6 Testiranje statističkih hipoteza

Mnoge praktične situacije u vezi sa slučajnim pojavama zahtijevaju da se donesu odluke tipa DA ili NE. Npr. pri praćenju procesa proizvodnje nekog proizvoda treba, na temelju rezultata mjerenja $x_1, \ldots, x_n$ statističkog obilježja X, donijeti odluku o tome da li proces proizvodnje osigurava ili ne osigurava zahtjevanu kvalitetu. Pretpostavlja se, dakako, da obilježje X, koje karakterizira kvalitetu pojedinog proizvoda (količina određenog sastojka npr.) ima slučajni karakter.

Teorijski gledano, riječ je o tome da se na temelju n mjerenja slučajne varijable X, odnosno na temelju vrijednosti $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$, donese odluka o prihvaćanju (DA) ili odbacivanju (NE) određene pretpostavke o svojstvima slučajne varijable X. Takva pretpostavka zove se *statistička hipoteza*, a postupak donošenja odluke o prihvaćanju ili odbacivanju statističke hipoteze zove se *testiranje*.

Primjer 56 *Želimo testirati da li je očekivanje trajanja neke vrste žarulja jednako npr. 1000h.*

Definiramo

$$H_0 : \mu = 1000h$$
$$H_1 : \mu \neq 1000h$$

H_0 je **nulta hipoteza**, a H_1 **alternativna hipoteza**. Budući iz alternativne hipoteze slijedi da može biti $\mu > 1000h$ ili $\mu < 1000h$, kažemo da je H_1 **dvostrana alternativna hipoteza**.

Ponekad je zgodnije imati **jednostranu alternativnu hipotezu**. Npr.

$$H_0 : \mu = 1000h$$
$$H_1 : \mu > 1000h$$

ili

$$H_1 : \mu < 1000h$$

Ukratko, nulta hipoteza u testu je na neki način "fiksna", dok je alternativna ona kod koje imamo mogućnost izbora.

Testiranje hipoteze (odnosno provjeru da li je ona istinita ili nije) provodimo na sljedeći način: uzmemo slučajni uzorak, izračunamo vrijednost odgovarajuće test-statistike, te na osnovu njene vrijednosti odlučujemo o istinitosti hipoteze.

Prilikom donošenja odluke o istinitosti hipoteze, postoji mogućnost pogreške, tj. krive odluke. Dvije su vrste mogućih pogrešaka:

→ *pogreška 1.vrste*: odbacili smo nultu hipotezu ako je ona istinita

→ *pogreška 2.vrste*: prihvatili smo nultu hipotezu ako je ona neistinita

	H_0 istinita	H_0 neistinita
prihvaćamo H_0	$\surd$	pogreška 2.vrste
odbacujemo H_0	pogreška 1.vrste	$\surd$

α = P(pogreška 1.vrste)= P(odbacujemo H_0 | H_0 istinita) $\Rightarrow$ **nivo signifikantnosti** ili **razina značajnosti**

β = P(pogreška 2.vrste)= P(prihvaćamo H_0 | H_0 neistinita)

1-β=P(odbacujemo H_0 | H_0 neistinita) $\Rightarrow$ **snaga testa**

Testiranja hipoteza (koja su ovdje obrađena) baziraju se na odgovarajućim *pouzdanim intervalima.* Ako izračunata vrijednost odgovarajuće test-statistike upadne u pouzdan interval tražene pouzdanosti, tada nultu hipotezu ne možemo odbaciti; ukoliko ona ne upadne u isti interval, nultu hipotezu odbacujemo!

6.1 Test o očekivanju normalno distribuirane populacije

6.1.1 Varijanca poznata

- neka je $X \sim N(\mu, \sigma^2)$, σ poznata

- imamo slučajni uzorak veličine n : $(X_1, \cdots, X_n)$
- želimo testirati da li je očekivanje μ jednako nekom unaprijed zadanom broju μ_0. Nulta hipoteza je $H_0 : \mu = \mu_0$. Za alternativnu možemo uzeti bilo koju od sljedeće tri:

$$H_1 : \mu \neq \mu_0 \quad \text{ili} \quad H_1 : \mu > \mu_0 \quad \text{ili} \quad H_1 : \mu < \mu_0$$

- u sva 3 slučaja koristimo istu test-statistiku:

$$Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Ako je nulta hipoteza $H_0 : \mu = \mu_0$ istinita, tada je $E[\bar{X}] = \mu_0$, odnosno $Z \sim N(0,1)$

Promotrimo redom slučajeve različitog izbora alternativne hipoteze:

1.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Ako je $H_0 : \mu = \mu_0$ istinita, tada

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

što je vjerojatnost da prihvatimo H_0 ako je ona istinita. S druge strane,

$$P((Z < -z_{\frac{\alpha}{2}}) \cup (Z > z_{\frac{\alpha}{2}})) = \alpha$$

je vjerojatnost da *ne* prihvatimo H_0 ako je one istinita.
Dakle,

ako je	$Z < -z_{\frac{\alpha}{2}}$ ili $Z > z_{\frac{\alpha}{2}}$	$\Rightarrow$	odbacujemo H_0
Ako je	$-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}$	$\Rightarrow$	ne možemo odbaciti H_0

2.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

H_0 odbacujemo ako je $\boxed{Z > z_\alpha}$
(ne $z_{\frac{\alpha}{2}}$, nego z_α!!! Kritično područje površine α je svo na desnoj strani)

3.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

H_0 odbacujemo ako je $\boxed{Z < -z_\alpha}$

Napomena: Treba paziti na terminologiju: ne kaže se "prihvaćamo hipotezu", nego "ne možemo ju odbaciti".

Zadatak 22 *Poznato je da napon u električnoj mreži od 220 volti ima normalnu distribuciju sa standarnom devijacijom od 6 volti. Ako je 16 nezavisnih mjerenja dalo rezultate:*

208, 216, 215, 228, 210, 224, 212, 213, 224, 218, 206, 209, 208, 218, 220, 206,

s razinom značajnosti 0.01 provjerite pretpostavku da je došlo do pada srednjeg napona u elektičnoj mreži.

Rješenje:

$$X \sim N(\mu, 6^2), \quad n = 16$$

Postavljamo hipoteze:

$$H_0 : \mu = 220$$
$$H_1 : \mu < 220$$

Nulta hipoteza je da je srednja vrijednost napona jednaka 220 (odnosno da je veća od te vrijednosti), dakle da *nije došlo* do pada napona, dok je alternativna da je srednja vrijednost napona manja od 220, odnosno da *je došlo* do pada napona, što je tvrdnja za koju želimo provjeriti da li vrijedi. Kad bismo kao alternativnu hipotezu uzeli $H_1 : \mu \neq 220$, u slučaju odbacivanja nulte hipoteze $H_0 : \mu = 220$, mogli bismo zaključiti samo da srednji napon *nije jednak* 220, no ne bismo znali je li on veći ili manji od te vrijednosti.

Računamo vrijednost test-statistike: $\quad Z = \dfrac{\bar{X}_n - \mu_0}{\sigma}\sqrt{n}$

$$\mu_0 = 220, \qquad \bar{x}_{16} = 214.6875$$
$$\Rightarrow \; z = \frac{214.6875 - 220}{6}\sqrt{16} = -3.54167$$
$$z_\alpha = z_{0.01} = 2.325$$
$$\Longrightarrow \; z < -z_{0.01}$$

$\Longrightarrow$ odbacujemo nultu hipotezu H_0, tj. došlo je do pada napona! $\square$

6.1.2 Varijanca nepoznata

- neka je $X \sim N(\mu, \sigma^2), \quad \sigma$ nepoznata
- imamo njen slučajni uzorak veličine n : $(X_1, \cdots, X_n)$
- želimo testirati da li je očekivanje μ jednako nekom unaprijed zadanom broju μ_0
- koristimo test-statistiku:

$$T = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$$

Ako je nulta hipoteza $H_0 : \mu = \mu_0$ istinita, tada je $T \sim t(n-1)$

1.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Nultu hipotezu H_0 odbacujemo ako je

$$T > t_{\frac{\alpha}{2}}(n-1) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n-1)$$

2.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

H_0 odbacujemo ako je

$$T > t_\alpha(n-1)$$

3.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

H_0 odbacujemo ako je

$$T < -t_\alpha(n-1)$$

Zadatak 23 *Tvornica tvrdi da je prosječan vijek trajanja baterija iz te tvornice 21.5 sati. Na slučajnom uzorku od 6 baterija iz te tvornice laboratorijskim mjerenjima vijeka trajanja dobivene su vrijednosti od 19, 18, 22, 20, 16, 25 sati. S razinom značajosti* $\alpha = 0.05$*, testirajte da li dobiveni uzorak indicira kraći prosječan vijek trajanja baterija.*

Rješenje:

$$\mu_0 = 21.5, \quad n = 6, \quad \alpha = 0.05$$

$$H_0 : \mu = 21.5$$
$$H_1 : \mu < 21.5$$

Treba nam vrijednosti test-statistike: $T = \dfrac{\bar{X}_n - \mu_0}{S_n}\sqrt{n} \sim t(n-1)$

