RAD SA PODATCIMA: DESKRIPTIVNA STATISTIKA

Hrvatski studiji

dr.sc. Luka Šikić

Preddiplomski studij sociologije

21 listopad, 2019

CILJEVI PREDAVANJA

- Podatci
- Mjere centralne tendencije
- Mjere varijabilnosti
- Mjere asimetrije i zaobljenosti
- Pregled varijabli i podatatkovnih okvira
- Standardizirane vrijednosti
- Korelacija

UČITAVANJE PODATAKA

```
# Učitaj paket
library(lsr)
# Definiraj put do podataka setwd()
# Provjera getwd()
load("aflsmall.Rdata") # Učitaj podatke u radni prostor
who() # Pregledaj učitane podatke
```

```
##     -- Name --       -- Class --    -- Size --
##     afl.finalists    factor         400
##     afl.margins      numeric        176
```

PREGLED PODATAKA

```
print(afl.margins[1:11])
```

```
##  [1] 56 31 56  8 32 14 36 56 19  1  3
```

```
print(afl.finalists[1:5])
```

```
## [1] Hawthorn  Melbourne Carlton   Melbourne Hawthorn
## 17 Levels: Adelaide Brisbane Carlton Collingwood Essendo
```

VIZUALIZACIJA PODATAKA

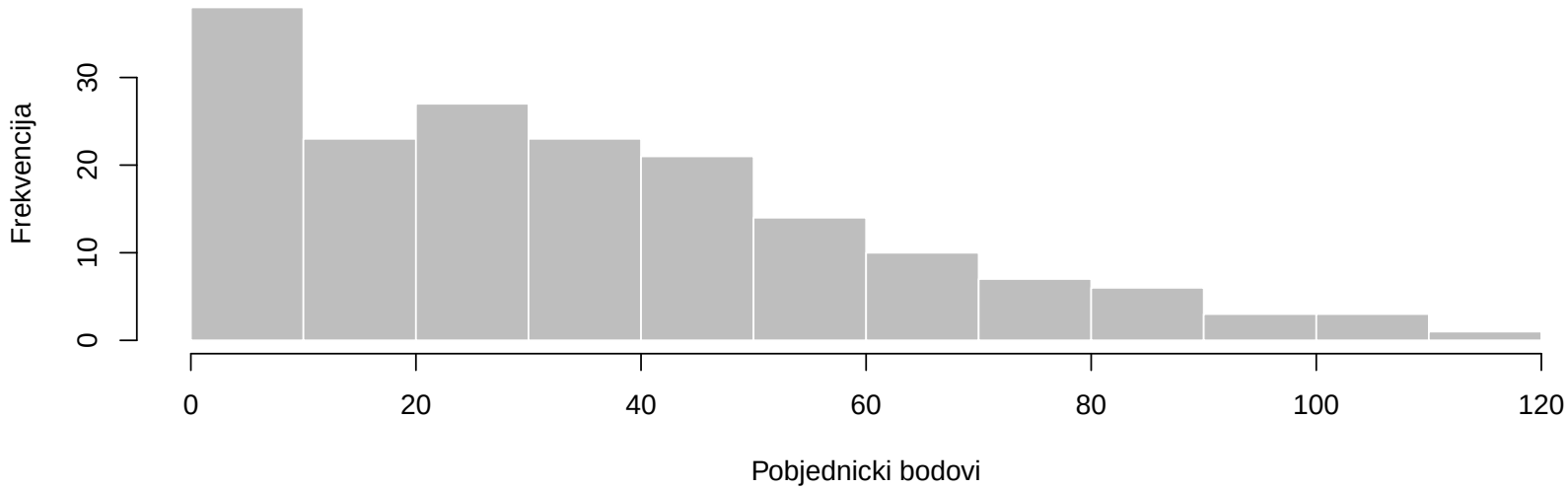

Figure 1: Histogram pobjedničkih bodova(`afl.margins`) iz AFL 2010 lige američkog nogometa. Grafikon prikazuje da se broj pobjeda uz veću razliku rijeđe pojavljuje.

MJERE CENTRALNE TENDENCIJE

- Aritmetička sredina
- Medijan
- Mod

ARITMETIČKA SREDINA

- Definicija

$$\bar{X} = \frac{X_1 + X_2 + ... + X_{N-1} + X_N}{N}$$

- Sumiranje

$$\sum_{i=1}^{5} X_i$$

- Skraćeni zapis

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

ARITMETIČKA SREDINA

- Izračun rukom

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36.60$$

- Kalkulator

```
(56 + 31 + 56 + 8 + 32) / 5
```

```
## [1] 36.6
```

- Funkcija

```
sum( afl.margins[1:5]) / 5
```

```
## [1] 36.6
```

MEDIJAN

- za neparni niz

$$8, 31, \mathbf{32}, 56, 56$$

- za parni niz

$$8, 14, \mathbf{31}, \mathbf{32}, 56, 56$$

```
#  Izračunaj median putem funkcije
median( x = afl.margins ) # Cijeli podatkovni skup
```

```
## [1] 30.5
```

