Vjerojatnost i statistika

Statističko zaključivanje - dvije varijable

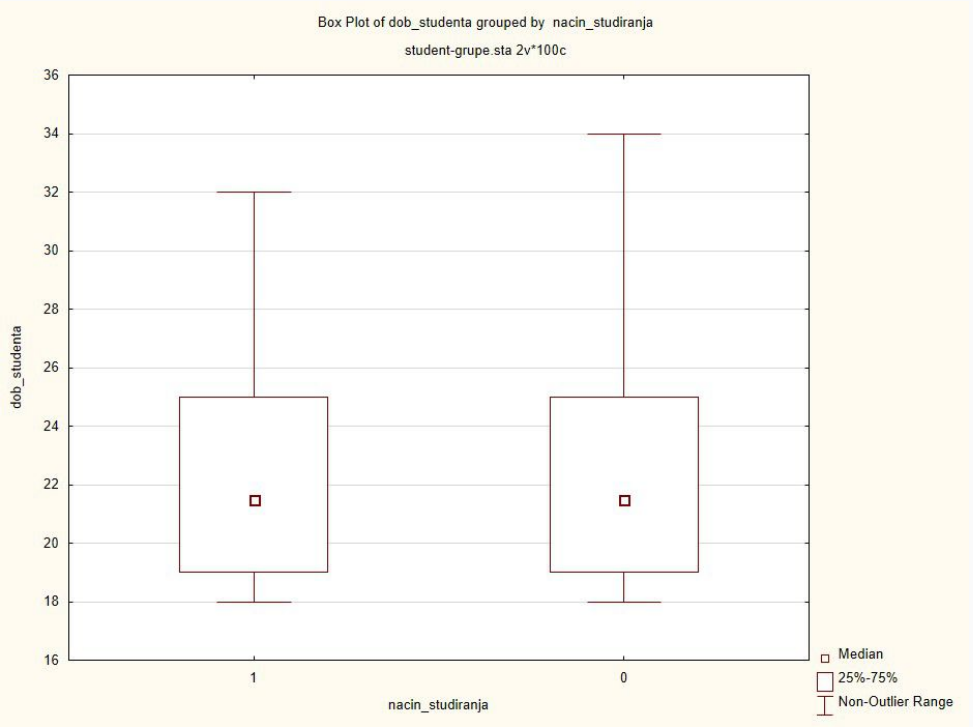
23. siječnja 2024.

Nevezani i vezani uzorci

- u praksi nas često zanima dolazi li do promjene obilježja koje proučavamo zbog provođenja neke aktivnosti, u nekom drugom trenutku ili općenito u nekim drugim uvjetima
- **nevezani uzorci** - uzorci koji nemaju zajedničkih jedinki, ali proučavaju isto obilježje
- **vezani uzorci** - uzorci koji sadrže iste jedinke, proučavaju isto obilježje, ali prije i poslije nekog **tretmana**
- zanima nas jesu li uočene razlike među ovakvim uzorcima statistički značajne

Primjer - student.sta

Neko sveučilište osim klasičnog načina studiranja nudi i studiranje temeljeno na konceptu e-learninga. Povjerenstvo za praćenje kvalitete studiranja želi vidjeti postoji li razlika u dobi između studenata koji studiraju na klasičan način i onih koji studiraju putem e-learninga. Podaci o dobi studenata nalaze se u bazi `student.sta`.



Primjer - djelatnici.sta

Baza podataka `djelatnici.sta` za svakog djelatnika iz uzorka zaposlenika nekog poduzeća sadrži iznos godišnje plaće u eurima prije i nakon restrukturiranja poduzeća (varijable *placa_prije* i *placa_poslije*).

Variable	Descriptive Statistics (djelatnici)					
	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
placa_prije	24522,00	23650,00	16000,00	42400,00	26069208	5105,801

Variable	Descriptive Statistics (djelatnici)					
	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
placa_poslije	24986,85	23950,40	16289,04	42496,65	26252790	5123,748

Nevezani uzorci

Usporedba očekivanja

- testiramo jednakost očekivanja slučajnih varijabli kojima modeliramo obilježje dvaju nevezanih jednostavnih slučajnih uzoraka $\mathbb{X} = (X_1, \ldots, X_{n_1})$ i $\mathbb{Y} = (Y_1, \ldots, Y_{n_2})$
- neka su μ_1 i σ_1 očekivanje i standardna devijacija slučajne varijable kojom modeliramo obilježje u prvom uzorku, a μ_2 i σ_2 iste karakteristike u drugom uzorku

Veliki uzorci: z-test

Nul-hipoteza:

$$H_0 : \mu_1 = \mu_2$$

Test-statistika:

$$Z' = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

U slučaju velikih uzoraka i u uvjetima istinitosti H_0 test-statistika Z' ima približno standardnu normalnu distribuciju.

