# Poslovna statistika

Poslovna ekonomija i finansije, Inženjerski menadžment u agrobiznisu

školska 2022/23

## Literatura


[1] Zagorka Lozanov-Crvenković, Marko Carić, Borivoj Subotić, Poslovna statistika, Fakultet za ekonomiju i inženjerski menadžment : Privredna akademija, Novi Sad, 2009

[2] Ljiljana Cvetković, Poslovna statistika, Folije za predavanja, Edukons

[3] Ghilezan et. al., Zbirka rešenih zadataka iz Verovatnoće i statistike, CMS, NS, 2009.


## Bodovi

| | Predavanja | Vežbe | Kol 1 | Kol 2 | Test | Σ |
|---|---|---|---|---|---|---|
| MAX | 10 | 10 | 40 | 40 | 10 | 110 |
| MIN | 0 | 0 | 15 | 15 | -50 | 51 |
| Datumi | sreda 9:00 - 11:45 | ponedeljak 15:00 - 16:45 | 19. IV 2023. 17. V 2023. | 31. V 2023. 9:00 | dogovor | |

# 1 Statistika, osnovni pojmovi

## 1.1 Populacija, obeležje, uzorak, statistika

**Populacija** je skup svih elemenata koje ispitujemo.
**Obeležje** je numerička karakteristika elementa. Modeliramo ga slučajnom promenljivom.
**Uzorak** je odabrani deo populacije na kojem ispitujemo realizovanu vrednost obeležja $X$.
**Prost slučajni uzorak** je $n$-dimenzionalna slučajna promenljiva čije komponente su nezavisne i imaju raspodelu posmatranog obeležja $(X_1, X_2, \ldots, X_n)$.
Realizovane vrednosti slučajnih promenljivih obeležavamo malim slovima $X_i \to x_i$.
**Realizovana vrednost prostog slučajnog uzorka** $(X_1, X_2, \ldots, X_n) \to (x_1, x_2, \ldots, x_n)$
**Statistika** je funkcija uzorka. $Y = h(X_1, X_2, \ldots, X_n)$
**Realizovana vrednost statistike** je $y = h(x_1, x_2, \ldots, x_n)$

## 1.2 Obeležje, atribut, parametar, promenljiva, varijabla

Obeležja nekad zovemo atributi ili parametri ili promenljive, odnosno, varijable. Često se na elementu populacije posmatraju vrednosti više atributa, odnosno parametara. Vrednosti više atributa jednog elementa populacije smeštamo u istu vrstu tabele. Kolone tabele možemo zvati promenljivama ili varijablama.

Na primer, podatke iz ankete studenata možemo smestiti u tabelu:

```
Datum   , Pol, Visina, Masa, Kafa   , TV     , Dužina sna, Boja kose   , Prosek ocena
20.03.82,   ž,    177,   63, ponekad, ponekad,          8, tamno smeđa ,          4.5
20.09.81,   ž,    170,   62, često  , ponekad,          8, svetlo smeđa,          4
26.06.83,   ž,    174,   63, često  , često  ,          5, smeđa       ,          4.5
20.11.77,   m,  171.5,   65, često  , ponekad,          8, svetlo smeđa,          4.2
```

Jednom podelom obeležja delimo na:

- **Nominalna:** Obeležja koja određujemo imenom bez bitnog redosleda. Na primer: crvena / zelena / plava ili da / ne / ne znam. Obeležje može biti **binarno:** mrtav / živ.
- **Ordinalna:** Obeležja redne klasifikacije, kod kojih postoji neki prirodni poredak ili rangiranje između kategorija. Na primer: junior / senior / veteran.
- **Kvantitativna:** Obeležja tipa merenja, kod kojih vrednosti mogu biti, barem u teoriji, realni brojevi. Na primer: težina, dužina ili vreme.