EKSTREMNE VRIJEDNOSTI I

```
# Definiraj vektor od 10 brojeva
vektor_10 <- c( -15,2,3,4,5,6,7,8,9,12 )
mean( x = vektor_10 ) # Izračunaj AS
```

```
## [1] 4.1
```

```
median( x = vektor_10 ) # Izračunaj medijan
```

```
## [1] 5.5
```

EKSTREMNE VRIJEDNOSTI II

```
# Ukloni 10% ekstremnih vrijednosti
mean( x = vektor_10, trim = .1)
```

```
## [1] 5.5
```

```
# Ukloni 5% ekstremnih vrijednosti
mean( x = afl.margins, trim = .05)
```

```
## [1] 33.75
```

MOD I

```
# Pogledaj frekvenciju podataka
table(afl.finalists)
```

```
## afl.finalists
##          Adelaide          Brisbane           Carlton
##                26                25                26
##          Essendon           Fitzroy         Fremantle
##                32                 0                 6
##          Hawthorn         Melbourne North Melbourne        Po
##                27                28                28
##          Richmond         St Kilda            Sydney
##                 6                24                26
## Western Bulldogs
##                24
```

MOD II

```
# Izračunaj modalnu vrijednost
modeOf( x = afl.finalists )
```

```
## [1] "Geelong"
```

```
# Izračunaj modalnu frekvenciju
maxFreq(x = afl.finalists)
```

```
## [1] 39
```

MOD III

```
# Izaračun za afl.margins podatke
modeOf(afl.margins)  # Mod
```

```
## [1] 3
```

```
maxFreq(afl.margins) # Modalna frekvencija
```

```
## [1] 8
```

MJERE VARIJABILNOSTI

- Raspon/Min-Max
- Kvartili
- Prosječno apsolutno odstupanje
- Varijanca
- Standardna devijacija
- Srednje apsolutno odstupanje

RASPON/MIN-MAX

```
# Maksimalna vrijednost
max(afl.margins)
```

```
## [1] 116
```

```
# Minimalna vrijednost
min(afl.margins)
```

```
## [1] 0
```

```
# Raspon podataka
range(afl.margins)
```

```
## [1]   0 116
```

KVARTILI

```
# Izračunaj pedeseti (50i) kvartil/percentil
quantile(x = afl.margins, probs = .5)
```

```
##  50%
## 30.5
```

```
# Izračunaj 25i i 75i kvartil/percentil
quantile(afl.margins, probs = c(.25,.75))
```

```
##    25%    75%
## 12.75 50.50
```

INTERKVARTILNI RASPON

```
# Izračunaj interkvartilni raspon
IQR(x = afl.margins)
```

```
## [1] 37.75
```

PROSJEČNO APSOLUTNO ODSTUPANJE I

$$(X) = \frac{1}{N} \sum_{i=1}^{N} |X_i - \bar{X}|$$

PROSJEČNO APSOLUTNO ODSTUPANJE II

Table 1: Tablica za ručni izračun prosječnog apsolutnog odstupanja.

i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})$
1	56	19.4	19.4
2	31	-5.6	5.6
3	56	19.4	19.4
4	8	-28.6	28.6
5	32	-4.6	4.6

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

PROSJEČNO APSOLUTNO ODSTUPANJE III

- izračun pomoću funkcija

```
X <- c(56, 31,56,8,32) # Napravi vektor
X.bar <- mean( X )     # Korak 1. Izračunaj AS
AD <- abs( X - X.bar ) # Korak 2. Uzmi aps vrijednost
AAD <- mean( AD )      # Korak 3. Izračunaj AS devijacija
print( AAD )           # Pogledaj rezultate
```

```
## [1] 15.52
```

VARIJANCA I

$$\text{Var}(X) = \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2$$

$$\text{Var}(X) = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2}{N}$$

VARIJANCA II

Table 2: Tablica za ručni izračun varijance.

i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	56	19.4	376.36
2	31	-5.6	31.36
3	56	19.4	376.36
4	8	-28.6	817.96
5	32	-4.6	21.16

```
# Kalkulator
(376.36 + 31.36 + 376.36 + 817.96 + 21.16 ) / 5
```

```
## [1] 324.64
```

VARIJANCA III

```r
# Izračunaj varijancu pomoću funkcija
mean( (X - mean(X) )^2)
```

```
## [1] 324.64
```

```r
var(X) # Skrati postupak
```

```
## [1] 405.8
```

VARIJANCA IV

```
## Isti primjer sa svim podatcima
# Izračunaj varijancu pomoću funkcija
mean( (afl.margins - mean(afl.margins) )^2)
```

```
## [1] 675.9718
```

```
var( afl.margins ) # Skrati postupak
```

```
## [1] 679.8345
```

STANDARDNA DEVIJACIJA I

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

```
# Izračunaj pomoću funkcije
sd( afl.margins )
```

```
## [1] 26.07364
```

