# Statistička obrada podataka

Ana Anušić Ervin Duraković Hrvoje Maltarić Ivan Pažin

**Sažetak**

U ovom članku provodimo statističko istraživanje koje se bazira na zavisnosti uspjeha na prijamnom ispitu i prve godine studiranja. U tu svrhu, definiramo i objašnjavamo osnovne statističke pojmove i općeniti tijek statističkog istraživanja. Grafičkim prikazima podataka objašnjavamo što je i čemu služi opisna statistika te zašto ona nije dovoljna za formiranje formalnih zaključaka. Naravno, analiziramo i matematički alat s pomoću kojeg se zaključci mogu smatrati valjanima. Posebnim komentarima ističemo kako interpretirati dobivene rezultate i na što, prilikom toga, treba obratiti posebnu pozornost.

## Sadržaj

# 1 Uvod

## 1.1 Čime se bavi statistika

Recimo da nam treba prosječna visina svih ljudi na Zemlji. Sasvim je jasno da bi pojedinačnim prikupljanjem podataka taj posao zaista dugo trajao. Zato nećemo mjeriti visinu svih ljudi, već ćemo odabrati neki broj ljudi i s pomoću njihovih visina *procijeniti* visinu svih. Upravo na taj način počinje svaka statistička analiza – ispitivanjem uzorka procjenjujemo svojstvo cijele populacije.

## 1.2 Što je uzorak i kako ga odabrati?

Naravno, nema smisla mjeriti visinu svih osnovnoškolaca jedne škole na svijetu ili košarkaške reprezentacije i na temelju toga procijenjivati visinu svih ljudi na Zemlji! Uzorak mora biti slučajno odabran. Ljudi ne smiju biti birani npr. prema spolu, boji kose, političkoj opredijeljenosti itd.

## 1.3 Istraživanje

Početak svakog istraživanja je formiranje hipoteza, pretpostavki koje želimo dokazati (ili opovrgnuti). Obradu podataka započinjemo njihovim vizualnim prikazom. Najčešće histogramima i box-plotovima. Na taj način odmah možemo uočiti u kojim granicama se podaci nalaze, kako su raspoređeni te u kojem smjeru uostalom naši zaključci idu. Međutim, vizualna reprezentacija podataka nije dovoljna da bismo neku hipotezu smatrali dokazanom ili opovrgnutom. Tek primjenom različitih statističkih testova možemo s određenom, unaprijed pretpostavljenom vjerojatnošću smatrati da je istraživanje završeno.

# 2 Analiza prikupljenih podataka

## 2.1 Hipoteze

- prosjek ocjena ne ovisi o spolu
- prosjek ocjena ne ovisi o godini upisa na fakultet
- prolaznost na 1. godini studija ovisi o mjestu na rang listi prijamnog ispita
- rang i prosjek linearno ovise i moguće je na temelju ranga procijeniti prosjek

## 2.2 O prikupljanju podataka

Podatke smo prikupljali od studenata s PMF– Matematičkog odjela, prediplomski studij matematike, koji su fakultet upisali 2007. i 2008. godine, a polagali su prijamni ispit. To smo ostvarili s pomoću anonimne ankete u kojoj smo tražili da napišu koje godine su upisali fakultet, kojeg su spola, njihov rang (mjesto) na prijamnom ispitu (uz uvjet da nisu imali izravan upis) te njihove ocjene iz svih kolegija na prvoj godini. Zanima nas njihova prva godina studiranja, pa smo zamolili da napišu i ako su neki predmet pali, s čime ćemo poslije baratati kao s ocjenom 1. Na taj način zaista uočavamo kakav je uspjeh student ostvario godinu dana nakon što se upisao na fakultet i registriramo razliku između studenata koji su neki kolegij položili u roku i onih koji su pali, ali možda položili sljedeće godine s boljom ocjenom.

Ukupna populacija studenata upisanih 2007. i 2008. godine je 500, a mi smo prikupili uzorak od 94, procijenivši da će to biti dovoljno za statističku analizu. Nakon prikupljenih podataka izračunali smo prosjek ocjena svakog studenta (računajući i ocjene 1), te posebno označili je li student prvu godinu prošao ili pao. Zbog anonimnosti, nismo tražili studente da napišu točno mjesto na prijamnom ispitu, nego u razredima od 10. Dakle, ako je netko ostvario npr. 103. mjesto, zapisao je da mu je rang 101– 110. Na taj način anonimnost je sačuvana, a razredi su dovoljno mali da bismo mogli dovoljno dobro provjeriti svoje hipoteze.

## 2.3 O prosjeku ocjena slučajnog uzorka

Promatrano statističko obilježje (spol, prosjek, ...) u idućim analizama označavat ćemo s $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots$. Kroz $n$ mjerenja dobivamo niz (tj. uzorak) $x_1, x_2, \ldots, x_n$ kojim ćemo procijeniti statističko obilježje. U najjednostavnijem slučaju, ako obilježje $\mathbf{X}$ poprima samo vrijednost iz nekog diskretnog (konačnog ili prebrojivog) skupa $A$, onda se kaže da je $\mathbf{X}$ diskretno obilježje. U uzorku možemo uočiti ponavljanje nekih veličina. Neka u uzorku $x_1 \ldots, x_n$ ima $k$ ($k \leq n$) različitih izmjerenih veličina $x'_1, \ldots, x'_k$. S $f_i$ označavamo broj ponavljanja veličine $x'_i$ u uzorku, $i \in 1, \ldots, k$. Taj broj $f_i$ zovemo frekvencija veličine $x'_i$. Još jedna korisna veličina usko vezana uz frekvenciju je relativna frekvencija veličine $x'_i$, koja se jednostavno definira kao $p_i := \frac{f_i}{n}, i = 1, \ldots, k$.

Dobivene podatke jednostavnije prikazujemo tablično.

Tablica 1: Tablica frekvencija

| $i$ | $x'_i$ | $f_i$ | $p_i$ |
|---|---|---|---|
| 1 | $x'_1$ | $f_1$ | $p_1$ |
| 2 | $x'_2$ | $f_2$ | $p_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $k$ | $x'_k$ | $f_k$ | $p_k$ |

U slučaju da obilježje $\mathbf{X}$ ne poprima diskretne vrijednosti, već iz nekog intervala iz $\mathbb{R}$ ne možemo prebrojiti ponavljanja, pa vrijednosti svrstavamo u razrede. Razredi su disjunktni, jednake širine i prekrivaju cijeli interval (biramo ih proizvoljno ovisno o praktičnim potrebama).

Prilikom grupiranja u razrede sve vrijednosti $i$-tog razreda aproksimiraju se sredinom tog razreda, čime se gubi određeni dio informacija, ali se mogu razlučiti bitna svojstva promatranog kontinuiranog obilježja $\mathbf{X}$.

U ovom slučaju broj veličina unutar nekog razreda predstavlja frekvenciju razreda.

