnosno smanjujemo interval pouzdanosti). Dakle, (95% interval pouzdanosti većeg skupa podataka) $\subset$ (95% interval pouzdanosti manjeg skupa podataka). (Ovdje smo priznavali i argument da sa povećanjem broja podataka zapravo ne znamo što se dešava sa samim podatcima, i njihova distribucija se može skroz promijeniti iz čega bi zaključili da vrijedi "ništa od navedenog" za takva dva intervala pouzdanosti).

5.

a)

$$\overline{x} = 18.1428, \quad \overline{y} = 56.5714, \quad n = 7$$

$$S_{xx} = \sum (x_i - \overline{x})^2 = 1374.8571, \quad S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = 3824.4285$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = 2.78169, \quad b_0 = \overline{y} - b_1\overline{x} = 6.1035$$

b)

$$S_{yy} = \sum (y_i - \overline{y})^2 = 11127.7142$$

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2} = 97.866 \Rightarrow s = 9.8927$$

$$x_0 = 35, \quad \hat{y_0} = 103.46, \quad t_{0.025,5} = 2.571$$

$$\hat{y_0} - t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y_0} + t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

$$88.42550 < \mu_{Y|x_0} < 118.5$$

c)

$$\hat{y_0} - t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} < y_0 < \hat{y_0} + t_{\frac{\alpha}{2},n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

$$73.91 < y_0 < 133.009$$

U b) zadatku tražimo interval pouzdanosti srednje vrijednosti $\mu_{Y|x_0}$, dok u c) tražimo interval pouzdanosti predikcije y_0. Očekujemo da će greška predikcije jedne vrijednosti biti veća nego greška predikcije srednje vrijednosti što i utječe na same intervale pouzdanosti. Dok $\mu_{Y|x_0}$ predstavlja interval pouzdanosti parametra populacije, interval pouzdanosti predikcije y_0 predstavlja interval koji sa vjerjoatnosti od $1 - \alpha$ sadrži buduću vrijednost y_0 slučajne varijable Y_0.