APSOLUTNO ODSTUPANJE OD MEDIJANA

```
# Prosječno apsolutno odstupanje od prosjeka
mean( abs(afl.margins - mean(afl.margins)) )
```

```
## [1] 21.10124
```

```
# *Medijansko*  apsolutno odstupanje od *medijana*:
median( abs(afl.margins - median(afl.margins)) )
```

```
## [1] 19.5
```

```
# Izračun putem funkcije
mad( x = afl.margins, constant = 1 )
```

```
## [1] 19.5
```

KOEFICIJENT ASIMETRIJE I

Figure 2: Asimetričnost: na lijevoj strani (skew $= -.93$), u sredini nema zakrivljenosti (skew $= -.006$), na desnoj strani (skew $= .93$).

KOEFICIJENT ASIMETRIJE II

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

```
# Izračunaj na stvarnim podatcima
skew( x = afl.margins )
```

```
## [1] 0.7671555
```

KOEFICIJENT ZAOBLJENOSTI I

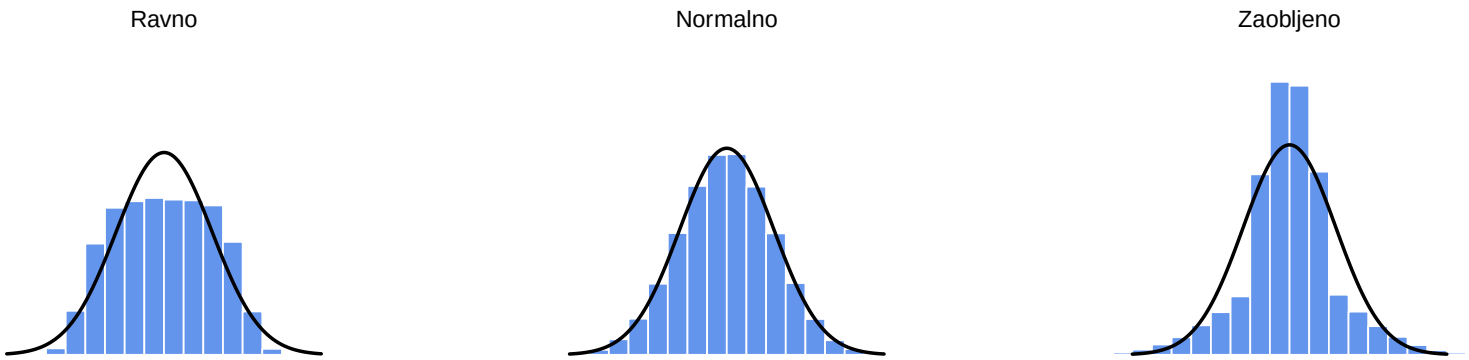

Figure 3: Zaobljenost: na lijevoj strani ravno (kurtosis $= -.95$),u sredini normalna zaobljenost (kurtosis ~ 0), na desnoj strani zaobljeno (kurtosis $= 2.12$). Zaobljenost se mjeri u odnosu na crnu liniju.

KOEFICIJENT ZAOBLJENOSTI II

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^4 - 3$$

```
# Izračunaj na stvarnim podatcima
kurtosi( x = afl.margins )
```

```
## [1] 0.02962633
```

DESKRIPTIVNA STATISTIKA NA VARIJABLI I

```
# Pregled numeričke varijable
summary( object = afl.margins ) # Deskriptivna stat
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.75   30.50   35.30   50.50  116.00
```

DESKRIPTIVNA STATISTIKA NA VARIJABLI II

```
# Pregled logičke varijable
ekstremi <-  afl.margins > 50 # Stvori log varijablu
head(ekstremi,5) # Pogledaj podatke
```

```
## [1]  TRUE FALSE  TRUE FALSE FALSE
```

```
summary(ekstremi) # Deskriptivna stat
```

```
##    Mode   FALSE    TRUE
## logical     132      44
```

DESKRIPTIVNA STATISTIKA NA VARIJABLI III

```
# Pregled faktorske varijable
summary(object = afl.finalists) # Deskriptivna stat
```

```
##           Adelaide           Brisbane             Carlton
##                 26                 25                  26
##           Essendon            Fitzroy           Fremantle
##                 32                  0                   6
##           Hawthorn          Melbourne  North Melbourne     Po
##                 27                 28                  28
##           Richmond          St Kilda             Sydney
##                  6                 24                  26
## Western Bulldogs
##                 24
```

DESKRIPTIVNA STATISTIKA NA VARIJABLI IV

```
# Pregled tekstualne varijable
txt <- as.character( afl.finalists ) # Stvori txt var
summary( object = txt ) # Deskriptivna stat
```

```
##    Length     Class      Mode
##       400 character character
```

NOVI PODATKOVNI SKUP

```r
rm(list = ls()) # Očisti radni prostor
load("clinicaltrial.Rdata") # Učitaj podatke
who(TRUE) # Pregled podataka
```

```
##     -- Name --      -- Class --    -- Size --
##     clin.trial      data.frame     18 x 3
##      $drug          factor         18
##      $therapy       factor         18
##      $mood.gain     numeric        18
```