$\bar{x}_6 = \dfrac{1}{6}(19 + 18 + 22 + 20 + 16 + 25) = 20$

$s_6^2 = \dfrac{1}{5}\displaystyle\sum_{i=1}^{6}(x_i - \bar{x}_6)^2 = \dfrac{1}{5}\left(\sum_{i=1}^{6} x_i^2 - 6 \cdot \bar{x}_6^2\right) = \dfrac{50}{5} = 10$

$\Rightarrow \ t = \dfrac{20 - 21.5}{\sqrt{10}}\sqrt{6} = -1.162$

$t_{0.05}(5) = 2.015$

$\Rightarrow \ t > -t_{0.05}(5)$

Nultu hipotezu H_0 ne možemo odbaciti, tj. uzorak ne indicira kraći prosječni vijek trajanja baterija. $\square$

6.2 Testovi o očekivanju na osnovi velikih uzoraka

- NE pretpostavljamo da slučajni uzorak uzimamo iz normalno distribuirane populacije
- iz Centralnog graničnog teorema za $n \to \infty$ slijedi da test-statistika

$$Z = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n} \overset{H_0}{\approx} N(0,1)$$

- osnovna hipoteza je ponovo oblika $H_0 : \mu = \mu_0$ za neki unaprijed zadani broj μ_0
- svodi se na testiranje očekivanja normalno distribuirane populacije uz $\sigma \approx S_n$ jer $S_n^2 \to \sigma^2$ kad $n \to \infty$

6.2.1 Test o proporciji

Pogledajmo kako izgleda test za očekivanje na osnovi velikih uzoraka u slučaju kada imamo binomno distribuiranu populaciju.

- promatramo statističko obilježje $X \sim B(n, p)$
- želimo testirati da li je proporcija p jednaka nekom unaprijed zadanom broju p_0. Nulta hipoteza je

$$H_0 : p = p_0.$$

  Za alternativnu možemo uzeti bilo koju od sljedeće tri:

$$H_1 : p \neq p_0 \quad \text{ili} \quad H_1 : p > p_0 \quad \text{ili} \quad H_1 : p < p_0$$

- u sva 3 slučaja koristimo istu test-statistiku:

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \sim N(0,1)$$

  gdje je $\bar{X} = \hat{P}$

Promotrimo redom slučajeve različitog izbora alternativne hipoteze:

1.

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

Nultu hipotezu H_0 odbacujemo ako je $\quad Z > z_{\frac{\alpha}{2}} \quad$ ili $\quad Z < -z_{\frac{\alpha}{2}}$

2.

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

H_0 odbacujemo ako je $\quad Z > z_\alpha$

3.

$$H_0 : p = p_0$$
$$H_1 : p < p_0$$

H_0 odbacujemo ako je $\quad Z < -z_\alpha$

Zadatak 24 *Proizvođač tvrdi da njegove pošiljke sadrže najviše* 7% *defektnih proizvoda. Uzet je slučajni uzorak od 200 komada iz jedne pošiljke i bilo je 11 defektnih. Da li biste prihvatili tvrdnju proizvođača uz razinu značajnosti* 0.05*?*

Rješenje: Postavljamo hipoteze:

$$H_0 : p = 0.07$$
$$H_1 : p > 0.07$$

Kada bi za alternativnu hipotezu postavili $H_1 : p \neq 0.07$, u slučaju odbacivanja nulte hipoteze mogli bi zaključiti samo da proporcija defektnih nije

0.07, a to može značiti da je veća, ali i da je manja od te vrijednosti što je još bolje. Izračunajmo vrijednost odgovarajuće test-statistike:

$$\bar{x}_{200} = \hat{p} = \frac{11}{200} = 0.055 \implies z = \frac{0.055 - 0.07}{\sqrt{0.07 \cdot 0.93}}\sqrt{200} = -0.83$$
$$z_\alpha = z_{0.05} = 1.65$$
$$\Rightarrow \; z < z_{0.05}$$

Nultu hipotezu H_0 ne možemo odbaciti, tj. možemo zaključiti da pošiljke sadrže najviše 7% defektnih proizvoda. □

6.3 Usporedba očekivanja dviju normalno distribuiranih populacija (t-test)

- pretpostavimo da mjerimo isto statističko obilježje X na dvije različite populacije
- pretpostavimo da je u obje populacije X normalno distribuirana slučajna varijabla **s jednakom varijancom** σ

$$X^{(1)} : \text{statističko obilježje } X \text{ za populaciju 1}, \quad X^{(1)} \sim N(\mu_1, \sigma^2)$$
$$X^{(2)} : \text{statističko obilježje } X \text{ za populaciju 2}, \quad X^{(2)} \sim N(\mu_2, \sigma^2)$$

- iz svake populacije uzimamo uzorak:

$$X_1^{(1)},\ X_2^{(1)}, \ldots, X_{n_1}^{(1)} \ \text{ za } X^{(1)} \text{ duljine } n_1$$
$$X_1^{(2)},\ X_2^{(2)}, \ldots, X_{n_2}^{(2)} \ \text{ za } X^{(2)} \text{ duljine } n_2$$

- želimo testirati sljedeću nultu hipotezu

$$H_0 : \mu_1 = \mu_2$$

u odnosu na neku od jednostranih alternativa

$$H_1 : \mu_1 < \mu_2 \quad \text{ili} \quad H_1 : \mu_1 > \mu_2$$

ili u odnosu na dvostranu alternativu

$$H_1 : \mu_1 \neq \mu_2$$

- u svim slučajevima koristimo istu test-statistiku
$$T = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
gdje su
$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^{(1)}, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^{(2)},$$
$$S^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right)$$
za S_1^2, S_2^2 uzoračke varijance uzoraka 1 i 2. S^2 se interpretira kao **zajednička varijanca uzoraka** 1 i 2. Ako je H_0 istinita, tada je
$$T \sim t(n_1 + n_2 - 2)$$

1.
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$
Nultu hipotezu H_0 odbacujemo ako
$$T > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \quad \text{ili} \quad T < -t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

2.
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$
Nultu hipotezu H_0 odbacujemo ako
$$T > t_{\alpha}(n_1 + n_2 - 2)$$

3.
$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$
Nultu hipotezu H_0 odbacujemo ako
$$T < -t_{\alpha}(n_1 + n_2 - 2)$$

Zadatak 25 *Ista vrsta jabuka uzgaja se u Slavoniji i u Zagorju. Na slučajan način izabrano je 7 slavonskih stabala te je izmjeren njihov prinos (u kg): 28, 26, 33, 29, 31, 27, 28; prinos sa 10 zagorskih stabala bio je: 36, 25, 21, 29, 30, 36, 27, 28, 30, 37. Uz razinu značajnosti 0.01, testirajte hipotezu da jabuke u Zagorju daju veći prinos, ako je poznato da je prinos normalna slučajna varijabla. Možemo li, uz istu razinu značajnosti, zaključiti da se prinosi jabuka u Slavoniji i Zagorju razlikuju?*

Rješenje:

$$n_1 = 7, \quad n_2 = 10$$

Postavljamo hipoteze

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

Koristimo test-statistiku

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\begin{aligned}
&\bar{x}_1 = \frac{1}{7}(28 + 26 + 33 + 29 + 31 + 27 + 28) = 28.857 \\
&\bar{x}_2 = \frac{1}{10}(36 + 25 + 21 + 29 + 30 + 36 + 27 + 28 + 30 + 37) = 29.9 \\
&s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\
&\Rightarrow \ s_1^2 = \frac{1}{6} \cdot 34.855 = 5.81, \quad s_2^2 = \frac{1}{9} \cdot 240.9 = 26.767 \\
&s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{6 \cdot 5.81 + 9 \cdot 26.767}{7 + 12 - 2} = 18.3842 \\
&\Rightarrow \ s = 4.2877 \\
&t = \frac{28.857 - 29.9}{4.2877\sqrt{\frac{1}{7} + \frac{1}{10}}} = -0.4936 \\
&t_\alpha(n_1 + n_2 - 2) = t_{0.01}(15) = 2.602 \\
&\Rightarrow t > -t_{0.01}(15)
\end{aligned}$$

Ne možemo odbaciti H_0, tj. ne možemo zaključiti da jabuke u Zagorju daju veći prinos.
Ako želimo testirati da li su prinosi različiti, moramo postaviti hipoteze

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

Tada nam treba

$$t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) = t_{0.005}(15) = 2.949$$

Kako je

$$t > -t_{0.005}(15)$$

(i očito $t < t_{0.005}(15)$) ponovo ne možemo odbaciti nultu hipotezu, tj. ne možemo zaključiti da se prinosi jabuka razlikuju. □

6.4 Usporedba proporcija

- promatramo dvije populacije i neko njihovo Bernoullijevo statističko obilježje X
  $X^{(1)}$: slučajna varijabla koja reprezentira X u populaciji 1
  $X^{(2)}$: slučajna varijabla koja reprezentira X u populaciji 2
- pripadni parametri (vjerojatnosti uspjeha): p_1, p_2
- sa $\hat{p}_1$ i $\hat{p}_2$ označimo procjenitelje od p_1 i p_2 na bazi uzorka iz svake od populacija duljine n_1 i n_2 (uzorci su međusobno nezavisni), te sa

  $$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

  procjenu zajedničke vjerojatnosti uspjeha
- koristimo test-statistiku

  $$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})}} \cdot \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- za velike uzorke, tj. kada $\min(n_1, n_2) \to +\infty$, vrijedi $Z \approx N(0,1)$

1.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_{\frac{\alpha}{2}} \quad \text{ili} \quad Z < -z_{\frac{\alpha}{2}}$$

2.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z > z_\alpha$$

3.