- Tablica prosjeka ocjena prikupljenog slučajnog uzorka:

| $i$ | prosjek | $f_i$ | $p_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | $[1.0, 2.0\rangle$ | 30 | 0.31914887 | 0.31914887 |
| 1 | $[2.0, 3.0\rangle$ | 30 | 0.31914887 | 0.63829774 |
| 2 | $[3.0, 4.0\rangle$ | 27 | 0.28723404 | 0.92553178 |
| 3 | $[4.0, 5.0]$ | 7 | 0.07446822 | 1 |

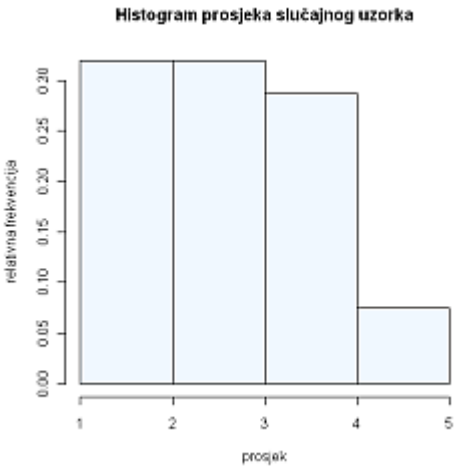
*Podsjetimo, prosjek smo računali tako da smo za pad kolegija uzimali ocjenu nedovoljan (1). Zato interval* $[1.0, 2.0\rangle$ *ima smisla.*

*Napomena* 2.3.1. Pri tome su $F_i$ kumulativne relativne frekvencije definirane rekurzivno, tj:

$$F_0 = p_0$$

$$F_i = F_{i-1} + p_i, i = 1, 2, \ldots, n.$$

- Histogram prosjeka ocjena slučajnog uzorka:




*Napomena* 2.3.2. Na osi apscisa nalaze se razredi, dok nam os ordinata predstavlja relativne frekvencije. Primijetimo, ukupna površina dobivenog histograma jednaka je 1. Histogramom relativno jednostavno možemo uočiti distribuciju statističkog obilježja $\mathbf{X}$.

Spomenimo još neke korisne veličine kojima se koristimo u opisnoj statistici. Prije svega poredajmo slučajni uzorak po veličini, tj.:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots, \leq x_{(n)}$$

- *Raspon* slučajnog uzorka:
$$d = x_{(n)} - x_{(1)}$$

- *Medijan* slučajnog uzorka je vrijednost koja ima svojstvo da je 50% podataka veće, a 50% manje od nje, tj. uzimamo za medijan:
$$m = \begin{cases} \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{ako je } n \text{ paran} \\ x_{(\frac{n+1}{2})}, & \text{ako je } n \text{ neparan} \end{cases}$$

- *Donji kvartil* je vrijednost koja ima svojstvo da je 25% podataka manje od nje, tj. uzimamo:
$$q_l = x_{(\frac{n+1}{4})}$$

- *Gornji kvartil* je vrijednost koja ima svojstvo da je 25% podataka veće od nje, tj. uzimamo:
$$q_u = x_{(\frac{3(n+1)}{4})}$$

- *Interkvartil*: $\mathbf{IQR} = q_u - q_r$

*Karakteristična petorka uzorka* : $(x_{(1)}, q_l, m, q_u, x_{(n)})$.
S pomoću karakteristične petorke uzorka formiramo *box–plot* (eng. box and whisker plot).

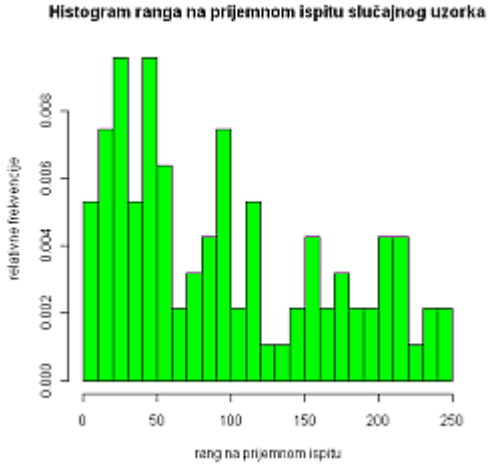
*Napomena* 2.3.3. Outlieri su sve vrijednosti koje su od gornjeg i donjeg kvantila udaljene za više od $\frac{3}{2}\mathbf{IQR}$. Brkovi su najveća i najmanja vrijednost koje nisu outlieri.

*Box-plot* prosjeka ocjena slučajnog uzorka:


*Napomena* 2.3.4. Budući da je u slučajnom uzorku prosjek ocjena u razredima, ne možemo točno odrediti gornji, donji kvartil i medijan već ih procjenjujemo linearnom interpolacijom iz grafa kumulativnih frekvencija.

"Box" predstavlja podatke koji se po vrijednosti nalaze u rasponu $25\% - 75\%$ ukupne veličine (tj. donja linija pravokutnika određena je donjim kvartilom, a gornja gornjim). Medijan je u pravokutniku posebno naznačen debljom linijom. Najmanja i najveća vrijednost koje nisu outlieri na grafu su označeni linijom i s pravokutnikom spojeni izlomljenom linijom (zato se i zovu brkovi). Primijetimo da na ovom box-plotu nema outliera (općenito, ako ih ima, posebno se naznače npr. kružićem).

## 2.4 Rang uzorka na prijamnom ispitu




*Napomena* 2.4.1. Primijetimo da podaci nisu jednako raspoređeni, što objašnjavamo činjenicom da dio lošije rangiranih studenata sigurno nije više prisutan na fakultetu. No to nam

Tablica 2: Rang na prijamnom ispitu slučajnog uzorka

| $i$ | razred | $f_i$ | $p_i$ | sredina | $F_i$ |
|---|---|---|---|---|---|
| 1 | $1-10$ | 5 | 0.0531914893617021 | 5.5 | 0.053191489317021 |
| 2 | $11-20$ | 7 | 0.074468085106383 | 15.5 | 0.127659574468085 |
| 3 | $21-30$ | 9 | 0.0957446808510638 | 25.5 | 0.223404255319149 |
| 4 | $31-40$ | 5 | 0.0531914893617021 | 35.5 | 0.27659595744680851 |
| 5 | $41-50$ | 9 | 0.0957446808510638 | 45.5 | 0.372340425531915 |
| 6 | $51-60$ | 6 | 0.0638297872340425 | 55.5 | 0.436170212765957 |
| 7 | $61-70$ | 2 | 0.0212765957446809 | 65.5 | 0.457446808510638 |
| 8 | $71-80$ | 3 | 0.0319148936170213 | 75.5 | 0.489361702127660 |
| 9 | $81-90$ | 4 | 0.0425531914893617 | 85.5 | 0.531914893617021 |
| 10 | $91-100$ | 7 | 0.074468085106383 | 95.5 | 0.606382978723404 |
| 11 | $101-110$ | 2 | 0.0212765957446809 | 105.5 | 0.627659574468085 |
| 12 | $111-120$ | 5 | 0.0531914893617021 | 115.5 | 0.680851063829787 |
| 13 | $121-130$ | 1 | 0.0106382978723404 | 125.5 | 0.691489361702128 |
| 14 | $131-140$ | 1 | 0.0106382978723404 | 135.5 | 0.702127659574468 |
| 15 | $141-150$ | 2 | 0.0212765957446809 | 145.5 | 0.723404255319149 |
| 16 | $151-160$ | 4 | 0.0425531914893617 | 155.5 | 0.76595744680851 |
| 17 | $161-170$ | 2 | 0.0212765957446809 | 165.5 | 0.787234042553192 |
| 18 | $171-180$ | 3 | 0.0319148936170213 | 175.5 | 0.819148936170213 |
| 19 | $181-190$ | 2 | 0.0212765957446809 | 185.5 | 0.80425531914294 |
| 20 | $191-200$ | 2 | 0.0212765957446809 | 195.5 | 0.861702127659575 |
| 21 | $201-210$ | 4 | 0.0425531914893617 | 205.5 | 0.904255319148936 |
| 22 | $211-220$ | 4 | 0.0425531914893617 | 115.5 | 0.946808510638298 |
| 23 | $221-230$ | 1 | 0.0106382978723404 | 225.5 | 0.957446808510638 |
| 24 | $231-240$ | 2 | 0.0212765957446809 | 235.5 | 0.978723404255319 |
| 25 | $241-250$ | 2 | 0.0212765957446809 | 245.5 | 1 |

ne smeta pri obradi, budući da ćemo sve potrebno dobiti linearnom interpolacijom.