Nominalna i ordinalna obeležja se nekad nazivaju **kvalitativnim**. Kvantitativna obeležja mogu biti **diskretna:** broj braće i sestara ili **neprekidna:** vreme stizanja u cilj na trci.
Kvalitativna obeležja nekad nazivanom **kategorijalnim** ili **faktorima**. Možemo zapisivati kategoriju + **frekvenciju** (broj) pojavljivanja. Vrednosti koje kategorijalna promenljiva može uzeti se zovu **nivoi**.

## 1.3 Anketa

Izvršena je anketa 2002. godine, studenti su popunili anketni listić i podaci su smešteni u tabelu.

```
Datum rođenja:      Pol:
Mesto rođenja:      Mesto završetka srednje škole:
Prosek proseka srednje škole:   Preosek ocena iz matematike u srednjoj:
Visina [cm]:        Masa [kg]:
Boja očiju (plava, smeđa, tamno smeđa, zelena):
Boja kose (plava, riđa, smeđa, tamno smeđa, crna):
Pre koliko godina si završio srednju školu [god]:
Broj godina provedenih na drugom fakultetu:
Broj braće i sestara:
Pijem kafu (nikad, ponekad, retko, često):
Pušim (nikad, ponekad, retko, često):
Gledam TV (nikad, ponekad, retko, često):
Prosečna dužina spavanja:
```

Slika 1: Anketni listić koji su popunili studenti

Za atribute označiti znakom ✓ da li su odgovarajućeg tipa.

| | | Kval. | Nom. | Ord. | Kvan. | Diskr. | Nepr. | Bin. |
|---|---|---|---|---|---|---|---|---|
| 1. | Datum rođenja | | | | ✓ | ✓ | | |
| 2. | Pol | ✓ | ✓ | | | | | ✓ |
| 3. | Mesto rođenja | ✓ | ✓ | | | | | |
| 4. | Mesto zavr. sr. šk. | ✓ | ✓ | | | | | |
| 5. | Prosek proseka srednje škole | | | | ✓ | | ✓ | |
| 6. | Prosek ocena iz matematike u srednjoj | | | | ✓ | | ✓ | |
| 7. | Visina [cm] | | | | ✓ | | ✓ | |
| 8. | Masa [kg] | | | | ✓ | | ✓ | |
| 9. | Boja očiju | ✓ | ✓ | | | | | |
| 10. | Boja kose | ✓ | ✓ | | | | | |
| 11. | Pre koliko sr. šk. | | | | ✓ | ✓ | | |
| 12. | Broj godina na dr. fak. | | | | ✓ | ✓ | | |
| 13. | Broj braće i sestara | | | | ✓ | ✓ | | |
| 14. | Kafa: nikad / ponekad / retko / često | ✓ | | ✓ | | | | |
| 15. | Pušim: nikad / ponekad / retko / često | ✓ | | ✓ | | | | |
| 16. | TV: nikad / ponekad / retko / često | ✓ | | ✓ | | | | |
| 17. | Prosečna dužina spavanja | | | | ✓ | | ✓ | |

## 1.4 Važne statistike

**Aritmetička sredina uzorka**

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

**Uzoračka disperzija (varijansa)**

$$\bar{S}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}_n^2$$

**Uzoračka standardna devijacija**

$$\bar{S}_n = \sqrt{\bar{S}_n^2}$$

**Uzoračka korigovana disperzija**

$$\bar{S}_n^{2\prime} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 = \frac{n}{n-1}\bar{S}_n^2$$

**Uzoračka korigovana standardna devijac.**

$$\bar{S}_n' = \sqrt{\bar{S}_n^{2\prime}}$$

**Za obeležje sa Normalnom raspodelom** $X : \mathcal{N}(m,\sigma)$

Statistika $\bar{X}_n$ ima $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$ raspodelu.
Statistika $T = \frac{\bar{X}_n - m}{\bar{S}_n}\sqrt{n-1}$ ima Studentovu $t_{n-1}$ raspodelu.