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 < p_2$$

Nultu hipotezu H_0 odbacujemo ako

$$Z < -z_\alpha$$

Zadatak 26 *Uzorci od 300 glasača iz županije A i 200 glasača iz županije B pokazali su da će* 56% *i* 48% *ljudi, redom, glasati za nekog određenog kandidata. S razinom značajnosti 0.05, testirajte hipotezu da*
a) postoji razlika među županijama
b) tog kandidata više "vole" u županiji A.

Rješenje:

$$n_1 = 300, \quad \hat{p}_1 = 0.56$$

$$n_2 = 200, \quad \hat{p}_2 = 0.48$$

$$a) \qquad H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{300 \cdot 0.56 + 200 \cdot 0.48}{500} = 0.528$$

$$z = \frac{0.56 - 0.48}{\sqrt{0.528 \cdot 0.472}} \cdot \frac{1}{\sqrt{\frac{1}{300} + \frac{1}{200}}} = 1.75$$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$
$$\Rightarrow \; z < z_{0.025}$$

$\Longrightarrow$ Ne možemo odbaciti nultu hipotezu, tj. ne možemo zaključiti da postoji razlika među županijama.

$$b) \qquad H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

$$z_\alpha = z_{0.05} = 1.64 \quad \Rightarrow \; z > z_{0.05}$$

$\Longrightarrow$ Odbacujemo nultu hipotezu, tj. možemo zaključiti da kandidata više "vole" u županiji A. $\square$

6.5 Usporedba varijanci dviju normalno distribuiranih populacija (F-test)

- neka je $X^{(1)} \sim N(\mu_1, \sigma_1^2), \quad X^{(2)} \sim N(\mu_2, \sigma_2^2)$
- imamo slučajne uzorke veličine n_i od X_i, $i = 1, 2$

$$X_1^{(1)},\ X_2^{(1)}, \ldots, X_{n_1}^{(1)} \ \text{ za } X^{(1)} \text{ duljine } n_1$$
$$X_1^{(2)},\ X_2^{(2)}, \ldots, X_{n_2}^{(2)} \ \text{ za } X^{(2)} \text{ duljine } n_2$$

- test- statistika
$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$
ima **Fisherovu** ili **F-distribuciju** sa parom stupnjeva slobode $(n_1 - 1, n_2 - 1)$.

- Vrijedi
$$f_{1-\frac{\alpha}{2}}(n_1, n_2) = \frac{1}{f_{\frac{\alpha}{2}}(n_2, n_1)}$$

1.
$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako
$$F > f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) \quad \text{ili} \quad F < f_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$$

2.
$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako
$$F > f_{\alpha}(n_1 - 1, n_2 - 1)$$

3.
$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$

Nultu hipotezu H_0 odbacujemo ako
$$F < f_{1-\alpha}(n_1 - 1, n_2 - 1)$$

Zadatak 27 *Iz dva 3.razreda neke srednje škole izabrano je, na slučajan način, po 10 učenika i izmjerena je njihova težina (zna se da je težina normalno distribuirana), a podaci su dani u tablici. Uz razinu značajnosti 0.02, testirajte hipotezu da su varijance jednake.*

3a:	57	60	63	59	62	60	58	56	54	62
3b:	58	62	60	56	63	58	61	57	53	61

Rješenje:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$\bar{x}_1 = 59.1, \quad \bar{x}_2 = 58.9$$
$$s_1^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - n\bar{x}^2\right) = 8.322, \quad s_2^2 = 9.433$$
$$\Rightarrow\ f = \frac{s_1^2}{s_2^2} = \frac{8.322}{9.433} = 0.8822$$
$$f_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = f_{0.01}(9, 9) = 5.35$$
$$f_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1) = f_{0.99}(9, 9) = \frac{1}{f_{\frac{\alpha}{2}}(n_2 - 1, n_1 - 1)} = \frac{1}{f_{0.01}(9, 9)} = 0.1869$$
$$\Rightarrow\ f_{0.99}(9, 9) < f < f_{0.01}(9, 9)$$

Ne možemo odbaciti nultu hipotezu, tj. možemo zaključiti da su varijance u ova dva uzorka jednake. □

6.6 χ^2 - test o prilagodbi modela podacima

- test-statistika je općenito

$$H = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'}$$

gdje su f_i eksperimentalne, a $f_i' = np_i$ teorijske frekvencije.

- ako vrijedi H_0, tada za velike n ($n \to \infty$)

$$H \sim \chi^2(k-r-1)$$

gdje $\chi^2(m)$ označava χ^2-**razdiobu s m stupnjeva slobode**.

- pritom je

$$\begin{aligned} k &= \text{(konačan) broj razreda u tablici} \\ r &= \text{broj nepoznatih parametara} \end{aligned}$$

- nultu hipotezu da se radi o određenoj razdiobi odbacujemo ako

$$H \geq \chi^2_\alpha(k-r-1)$$

Zadatak 28 *Proizvođač tvrdi da je 5% njegovih proizvoda prve klase, 92% druge i 3% treće klase. U slučajnom uzorku od 500 proizvoda nađeno je 40 proizvoda prve, 432 druge i 28 treće klase. Uz razinu značajnosti 0.05, testirajte hipotezu da je proizvođač u pravu.*

Rješenje: Proizvođač tvrdi da njegovi proizvodi imaju neku distribuciju, odnosno razdiobu. Govori li istinu, provjerit ćemo χ^2- testom. Duljina uzorka je $n = 500$. Kako bismo izračunali vrijednost odgovarajuće test-statistike trebaju nam teorijske frekvencije. Njih računamo po formuli $f_i' = np_i$ gdje je p_i odgovarajuća vjerojatnost, odnosno u ovom slučaju odgovarajuća proporcija. Tako je

$$p_1 = \frac{5}{100}, \quad p_2 = \frac{92}{100}, \quad p_3 = \frac{3}{100}.$$

Formirajmo tablicu:

i	f_i	f_i'	$\frac{(f_i-f_i')^2}{f_i'}$
1	40	$500 \cdot \frac{5}{100} = 25$	9
2	432	$500 \cdot \frac{92}{100} = 460$	1.7
3	28	$500 \cdot \frac{3}{100} = 15$	11.27
Σ	500	500	21.97

Suma poslijednjeg stupca u tablici daje nam vrijednost tražene test-statistike:

$$h = \sum_{i=1}^{3} \frac{(f_i - f_i')^2}{f_i'} = 21.97$$

Tablična vrijednost s kojom ju moramo usporediti kako bismo donijeli odluku o istinitosti nulte hipoteze je $\chi^2_\alpha(k-r-1)$. α je zadana (=0.05), $k = 3$ (ukupan broj razreda), a $r = 0$ (nije bilo nijednog nepoznatog parametra pa ništa nije bilo potrebno procijenjivati). Dakle,

$$\chi^2_\alpha(k-r-1) = \chi^2_{0.05}(2) = 6.0$$

Kako je

$$h > \chi^2_{0.05}(2),$$

što znači da je vrijednost test-statistike upala u kritično područje, moramo odbaciti nultu hipotezu. Drugim riječima, odbacujemo tvrdnju proizvođača, tj. on nije u pravu. □

Zadatak 29 *Pet novčića, s istom ali nepoznatom vjerojatnošću p da padne pismo, bacaju se 100 puta (rezultati su dani u tablici). Uz razinu značajnosti 0.01, testirajte hipotezu da broj pisama koji se dobije u jednom bacanju predstavlja binomnu slučajnu varijablu.*

broj pisama x_i	0	1	2	3	4	5
frekvencija f_i	3	16	36	32	11	2

Rješenje: Potrebno je provjeriti imaju li dani podaci binomnu distribuciju. Pokus koji izvodimo (ponavljamo ga 100 puta, dakle $n = 100$) je bacanje novčića 5 puta a "uspjeh" je "palo je pismo". Slučajna varijabla X broji pisma. Parametar n binomne distribucije je stoga jednak 5. Parametar p nije zadan te moramo ga procijeniti. Oprez! mali n sada označava i duljinu uzorka i parametar distribucije, no to su različite stvari i različite vrijednosti pa treba na to pripaziti.