Tablica 3: Tablica kvartila ranga slučajnog uzorka

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 55 | 35.5 | 83 | 156.75 | 245.5 |


# 3 Testiranje nezavisnosti statističkih obilježja

## 3.1 Testiranje hipoteza

Nakon formiranja hipoteze, moramo osmisliti način na koji ćemo je provjeriti, odnosno postupak donošenja odluke o njenom prihvaćanju ili odbacivanju. Taj postupak zove se *testiranje*. Općenito se problem testiranja sastoji od definiranja područja $C \in \mathbb{R}^n$, koje zovemo *kritično područje* hipoteze $H$. Ako se izmjereni uzorak shvati kao $(x_1, x_2, x_3, \ldots, x_n) \in \mathbb{R}^n$ može vrijediti $(x_1, x_2, x_3, \ldots, x_n) \in C$ ili $(x_1, x_2, x_3, \ldots, x_n) \notin C$. Ako vrijedi prvo, hipoteza se odbacuje, a u suprotnom se prihvaća. Ovako definirani postupak zove se *statistički test*.

U statističkom testu ključnu ulogu ima kritično područje. Njega treba odrediti tako da sadržava one točke $(x_1, x_2, x_3, \ldots, x_n) \in C$ u kojima dolazi do značajnog odstupanja od hipoteze koju testiramo. Naravno, javlja se problem određivanja što je značajno, a što tolerantno odstupanje.

Neka je $H_0$ hipoteza koju testiramo. Vidimo da je zapravo riječ o dvije hipoteze, $H_0$ i $H_1$ , tj. odbacivanjem hipoteze $H_0$ prihvaćamo hipotezu $H_1$, a prihvaćanjem $H_0$ odbacujemo $H_1$. $H_0$ zove se *nul–hipoteza* , a $H_1$ *alternativna hipoteza*.

Budući da na temelju uzorka procjenjujemo svojstvo cijele populacije, odbacivanjem hipoteze $H_0$ uvijek postoji određeni rizik da je hipoteza odbačena kada je zapravo trebala biti prihvaćena. Taj rizik označava se brojem $\alpha$ $(0 \leq \alpha \leq 1)$ i kaže se da test ima razinu značajnosti $\alpha$. Naravno, želimo da $\alpha$ bude što manji, najčešće se uzima $0,05$ $(5\%)$. Problem nalaženja najboljeg testa svodi se na određivanje kritičnog područja $C$ tako da razina značajnosti iznosi zadani broj $\alpha$, te da vjerojatnost prihvaćanja nul–hipoteze kada je stvarno neistinita (pogreška druge vrste) bude minimalna.

## 3.2 Dvodimenzionalna statistička obilježja

Istodobno promatramo više veličina i želimo ustanoviti njihovu ovisnost. Primjerice, promatramo dva statistička obilježja, $\mathbf{X}$ i $\mathbf{Y}$. Višestrukim ponavljanjem mjerenja dobiva se niz uređenih parova:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n). \tag{1}$$

U tom slučaju kažemo da promatramo *dvodimenzionalno statističko obilježje* $(\mathbf{X}, \mathbf{Y})$.

Neka obilježje $\mathbf{X}$ poprima vrijednosti iz nekog diskretnog skupa $\mathbf{A} = \{a_1, a_2, \ldots, a_r\}$, a obilježje $\mathbf{Y}$ iz skupa $\mathbf{B} = \{b_1, b_2, \ldots, b_c\}$. Analogno jednodimenzionalnom slučaju, za svaki par $(a_i, b_j), i = 1, \ldots, r,\ j = 1, \ldots, c$, možemo uočiti njegovu frekvenciju u (1), označimo je s $f_{ij}$. Frekvenije nam omogućuju da formiramo sljedeću tablicu, koja se zove *kontingencijska tablica frekvencija*:

Tablica 4: Kontingencijska frekvencijska tablica

| $\mathbf{X} \diagdown \mathbf{Y}$ | $b_1$ | $b_2$ | $\ldots$ | $b_c$ | $\Sigma$ |
|---|---|---|---|---|---|
| $a_1$ | $f_{11}$ | $f_{12}$ | $\ldots$ | $f_{1c}$ | $f_1$ |
| $a_2$ | $f_{21}$ | $f_{22}$ | $\ldots$ | $f_{2c}$ | $f_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $a_r$ | $f_{r1}$ | $f_{r2}$ | $\ldots$ | $f_{rc}$ | $f_r$ |
| $\Sigma$ | $g_1$ | $g_2$ | $\ldots$ | $g_c$ | $n$ |

*Napomena* 3.2.1. Brojevi $f_i, i = 1, \ldots, r$, nazivaju se *marginalne frekvencije* od $a_i$ u (1), a brojevi $g_j, j = 1, \ldots, c$, marginalne frekvencije od $b_j$ u (1).

Uz ovisnost dvodimenzionalnih statističkih obilježja usko je vezan tzv. *Pearsonov koeficijent korelacije*, koji pokazuje stupanj afine povezanosti među podacima u uzorku, a definira se kao:

$$r_{\mathbf{XY}} := \frac{S_{\mathbf{XY}}}{\sqrt{S_{\mathbf{XX}} S_{\mathbf{YY}}}}, \tag{2}$$

gdje su

$$S_{\mathbf{XX}} := \sum_{k=1}^{n} x_k^2 - n\bar{x}^2, S_{\mathbf{YY}} := \sum_{k=1}^{n} y_k^2 - n\bar{y}^2, S_{\mathbf{XY}} := \sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}, \bar{x} := \frac{1}{n}\sum_{k=1}^{n} x_k, \bar{y} := \frac{1}{n}\sum_{k=1}^{n} y_k. \tag{3}$$

*Napomena* 3.2.2. Za Pearsonov koeficijent korelacije vrijedi:

$$-1 \leq r_{\mathbf{XY}} \leq 1.$$

Ako vrijedi:

- $r_{\mathbf{XY}} < 0.5$, kažemo da su obilježja $\mathbf{X}$ i $\mathbf{Y}$ slabo korelirana,
- $r_{\mathbf{XY}} \geq 0.5$, kažemo da su obilježja $\mathbf{X}$ i $\mathbf{Y}$ značajno korelirana,
- $r_{\mathbf{XY}} = 1$ ili $r_{\mathbf{XY}} = -1$, kažemo da je veza potpuno linearna,
- $r_{\mathbf{XY}} = 0$, veza nije linearna (to ne mora značiti da ne postoji!).

*Napomena* 3.2.3. Prethodne definicije i svojstva potpuno vrijede i ako su $\mathbf{X}$ i $\mathbf{Y}$ neprekidna statistička obilježja. Potrebno je samo napraviti podjelu podataka u razrede.