# 2 Intervali poverenja $(U_1, U_2)$ za srednju vrednost obeležja

## 2.1 $\sigma$ poznato

Ako $X : \mathcal{N}(m,\sigma)$ onda $\bar{X}_n : \mathcal{N}(m,\sigma/\sqrt{n})$, odnosno, onda $Z = \dfrac{\bar{X}_n - m}{\sigma}\sqrt{n} : \mathcal{N}(0,1)$.
Označimo sa $z_{1-\alpha/2}$ vrednost za koju je $P(|Z| < z_{1-\alpha/2}) = 1-\alpha$, gde je $1-\alpha$ širina intervala.
$\left(z_{1-\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right.$ je $1-\frac{\alpha}{2}$ kvantil $\left.\right)$.
$U_1 = \bar{X}_n - z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$, $U_2 = \bar{X}_n + z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$. Izraz $\dfrac{\sigma}{\sqrt{n}}$ nazivamo **standardna greška**.

## 2.2 $\sigma$ nepoznato

Ako $X : \mathcal{N}(m,\sigma)$ onda $T = \dfrac{\bar{X}_n - m}{\bar{S}_n}\sqrt{n-1} = \dfrac{\bar{X}_n - m}{\bar{S}'_n}\sqrt{n} : t_{n-1}$.
Označimo sa $t_{1-\alpha/2}$ vrednost za koju je $P(|T| < t_{1-\alpha/2}) = 1-\alpha$.
($t_{1-\alpha/2}$ je $1-\alpha/2$ kvantil raspodele $t_{n-1}$.)
$U_1 = \bar{X}_n - t_{1-\alpha/2}\dfrac{\bar{S}_n}{\sqrt{n-1}}$, $U_2 = \bar{X}_n + t_{1-\alpha/2}\dfrac{\bar{S}_n}{\sqrt{n-1}}$. Standardna greška je $\dfrac{\bar{S}_n}{\sqrt{n-1}} = \dfrac{\bar{S}'_n}{\sqrt{n}}$.

**Zadatak**

Naći 90% interval poverenja za srednju vrednost $m$ obeležja sa normalnom $\mathcal{N}(m,\sigma)$ raspodelom

(a) Ako je poznato $\sigma = 3$, (b) ako je $\sigma$ nepoznato,

za uzorak $(17.3, 12.9, 10.4, 11.9, 9.9, 8.9, 9.9, 6.3, 12.9, 9.4)$.

**Rešenje:**

$n = 10$, $\bar{x}_n = 10.98$, $\bar{s}'_n = 2.973139$, $z_{0.95} = 1.645$, $t_{0.95} = 1.833$

(a) $U_1 = 9.4195$, $U_2 = 12.5404$ (b) $U_1 = 9.2565$, $U_2 = 12.7035$

# 3 Testiranje hipoteza

## Statistički testovi

Hipoteza $H_0$ protiv $H_1$

| | Usvojena $H_0$ | Usvojena $H_1$ |
|---|---|---|
| Tačna $H_0$ | OK | Greška I vrste |
| Tačna $H_1$ | Greška II vrste | OK |

## Parametarske hipoteze

- Zadaje se prag značajnosti $\alpha$ (recimo $\alpha = 5\% = 0.05$)
- Bira se parametar raspodele obeležja ($\theta$) i ocenjivač parametra $\theta = h(x_1, \ldots, x_n)$.
- Nalazi se kritična oblast $C$ (koja daje nedozvoljene vrednosti) ocene parametra, takva da je $P_{H_0}(\hat{\theta} = h(X_1, \ldots, X_n) \in C) = \alpha$. Vrednosti parametra koje ne pripadaju intervalu poverenja širine $1 - \alpha$ daju ocene koje pripadaju kritičnoj oblasti.
- Računa se statistika uzorka $\theta = h(x_1, x_2, \ldots, x_n)$ i ako $\theta \in C$, odbacujemo $H_0$.
- Ili, ekvivalentno, računa se p-vrednost($\alpha^*$): verovatnoća da se dobije ocena izračunata iz uzorka ili ekstremnija. Ako je p-vrednost manja od praga značajnosti, $\alpha^* < \alpha$, odbacujemo $H_0$.