DESKRIPTIVNA STATISTIKA NA PODATKOVNOM OKVIRU I

```
# Deksriptivna statistika na podatkovnom okviru
summary(clin.trial) # Desktiptivna stat
```

```
##         drug           therapy    mood.gain
##  placebo :6   no.therapy:9   Min.   :0.1000
##  anxifree:6   CBT        :9   1st Qu.:0.4250
##  joyzepam:6                   Median :0.8500
##                               Mean   :0.8833
##                               3rd Qu.:1.3000
##                               Max.   :1.8000
```

DESKRIPTIVNA STATISTIKA NA PODATKOVNOM OKVIRU II

```
# Deksriptivna statistika na podatkovnom okviru
describe(clin.trial) # Desktiptivna stat/ druga funkcija
```

```
##              vars  n mean   sd median trimmed  mad min max
## drug*           1 18 2.00 0.84   2.00    2.00 1.48 1.0 3.0
## therapy*        2 18 1.50 0.51   1.50    1.50 0.74 1.0 2.0
## mood.gain       3 18 0.88 0.53   0.85    0.88 0.67 0.1 1.8
##              kurtosis   se
## drug*           -1.66 0.20
## therapy*        -2.11 0.12
## mood.gain       -1.44 0.13
```

DESKRIPTIVNA STATISTIKA NA PODATKOVNOM OKVIRU III

```
# Pregledaj grupirano prema terapiji
by(data = clin.trial, # Izvor podataka
   INDICES = clin.trial$therapy, # Odredi grupiranje
   FUN = summary) # Odredi funkciju
```

```
## clin.trial$therapy: no.therapy
##        drug           therapy    mood.gain
##  placebo :3   no.therapy:9   Min.   :0.1000
##  anxifree:3   CBT        :0   1st Qu.:0.3000
##  joyzepam:3                   Median :0.5000
##                               Mean   :0.7222
##                               3rd Qu.:1.3000
##                               Max.   :1.7000
## ------------------------------------------------------------
```

DESKRIPTIVNA STATISTIKA NA PODATKOVNOM OKVIRU IV

```r
# Pregledaj grupirano prema razlici u raspoloženju
aggregate(formula = mood.gain ~ drug + therapy, # Prikaz
          data = clin.trial, # Podatci
          FUN = mean) # AS
```

```
##       drug    therapy mood.gain
## 1  placebo no.therapy  0.300000
## 2 anxifree no.therapy  0.400000
## 3 joyzepam no.therapy  1.466667
## 4  placebo        CBT  0.600000
## 5 anxifree        CBT  1.033333
## 6 joyzepam        CBT  1.500000
```

DESKRIPTIVNA STATISTIKA NA PODATKOVNOM OKVIRU V

```
# Pregledaj grupirano prema razlici u raspoloženju
aggregate(mood.gain ~ drug + therapy,  # Prikaz
          clin.trial, # Podatci
          sd) # Standardna devijacija
```

```
##       drug    therapy mood.gain
## 1  placebo no.therapy 0.2000000
## 2 anxifree no.therapy 0.2000000
## 3 joyzepam no.therapy 0.2081666
## 4  placebo        CBT 0.3000000
## 5 anxifree        CBT 0.2081666
## 6 joyzepam        CBT 0.2645751
```

STANDARDNE VRIJEDNOSTI

$$\text{standardna vrijednost} = \frac{\text{vrijednost opservacije} - \text{prosjek}}{\text{standardna devijacija}}$$

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

$$z = \frac{35 - 17}{5} = 3.6$$

```
# Vidi dio u distribuciji
pnorm( 3.6 )
```

```
## [1] 0.9998409
```

NOVI PODATKOVNI SKUP

```
rm(list = ls()) # Očisti radni prostor
 # Učitaj podatke
load("parenthood.Rdata")
who(TRUE) # Pregled podataka
```

```
##     -- Name --      -- Class --    -- Size --
##     parenthood      data.frame     100 x 4
##       $dan.sleep    numeric        100
##       $baby.sleep   numeric        100
##       $dan.grump    numeric        100
##       $day          integer        100
```

NOVI PODATKOVNI SKUP

```
# Pregledaj podatke
head(parenthood, 7) # Prvih 7 redova
```

```
##   dan.sleep baby.sleep dan.grump day
## 1      7.59      10.18        56   1
## 2      7.91      11.66        60   2
## 3      5.14       7.92        82   3
## 4      7.71       9.61        55   4
## 5      6.68       9.75        67   5
## 6      5.99       5.04        72   6
## 7      8.19      10.45        53   7
```