## 3.3 Pearsonov $\chi^2$-test o nezavisnosti

Neka je $(\mathbf{X}, \mathbf{Y})$ dvodimenzionalno statističko obilježje, te neka je (1) prikupljeni slučajni uzorak. Prirodno se postavlja pitanje što možemo reći o (ne)zavisnosti obilježja $\mathbf{X}$ i $\mathbf{Y}$ na temelju prikupljenog uzorka. Dakle, moramo konstruirati najbolji statistički test za testiranje hipoteze $H_0$ : $\mathbf{X}$ i $\mathbf{Y}$ su nezavisna obilježja, u odnosu na alternativu $H_1$ : $\mathbf{X}$ i $\mathbf{Y}$ su zavisna statistička obilježja.

Prvo moramo definirati veličinu koja će nam predstavljati "udaljenost" od nezavisnosti obilježja na temelju $n$-članog niza mjerenja, tako da ta "udaljenost" predstavlja značajno odstupanje od hipoteze $H_0$. U tu svrhu definiramo sljedeću veličinu (uz iste oznake kao u Tablici 4):

$$H_n := \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - n\hat{p_i}\hat{q_j})^2}{n\hat{p_i}\hat{q_j}}, \tag{4}$$

gdje su $\hat{p_i} = \frac{f_i}{n}$, $\hat{q_j} = \frac{g_j}{n}$. Može se pokazati da je upravo $H_n$ veličina kojom najbolje procjenjujemo ono što intuitivno shvaćamo kao "značajnu udaljenost od nezavisnosti".

Također, moramo odrediti kritično područje $C$, tako da razina značajnosti testa iznosi zadani broj $\alpha \in \langle 0, 1\rangle$. Od sada nadalje, $\alpha$=0.05=5%. Za $C$ mora vrijediti da ako $H_n \in C$, hipotezu $H_0$ odbacujemo u korist $H_1$ (s rizikom od 5%). Za rješenje ovog problema koristan je sljedeći teorem:

**Teorem 3.3.1** (Pearsonov teorem)**.** *Ako je $H_0$ točna hipoteza, za $n \to \infty$ vrijedi*

$$H_n \sim \chi^2((r-1)(c-1)), \tag{5}$$

*tj. za velike $n$, $H_n$ ima $\chi^2$-razdiobu[1] s $(r-1)(c-1)$ stupnjeva slobode.*

Sada možemo izračunati kritično područje $[x_0, +\infty\rangle$, tj. tražimo točku $x_0 \in \mathbb{R}$ za koju vrijedi: $\mathbb{P}(H_n < x_0) \geq 0.95$, tj. $\mathbb{P}(H_n \geq x_0) \leq 0.05$. Budući da znamo distribuciju $H_n$, vrijednost točke $x_0$ iščitavamo iz tablice, a budući da ovisi o $\alpha$, $r$ i $c$, označava se s $\chi^2_\alpha((r-1)(c-1))$.

Sada lako možemo testirati nezavisnost obilježja $\mathbf{X}$ i $\mathbf{Y}$. S pomoću vrijednosti u kontingencijskoj tablici izračunamo $H_n$ i ako vrijedi $H_n \geq \chi^2_\alpha((r-1)(c-1))$, odbacujemo $H_0$ u korist $H_1$ uz rizik od 5%. U suprotnom, prihvaćamo $H_0$.

## 3.4 Hipoteza: *Prosjek ocjena ne ovisi o spolu*

Unaprijed ne očekujemo razliku prosjeka na prvoj godini studiranja između žena i muškaraca. Za početak pogledajmo opisni prikaz odnosa prosjeka slučajnog uzorka obaju spolova.

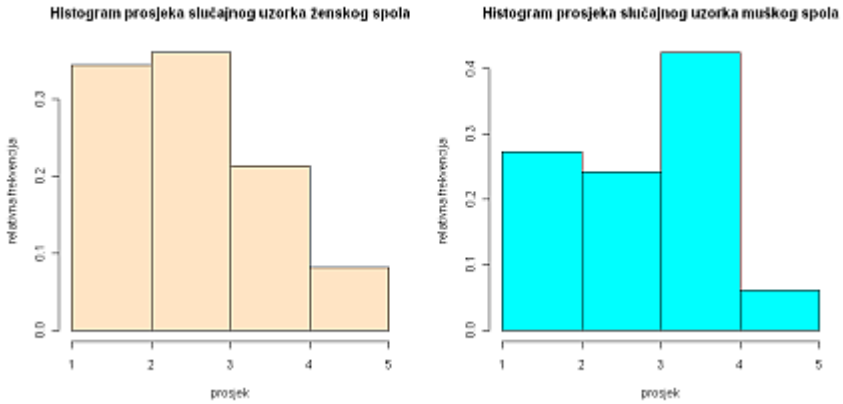
Tablica 5: Frekvencijska tablica prosjeka ocjena uzorka ženskog spola

| prosjek | frekvencija | relativna frekvencija |
|---|---|---|
| $[1.0, 2.0\rangle$ | 21 | 0.34426223 |
| $[2.0, 3.0\rangle$ | 22 | 0.36065574 |
| $[3.0, 4.0\rangle$ | 13 | 0.21311475 |
| $[4.0, 5.0]$ | 5 | 0.08196721 |

[1] U teoriji vjerojatnosti i statistici, hi-kvadrat razdioba ($\chi^2$-razdioba) jedna je od najčešće korištenih teorijskih razdioba u statističkim testovima, a ovisi o jednom parametru $k$ (stupnjevi slobode). Ako $X$ ima $\chi^2$-razdiobu s $k$ stupnjeva slobode koristimo oznaku $X \sim \chi^2(k)$. Kada je parametar $k$ dovoljno velik, $\chi^2$-razdiobu možemo aproksimirati normalnom razdiobom $N(k, 2k)$.

Tablica 6: Frekvencijska tablica prosjeka ocjena uzorka muškog spola

| prosjek | frekvencija | relativna frekvencija |
|---|---|---|
| $[1.0, 2.0\rangle$ | 9 | 0.27272722 |
| $[2.0, 3.0\rangle$ | 8 | 0.24242424 |
| $[3.0, 4.0\rangle$ | 14 | 0.42424242 |
| $[4.0, 5.0]$ | 2 | 0.06060606 |




Primijetimo, najveći broj žena iz uzorka ima prosjek između 2.00 i 3.00, dok najveći dio muškaraca iz uzorka ima prosjek između 3.00 i 4.00. Međutim, gotovo dva puta više žena ima prosjek 4.00 do 5.00. Prirodno se pitamo što od toga može utjecati na nezavisnost danih obilježja i na koji način.

Promotrimo box-plot slučajnog uzorka obaju spolova:

Tablica 7: Tablica kvartila prosjeka slučajnog uzorka

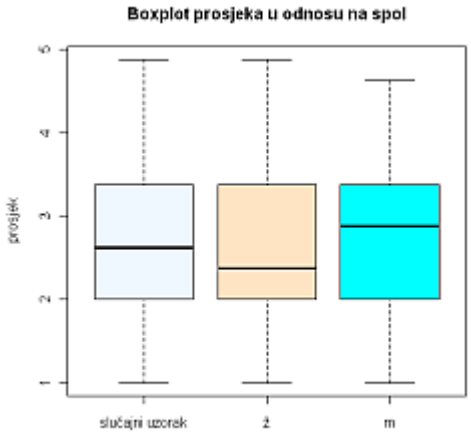
| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 1.000 | 2.000 | 2.625 | 3.375 | 4.875 |

Tablica 8: Tablica kvartila prosjeka slučajnog uzorka ženskog spola

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 1.000 | 2.000 | 2.375 | 3.375 | 4.875 |

Tablica 9: Tablica kvartila prosjeka slučajnog uzorka muškog spola

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 1.000 | 2.000 | 2.875 | 3.375 | 4.625 |




Za razliku od histograma, na box-plotu se ne mogu vidjeti značajne razlike između spolova.