### 3.1 $H_0(m = m_0)$ **protiv** $H_1(m \neq m_0)$ **za** $X : \mathcal{N}(m,\sigma)$, $\sigma$ **poznato**

Koristimo interval poverenja za nepoznato očekivanje $m$ obeležja sa normalnom raspodelom, $\sigma$ poznato, širine $\beta = 1 - \alpha$. $(X^* : \mathcal{N}(0,1))$

$$m_0 \in \mathbb{R} \setminus \left( \bar{x}_n \mp z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \Leftrightarrow z := \frac{|\bar{x}_n - m_0|}{\sigma} \sqrt{n} > z_{1-\alpha/2} \Leftrightarrow \alpha^* := P_{H_0} \left( |X^*| > \frac{|\bar{x}_n - m_0|}{\sigma} \sqrt{n} \right) < \alpha$$

### 3.2 $H_0(m = m_0)$ **protiv** $H_1(m \neq m_0)$ **za** $X : \mathcal{N}(m,\sigma)$, $\sigma$ **nepoznato**

Koristimo interval poverenja za nepoznato očekivanje $m$ obeležja sa normalnom raspodelom, $\sigma$ nepoznato, širine $1 - \alpha$. $(T : t_{n-1})$

$$m_0 \in \mathbb{R} \setminus \left( \bar{x}_n \mp t_{1-\alpha/2} \frac{\bar{s}'_n}{\sqrt{n}} \right) \Leftrightarrow t := \frac{|\bar{x}_n - m_0|}{\bar{s}'_n} \sqrt{n} > t_{1-\alpha/2} \Leftrightarrow \alpha^* := P_{H_0} \left( |T| > \frac{|\bar{x}_n - m_0|}{\bar{s}'_n} \sqrt{n} \right) < \alpha$$

**Zadatak**

Za uzorak $(17.3, 12.9, 10.4, 11.9, 9.9, 8.9, 9.9, 6.3, 12.9, 9.4)$ obeležja sa normalnom $\mathcal{N}(m,\sigma)$ raspodelom, sa pragom značajnosti $\alpha = 0.10$, testirati hipotezu $H_0(m = 13)$

(a) ako je poznato $\sigma = 3$, (b) ako je $\sigma$ nepoznato.

Dati tumačenja rešenja preko intervala poverenja, preko vrednosti statistike i preko p-vrednosti.

**Rešenje:**
$\alpha = 0.05 \Rightarrow 1 - \alpha/2 = 0.95$
Iz uzorka računamo $n = 10$, $\bar{x}_n = (17.3 + 12.9 + \cdots + 9.4)/10 = 10.98$,
$\bar{s}_n^{2\prime} = ((17.3 - 10.98)^2 + (12.9 - 10.98)^2 + \cdots + (9.4 - 10.98)^2)/9 = 8.839556$,
$\bar{s}_n' = \sqrt{8.839556} = 2.973139$.
Iz tablica očitavamo kvantile $z_{0.95} = 1.645$ i $t_{0.95,9} = 1.833$.
Interval poverenja:

(a) $13 \notin (x_n \pm z_{0.95}\frac{3}{\sqrt{10}}) = (9.419416, 12.54058)$, odbacujemo $H_0$.

(b) $13 \notin (x_n \pm t_{0.95,9}\frac{2.973139}{\sqrt{10}}) = (9.256527, 12.70347)$, odbacujemo $H_0$.

Vrednost statistike

(a) $z = \frac{|\bar{x}_n - m_0|}{\sigma}\sqrt{n} = 2.129 > 1.645$, odbacujemo $H_0$.

(b) $t = \frac{|\bar{x}_n - m_0|}{\bar{s}_n'}\sqrt{n} = 2.148 > 1.833$, odbacujemo $H_0$.