Parametar p jednak je vjerojatnosti "uspjeha" u jednom bacanju novčića. Njegovu procjenu dobijemo tako da ukupan broj palih pisama podijelimo sa ukupnim brojem bacanja novčića. Novčić je ukupno bačen $5 \cdot 100 = 500$ puta (100 pokusa a svaki se sastoji od 5 bacanja). Ukupan broj pisama računamo pomoću dane tablice:

$$0 \cdot 3 + 1 \cdot 16 + 2 \cdot 36 + 3 \cdot 32 + 4 \cdot 11 + 5 \cdot 2 = 238.$$

Konačno,

$$\hat{p} = \frac{238}{500} = 0.476$$

Sljedeći korak je izračunati teorijske frekvencije $f_i' = np_i$. Funkcija gustoće slučajne varijable $X \sim B(5, 0.476)$ je

$$p_i := p_X(i) = P(X = i) = \binom{5}{i}(0.476)^i \cdot (0.524)^{5-i},$$

pa dobivamo

$$\begin{aligned}
f_0' &= 100 \cdot p_0 = 100 \cdot \binom{5}{0}(0.476)^0 \cdot (0.524)^5 = 3.95054 \\
f_1' &= 100 \cdot p_1 = 100 \cdot \binom{5}{1}(0.476)^1 \cdot (0.524)^4 = 17.9433 \\
f_2' &= 100 \cdot p_2 = 100 \cdot \binom{5}{2}(0.476)^2 \cdot (0.524)^3 = 32.6 \\
f_3' &= 100 \cdot p_3 = 100 \cdot \binom{5}{3}(0.476)^3 \cdot (0.524)^2 = 29.613 \\
f_4' &= 100 \cdot p_4 = 100 \cdot \binom{5}{4}(0.476)^4 \cdot (0.524)^1 = 13.45 \\
f_5' &= 100 \cdot p_5 = 100 \cdot \binom{5}{5}(0.476)^5 \cdot (0.524)^0 = 2.4436
\end{aligned}$$

Uočimo da je teorijska frekvencija prvog i poslijednjeg razreda < 5. Stoga ćemo te razrede spojiti s njima susjednim razredima. Ukoliko bi tako opet dobili razred čija je teorijska frekvencija stogo manja od 5, postupak bi ponovljali dok ne dobijemo razred s (ukupnom) teorijskom frekvencijom > 5. Sada formiramo tablicu:

i	f_i	f'_i	$\frac{(f_i-f'_i)^2}{f'_i}$
1	$3+16=\mathbf{19}$	$3.95054+17.9433=\mathbf{21.89384}$	0.3825
2	36	32.6	0.3546
3	32	29.613	0.1924
4	$11+2=\mathbf{13}$	$13.45+2.4436=\mathbf{15.8936}$	0.5268
Σ	100	100	1.4563

Vrijednost test-statistike je dakle

$$h = 1.4563.$$

Konačan broj razreda $k = 4$, a broj procijenjenih parametara $r = 1$. Iz tablice očitavamo

$$\chi^2_\alpha(k-r-1) = \chi^2_{0.01}(2) = 9.2$$

Kako je

$$h < \chi^2_{0.01}(2),$$

dakle vrijednost test-statistike nije ušla u kritično područje, ne možemo odbaciti nultu hipotezu, odnosno možemo zaključiti da se radi o binomnoj distribuciji. □

Zadatak 30 *Anketirano je 100 studenata i dobiven je prosječan broj njihovih odlazaka u kazalište tijekom godine. S nivoom signifikantnosti 0.05, testirajte hipotezu da se radi o uzorku iz populacije s normalnom distribucijom.*

broj posjeta	$[0,2\rangle$	$[2,4\rangle$	$[4,6\rangle$	$[6,8\rangle$	$[8,10\rangle$	$[10,12\rangle$	$[12,14\rangle$
broj studenata	5	10	20	33	18	10	4

Rješenje: Normalna distribucija ima 2 parametra - očekivanje μ i varijancu σ^2. Kako nijedan od njih nije zadan, moramo ih procijeniti, pa odmah slijedi da je $r = 2$. Procjenitelj za očekivanje je $\hat{\mu} = \bar{x}$ a za varijancu $\hat{\sigma}^2 = s_n^2$.

U tablici su dani sortirani podaci. Vidimo da je 5 studenata išlo u kazalište 0 ili 1 put ali ne znamo koliko točno od tih 5 je išlo 0 a koliko

1 put. Treba nam "predstavnik" tog razreda - uzimamo sredinu razreda. Sada

$$\hat{\mu} = \bar{x} = \frac{1 \cdot 5 + 3 \cdot 10 + 5 \cdot 20 + 7 \cdot 33 + 9 \cdot 18 + 11 \cdot 10 + 13 \cdot 4}{100} = 6.9$$

$$\hat{\sigma}^2 = s_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} a_i^2 \cdot f_i - n\bar{x}^2\right)$$

no kako je $n = 100$ velik možemo umjesto s $n-1$ dijeliti s n:

$$\Rightarrow \hat{\sigma}^2 = \frac{1^2 \cdot 5 + 3^2 \cdot 10 + 5^2 \cdot 20 + 7^2 \cdot 33 + 9^2 \cdot 18 + 11^2 \cdot 10 + 13^2 \cdot 4}{100} - 6.9^2 = 7.95$$

Postavljamo (nultu) hipotezu da slučajna varijabla X koja broji odlaske u kazalište ima distribuciju

$$X \sim N(6.9, 7.95)$$

Sljedeći korak je odrediti teorijske frekvencije $f_i' = 100 \cdot p_i$. Imamo

$$\begin{aligned}
p_1 &= P(0 \leq X < 2) = P\left(\frac{0-6.9}{\sqrt{7.95}} \leq X^* < \frac{2-6.9}{\sqrt{7.95}}\right)\\
&= \Phi_0(-1.74) - \Phi_0(-2.45) = \Phi_0(2.45) - \Phi_0(1.74)\\
&= 0.4928572 - 0.4591 = 0.0338 \quad \Rightarrow \quad f_1' = 3.38\\
p_2 &= P(2 \leq X < 4) = P\left(\frac{2-6.9}{2.82} \leq X^* < \frac{4-6.9}{2.82}\right)\\
&= \Phi_0(-1.03) - \Phi_0(-1.74) = \Phi_0(1.74) - \Phi_0(1.03)\\
&= 0.4591 - 0.3485 = 0.1106 \quad \Rightarrow \quad f_2' = 11.06\\
p_3 &= P(4 \leq X < 6) = P\left(-1.03 \leq X^* < -0.32\right)\\
&= \Phi_0(-0.32) - \Phi_0(-1.03) = 0.223 \quad \Rightarrow \quad f_3' = 22.3\\
p_4 &= P(6 \leq X < 8) = P\left(-0.32 \leq X^* < 0.39\right)\\
&= \Phi_0(0.39) - \Phi_0(-0.32) = 0.2772 \quad \Rightarrow \quad f_4' = 27.72\\
p_5 &= P(8 \leq X < 10) = P\left(0.39 \leq X^* < 1.10\right)\\
&= \Phi_0(1.10) - \Phi_0(0.39) = 0.2126 \quad \Rightarrow \quad f_5' = 21.26\\
p_6 &= P(10 \leq X < 12) = P\left(1.1 \leq X^* < 1.8\right)\\
&= \Phi_0(1.8) - \Phi_0(1.1) = 0.09974 \quad \Rightarrow \quad f_6' = 9.97
\end{aligned}$$

$$
\begin{aligned}
p_7 &= P(12 \leq X < 14) = P\left(1.8 \leq X^* < 2.52\right) \\
&= \Phi_0(2.52) - \Phi_0(1.8) = 0.03006 \quad \Rightarrow \quad f_7' = 3
\end{aligned}
$$

Budući je $f_1' < 5$ i $f_7' < 5$, spojit ćemo prva dva i posljednja dva razreda, pa će tako ostati ukupno 5 razreda. Dakle, $k = 5$. Formiramo tablicu:

i	1	2	3	4	5	Σ
f_i	15	20	33	18	14	100
f_i'	14.44	22.3	27.72	21.26	12.97	
$\frac{(f_i-f_i')^2}{f_i'}$	0.022	0.237	1.006	0.499	0.082	1.846

Vrijednost test-statistike je prema tome

$$
h = \sum_{i=1}^{5} \frac{(f_i - f_i')^2}{f_i'} = 1.846,
$$

a

$$
\chi^2_\alpha(k - r - 1) = \chi^2_{0.05}(2) = 6,
$$

pa kako je $h < \chi^2_{0.05}(2)$, nultu hipotezu ne možemo odbaciti, odnosno zaključujemo da se radi o uzorku iz normalno distribuirane populacije. $\square$

Zadatak 31 *(DZ) Bilježen je broj četvorki rođenih u nekoj županiji tijekom 70 godina. Podaci su dani u tablici. Uz razinu značajnosti 0.05, testirajte hipotezu da su podaci uzeti iz populacije s Poissonovom distribucijom.*

broj rođenih četvorki	0	1	2	3	4	5	6
broj godina	14	24	17	10	2	2	1

Napomena: $\hat{\lambda} = \bar{x}$

6.7 χ^2 - test nezavisnosti dviju varijabli

Neka je $(X_1, Y_1),\ (X_2, Y_2),\ \ldots\ (X_n, Y_n)$ slučajni uzorak za dvodimenzionalno diskretno statističko obilježje (X, Y) i neka je pritom:

$$
\begin{aligned}
&\mathrm{Im}X = \{a_1, \ldots, a_r\} \\
&\mathrm{Im}Y = \{b_1, \ldots, b_s\} \\
&\Rightarrow\ \mathrm{Im}(X, Y) = \{(a_i, b_j) : 1 \leq i \leq r,\ \ 1 \leq j \leq s\}
\end{aligned}
$$

Nadalje,

f_{ij} : frekvencija od (a_i, b_j) u uzorku

f_i : (marginalna) frekvencija od a_i u uzorku

g_j : (marginalna) frekvencija od b_j u uzorku

Vrijedi:

$$f_i = \sum_{j=1}^{s} f_{ij}, \quad g_j = \sum_{i=1}^{r} f_{ij}$$

Kontingencijska frekvencijska tablica:

$X \setminus Y$	b_1	b_2	$\dots$	b_s	Σ
a_1	f_{11}	f_{12}	$\dots$	f_{1s}	f_1
a_2	f_{21}	f_{22}	$\dots$	f_{2s}	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\dots$	f_{rs}	f_r
Σ	g_1	g_2	$\dots$	g_s	n

Označimo:

$$p_{ij} = P(X = a_i,\ Y = b_j)$$
$$p_i = P(X = a_i)$$
$$q_j = P(X = b_j)$$

Hipoteze su:

H_0 : $p_{ij} = p_i \cdot q_j, \quad \forall\, i, j$

tj. X i Y su nezavisne slučajne varijable

H_1 : $\exists\, i, j$ takvi da $p_{ij} \neq p_i \cdot q_j$

Uz H_0, procjene za p_i i q_j su:

$$\hat{p}_i = \frac{f_i}{n}, \quad \hat{q}_j = \frac{g_j}{n}$$

Očekivane vrijednosti f'_{ij} od f_{ij} uz H_0 su:

$$f'_{ij} = n\,\hat{p}_i\,\hat{q}_j = n \cdot \frac{f_i}{n} \cdot \frac{g_j}{n} = \frac{f_i \cdot g_j}{n}$$

Koristimo test-statistiku

$$H = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

Ako je H_0 istinita, tada

$$H \sim \chi^2((r-1)(s-1))$$

Hipotezu o nezavisnosti odbacujemo ako

$$H \geq \chi^2_\alpha((r-1)(s-1))$$

Zadatak 32 *U cilju ispitivanja sklonosti potrošača proizvodu A uzet je uzorak na temelju kojeg su dobiveni podaci dani u tablici. Možete li na osnovu ovih podataka zaključiti da sklonost potrošača proizvodu A NE ovisi o njihovom dohotku, uz razinu značajnosti 0.05?*

mjesečni dohodak anketiranih kupaca u kn	*sklonost potrošnji*		
	stalno kupuju	*povremeno kupuju*	*ne kupuju*
−3000	70	17	21
3000 − 5000	165	56	28
5000 − 7000	195	85	26
7000−	170	42	25

Rješenje: Označimo s X slučajnu varijablu koja mjeri visinu dohotka, a s Y onu koja mjeri sklonost potrošnji. Postavljamo hipoteze:

H_0 : X i Y su nezavisne slučajne varijable

H_1 : X i Y su zavisne slučajne varijable

Provest ćemo χ^2-test o nezavisnosti dviju varijabli. Potrebno je izračunati teorijske frekvencije f'_{ij} za $i = 1, 2, 3, 4$, $j = 1, 2, 3$, no pogledajmo najprije kolike su marginalne frekvencije f_i i g_j:

mjesečni dohodak	stalno kupuju	povremeno kupuju	ne kupuju	$\sum$
-3000	70	17	21	$f_1 = 108$
$3000 - 5000$	165	56	28	$f_2 = 249$
$5000 - 7000$	195	85	26	$f_3 = 306$
$7000-$	170	42	25	$f_4 = 237$
$\sum$	$g_1 = 600$	$g_2 = 200$	$g_3 = 100$	$n = 900$

Sada dobivamo:

$$f'_{11} = \frac{f_1 \cdot g_1}{n} = \frac{108 \cdot 600}{900} = 72 \qquad f'_{31} = \frac{f_3 \cdot g_1}{n} = \frac{306 \cdot 600}{900} = 204$$

$$f'_{12} = \frac{f_1 \cdot g_2}{n} = \frac{108 \cdot 200}{900} = 24 \qquad f'_{32} = \frac{f_3 \cdot g_2}{n} = \frac{306 \cdot 200}{900} = 68$$

$$f'_{13} = \frac{f_1 \cdot g_3}{n} = \frac{108 \cdot 100}{900} = 12 \qquad f'_{33} = \frac{f_3 \cdot g_3}{n} = \frac{306 \cdot 100}{900} = 34$$

$$f'_{21} = \frac{f_2 \cdot g_1}{n} = \frac{249 \cdot 600}{900} = 166 \qquad f'_{41} = \frac{f_4 \cdot g_1}{n} = \frac{237 \cdot 600}{900} = 158$$

$$f'_{22} = \frac{f_2 \cdot g_2}{n} = \frac{249 \cdot 200}{900} = 55.3 \qquad f'_{42} = \frac{f_4 \cdot g_2}{n} = \frac{237 \cdot 200}{900} = 52.67$$

$$f'_{23} = \frac{f_2 \cdot g_3}{n} = \frac{249 \cdot 100}{900} = 27.67 \qquad f'_{43} = \frac{f_4 \cdot g_3}{n} = \frac{237 \cdot 100}{900} = 26.3$$

Da bismo lakše izračunali vrijednost test-statistike, zgodno je, radi preglednosti, u tablici eksperimentalnim frekvencijama pridružiti odgovarajuće teorijske:

mjesečni dohodak	stalno kupuju	povremeno kupuju	ne kupuju
-3000	70/*72*	17/*24*	21/*12*
$3000 - 5000$	165/*166*	56/*55.3*	28/*27.67*
$5000 - 7000$	195/*204*	85/*68*	26/*34*
$7000-$	170/*158*	42/*52.67*	25/*26.3*

Preostalo je izračunati vrijednost test-statistike:

$$h = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}} = 18.532$$

Iz tablice očitavamo:

$$\chi^2_\alpha((r-1)(s-1)) = \chi^2_{0.05}((4-1)(3-1)) = \chi^2_{0.05}(6) = 12.6,$$

pa kako je

$$h > \chi^2_{0.05}(6)$$

vidimo da je vrijednost test-statistike upala u kritično područje. Nultu hipotezu o nezavisnosti stoga odbacujemo i zaključujemo da su visina mjesečnog dohotka (slučajna varijabla X) i sklonost potrošnji (slučajna varijabla Y) međusobno zavisne. □

6.8 χ^2 - test homogenosti populacija

- zanima nas razdioba istog diskretnog statističkog obilježja u raznim populacijama
- na osnovi nezavisnih uzoraka uzetih iz tih populacija, testiramo osnovnu hipotezu da su razdiobe od X u tim populacijama jednake, tj. da su populacije *homogene* obzirom na X
- m : broj populacija koje promatramo
  $X^{(i)}$: slučajna varijabla koja predstavlja X u i-toj populaciji ($i = 1, \ldots, m$); vrijedi

$$X^{(i)} \sim \begin{pmatrix} a_1 & a_2 & \ldots & a_k \\ p_1^{(i)} & p_2^{(i)} & \ldots & p_k^{(i)} \end{pmatrix}$$

- nulta hipoteza je da su sve $X^{(i)}$ jednake po distribuciji, a alternativna je da postoji bar jedna koja se po distribuciji razlikuje od ostalih, odnosno:

$$
\begin{aligned}
H_0 &: \quad X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} \ldots \stackrel{D}{=} X^{(m)} \\
H_1 &: \quad \exists\, i, j \quad \text{tako da} \quad X^{(i)} \stackrel{D}{\neq} X^{(j)}
\end{aligned}
$$

- H_0 se može zapisati i ovako

$$H_0 \,:\, p_j^{(i)} = p_j, \quad j = 1, \ldots, k, \quad i = 1, \ldots, m$$

gdje p_j predstavljaju zajedničke (tj. po populacijama jednake) vjerojatnosti od a_j

Frekvencijska tablica:

X	a_1	a_2	$\ldots$	a_k	$\sum$
populacija 1	f_{11}	f_{12}	$\ldots$	f_{1k}	n_1
populacija 2	f_{21}	f_{22}	$\ldots$	f_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
populacija m	f_{m1}	f_{m2}	$\ldots$	f_{mk}	n_m
$\sum$	f_1	f_2	$\ldots$	f_k	n

- n_i : duljina uzroka iz i-te populacije,
  f_{ij} : frekvencija od a_j u uzorku iz i-te populacije
  $f_j = \sum_{i=1}^{m} f_{ij}$: frekvencija od a_j u svim uzorcima zajedno

- vrijedi: $n_i = \sum_{j=1}^{k} f_{ij}$

- procjena zajedničkih vrijednosti p_j ako vrijedi H_0:

$$\hat{p}_j = \frac{f_j}{n}, \quad j = 1, \ldots, k$$

- očekivane frekvencije (ako vrijedi H_0):

$$f'_{ij} = n_i \cdot \hat{p}_j = \frac{n_i \cdot f_j}{n}$$

- koristimo test-statistiku:

$$H = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

Ako je H_0 istinita, tada

$$H \sim \chi^2((m-1)(k-1))$$

- hipotezu o homogenosti populacija odbacujemo ako

$$H \geq \chi^2_\alpha((m-1)(k-1))$$

Zadatak 33 *U tvorničkom pogonu proizvode se televizori. Svakog radnog dana u tjednu registrira se broj neispravnih televizora. Provedena su opažanja tijekom 753 dana i rezultati su prikazani u tablici. Može li se, uz razinu značajnosti 0.05, zaključiti da nema značajne razlike u pojavi neispravnih televizora tijekom tjedna?*

broj neispravnih televizora	*PON*	*UTO*	*SRI*	*ČET*	*PET*
$0-2$	60	63	61	70	50
$3-5$	72	62	60	53	69
$6->$	20	26	28	31	28