Veliki uzorci: z-test

Realizacija test statistike:

$$\hat{z} = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Računanje p-vrijednosti ovisi o obliku alternativne hipoteze:

- $p = P(|Z'| \geq |\hat{z}|)$ ako je $H_1 : \mu_1 \neq \mu_2$
- $p = P(Z' \geq \hat{z})$ ako je $H_1 : \mu_1 > \mu_2$
- $p = P(Z' \leq \hat{z})$ ako je $H_1 : \mu_1 < \mu_2$

Primjer

Ekonomisti u nekoj zemlji odlučili su provjeriti jesu li očekivane cijene u eurima uvoznih automobila više u njihovoj zemlji nego u matičnoj zemlji određenog proizvođača. Prikupljen je uzorak od 50 cijena u promatranoj zemlji i 30 cijena u matičnoj zemlji za isto razdoblje. Na temelju tih uzoraka, procijenjena očekivanja i standardne devijacije slučajnih varijabli kojima se modelira cijena tog tipa automobila su

promatrana zemlja:	$n_1 = 50,\ \overline{x}_{n_1} = 16\,545,\ s_{n_1} = 1\,989$
matična zemlja proizvođača:	$n_2 = 30,\ \overline{y}_{n_2} = 17\,243,\ s_{n_2} = 1\,843$

Možemo li na razini značajnosti $\alpha = 0.05$ potvrditi da je očekivana cijena u promatranoj zemlji manja nego u matičnoj?

Primjer

$$H_0 \ : \ \mu_1 = \mu_2$$
$$H_1 \ : \ \mu_1 < \mu_2$$

$$\hat{z} = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{16\,545 - 17\,243}{\sqrt{\frac{1\,989^2}{50} + \frac{1\,843^2}{30}}} = -1.592$$

- p-vrijednost: $p = P(Z \leq \hat{z}) = P(Z \leq -1.592) = 0.056$
- na razini značajnosti $\alpha = 0.05$ ne odbacujemo H_0, tj. ne možemo tvrditi da je očekivana cijena automobila u promatranoj zemlji statistički značajno manja od očekivane cijene automobila u matičnoj zemlji

Zadatak - burza.sta

U bazi podataka `burza.sta` zabilježene su cijene nekih dionica na dvije burze smještene u dva različita grada - A i B. U jednom financijskom časopisu, pročitali smo da je očekivana cijena dionice viša na burzi u gradu A u odnosu na očekivanu cijenu na burzi u gradu B. Možemo li na razini značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanoj cijeni dionice na promatranim burzama?

Rješenje: $H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2, \quad p \approx 0.0056 < \alpha$

Mali uzorci: t-test

- testiramo jednakost očekivanja slučajnih varijabli kojima modeliramo obilježje dvaju nevezanih uzoraka
- pretpostavke: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ i $Y_1 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $\sigma_1^2 = \sigma_2^2$

Nul-hipoteza:

$$H_0 : \mu_1 = \mu_2$$

Test-statistika:

$$T' = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p^2 = \frac{(n_1 - 1)s_{n_1}^2 + (n_2 - 1)s_{n_2}^2}{n_1 + n_2 - 2}$$

U uvjetima istinitosti nul-hipoteze, test-statistika T' ima Studentovu t-distribuciju s $(n_1 + n_2 - 2)$ stupnjeva slobode.

Mali uzorci: t-test

Realizacija test statistike:

$$\hat{t} = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Računanje p-vrijednosti ovisi o obliku alternativne hipoteze:

- $p = P(|T| \geq |\hat{t}|)$ ako je $H_1 : \mu_1 \neq \mu_2$
- $p = P(T \geq \hat{t})$ ako je $H_1 : \mu_1 > \mu_2$
- $p = P(T \leq \hat{t})$ ako je $H_1 : \mu_1 < \mu_2$

F-test

Za primjenu prethodnog testa, nužna je jednakost varijanci varijabli pa je najprije potrebno testirati hipotezu o jednakosti varijanci. U tu svrhu koristimo **F-test**.

- hipoteze:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Provođenje t-testa

- u Statistici:

  Statistics → Basic Statistics → t-test, independent, by variables → Summary

- u slučaju da nije ispunjena pretpostavka o jednakosti varijanci, provodi se Welchova inačica t-testa:

  Statistics → Basic Statistics → t-test, independent, by variables → Options → označiti "t-test with separate variance estimates"

Primjer - potrosac.sta

Marketinški stratezi željeli bi predvidjeti prijem nove vrste paste za zube kod potrošača prema njihovoj dobi. U bazi podataka `potrosac.sta` raspolažemo podacima o dobi u godinama za 20 potrošača koji su kupili novu pastu za zube (varijabla *korisnici*) i 20 potrošača koji ju još uvijek nisu kupili (varijabla *nisu_korisnici*). Možemo li na razini značajnosti $\alpha = 0.01$ potvrditi postojanje razlike u očekivanoj dobi potrošača iz te dvije grupe?