P-vrednost se može dobiti pomoću statističkog softvera R:

(a) `2*(1-pnorm(abs(10.98-13)/3*sqrt(10))) = 0.033 < 0.100`, odbacujemo $H_0$.

(b) `2*(1-pt(abs(10.98-13)/2.973*sqrt(10),9)) = 0.060 < 0.100`, odbacujemo $H_0$.

**Zadatak**

Da li bi u prethodnom zadatku $H_0$ bila odbačena sa pragom značajnosti $\alpha = 0.05$?
**Rešenje:**
Koristimo p-vrednost iz prethodnog zadatka.

(a) `0.033 < 0.050`, odbacujemo $H_0$.

(b) `0.060 > 0.050`, **ne** odbacujemo $H_0$.

## 3.3 $H_0(m_1 = m_2)$ **protiv** $H_1(m_1 \neq m_2)$ **(T-test)**

Koristimo statistiku $T := \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\bar{S}_1^{2\prime}}{n_1} + \frac{\bar{S}_2^{2\prime}}{n_2}}}$, koja približno ima $t_\nu$ raspodelu,

gde se za $\nu$ uzima $\nu = n_1 + n_2 - 2$ ili procena Welcha $\nu = \dfrac{\left(\frac{s_1^{2\prime}}{n_1} + \frac{s_2^{2\prime}}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^{2\prime}}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^{2\prime}}{n_2}\right)^2}$.

**Zadatak**

Merena je visina glukoze u krvi za grupu koja je dobijala lek i grupu koja je dobijala placebo:

| | |
|---|---|
| lek | 9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95 |
| placebo | 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78 |

Testirati hipotezu da su srednje vrednosti glukoza jednake.

**Rešenje:**

Primenom statističkog softvera R dobijamo:

```
> t.test(c(9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95),
         c( 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78))
  Welch Two Sample t-test
t = 1.7536, df = 16.766, p-value = 0.09776
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1439427 1.5522760
sample estimates:
mean of x mean of y
10.032500 9.328333
```

$H_0$ ne odbacujemo jer $t = 1.7536 < t_{1-\alpha/2,\nu} = 2.074$, gde je $1-\alpha/2 = 0.975$ i $\nu = 22$.

## 3.4 T-test parova

Koristi se kad imamo dva obeležja sa uparenim vrednostima ("pre" i "posle"): $x_1, x_2, \ldots, x_n$ i $y_1, y_2, \ldots, y_n$. Nalazimo $z_1 = x_1 - y_1, z_2 = x_2 - y_2, \ldots, z_n = x_n - y_n$ i testiramo $H_0(m = 0)$ protiv $H_1(m \neq 0)$ ili $H_1(m < 0)$ ili $H_1(m > 0)$, sa $\sigma$ nepoznato za uzorak $z_1, z_2, \ldots, z_n$

**Zadatak**

Merena je visina glukoze u krvi za 12 pacijenata pre i posle terapije:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pre | 9.81 | 9.83 | 10.43 | 11.13 | 9.70 | 9.59 | 10.88 | 10.97 | 9.35 | 9.34 | 9.41 | 9.95 |
| posle | 9.85 | 9.30 | 9.08 | 8.07 | 9.22 | 9.55 | 7.88 | 7.84 | 8.50 | 11.95 | 10.92 | 9.78 |
| razlika | −0.04 | 0.53 | 1.35 | 3.06 | 0.48 | 0.04 | 3.00 | 3.13 | 0.85 | −2.61 | −1.51 | 0.17 |

Testirati hipotezu da su srednje vrednosti glukoze jednake.

**Rešenje:**

Primenom statističkog softvera R dobijamo:

```
> t.test(c(9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95),
         c( 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78),
         paired = T)
  Paired t-test
t = 1.3796, df = 11, p-value = 0.1951
```

$H_0$ ne odbacujemo jer $t = 1.3796 < t_{1-\alpha/2,11} = 2.201$, gde je $1 - \alpha/2 = 0.975$.