NOVI PODATKOVNI SKUP

```
# Pogledaj deskriptivnu statistiku
describe(parenthood)
```

```
##                vars   n  mean    sd median trimmed   mad    r
## dan.sleep         1 100  6.97  1.02   7.03    7.00  1.09  4.
## baby.sleep        2 100  8.05  2.07   7.95    8.05  2.33  3.
## dan.grump         3 100 63.71 10.05  62.00   63.16  9.64 41.
## day               4 100 50.50 29.01  50.50   50.50 37.06  1.
##                skew kurtosis   se
## dan.sleep  -0.29    -0.72 0.10
## baby.sleep -0.02    -0.69 0.21
## dan.grump   0.43    -0.16 1.00
## day         0.00    -1.24 2.90
```

NOVI PODATKOVNI SKUP

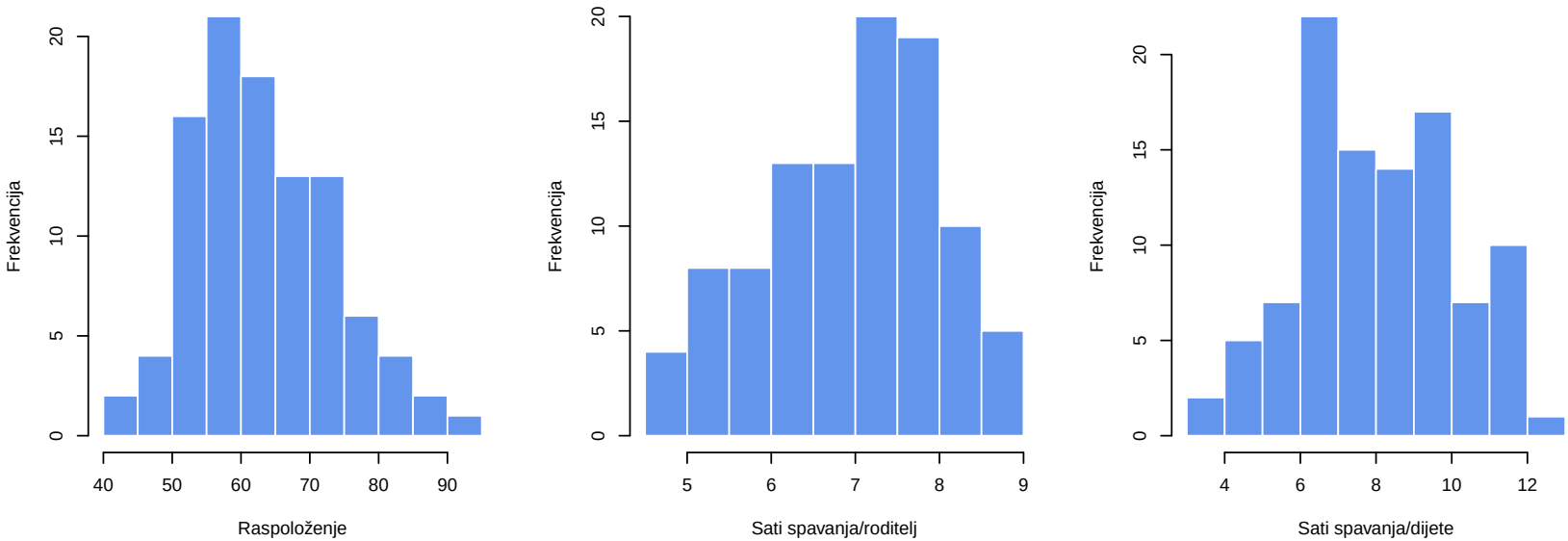

Figure 4: Grafički prikaz varijabli u `parenthood` podatkovnom skupu.