Provedimo Pearsonov $\chi^2$-test o nezavisnosti:

Neka obilježje $\mathbf{X}$ poprima vrijednosti u razredima: $[1.0, 2.0\rangle$, $[2.0, 3.0\rangle$, $[3.0, 4.0\rangle$, $[4.0, 5.0]$, a obilježje $\mathbf{Y}$ neka poprima vrijednosti: muško, žensko. Testiramo:

$H_0$: $\mathbf{X}$ i $\mathbf{Y}$ su nezavisna obilježja,
$H_1$: $\mathbf{X}$ i $\mathbf{Y}$ su zavisna obilježja.

Tablica 10: Kontingencijska tablica

| $\mathbf{X} \setminus \mathbf{Y}$ | žensko | muško | $\Sigma$ |
|---|---|---|---|
| $[1.0, 2.0\rangle$ | 21 | 9 | 30 |
| $[2.0, 3.0\rangle$ | 22 | 8 | 30 |
| $[3.0, 4.0\rangle$ | 13 | 14 | 27 |
| $[4.0, 5.0]$ | 5 | 2 | 7 |
| $\Sigma$ | 61 | 33 | 94 |

Iz tablice i (4) računamo: $H_{94} = 14.003328$. Broj stupnjeva slobode je $(2-1)(4-1) = 3$, pa je kritično područje $[\chi^2_{0.05}(3), +\infty\rangle$. Još je samo preostalo odrediti $\chi^2_{0.05}(3)$, a to čitamo iz tablice. Dakle, kritično područje je $[7.8147, +\infty\rangle$, a budući da je $14.003328 > 7.8147$, odbacujemo hipotezu $H_0$ u korist alternativne hipoteze $H_1$.

Iako posve neočekivano, provedenim testom zaključujemo da prosjek ocjena na prvoj godini studija ovisi o spolu, na nivou značajnosti od 5%, tj. rizik da je naš zaključak pogrešan je 5%.

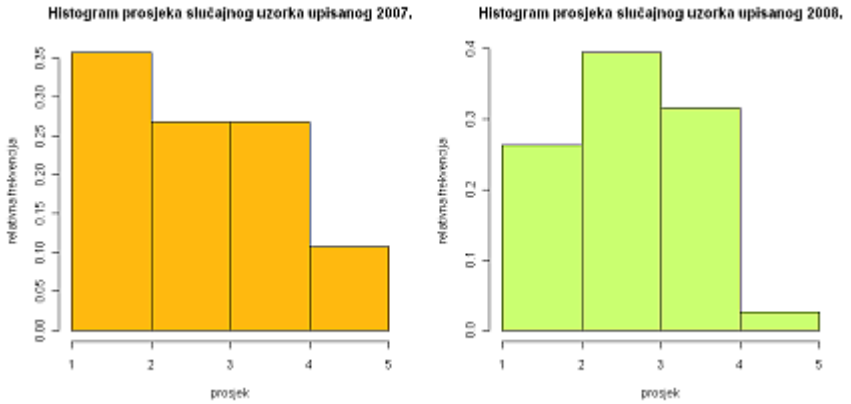
Tablica 11: Frekvencijska tablica prosjeka ocjena uzorka upisanog 2007. godine

| prosjek | frekvencija | relativna frekvencija |
|---|---|---|
| $[1.0, 2.0\rangle$ | 20 | 0.3571428 |
| $[2.0, 3.0\rangle$ | 15 | 0.2678571 |
| $[3.0, 4.0\rangle$ | 15 | 0.2678571 |
| $[4.0, 5.0]$ | 6 | 0.1071429 |

Tablica 12: Frekvencijska tablica prosjeka ocjena uzorka upisanog 2008. godine

| prosjek | frekvencija | relativna frekvencija |
|---|---|---|
| $[1.0, 2.0\rangle$ | 10 | 0.26315784 |
| $[2.0, 3.0\rangle$ | 15 | 0.39473684 |
| $[3.0, 4.0\rangle$ | 12 | 0.31578947 |
| $[4.0, 5.0]$ | 1 | 0.02631579 |

## 3.5 Hipoteza: *Prosjek ocjena ne ovisi o godini upisa na fakultet*




Primijetimo, najveći postotak uzorka upisanog 2007. godine ima prosjek od 1.00 do 2.00, za razliku od uzorka upisanog 2008. godine, gdje se taj prosjek kreće od 2.00 do 3.00. Međutim, relativna frekvencija prosjeka 4.00-5.00 čak je četiri puta veća u korist upisanih 2007. godine. Provedimo $\chi^2$-test o nezavisnosti.

Neka obilježje $\mathbf{X}$ poprima vrijednosti u razredima: $[1.0, 2.0\rangle$, $[2.0, 3.0\rangle$, $[3.0, 4.0\rangle$, $[4.0, 5.0]$, a obilježje $\mathbf{Y}$ neka poprima vrijednosti godine upisa: 2007., 2008. Testiramo:

$H_0$: $\mathbf{X}$ i $\mathbf{Y}$ su nezavisna obilježja,
$H_1$: $\mathbf{X}$ i $\mathbf{Y}$ su zavisna obilježja.

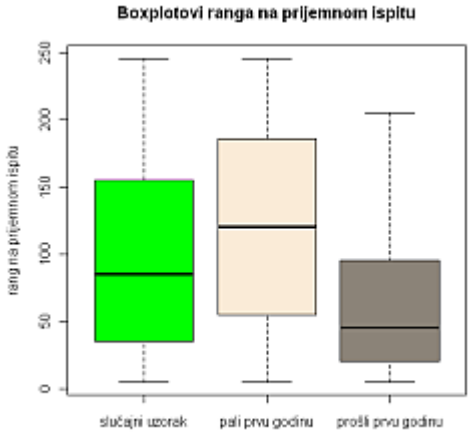
Iz kontingencijske tablice i (4) računamo: $H_{94} = 3.935598$. Broj stupnjeva slobode je kao i u prethodnom, $(4-1)(2-1) = 3$, pa je na isti način kritično područje $[7.8147, +\infty\rangle$. Budući da je $H_{94} = 3.935598 < 7.8147$, ne odbacujemo hipotezu $H_0$.

Dakle, $\chi^2$-test potvrdio je naša očekivanja na nivou značajnosti od 5%. Odnosno, zaključujemo da prosjek ocjena na prvoj godini fakulteta na ovisi o godini upisa.

Tablica 13: Kontingencijska tablica

| $\mathbf{X} \setminus \mathbf{Y}$ | 2007. | 2008. | $\Sigma$ |
|---|---|---|---|
| $[1.0, 2.0\rangle$ | 20 | 10 | 30 |
| $[2.0, 3.0\rangle$ | 15 | 15 | 30 |
| $[3.0, 4.0\rangle$ | 15 | 12 | 27 |
| $[4.0, 5.0]$ | 6 | 1 | 7 |
| $\Sigma$ | 56 | 38 | 94 |

## 3.6 Hipoteza: *Prolaznost na prvoj godini studiranja ovisi o mjestu na rang listi prijamnog ispita*




U prethodnim box-plotovima možemo uočiti niz zanimljivih činjenica. Promatrajući studente iz uzorka koji su prošli prvu godinu, uočavamo da nitko nije bio niže od 200. mjesta na rang listi, dok ih je čak 75% bilo rangirano iznad 100. mjesta, te 50% iznad 50. mjesta na prijamnom ispitu. Za razliku od njih, 75% studenata iz uzorka koji su pali prvu godinu bili su rangirani ispod 50. mjesta i čak 25% ispod 170. mjesta na prijamnom ispitu.