# 4 Linearna regresija

## Zadatak

U šest prolećnih dana je zabeležena srednja dnevna temperatura u Budimpešti i Beogradu:

| | | | | | |
|---|---|---|---|---|---|
| $BU$ | 18.5 | 19.5 | 20 | 20.5 | 22.5 |
| $BG$ | 19.4 | 19.8 | 20 | 20.2 | 21.2 |

Napraviti linearni model, naći koeficijent korelacije i naći procenat varijacije koji se objašnjava linearnim modelom.

Naći predikciju temperature u BG za BU $= 21^{\circ}$C.

## 4.1 Uzorački koeficijent korelacije

Ako posmatramo dvodimenzionalni uzorak $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, odnosno, za realizovanu vrednost $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, definišemo **uzorački koeficijent korelacije**:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\bar{x}_n\bar{y}_n}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}_n^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}_n^2}} =$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}} = \frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}_n\bar{y}_n}{\sqrt{\sum_{i=1}^{n}x_i^2-n\bar{x}_n^2}\sqrt{\sum_{i=1}^{n}y_i^2-n\bar{y}_n^2}}$$

Prva formula: $r = \frac{3.980}{\sqrt{8.80}\sqrt{1.8080}} = 0.9977978$, $r^2 = 0.9956004$.

Četvrta formula: $r = \frac{2036.1-5\cdot 20.2\cdot 20.12}{\sqrt{2049-5\cdot 20.2^2}\sqrt{2025.88-5\cdot 20.12^2}} = 0.9977978$, $r^2 = 0.9956004$.

## 4.2 Linearna regresija najmanjih kvadrata

Za $n$ parova tačaka $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, tražimo vezu između $x$ i $y$ u obliku prave linije $y = a + bx$.

Tražimo vrednosti $a$ i $b$ za koje funkcija $g(a,b) = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$ ostvaruje minimum.

Rešenje:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\,(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}_n\,\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}_n^2} = r\frac{s_y}{s_x} = r\frac{s'_y}{s'_x},\ a = \bar{y}_n - b\bar{x}_n,$$

gde su $s_x$ i $s_y$ standardne devijacije uzorka $x$ i $y$, $s'_x$ i $s'_y$ korigovane standardne devijacije.
Za tako izračunate $a$ i $b$ funkciju $\hat{y} = ax + b$ zovemo **prava najmanjih kvadrata**. Vrednosti $\hat{y}_i = a + bx_i$ su **predikcije**. Važi $\hat{\bar{y}}_n = \bar{y}_n$. Prava najmanjih kvadrata prolazi kroz $(\bar{x}_n,\bar{y}_n)$.
$ss_x = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$, $ss_y = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2$, $s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x}_n)\,(y_i - \bar{y}_n)$, $\Rightarrow\ r = \frac{s_{xy}}{\sqrt{ss_x\,ss_y}}$, $b = \frac{s_{xy}}{ss_x}$.
Sa tim oznakama imamo: $s_{xy} = r\sqrt{ss_x}\sqrt{ss_y}$, takođe: $b = r\frac{\sqrt{ss_x}\sqrt{ss_y}}{ss_x} = r\frac{\sqrt{ss_y}}{\sqrt{ss_x}}$.
$\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}_n))^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$.

Takođe: $\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2=\sum_{i=1}^{n}(a+bx_i-(a+b\bar{x}_n))^2=b^2\sum_{i=1}^{n}(x_i-\bar{x}_n)^2=b^2ss_x=r^2ss_y.$

$$r^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_n)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}=\frac{\text{varijansa predikcija } y}{\text{varijansa realizovanih } y}$$

odnosno, $r^2\cdot 100\%$ je **procenat varijanse objašnjene pravom linijom** najmanjih kvadrata. Vrednosti $\epsilon_i=y_i-\hat{y}_i$ zovemo **reziduali**.