KORELACIJA I

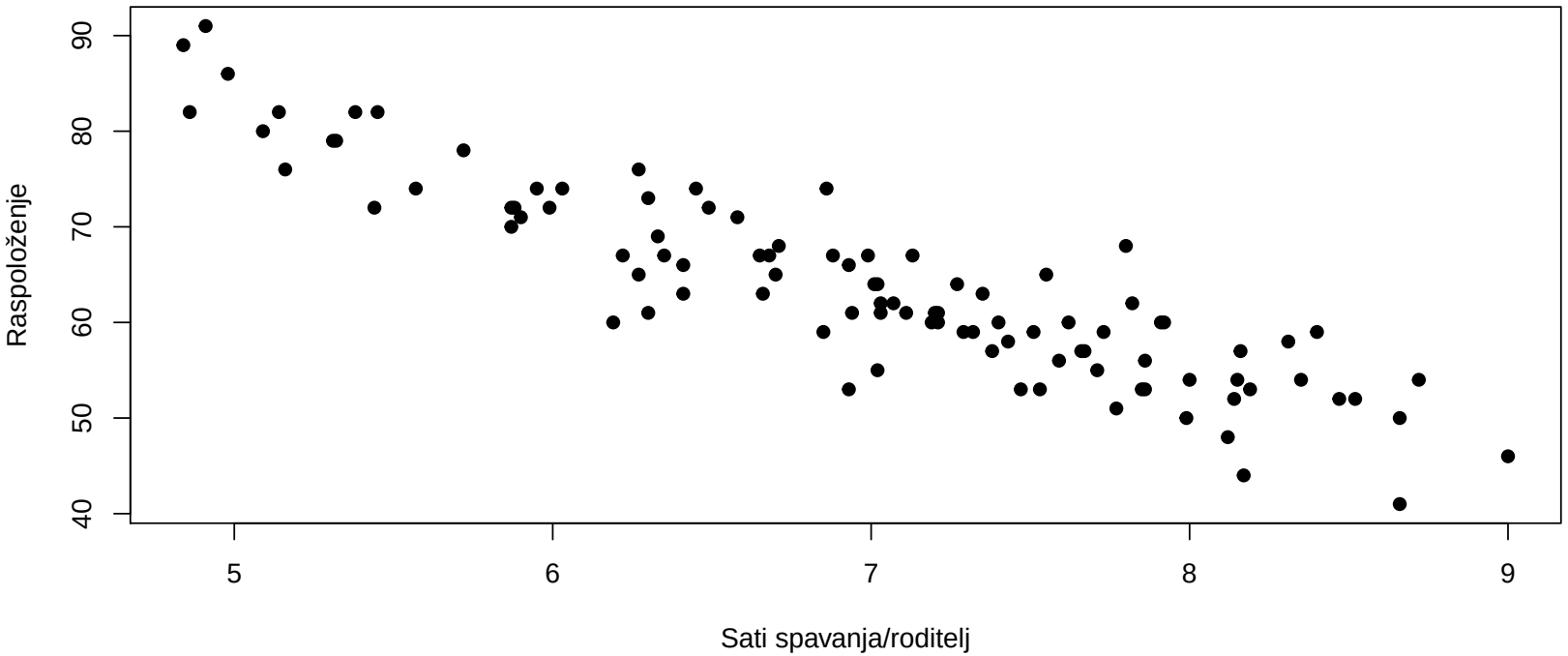

Figure 5: Dijagram rasipanja za varijable `Sati spavanja/roditelj` i `Raspoloženje`.

KORELACIJA II

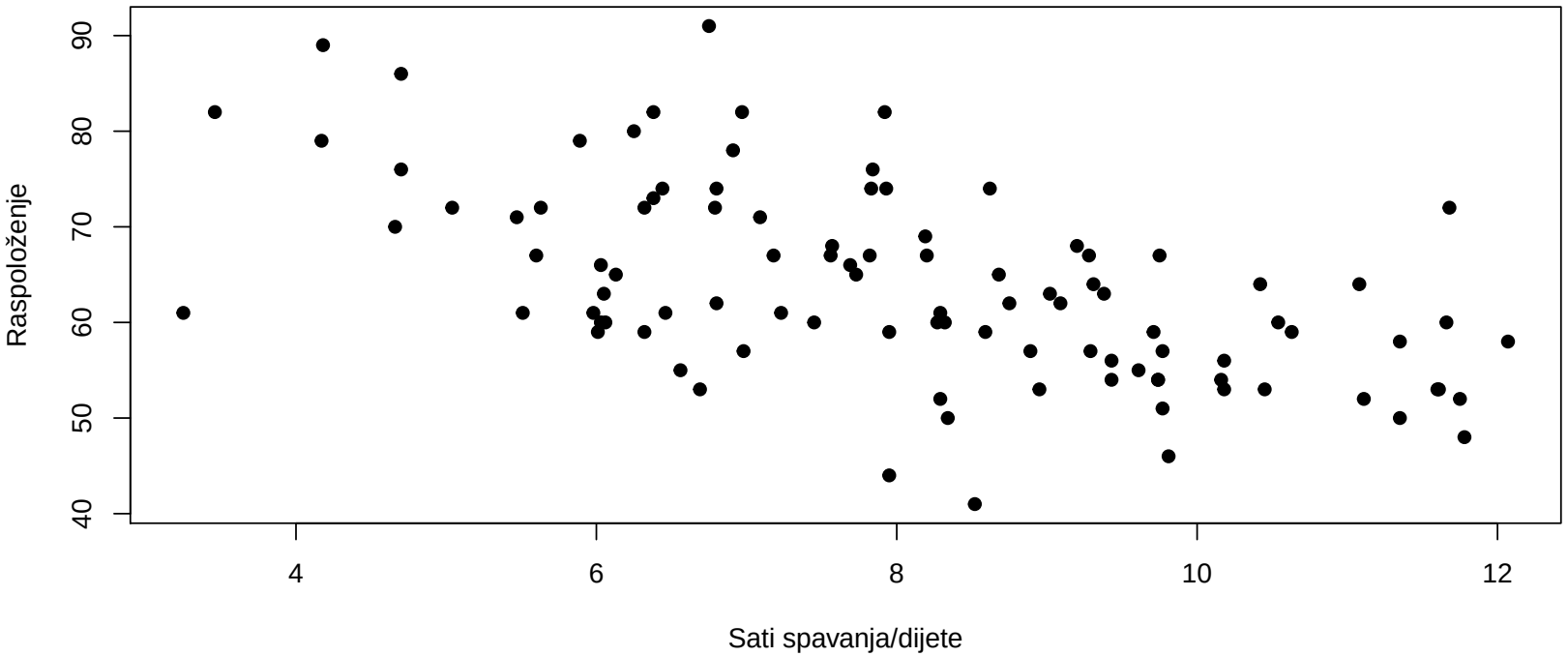

Figure 6: Dijagram rasipanja za varijable `Sati spavanja/dijete` i `Raspoloženje`.