Ako studente koji su prošli prvu godinu označimo s 1, a one koji su pali s 0, iz (2) dobivamo da su obilježja prolaznost i rang na prijamnom ispitu negativno korelirana (Pearsonov koeficijent korelacije iznosi $-0.4356049$).

Dakle, sve upućuje na zavisnost prolaznosti i ranga na prijamnom ispitu. Provedimo $\chi^2$-test o nezavisnosti:

Neka obilježje $\mathbf{X}$ poprima vrijednosti ranga na prijamnom ispitu u razredima: $1-10$, $11-20$, $21-30, \ldots, 241-250$, a obilježje $\mathbf{Y}$ neka poprima vrijednosti: prolaz, pad. Testiramo:

$$H_0\text{: } \mathbf{X} \text{ i } \mathbf{Y} \text{ su nezavisna obilježja,}$$
$$H_1\text{: } \mathbf{X} \text{ i } \mathbf{Y} \text{ su zavisna obilježja.}$$

Broj stupnjeva slobode je $(25-1)(2-1) = 24$, dakle za kritično područje trebamo odrediti $\chi^2_{0.05}(24)$, a tu vrijednost čitamo iz tablice. Dakle, kritično područje je $[36.415, +\infty\rangle$. Iz kontingencijske tablice (Tablica 14) i (4) računamo $H_{94} = 36.724$, a budući da je $36.724 > 36.415$, odbacujemo hipotezu $H_0$ u korist alternative $H_1$ na nivou značajnosti od 5%.

Dakle, zaključujemo da su prolaznost na prvoj godini studiranja i mjesto na rang listi prijamnog ispita zavisna obilježja na nivou značajnosti od 5%, kao što smo i očekivali.

Tablica 14: Kontingencijska tablica

| $\mathbf{X} \setminus \mathbf{Y}$ | pad | prolaz | $\Sigma$ |
|---|---|---|---|
| $1-10$ | 1 | 4 | 5 |
| $11-20$ | 1 | 6 | 7 |
| $21-30$ | 4 | 5 | 9 |
| $31-40$ | 2 | 3 | 5 |
| $41-50$ | 6 | 3 | 9 |
| $51-60$ | 3 | 3 | 6 |
| $61-70$ | 2 | 0 | 2 |
| $71-80$ | 1 | 2 | 3 |
| $81-90$ | 3 | 1 | 4 |
| $91-100$ | 2 | 5 | 7 |
| $101-110$ | 0 | 2 | 2 |
| $111-120$ | 3 | 2 | 5 |
| $121-130$ | 1 | 0 | 1 |
| $131-140$ | 1 | 0 | 1 |
| $141-150$ | 2 | 0 | 2 |
| $151-160$ | 4 | 0 | 4 |
| $161-170$ | 2 | 0 | 2 |
| $171-180$ | 2 | 1 | 3 |
| $181-190$ | 2 | 0 | 2 |
| $191-200$ | 1 | 1 | 2 |
| $201-210$ | 4 | 0 | 4 |
| $211-220$ | 4 | 0 | 4 |
| $221-230$ | 1 | 0 | 1 |
| $231-240$ | 2 | 0 | 2 |
| $241-250$ | 2 | 0 | 2 |
| $\Sigma$ | 56 | 38 | 94 |

# 4 Regresijska analiza

Podsjetimo se, jedna od početnih hipoteza bila je da su prosjek ocjena na prvoj godini i rang na prijamnom ispitu u linearnoj vezi. Zadatak ovog poglavlja je tu hipotezu potvrditi ili opovrgnuti. Provest ćemo vrlo česti statistički model, koji se zove *linearni regresijski model*.

Na temelju podataka koje smo prikupili, ovaj model provodimo za obilježja koja ćemo označiti na sljedeći način: $\mathbf{X}$ predstavlja mjesto na rang listi prijamnog ispita, dok $\mathbf{Y}$ predstavlja prosjek ocjena na prvoj godini studiranja.

## 4.1 Metoda najmanjih kvadrata

Prikupljene podatke, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, prvo prikazujemo u koordinatnoj ravnini. Taj prikaz omogućuje nam da zapazimo moguću funkcijsku ovisnost između podataka.

*Metoda najmanjih kvadrata* unaprijed pretpostavlja *linearnu* funkcijsku ovisnost te pronalazi pravac $y = \hat{a}x + \hat{b}$ koji najbolje aproksimira vezu između prikupljenih podataka. Procjene $\hat{a}$ i $\hat{b}$ treba odrediti tako da vrijedi:

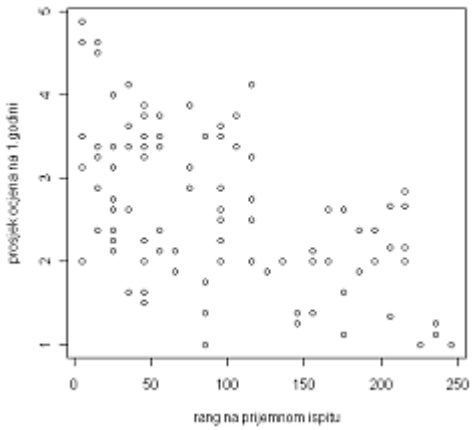
$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - \hat{a}x_i - \hat{b})^2.$$

Pokazuje se da ta jednadžba ima jedinstveno rješenje:

$$\hat{a} = \frac{S_{\mathbf{XY}}}{S_{\mathbf{XX}}}, \hat{b} = \bar{y} - \hat{a}\bar{x}, \tag{6}$$

gdje su $S_{\mathbf{XY}}$, $S_{\mathbf{XX}}$, $\bar{x}$ i $\bar{y}$ kao u (3).

Sada kada znamo koji pravac najbolje aproksimira prikupljene podatke, pogledajmo kako izgleda na prikupljenom uzorku. Za početak, pogledajmo kako originalni podaci izgledaju u koordinatnom sustavu:



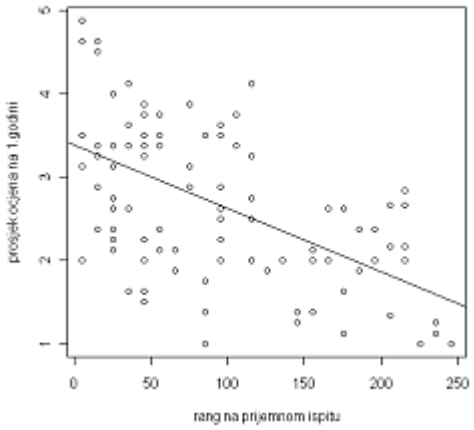

Sa slike možemo uočiti funkcijsku zavisnost ranga na prijamnom ispitu i prosjeka ocjena na prvoj godini. Metodom najmanjih kvadrata odredimo koji pravac najbolje opisuje primijećenu zavisnost. Potrebno je:

$$n = 94, \bar{x} = 99.80280264, \bar{y} = 2.62235, S_{\mathbf{XY}} = -3236.216722, S_{\mathbf{XX}} = 425137.155.$$