| | $x_i$ | $y_i$ | $x_iy_i$ | $x_i-\bar{x}$ | $y_i-\bar{y}$ | $(x_i-\bar{x})^2$ | $(y_i-\bar{y})^2$ | $(x_i-\bar{x})(y_i-\bar{y})$ | $x_i^2$ | $y_i^2$ | $x_iy_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18.5 | 19.4 | 358.90 | −1.7 | −0.72 | 2.89 | 0.5184 | 1.224 | 342.25 | 376.36 | 358.9 |
| 2 | 19.5 | 19.8 | 386.10 | −0.7 | −0.32 | 0.49 | 0.1024 | 0.224 | 380.25 | 392.04 | 386.1 |
| 3 | 20.0 | 20.0 | 400.00 | −0.2 | −0.12 | 0.04 | 0.0144 | 0.024 | 400.00 | 400.00 | 400.0 |
| 4 | 20.5 | 20.2 | 414.10 | 0.3 | 0.08 | 0.09 | 0.0064 | 0.024 | 420.25 | 408.04 | 414.1 |
| 5 | 22.5 | 21.2 | 477.00 | 2.3 | 1.08 | 5.29 | 1.1664 | 2.484 | 506.25 | 449.44 | 477.0 |
| $\Sigma$ | 101.0 | 100.6 | 2036.10 | 0.0 | 0.00 | 8.80 | 1.8080 | 3.980 | 2049.00 | 2025.88 | 2036.1 |
| $\Sigma/n$ | 20.2 | 20.12 | 407.22 | 0.0 | 0.00 | 1.76 | 0.3616 | 0.796 | 409.80 | 405.176 | 407.22 |

Prva formula: $b=\dfrac{3.980}{8.80}=0.45227$, $a=20.12-0.45227\cdot 20.2=10.9841$.

Linearni model: BG = 10.9841 + 0.45227 · BU

Linearni model objašnjava $r^2=99.56\%$ varijacije.

Predikcija: BR ( BU = 21 ) = 10.9841 + 0.45227 · 21 = 20.48177

## 4.3 Analiza varijanse ANOVA (jednofaktorska)

$H_0(m_1 = m_2 = \cdots = m_G)$, $H_1(\exists i,j, m_i \neq m_j)$

### Zadatak

Poljoprivredni proizvođač želi da testira kvalitet četiri vrste semena soje: A, B, C, D i u tom cilju je odabrao 30 parcela iste površine koje imaju sličan kvalitet zemljišta, drenažu i izloženost suncu. Dobijeni su sledeći prinosi:

| Seme | Prinos |
|---|---|
| A | 46,43,43,46,44,42 |
| B | 51,58,62,49,53,51,50,59 |
| C | 37,39,41,38,39,37,42,36,40 |
| D | 42,43,42,45,47,50,48 |

Metodom analize varijanse ispitati da li postoje razlike u prosečnim prinosima soje kod semena A, B, C, D sa nivoom značajnosti $\alpha = 0.05$.

**Rešenje:**

Testira se $H_0$ : srednje vrednosti prinosa za četiri vrednosti semena se ne razlikuju
protiv: $H_1$ : Srednje vrednosti prinosa za barem dva semena se razlikuju.

Posmatra se kvantitativno obeležje Prinos i kategorijalno obeležje Seme. Obeležje Seme deli uzorak na četiri grupe: A, B, C, D. Pretpostavka je da je raspodela prinosa u grupama Normalna sa istom varijansom.

**Analiza varijanse Fišerovom statistikom**

| Grupa | Merenje | | | | | Grupna sredina |
|---|---|---|---|---|---|---|
| 1 | $Y_{11}$ | $Y_{12}$ | $\cdots$ | $Y_{1n_1}$ | | $\bar{Y}_{1\cdot}$ |
| 2 | $Y_{21}$ | $Y_{22}$ | $\cdots$ | | $Y_{2n_2}$ | $\bar{Y}_{2\cdot}$ |
| $\vdots$ | | | $\ddots$ | | | |
| $G$ | $Y_{G1}$ | $Y_{G2}$ | $\cdots$ | $Y_{Gn_G}$ | | $\bar{Y}_{G\cdot}$ |