KORELACIJA III

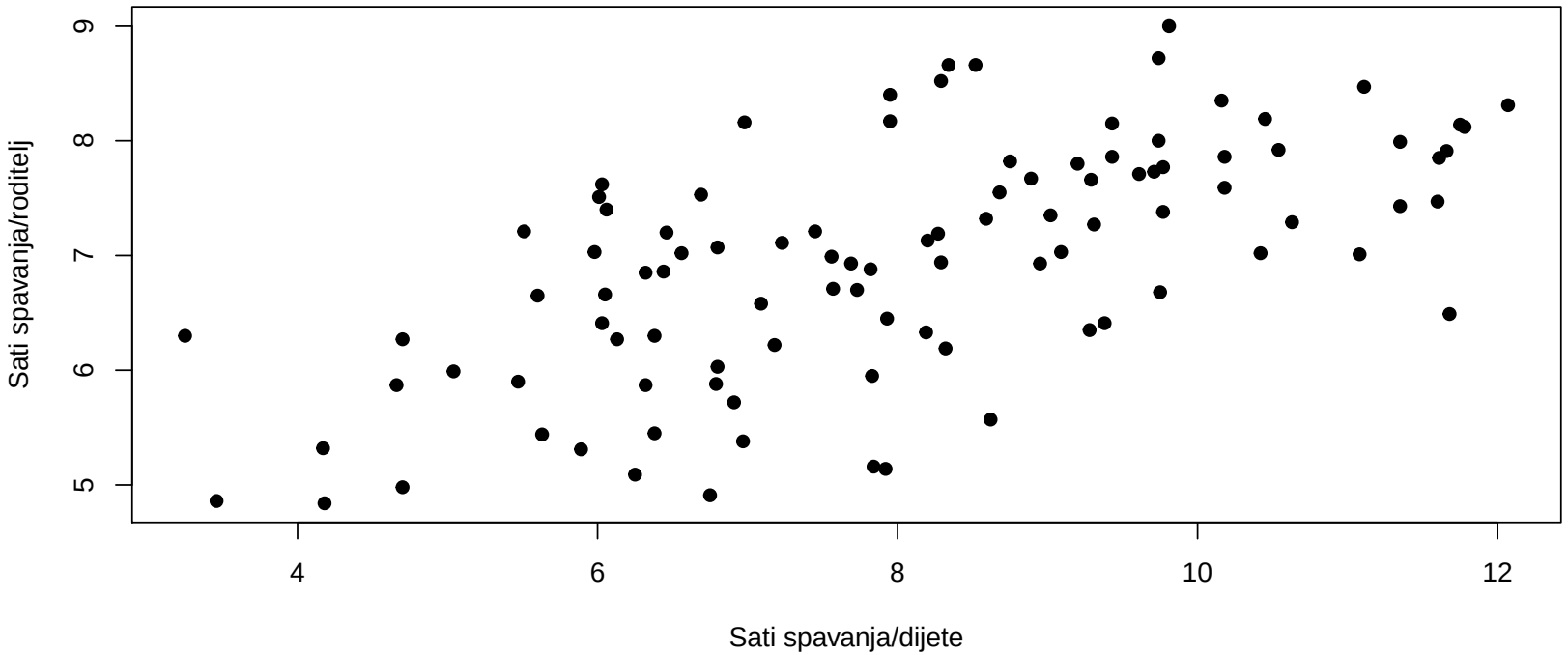

Figure 7: Dijagram rasipanja za varijable `Sati spavanja/dijete` i `Sati spavanja/roditelj`.

KORELACIJA IV

- Kovarijanca

$$\mathrm{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right)$$

- Personov korelacijski koeficijent(standardizacija kovarijance)

$$r_{XY} = \frac{\mathrm{Cov}(X, Y)}{\hat{\sigma}_X \ \hat{\sigma}_Y}$$

IZRAČUN KORELACIJE U R

```
# Izračunaj korelaciju između spavanja i raspoloženja
cor(x = parenthood$dan.sleep, y = parenthood$dan.grump)
```

```
## [1] -0.903384
```

```
# Izračunaj korelacijsku tablicu
cor(x = parenthood)
```

```
##                dan.sleep  baby.sleep   dan.grump           c
## dan.sleep    1.00000000  0.62794934 -0.90338404 -0.098407
## baby.sleep   0.62794934  1.00000000 -0.56596373 -0.010433
## dan.grump   -0.90338404 -0.56596373  1.00000000  0.076479
## day         -0.09840768 -0.01043394  0.07647926  1.000000
```

INTERPRETACIJA KORELACIJE

Table 3: Okvirne smjernice za interpretaciju korelacije.