Dakle, dobivamo $\hat{a} = -0.0076121712$, $\hat{b} = 3.382066$, tj. traženi pravac je

$$y = -0.007621712x + 3.382066.$$

Prikažimo dobiveni pravac i grafički:



## 4.2 Konstrukcija pouzdanih intervala za parametre linearne regresije

Metodom najmanjih kvadrata odredili smo pravac koji najbolje aproksimira vezu prikupljenih podataka o prosjeku i uspjehu na prijamnom ispitu. Međutim, mi ne želimo odrediti vezu tih 94 parova podataka, već cijele populacije. Tražimo pravac $y = ax+b$ koji najbolje aproksimira vezu ranga na prijamnom ispitu i prosjeka ocjena na prvoj godini cijele populacije. Budući da s pomoću uzorka aproksimiramo populaciju, pravac dobiven metodom najmanjih kvadrata poslužit će nam kao dobra osnova za procjenu parametara $a$ i $b$. Konstruirat ćemo tzv. *pouzdane intervale*, tj. s vjerojatnošću od 95% procijeniti ćemo u kojem se intervalu nalaze vrijednosti paramatara $a$ i $b$.

Formalno, $(1-\alpha)\cdot 100\%$ pouzdani interval za $a$ je interval $[L, D]$ za koji vrijedi:

$$\mathbb{P}(L \leq a \leq D) \geq 1-\alpha, \alpha \in \langle 0, 1\rangle.$$

Kao i do sada, uzimamo $\alpha = 0.05$. Potpuno analogno definira se i 95% pouzdani interval za $b$.

Pri konstrukciji pouzdanih intervala za $a$ i $b$ od iznimne su važnosti sljedeći teoremi:

**Teorem 4.2.1.** *Za sve prirodne brojeve $n$ vrijedi:*

$$\frac{\hat{b}-b}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}}{S_{\mathbf{xx}}}}} \sim t(n-2), \tag{7}$$

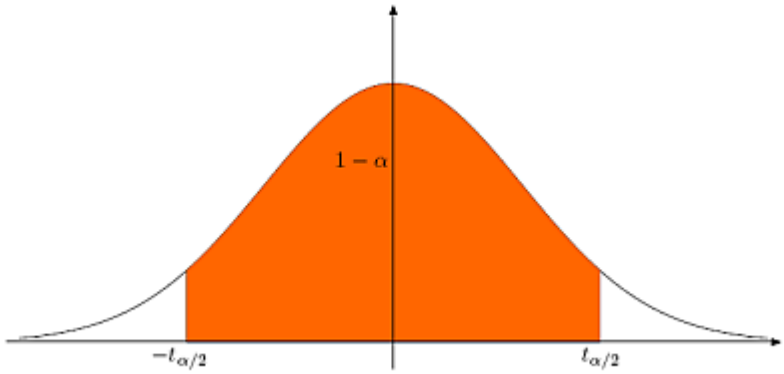
*tj. navedena testna statistika ima Studentovu t-razdiobu*[2] *s $n-2$ stupnja slobode.*

**Teorem 4.2.2.** *Za sve prirodne brojeve $n$ vrijedi:*

$$\frac{\hat{a}-a}{\hat{\sigma}\sqrt{\frac{1}{S_{\mathbf{xx}}}}} \sim t(n-2). \tag{8}$$

*Napomena* 4.2.1. U prethodnim teoremima $S_{\mathbf{XX}}$ i $\bar{x}$ su kao u (3), a $\hat{\sigma} := \sqrt{\frac{SSE}{n-2}}$, $SSE := S_{\mathbf{YY}} - \hat{a}^2 S_{\mathbf{XX}}$.

Sada lako pronađemo 95% pouzdane intervale za $a$ i $b$. Na sljedećoj slici prikazana je Studentova t-distribucija.




[2]Studentova $t$-razdioba (ili samo $t$-razdioba) je vjerojatnosna razdioba koja se primjenjuje kod procjene srednje vrijednosti normalno distribuirane populacije kada je uzorak mali. Isto tako primjenjuje se za testiranje razlike između dviju srednjih vrijednosti. Studentova razdioba ovisi o jednom parametru $k$ (stupnjevi slobode). Ako $X$ ima $t$-razdiobu s $k$ stupnjeva slobode, koristimo se oznakom $X \sim t(k)$. Za dovoljno veliku vrijednost parametra $k$, $t$-razdioba može se aproksimirati standardnom normalnom razdiobom $N(0,1)$.

Napomenimo da su vrijednosti $t_{\alpha/2}(n-2)$ tabelirane i u našem slučaju je $t_{0.05/2}(92) = 1.951$. Sada iz (7) uočavamo:

$$\mathbb{P}(-t_{0.05/2}(92) \leq \frac{\hat{b}-b}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}}{S_{\mathbf{xx}}}}} \leq t_{0.05/2}(92)) = 0.95,$$

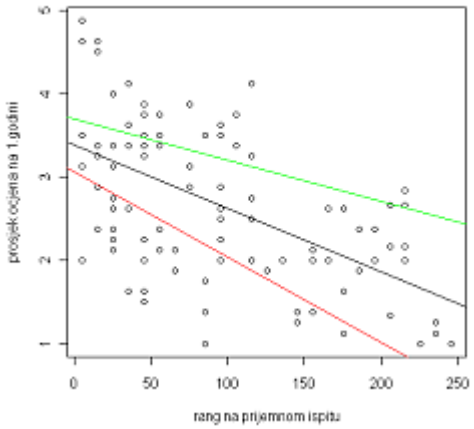
te iz (8)

$$\mathbb{P}(-t_{0.05/2}(92) \leq \frac{\hat{a}-a}{\hat{\sigma}\sqrt{\frac{1}{S_{\mathbf{xx}}}}} \leq t_{0.05/2}(92)) = 0.95.$$

Iz tih jednadžbi sad jednostavnim manipulacijama dobivamo 95% pouzdane intervale za $a$ i $b$.

*Napomena* 4.2.2. Iz prikupljenih podataka dobivamo $SSE = 71.61195282$ i $\hat{\sigma} = 0.882264581$.

Iz prethodnih zaključivanja i vrijednosti $SSE$ i $\hat{\sigma}$ dobivamo da je 95% pouzdani interval za $b$ $[3.06435966, 3.699772334]$ i za $a$ $[-0.010252096, -0.004972246]$.




## 4.3 Konstrukcija pouzdanih intervala za očekivani prosjek prve godine studiranja s obzirom na rang na prijamnom ispitu

Prisjetimose, ono što nas je također zanimalo bilo je može li student na temelju svog uspjeha na prijamnom ispitu unaprijed znati koliki je njegov očekivani prosjek na prvoj godini studiranja, pri čemu se pad računa kao ocjena 1. Formalno, želimo procijeniti $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x_0]$.

Budući da smo u prethodnim poglavljima već pretpostavili da je veza obilježja $\mathbf{X}$ i $\mathbf{Y}$ linearna te procijenili pravac $y = ax + b$ koji tu vezu najbolje aproksimira i pravac $y = \hat{a}x + \hat{b}$ koji najbolje aproksimira vezu podataka iz uzorka, na taj način postupit ćemo i ovdje. Dakle, $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x_0] = ax_0 + b$ procjenjujemo s $\hat{\mathbf{Y}} = \hat{a}x_0 + \hat{b}$.