Rješenje: Neka je X broj neispravnih televizora po danu. Ako dane u tjednu shvatimo kao 5 različitih populacija (iz kojih su uzeti uzorci), tada je potrebno provjeriti ima li X jednaku distribuciju u svih tih 5 populacija, odnosno dana. To ćemo provjeriti χ^2-testom o homogenosti populacija. Hipoteze su dakle:

H_0 : podaci iz svih 5 populacija potječu iz iste vjerojatnosne razdiobe, tj. $X^{(1)} \stackrel{D}{=} X^{(2)} \stackrel{D}{=} \ldots \stackrel{D}{=} X^{(5)}$

H_1 : ne potječu iz iste razdiobe

Da bismo izračunali vrijednost odgovarajuće test-statistike, potrebne su nam procjene frekvencija f'_{ij}, pa najprije pogledajmo kolike su duljine uzoraka n_i iz svake od populacija ($i = 1, 2, 3, 4, 5$) i kumulativne frekvencije f_j svake od mogućih vrijednosti koje X poprima ($j = 1, 2, 3$):

broj neispr.tv	PON	UTO	SRI	ČET	PET	$\sum$
$0-2$	60	63	61	70	50	$f_1 = 304$
$3-5$	72	62	60	53	69	$f_2 = 316$
$6->$	20	26	28	31	28	$f_3 = 133$
$\sum$	$n_1 = 152$	$n_2 = 151$	$n_3 = 149$	$n_4 = 154$	$n_5 = 147$	$n = 753$

Sada:

$$f'_{11} = \frac{n_1 \cdot f_1}{n} = \frac{152 \cdot 304}{753} = 61.365 \qquad f'_{33} = \frac{n_3 \cdot f_3}{n} = \frac{149 \cdot 133}{753} = 26.317$$

$$f'_{12} = \frac{n_1 \cdot f_2}{n} = \frac{152 \cdot 316}{753} = 63.788 \qquad f'_{41} = \frac{n_4 \cdot f_1}{n} = \frac{154 \cdot 304}{753} = 62.173$$

$$f'_{13} = \frac{n_1 \cdot f_3}{n} = \frac{152 \cdot 133}{753} = 26.847 \qquad f'_{42} = \frac{n_4 \cdot f_2}{n} = \frac{154 \cdot 316}{753} = 64.627$$

$$f'_{21} = \frac{n_2 \cdot f_1}{n} = \frac{151 \cdot 304}{753} = 60.962 \qquad f'_{43} = \frac{n_4 \cdot f_3}{n} = \frac{154 \cdot 133}{753} = 27.2$$

$$f'_{22} = \frac{n_2 \cdot f_2}{n} = \frac{151 \cdot 316}{753} = 63.368 \qquad f'_{51} = \frac{n_5 \cdot f_1}{n} = \frac{147 \cdot 304}{753} = 59.347$$

$$f'_{23} = \frac{n_2 \cdot f_3}{n} = \frac{151 \cdot 133}{753} = 26.671 \qquad f'_{52} = \frac{n_5 \cdot f_2}{n} = \frac{147 \cdot 316}{753} = 61.689$$

$$f'_{31} = \frac{n_3 \cdot f_1}{n} = \frac{149 \cdot 304}{753} = 60.154 \qquad f'_{53} = \frac{n_5 \cdot f_3}{n} = \frac{147 \cdot 133}{753} = 25.964$$

$$f'_{32} = \frac{n_3 \cdot f_2}{n} = \frac{149 \cdot 316}{753} = 62.529$$

Vrijednost test-statistike je:

$$H = \sum_{i=1}^{5} \sum_{j=1}^{3} \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}} = 9.277$$

Iz tablice za χ^2-razdiobu očitavamo

$$\chi^2_\alpha((m-1)(k-1)) = \chi^2_{0.05}(4 \cdot 2) = \chi^2_{0.05}(8) = 15.5$$

Kako je

$$h < \chi^2_{0.05}(8),$$

vidimo da vrijednost test-statistike nije upala u kritično područje pa nultu hipotezu ne možemo odbaciti. Dakle, možemo zaključiti da su populacije homogene što znači da promatrano statističko obilježje (= broj pokvarenih televizora) ima jednaku distribuciju u svim populacijama (= u svim danima).

$\square$

6.9 Usporedba očekivanja više normalno distribuiranih populacija (jednofaktorska analiza varijance ANOVA)

- ANOVA-u koristimo za usporedbu *više od dvije* normalno distribuirane populacije (za usporedbu *točno dvije* normalno distribuirane populacije koristimo `t-test`!)

- neka su

$$\begin{array}{lcl} X_{11},\ X_{12},\ldots,X_{1n_1} & \text{za} & X^{(1)} \sim N(\mu_1,\sigma^2) \\ X_{21},\ X_{22},\ldots,X_{2n_2} & \text{za} & X^{(2)} \sim N(\mu_2,\sigma^2) \\ \vdots & & \vdots \\ X_{k1},\ X_{k2},\ldots,X_{kn_k} & \text{za} & X^{(k)} \sim N(\mu_k,\sigma^2) \end{array}$$

  k nezavisnih slučajnih uzoraka, svaki za normalno distribuirano obilježje X reprezentirano s $X^{(i)}$ za i-tu populaciju iz koje je uzet uzorak duljine n_i $(i = 1, 2, \ldots, k)$

- pretpostavljamo da su varijance od $X^{(i)}$ jednake (u svim populacijama)
- želimo testirati nultu hipotezu

$$H_0 \; : \; \mu_1 = \mu_2 = \ldots = \mu_k,$$

tj. hipotezu da *nema razlike* u očekivanjima među populacijama; alternativna hipoteza je onda naravno da razlika postoji, odnosno da se bar dvije populacije razlikuju po očekivanjima

- za test-statistiku treba nam sljedeće, za $i = 1, 2, \ldots k$:

$$\bar{X}_i = \frac{1}{n_i}(X_{i1} + \ldots + X_{in_i})$$

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

- ukupna aritmetička sredina svih podataka:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i, \quad n = \sum_{i=1}^{k} n_i$$

- suma kvadrata odstupanja srednjih vrijednosti uzoraka od ukupne sredine (= suma kvadrata u odnosu na tretman)

$$SST = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{k} n_i \bar{X}_i^2 - n\bar{X}^2$$

- suma kvadrata pogrešaka

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i \bar{X}_i^2$$

- srednjekvadratno odstupanje među uzorcima (zbog razlike u tretmanima)

$$MST = \frac{SST}{k-1} \quad , \quad MSE = \frac{SSE}{n-k}$$

$$F = \frac{MST}{MSE}$$

- srednjekvadratna pogreška

$$MSE = \frac{SSE}{n-k}$$

- konačno, test-statistika je

$$F = \frac{MST}{MSE}$$

  Ako je H_0 istinita, tada je

$$F \sim F(k-1, n-k)$$

- nultu hipotezu odbacujemo ako

$$F \geq f_\alpha(k-1, n-k)$$

ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
zbog razlike među tretmanima	$k-1$	SST	MST	F
zbog greške	$n-k$	SSE	MSE	
$\sum$	$n-1$	SS		

pritom je

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Zadatak 34 *Pivovara koristi 3 različite linije punjenja limenki piva. Sumnja se da se srednji neto sadržaj limenki razlikuje od linije do linije. Na slučajan način bira se 5 limenki sa svake linije i mjeri se njihov neto sadržaj. Testirajte postoji li značajna razlika između sredina neto sadržaja po linijama uz razinu značajnosti 0.05.*

linija	*sadržaj u dcl*				
1	3.633	3.651	3.66	3.645	3.654
2	3.615	3.627	3.636	3.63	3.624
3	3.645	3.63	3.627	3.63	3.633

Rješenje: Potrebno je provjeriti postoji li razlika između sredina neto sadržaja po linijama. Budući imamo 3 populacije (=linije), t-test nam ne može pomoći, već moramo provesti ANOVA-u. Krenimo redom:

$$k = 3, \quad n_1 = n_2 = n_3 = 5, \quad n = \sum_{i=1}^{3} n_i = 15$$

$$\bar{x}_1 = \frac{3.633 + 3.651 + 3.66 + 3.645 + 3.654}{5} = 3.6486$$

$$\bar{x}_2 = \frac{3.615 + 3.627 + 3.636 + 3.63 + 3.624}{5} = 3.6264$$

$$\bar{x}_3 = 3.633$$

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij} = \frac{1}{15}\sum_{i=1}^{3} n_i \cdot \bar{x}_i = \frac{1}{3}\sum_{i=1}^{3} \bar{x}_i = 3.636$$

$$SST = \sum_{i=1}^{3} n_i \bar{X}_i^2 - n\bar{X}^2 = 5\sum_{i=1}^{3} \bar{x}_i^2 - 15\bar{x}^2 = 0.0013$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} n_i \bar{X}_i^2 = \sum_{i=1}^{3}\sum_{j=1}^{5} x_{ij}^2 - 5\sum_{i=1}^{3} \bar{x}_i^2 = 0.00086$$

$$MST = \frac{SST}{k-1} = \frac{0.0013}{2} = 0.00065$$

$$MSE = \frac{SSE}{n-k} = \frac{0.00086}{15-3} = 0.000072$$

i konačno dobivamo vrijednost test-statistike:

$$\Rightarrow \quad f = \frac{MST}{MSE} = \frac{0.00065}{0.000072} = 9.02778$$

Iz tablice za F-razdiobu potrebno je očitati:

$$f_\alpha(k-1, n-k) = f_{0.05}(2, 12) = 3.89$$

Kako je

$$f > f_{0.05}(2, 12)$$

vidimo da je vrijednost test-statistike upala u kritično područje što znači da nultu hipotezu o jednakosti očekivanja moramo odbaciti. Zaključujemo stoga da `postoji` značajna razlika među sredinama neto sadržaja po linijama.