$\left(\bar{Y}_{g\cdot} = \frac{1}{n_g}\sum_{k=1}^{n_g} Y_{gk}\right)$

$Y_{gk} = m_g + \epsilon_{gk}$, gde je $m_g = E(Y_{gk})$, $\quad \bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n_g} Y_{gk} = \frac{1}{n}\sum_{g=1}^{G} n_g \bar{Y}_{g\cdot}$, $\quad n = \sum_{k=1}^{G} n_g$

Treatment: $SSTR = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{g=1}^{G} n_g(\bar{Y}_{g\cdot} - \bar{Y}_{\cdot\cdot})^2$

Error: $SSE = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g\cdot})^2$, $\qquad$ Total: $SST = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{\cdot\cdot})^2$

$SST = SSTR + SSE$, $\quad \bar{S}_g^{2\prime} = \frac{1}{n_g - 1}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g\cdot})^2$, $\quad SSE = \sum_{g=1}^{G}(n_g - 1)\bar{S}_g^{2\prime}$

$$\frac{(n_1-1)\,\bar{S}_1^{2\prime} + (n_2-1)\,\bar{S}_2^{2\prime} + \cdots + (n_G-1)\,\bar{S}_G^{2\prime}}{(n_1-1) + (n_2-1) + \cdots + (n_G-1)} = \frac{\sum_{g=1}^{G}(n_g-1)\,\bar{S}_g^{2\prime}}{n-G} = \frac{SSE}{n-G}$$

Neka su $Y_{gk}$, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$ nezavisne slučajne promenljive sa istim očekivanjem u grupi: $E(Y_{gk}) = m_g$ i istom varijansom $D(Y_{gk}) = \sigma^2$ i neka $m = \sum_{g=1}^{G} n_g m_g / n$.

Može se dokazati da je $E\left(\frac{SSTR}{G-1}\right) = \sigma^2 + \frac{1}{G-1}\sum_{g=1}^{G} n_g (m_g - m)^2$ i $E\left(\frac{SSE}{n-G}\right) = \sigma^2$.

Takođe važi: Ako su $Y_{gk}$, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$ nezavisne slučajne promenljive sa normalnom raspodelom $Y_{gk} : \mathcal{N}(m_g,\sigma)$, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$, onda

1. $SSE$ i $SSTR$ su nezavisne
2. $SSE/\sigma^2$ ima $\chi^2_{n-G}$ raspodelu
3. Ako $m_1 = m_2 = \cdots = m_G$, onda $SSTR/\sigma^2$ ima $\chi^2_{G-1}$ raspodelu

Odatle sledi: ako $MSTR = \frac{SSTR}{G-1}$ i $MSE = \frac{SSE}{n-G}$, statistika $F = \frac{MSTR}{MSE}$ ima $F_{G-1,n-G}$ raspodelu

```
> anova(prinostab)
Analysis of Variance Table
Response: prinos
          Df  Sum Sq Mean Sq F value    Pr(>F)
seme       3 1015.51  338.50  32.614 5.781e-09 ***
Residuals 26  269.86   10.38
```

Vidimo rezultat primene statističkog softvera R na ulaz oblika

| Prinos | Seme |
|---|---|
| 46 | $A$ |
| 43 | $A$ |
| $\dots$ | $\dots$ |
| 48 | $D$ |

Preko vrednosti statistike (pročitane iz tabele):

Kako je realizovana vrednost $F$ statistike $f = 32.614 > 2.9715$ odbacujemo $H_0$.

Preko p-vrednosti (pročitane iz tabele):

Kako je p-vrednost koja odgovara realizovanoj vrednosti $F$ statistike $f = 32.614$ ili ekstremnijim vrednostima jednaka $\alpha^* = 5{,}781 \times 10^{-9} < 0.05 = \alpha$, zaključujemo da se $H_0$ odbacuje. Možemo tvrditi da se srednje vrednosti prinosa na barem dva ogledna polja razlikuju.