**Teorem 4.3.1.** *Za sve prirodne brojeve $n$ vrijedi:*

$$\sqrt{\frac{n-2}{SSE}} \cdot \frac{\hat{a}x + \hat{b} - (ax+b)}{\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{\mathbf{xx}}}}} \sim t(n-2), \tag{9}$$

*gdje su $SSE$, $S_{\mathbf{xx}}$, $\bar{x}$ kao u napomeni 4.2.1.*

*Napomena* 4.3.1. Prethodni teorem kaže da za prirodan broj $n$ dana testna statistika ima Studentovu t-distribuciju s n-2 stupnja slobode. Primijetimo da smo time dobili efektivan način za računanje pouzdanih intervala za $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x_0]$.

Iz (9) slijedi:

$$\mathbb{P}(-t_{0.05/2}(92) \leq \sqrt{\frac{92}{SSE}} \cdot \frac{\hat{a}x_0 + \hat{b} - (ax_0 + b)}{\sqrt{\frac{1}{94} + \frac{(x_0 - \bar{x})^2}{S_{\mathbf{xx}}}}} \leq t_{0.05/2}(92)) = 0.95,$$

odnosno procjena 95% pouzdanog intervala za $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x_0] = ax_0 + b$ je:

$$[\hat{a}x_0 + \hat{b} - t_{0.025}(92)\sqrt{\frac{SSE}{92}}\sqrt{\frac{1}{94} + \frac{(x_0 - \bar{x})^2}{S_{\mathbf{xx}}}}, \hat{a}x_0 + \hat{b} + t_{0.025}(92)\sqrt{\frac{SSE}{92}}\sqrt{\frac{1}{94} + \frac{(x_0 - \bar{x})^2}{S_{\mathbf{xx}}}}].$$

Pogledajmo koliko to iznosi za konkretne $x_0$. Dobivene rezultate prikazujemo tablicom (Tablica 15).

Tablica 15: Očekivani prosjek prve godine studiranja na temelju ranga na prijamnom ispitu

| rang na prijamnom ispitu | 95% pouzdani interval za očekivani prosjek prve godine studiranja |
|---|---|
| $1 - 10$ | $[3.335, 3.345]$ |
| $11 - 20$ | $[3.260, 3.268]$ |
| $21 - 30$ | $[3.184, 3.191]$ |
| $31 - 40$ | $[3.109, 3.115]$ |
| $41 - 50$ | $[3.033, 3.038]$ |
| $51 - 60$ | $[2.957, 2.962]$ |
| $61 - 70$ | $[2.882, 2.885]$ |
| $71 - 80$ | $[2.806, 2.809]$ |
| $81 - 90$ | $[2.731, 2.732]$ |
| $91 - 100$ | $[2.654, 2.655]$ |
| $101 - 110$ | $[2.578, 2.579]$ |
| $111 - 120$ | $[2.502, 2.504]$ |
| $121 - 130$ | $[2.425, 2.428]$ |
| $131 - 140$ | $[2.349, 2.352]$ |
| $141 - 150$ | $[2.272, 2.277]$ |
| $151 - 160$ | $[2.196, 2.201]$ |
| $161 - 170$ | $[2.119, 2.125]$ |
| $171 - 180$ | $[2.043, 2.050]$ |
| $181 - 190$ | $[1.966, 1.974]$ |
| $191 - 200$ | $[1.889, 1.898]$ |
| $201 - 210$ | $[1.813, 1.823]$ |
| $211 - 220$ | $[1.736, 1.747]$ |
| $221 - 230$ | $[1.659, 1.671]$ |
| $231 - 240$ | $[1.583, 1.596]$ |
| $241 - 250$ | $[1.506, 1.520]$ |

## 4.4 Test značajnosti linearnog regresijskog modela

Primijetimo, u slučaju $a = 0$ dobivamo $y = b = const.$, što nam govori da među promatranim obilježjima nema linearne ovisnosti. Dakle, naša je pretpostavka bila pogrešna. Zato ćemo provjeriti može li se dogoditi ova situacija. Formiramo sljedeće hipoteze:

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

Testiranje ovih hipoteza zove se *test značajnosti linearnog regresijskog modela.* Test značajnosti provodimo uz nivo značajnosti $\alpha = 0.05$. Budući da je procjena 95% pouzdanog intervala za $a$ jednaka $[-0.010252096, -0.004972246]$, a $0 \notin [-0.010252096, -0.004972246]$, odbacujemo $H_0$ u korist $H_1$ na nivou značajnosti od 0.05, tj. model je značajan.

Dakle, možemo biti 95% sigurni da je pretpostavka o linearnoj ovisnosti dobra, tj. da su naši rezultati valjani.

# 5 Zaključak

Rezimirajmo dobivene rezultate. Istraživanje smo započeli formiranjem sljedećih hipoteza:

- prosjek ocjena prve godine studiranja ne ovisi o spolu
- prosjek ocjena prve godine studiranja ne ovisi o godini upisa na fakultet
- prolaznost na prvoj godini studiranja ovisi o mjestu na rang listi prijamnog ispita
- prosjek ocjena na prvoj godini studiranja i rang na prijamnom ispitu u linearnoj su vezi i na temelju uspjeha na prijamnom ispitu možemo procijeniti prosjek prve godine studiranja

Da bismo ove hipoteze potvrdili ili opovrgnuli, prikupili smo potrebne podatke od 94 studenta iz populacije studenata PMF–MO koji su upisali 1. godinu studija akademske godine 2007./2008., 2008./2009. Na temelju toga, uz nivo značajnosti od 5%, dobili smo sljedeće rezultate:

- pomalo neočekivano, prosjek ocjena na 1. godini studiranja ovisi o spolu
- prosjek ocjena ne ovisi o godini upisa na fakultet, tj. ne postoji značajna razlika u prosjeku ocjena generacija upisanih 2007. i 2008. godine
- prolaznost na 1. godini studiranja ovisi o mjestu na rang listi prijamnog ispita, takav rezultat posve je očekivan
- potvrdili smo značajnost provedenog linearnog regresijskog modela, što nam dokazuje da su statistička obilježja rang na prijamnom ispitu i prosjek ocjena na 1. godini u linearnoj vezi $y = ax + b$, pri čemu je $a \in [-0.010252096, -0,004972246]$ i $b \in [3,06435966, 3,699772334]$.

Možda je jedan od najzanimljivijih rezultata bilo kreiranje tablice (Tablica 15) iz koje se na temelju mjesta na rang listi može isčitati 95% pouzdani interval o prosjeku na 1. godini studiranja. Provođenjem dvaju testova ($\chi^2$–test o nezavisnosti i linearni regresijski model) dotakli smo mali dio onoga što nam zapravo statistika kao takva omogućuje. Glavni zadatak ovog članka je objasniti široj populaciji da statistika nije samo crtanje dijagrama, već da iza svake tvrdnje stoji matematički alat koji i nije uvijek tako jednostavan. Također, svaki dobiveni rezultat temelji se na malom dijelu ukupne populacije, i kao takav vrijedi tek s određenom vjerojatnošću. Dakle, ne možemo reći "prosjek ocjena ovisi o spolu", nego tek kad napomenemo da to vrijedi s 95% vjerojatnosti, dobivamo valjani zaključak. Glavni problem danas je predstavljanje statističkih zaključaka sa 100% vjerojatnošću kako se to široj populaciji predstavlja. Usprkos svemu, odabirom dobre teme i analizom kvalitetnih hipoteza mogu se dobiti vrlo zanimljivi i ponekad neočekivani rezultati.