ANOVA tablica:

izvor rasipanja	stupnjevi slobode	suma kvadrata	srednjekvadratno odstupanje	vrijednost test-statistike
zbog tretmana	2	0.0013	0.00065	
zbog greške	12	0.00086	0.000072	9.02778
$\sum$	14	0.00216		

□

6.10 Test koreliranosti dviju varijabli

- neka je

$$(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$$

slučajni uzorak za normalno distribuirani slučajni vektor (X, Y)

- $\bar{X}$, $\bar{Y}$: aritmetičke sredine uzoraka
- S_x^2, S_y^2 : uzoračke varijance
- kovarijanca od X i Y:

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Vrijedi:

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - \bar{X} \cdot \sum_{i=1}^{n} Y_i - \bar{Y} \cdot \sum_{i=1}^{n} X_i + n\bar{X}\bar{Y}$$
$$= \sum_{i=1}^{n} X_i Y_i - \bar{X} \cdot n\bar{Y} - \bar{Y} \cdot n\bar{X} + n\bar{X}\bar{Y} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

pa onda

$$S_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} \right)$$

- želimo testirati nultu hipotezu

$$H_0\ :\ \rho = 0 \quad (= \text{nema korelacije})$$

u odnosu na jednostranu alternativu

$$H_1 \ : \ \rho > 0 \quad (= \text{korelacija postoji i pozitivna je})$$

ili

$$H_1 \ : \ \rho < 0 \quad (= \text{korelacija postoji i negativna je})$$

ili u odnosu na dvostranu alternativu

$$H_1 \ : \ \rho \neq 0 \quad (= \text{korelacija postoji})$$

- *Pearsonov koeficijent korelacije* je statistika

$$R = \frac{S_{xy}}{S_x \cdot S_y}$$

- test-statistika je:

$$Z = \frac{R}{\sqrt{1-R^2}} \cdot \sqrt{n-2}$$

Ako je H_0 istinita, tada

$$Z \sim t(n-2)$$

1.

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Nultu hipotezu H_0 odbacujemo ako je

$$Z > t_{\frac{\alpha}{2}}(n-2) \qquad \text{ili} \qquad Z < -t_{\frac{\alpha}{2}}(n-2)$$

2.

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

H_0 odbacujemo ako je

$$Z > t_{\alpha}(n-2)$$

3.

$$H_0 : \rho = 0$$
$$H_1 : \rho < 0$$

H_0 odbacujemo ako je

$$Z < -t_\alpha(n-2)$$

Zadatak 35 *U jednom razredu od 30 učenika promatra se ocjena iz matematike (X) i ocjena iz fizike (Y). Uvidom u imenik dobiveni su ovi podaci:* $(1,3),(4,3),(2,2),(3,2),(1,2),(1,1),(2,2),(4,4),(2,2),(3,3),(4,4),(5,5),$ $(3,5),(2,1),(2,3),(2,2),(5,5),(3,3),(2,2),(2,2),(3,3),(3,2),(4,4),(2,2),$ $(3,3),(2,1),(3,2),(3,2),(3,2),(2,2).$
Uz razinu značajnosti 0.05, testirajte hipotezu da nema značajne korelacije između ocjena iz matematike i fizike.

Rješenje: Zanima nas postoji li korelacija između ocjena iz matematike i fizike. To ćemo ispitati pomoću testa o koreliranosti dviju varijabli - X (koja označava ocjene iz matematike) i Y (koja označava ocjene iz fizike). Budući nas zanima samo postoji li korelacije ili ne, a ne da li je (ako postoji) ona pozitivna ili negativna, dovoljno je za alternativnu hipotezu H_1 postaviti $\rho \neq 0$. Dakle,

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

Izračunajmo sada vrijednost odgovarajuće test-statistike:

$$\begin{aligned}
\bar{x} &= \frac{1}{30}(1+4+2+3+1+1+2+4+2+3+\ldots+3+2) = 2.7 \\
\bar{y} &= \frac{1}{30}(3+3+2+2+1+2+4+2+3+4+\ldots+2+2) = 2.63 \\
s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{29}(251 - 30\cdot 2.7^2) = 1.114 \quad\Rightarrow\quad s_x = 1.056 \\
s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) = \frac{1}{29}(245 - 30\cdot 2.63^2) = 1.293 \quad\Rightarrow\quad s_y = 1.137
\end{aligned}$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{29}(239 - 30 \cdot 2.7 \cdot 2.63) = 0.896$$

$$\Rightarrow \quad r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{0.896}{1.056 \cdot 1.137} = 0.746$$

Vrijednost test-statistike je

$$z = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} = \frac{0.746}{\sqrt{1-0.746^2}} \cdot \sqrt{28} = 5.927$$

Iz tablice očitavamo

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(28) = 2.048$$

Kako je

$$z > t_{0.025}(28)$$

vidimo da je vrijednost test-statistike upala u kritično područje, pa nultu hipotezu odbacujemo. Zaključujemo stoga da korelacija između ocjena iz matematike i fizike *postoji*, odnosno da su varijable X i Y korelirane. □

6.11 Linearni regresijski model

Imamo n parova podataka

$$(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$$

koji su dobiveni mjerenjem (opažanjem) nekog dvodimenzionalnog numeričkog statističkog obilježja (X, Y) promatrane populacije. Nezavisna varijabla X interpretira se kao neslučajna a zavisna varijabla Y kao slučajna. Da bi se to naglasilo, X se najčešće zapisuje kao "malo" x. Želimo odrediti linearnu vezu između x i Y:

$$Y = \alpha\, x + \beta + \varepsilon,$$

pri čemu su α, β parametri, x je broj (neslučajna varijabla), a ε slučajna varijabla za koju vrijedi $\mathrm{E}[\varepsilon] = 0$ i koja se najčešće interpretira kao slučajna greška ili šum.

- procjenitelji od (α, β) dobiveni metodom najmanjih kvadrata:

$$\hat{\alpha} := \frac{S_{xy}}{S_x^2}$$
$$\hat{\beta} := \bar{y} - \hat{\alpha}\ \bar{x}$$

- procjenitelj za varijancu σ^2 je:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

pri čemu je

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta} - \hat{\alpha}\ x_i)^2 = S_{yy} - \hat{\alpha}^2 S_{xx}$$

i dakle vrijedi: $\hat{Y}_i = \hat{\alpha}\ x_i + \hat{\beta}$

- $(1-\alpha)\cdot 100\%$ **pouzdan interval za** α:

$$\hat{\alpha} - t_{\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}} \quad\leq\quad \alpha \quad\leq\quad \hat{\alpha} + t_{\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}}$$

- $(1-\alpha)\cdot 100\%$ **pouzdan interval za** β:

$$\hat{\beta} - t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}} \quad\leq\quad \beta \quad\leq\quad \hat{\beta} + t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\ \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}$$

- test-statistike za testiranje sljedećih nul-hipoteza:

  1. $H_0\ :\ \alpha = \alpha_0\ \ (\alpha_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_\alpha = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}}\ \sqrt{(n-1)S_x^2}$$

     Ako je H_0 istinita tada je

$$T_\alpha \sim t(n-2)$$

2. $H_0 \,:\, \beta = \beta_0 \;\; (\beta_0 \in \mathbb{R})$ (u odnosu na razne alternative):

$$T_\beta = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \cdot \sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{(n-1)S_x^2}}}$$

Ako je H_0 istinita tada je

$$T_\beta \sim t(n-2)$$

Zadatak 36 *Izabrano je 5 osoba starih 35, 45, 55, 65 i 75 godina (x), kojima je izmjeren krvni tlak (Y), pri čemu su dobiveni podaci: 114, 124, 143, 158, 166 redom. Odredite:*

a) procjenu pravca regresije za ove podatke

b) 95% pouzdane intervale za α i β

c) testirajte hipotezu da je koeficijent smjera tog pravca jednak 0, tj. da između x i Y ne postoji linearna veza, uz razinu značajnosti 0.01.