Korelacija	Snaga	Smjer
-1.0 to -0.9	Izrazito jaka	Negativna
-0.9 to -0.7	Jaka	Negativna
-0.7 to -0.4	Umjerena	Negativna
-0.4 to -0.2	Slaba	Negativna
-0.2 to 0	Zanemariva	Negativna
0 to 0.2	Zanemariva	Pozitivna
0.2 to 0.4	Slaba	Pozitivna
0.4 to 0.7	Umjerena	Pozitivna
0.7 to 0.9	Jaka	Pozitivna
0.9 to 1.0	Izrazito jaka	Pozitivna

NOVI PODATKOVNI SKUP

```
rm(list=ls()) # Očisti radni prostor
load("effort.Rdata") # Učitaj podatke
who(TRUE) # Pregledaj podatke
```

```
##     -- Name --    -- Class --    -- Size --
##     effort        data.frame     10 x 2
##      $hours       numeric        10
##      $grade       numeric        10
```

```
head(effort, 3) #Pregledaj podatke
```

```
##   hours grade
## 1     2    13
## 2    76    91
## 3    40    79
```

NOVI PODATKOVNI SKUP

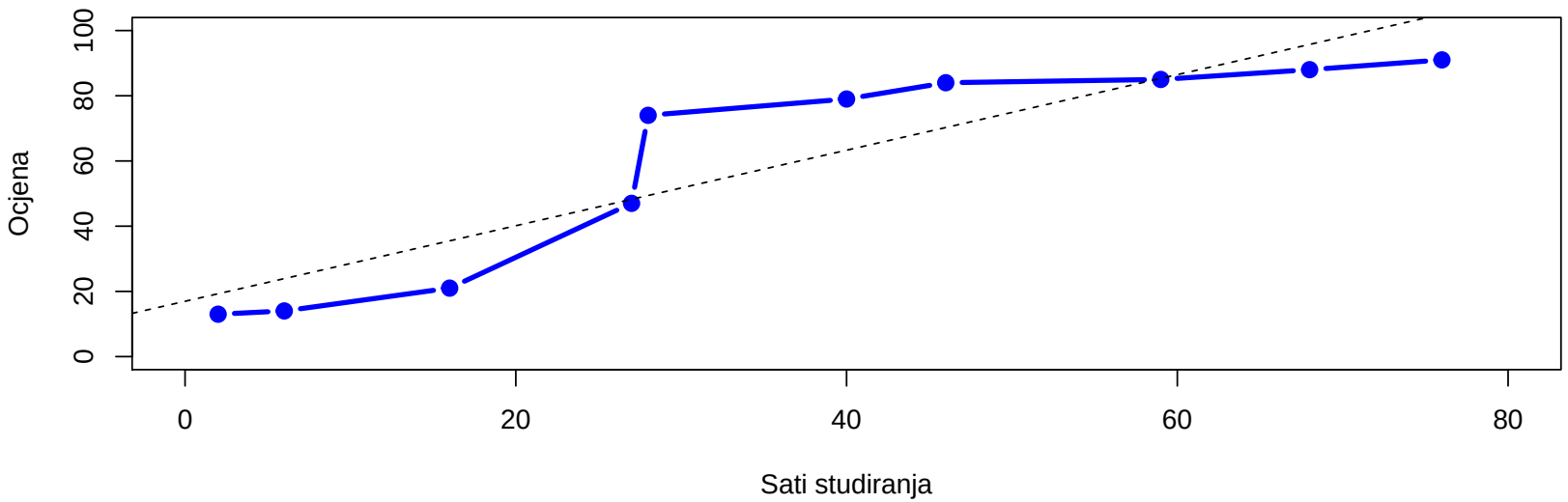

Figure 8: Odnos između sati studiranja i ocjene (svaka točka predstavlja jednog studenta). Isprekidana linija prikazuje linearni odnos. Korelacija između ove dvije varijable je visoka $r = .91$. Valja primjetiti da više sati učenja uvijek dodnosi veću ocjenu što se odražava u visokom Spearman koeficijentu korelacije of $rho = 1$.

SPEARMANOVA KORELACIJA I

```
sati_studiranja <- rank( effort$hours ) # Rang sati
ocjena <- rank( effort$grade ) # Rang ocjena
```

	Rang sati rada	Rang visine ocjene
student	1	1
student	2	10
student	3	6
student	4	2
student	5	3
student	6	5
student	7	4
student	8	8
student	9	7
student	10	9

SPEARMANOVA KORELACIJA II

```
cor(sati_studiranja,ocjena) # Izračunaj korelaciju
```

```
## [1] 1
```

```
# Dodaj argument "spearman"
cor(effort$hours, effort$grade, method = "spearman")
```

```
## [1] 1
```