Postavljamo hipoteze:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Primjer - potrosac.sta

Provjeravamo pretpostavke t-testa:

- normalnost podataka: nemamo razloga sumnjati u normalnu distribuiranost za obje varijable na razini značajnosti $\alpha = 0.01$ (Lillieforsov KS test i Shapiro-Wilk W test,)
- jednakost varijanci: nemamo razloga sumnjati u jednakost varijanci na razini značajnosti $\alpha = 0.01$ (F-test, $p = 0.195 > \alpha$)

Primjer - potrosac.sta

Provodimo t-test:

- računamo $\hat{t}$:

$$\hat{t} = \frac{\overline{x}_{n_1} - \overline{y}_{n_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{39.8 - 47.15}{11.95\sqrt{\frac{1}{20} + \frac{1}{20}}} = -1.94442$$

- računamo p-vrijednost:

$$p = P(\leq \hat{t}) = P(T \leq -1.94442) = 0.029637$$

- $p = 0.029637 > \alpha$
- na razini značajnosti $\alpha = 0.01$, ne odbacujemo H_0 tj. ne možemo tvrditi da je očekivana dob potrošača paste statistički značajno manja od očekivane dobi potrošača koji još nisu kupili pastu

Zadatak - manager.sta

Jedna grupa istraživača razvila je indeks koji mjeri uspjeh menadžera. Neki istraživač želi usporediti taj indeks za dvije grupe menadžera. Jedna grupa ima mnogo interakcija s ljudima izvan svog radnog okruženja (telefoniranje, razgovori, sastanci i sl.), dok druga grupa ima vrlo rijetke kontakte izvan svog okruženja. U bazi podataka `manager.sta` nalaze se indeksi za uzorak menadžera iz grupe koja ima mnogo interakcija (*mnogo_interakcija*) i indeksi za uzorak menadžera iz grupe koja ima malo interakcija (*malo_interakcija*). Možemo li na nivou značajnosti $\alpha = 0.05$ potvrditi postojanje razlika u očekivanim indeksima uspješnosti menadžera iz te dvije grupe pod uvjetima jednakosti varijanci i normalne distribuiranosti slučajnih varijabli kojima modeliramo indekse?

Rješenje: $H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2, \quad p \approx 0.00002 < \alpha$

Vezani uzorci

Usporedba očekivanja

- u praksi često uspoređujemo varijable u vezanim tretmanima, npr. rezultati nekog testa prije i poslije liječenja/terapije
- slučajevi se moraju pratiti u paru, a zaključci o postojanju razlika među tretmanima donose se na osnovu razlika varijabli u pojedinim tretmanima kao što je prikazano u tablici

ispitanik	tretman 1	tretman 2	razlika
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_1 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = x_1 - y_n$

Usporedba očekivanja

- definiramo slučajne varijable razlika $D_i = X_i - Y_i$, $i = 1, \ldots, n$, gdje su slučajne varijable $X_1, \ldots, X_n$ nezavisne i jednako distribuirane (isto vrijedi za slučajne varijable $Y_1, \ldots, Y_n$)
- pretpostavimo da su i $D_1, \ldots, D_n$ nezavisne i jednako distribuirane
- očekivanje μ_D slučajne varijable D_i, $i = 1, \ldots, n$, je razlika očekivanja μ_1 i μ_2 slučajnih varijabli X_i i Y_i, tj.

$$\mu_D = \mu_1 - \mu_2$$

- testiranje hipoteze

$$H_0 : \ \mu_1 - \mu_2 = 0$$

sada se svodi na testiranje ekvivalentne hipoteze

$$H_0 : \mu_D = 0$$

Usporedba očekivanja

Taj test u Statistici provodimo pomoću

Statistics → Basic Statistics → t-test, dependent samples → Summary

Primjer - igre.sta

U jednoj je školi napravljeno istraživanje o tome što djeca misle i osjećaju prema sebi. Test se sastojao od toga da na početku testiranja djeca ocjenom od 1 do 5 ocijene tvrdnju "imam mnogo dobrih osobina". Nakon toga, u razdoblju od šest tjedana djeca su igrala četiri igre koje potiču pozitivan stav prema sebi. Poslije tih igara ponovno im je postavljeno isto pitanje koje su na isti način ocijenili. U bazi podataka `igre.sta` nalaze se ocjene prije i nakon provođenja igara. Možemo li na razini značajnosti $\alpha = 0.05$ prihvatiti hipotezu o postojanju razlike u očekivanoj ocjeni djece prije i nakon tretmana igrama?

Primjer - igre.sta

- postavljamo hipoteze:

$$H_0 \ : \ \mu_1 = \mu_2$$

$$H_1 \ : \ \mu_1 < \mu_2$$

- p-vrijednost: $p \approx 0.0045 < \alpha$
- na razini značajnosti $\alpha = 0.05$ odbacujemo nul-hipotezu i prihvaćamo alternativnu, tj. možemo tvrditi da je očekivana ocjena djece poslije tretmana statistički značajno veća u odnosu na očekivanu ocjenu djece prije tretmana