Rješenje:

a) izračunajmo procjenu parametara α i β : $\hat{\alpha} = \dfrac{S_{xy}}{S_x^2}$, $\hat{\beta} = \bar{Y} - \hat{\alpha}\,\bar{x}$

$$\bar{x} = \frac{35+45+55+65+75}{5} = 55$$

$$\bar{y} = \frac{114+124+143+158+166}{5} = 141$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right) = \frac{1}{4}\left(16125 - 5\cdot 55^2\right) = 250$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\bar{y}\right) = \frac{1}{4}\left(40155 - 5\cdot 55\cdot 141\right) = 345$$

$$\Rightarrow \quad \hat{\alpha} = \frac{345}{250} = 1.38$$

$$\Rightarrow \quad \hat{\beta} = \bar{y} - \hat{\alpha}\,\bar{x} = 141 - 1.38\cdot 55 = 65.1$$

$\Longrightarrow y = 1.38x + 65.1$ je procjena pravca regresije za ove podatke

b) Zanimaju nas pouzdani intervali za α i β. Najprije moramo izračunati $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}, \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Znamo da je $\quad \hat{Y}_i = \hat{\alpha}\ x_i + \hat{\beta} \quad$ pa onda:

$$\hat{y_1} = \hat{\alpha}\ x_1 + \hat{\beta} = 1.38 \cdot 35 + 65.1 = 113.4$$
$$\hat{y_2} = \hat{\alpha}\ x_2 + \hat{\beta} = 1.38 \cdot 45 + 65.1 = 127.2$$
$$\hat{y_3} = \hat{\alpha}\ x_3 + \hat{\beta} = 1.38 \cdot 55 + 65.1 = 141$$
$$\hat{y_4} = \hat{\alpha}\ x_4 + \hat{\beta} = 1.38 \cdot 65 + 65.1 = 154.8$$
$$\hat{y_5} = \hat{\alpha}\ x_5 + \hat{\beta} = 1.38 \cdot 75 + 65.1 = 168.6$$

Formirajmo tablicu:

i	1	2	3	4	5	$\sum$
x_i	35	45	55	65	75	
y_i	114	124	143	158	166	
$\hat{y}_i$	113.4	127.2	141	154.8	168.6	
$(y_i - \hat{y}_i)^2$	0.36	10.24	4	10.24	6.76	31.6

Dobili smo: $\quad SSE = 31.6 \quad$ pa je onda

$$\hat{\sigma}^2 = \frac{31.6}{3} = 10.5\dot{3} \quad \Rightarrow \quad \sigma = 3.246$$

Pogledajmo sada kako izgleda 95% pouzdan interval za α, odnosno β:

$$\hat{\alpha} \pm t_{\frac{\alpha}{2}}(n-2) \cdot \frac{\hat{\sigma}}{\sqrt{(n-1)s_x^2}} = 1.38 \pm t_{0.025}(3) \cdot \frac{3.246}{\sqrt{4 \cdot 250}}$$
$$= 1.38 \pm 3.182 \cdot 0.103 = 1.38 \pm 0.33$$
$$\Longrightarrow \quad 1.05 \ \leq \ \alpha \ \leq \ 1.71$$

$$\hat{\beta} \ \pm \ t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = 65.1 \ \pm \ t_{0.025}(3) \cdot 3.246 \sqrt{\frac{1}{5} + \frac{55^2}{1000}}$$
$$= 65.1 \pm 18.55$$
$$\Longrightarrow \quad 46.55 \ \leq \ \beta \ \leq \ 83.65$$

c) Želimo, uz razinu značajnosti 0.01, testirati hipotezu da ne postoji linearna veza između x i Y. Linearna veza ne postoji jedino ako je koeficijent smjera pravca regresije jednak 0. Ako je on različit od 0, bez obzira da li je pozitivan (tj. > 0) ili negativan (tj. < 0), linearna veza postoji. Postavljamo stoga hipoteze:

$$H_0 \; : \; \alpha = 0$$
$$H_1 \; : \; \alpha \neq 0$$

Sljedeći korak je izračunati vrijednost odgovarajuće test-statistike:

$$T_\alpha = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{(n-1)S_x^2} \quad \sim \quad t(n-2)$$

Imamo:

$$t_\alpha = \frac{1.38 - 0}{3.246} \sqrt{1000} = 13.44$$

Iz tablice za t-razdiobu očitavamo

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.005}(3) = 5.841$$

Kako je

$$t_\alpha > t_{0.005}(3)$$

vrijednost test-statistike je upala u kritično područje, pa nultu hipotezu H_0 : $\alpha = 0$ moramo odbaciti. Zaključujemo stoga da koeficijent smjera pravca regresije nije jednak 0, pa onda linearna veza postoji. □

Linearni model najčešće se koristi u dvije svrhe:

1. za predviđanje (procjenu) vrijednosti *srednje tj. očekivane vrijednosti od* Y za neku danu vrijednost x_0 od x, tj. $\mathrm{E}\,[Y|x = x_0]$. U ovom slučaju, nastoji se procijeniti *srednja vrijednost* mjerenja *velikog broja pokusa* pri zadanoj vrijednosti od x.

   - procjenitelj od $\mathrm{E}\,[Y|x = x_0]$ je
   $$\mathrm{E}\,[\widehat{Y|x =} x_0] = \hat{\alpha}\; x_0 + \hat{\beta}$$
   - $(1-\alpha)$ 100% pouzdan interval za $\mathrm{E}\,[Y|x = x_0]$:
   $$\Bigg[\; \mathrm{E}\,[\widehat{Y|x =} x_0] \;-\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \;, \qquad (15)$$
   $$\mathrm{E}\,[\widehat{Y|x =} x_0] \;+\; t_{\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2}} \;\Bigg]$$

2. za predviđanje (procjenu) *vrijednosti* Y za neku danu vrijednost x_0 od x. U ovom slučaju, nastoji se procijeniti *rezultat jednog pokusa* provedenog pri zadanoj vrijednosti od x, dakle rezultat nekog budućeg mjerenja.

   - procjenitelj od Y za $x = x_0$ je

   $$\hat{Y} = \hat{\alpha}\, x_0 + \hat{\beta}$$

   - $(1-\alpha)\,100\%$ pouzdan interval za Y u $x = x_0$:

   $$\left[\hat{Y} \;-\; t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\,\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}\;, \right. \qquad (16)$$
   $$\left. \hat{Y} \;+\; t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\,\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}} \right]$$

Uočimo da je pouzdani interval (16) za Y širi, odnosno manje precizan od pouzdanog intervala (15) za $\mathrm{E}\,[Y|x = x_0]$, što je bilo prirodno za očekivati.

Zadatak 37 *Nađite* 95% *pouzdan interval za* Y *u* $x = 55$, *te* 95% *pouzdan interval za* $\mathrm{E}\,[Y|x = 55]$ *za podatke iz Zadatka 36.*

Rješenje: Pouzdane intervale za Y u $x = 55$ i $\mathrm{E}\,[Y|x = 55]$ dobit ćemo uvrštavanjem odgovarajućih vrijednosti u (16) i (15), redom. Većina parametara već je izračunata, treba nam još samo:

$$\hat{Y} = \mathrm{E}\,[\widehat{Y|x=}\,55] = \hat{\alpha}\cdot 55 + \hat{\beta} = 1.38\cdot 55 + 65.1 = 141$$

Sada:

$$\mathrm{E}\,[\widehat{Y|x=}\,x_0] \;\pm\; t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\,\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}$$
$$= \mathrm{E}\,[\widehat{Y|x=}\,55] \;\pm\; t_{0.025}(3)\cdot 3.246\,\sqrt{\frac{1}{5}+\frac{(55-55)^2}{4\cdot 250}} = 141 \pm 4.62$$

pa slijedi da je 95% pouzdan interval za $\mathrm{E}\,[Y|x = 55]$:

$$136.38 \;\leq\; \mathrm{E}\,[Y|x = 55] \;\leq\; 145.62$$

Slično dobivamo:

$$\begin{aligned}\hat{Y} \ \pm \ t_{\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)S_x^2}}\\ = 141 \ \pm \ t_{0.025}(3)\cdot 3.246\sqrt{1+\frac{1}{5}+\frac{(55-55)^2}{4\cdot 250}} = 141 \pm 11.3146\end{aligned}$$

pa je 95% pouzdan interval za procjenu (predviđanje) vrijednosti Y u $x=55$:

$$129.685 \ \leq \ Y \ \leq \ 152.315$$

□

Pokazatelji da li je predloženi linearni model dobar (prihvatljiv) model za dane podatke:

- **koeficijent determinacije**

$$R^2 := \frac{(n-1)S_y^2 - SSE}{(n-1)S_y^2} = 1 - \frac{SSE}{(n-1)S_y^2} \ \in \ [0,1]$$

- što je R^2 bliže vrijednosti 1, to je prilagodba linearnog modela podacima bolja

- koeficijent determinacije jednak je kvadratu koeficijenta korelacije

- **test značajnosti linearnog modela**

- svodi se na testiranje

$$\begin{aligned} H_0 &: \ \alpha = 0\\ H_1 &: \ \alpha \neq 0 \end{aligned}$$

Zadatak 38 *Izračunajte koeficijent determinacije za podatke iz Zadatka 36.*

Rješenje:

Znamo da je: $SSE = 31.6$

Treba nam još:

$$(n-1)\cdot s_y^2 = \sum_{i=1}^{n} y_i^2 - n\cdot\bar{y}^2 = 101341 - 5\cdot 141^2 = 1936$$

$$\Rightarrow\ R^2 = 1 - \frac{SSE}{(n-1)S_y^2} = 1 - \frac{31.6}{1936} = 0.984$$

Linearni model je dakle za ove podatke jako dobar.

$\square$