

Testiranje razdiobe bioloških podataka u R-u

Lucija Kanjer, e-mail: lucija.kanjer@biol.pmf.hr

2024-11-11

Sadržaj praktikuma

- Uvod u rad u programskom okruženju R i osnovne funkcije, instaliranje programskih paketa
- Unos podataka u programsko okruženje R, struktura objekata
- Rad s objektima i podacima te definiranje bioloških varijabli u R-u
- ***Grafički prikaz bioloških podataka i testiranje razdiobe podataka u R-u***
- Primjeri osnovnih statističkih analiza kategoričkih i numeričkih varijabli u biološkim istraživanjima u R-u
- Regresije i korelacije, linearni modeli bioloških podataka – primjeri u R-u
- Primjena parametrijskih statističkih testova bioloških podataka u R-u
- Primjena neparametrijskih statističkih testova bioloških podataka u R-u
- Primjeri multivarijatnih analize bioloških podataka u R-u - linearni modeli, klaster analize i ordinacijske analize

Sadržaj ove vježbe

Grafički prikazi

- Faktori u R-u - organizacija kategoričkih varijabli
- izvoz (eksportiranje) grafa u dokument

Testiranje razdiobe bioloških varijabli

- histogram
- Q-Q plot
- Shapiro-Wilk test
- deskriptivna statistika
- Cullen and Fray graf

Otvorite skriptu!

Materijali za ovu vježbu se nalaze na GitHub repoziroriju **APBI_2024** u mapi **05_Testiranje razdiobe**.

https://github.com/lucijakanjer/APBI_2024

Skripta 05_Testiranje razdiobe.R"

```
# Instalacija novih paketa  
# install.packages() - popunite za pakete koje nemate, pazite navodnike!
```

```
# Učitavanje potrebnih paketa  
library(ggplot2) # za crtanje grafova  
library(fitdistrplus) # za fit-atanje ditribucije
```

```
## Loading required package: MASS
```

```
## Loading required package: survival
```

```
library(patchwork) # za prikaz više grafova
```

```
##  
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:MASS':
```

```
##
```

```
##      area
```

Postavite radni direktorij

```
# Postavljanje radnog direktorija  
getwd()
```

```
## [1] "C:/Users/Hrvoje/Documents/APUBI/05_Testiranje_razdiobe"
```

Primjer:

```
setwd("C:/Users/Računalo/Documents/ime_prezime/vjezba_5")
```

Ribe - novi set podataka

```
# Učitavanje seta podataka o ribama u jezerima grada Zagreba  
ribe <- read.csv("ribe.csv", header = TRUE)
```

```
# Pregledajte set podataka!  
str(ribe)
```

```
## 'data.frame':   30 obs. of  9 variables:  
##  $ duljina_cm      : num  30.5 28.1 28.7 37.9 36.8 ...  
##  $ velicina_ribe   : chr   "subadultna" "subadultna" "subadultna" "adultna" ...  
##  $ masa_g          : num  65.8 53.1 35.7 73.1 73.1 ...  
##  $ brzina_plivanja : num  2.95 2.13 2.81 2.18 2.41 1.24 2.32 2.04 1.74 2.1 ...  
##  $ starost_riba    : num  6.46 51.61 8.52 3.27 10.35 ...  
##  $ broj_jaja        : int   129 96 87 105 96 108 109 117 83 98 ...  
##  $ broj_parazita    : int    6 4 4 4 6 7 2 5 3 5 ...  
##  $ temperatura_vode: num   11.9 12.8 18.1 15.9 15.9 ...  
##  $ jezero           : chr    "Bundek" "Bundek" "Maksimir" "Maksimir" ...
```

Ponovimo tipove varijabli

- **numeric** = numerička kontinuirana varijabla
- **chraracter** = tekstualna (kategorička) varijabla
- **integer** = numerička diskretna (cjelobrojna) varijabla
- **factor** = ???

Uvijek pogledajmo kako izgleda tablica!

```
print(ribe)
```

##	duljina_cm	velicina_ribe	masa_g	brzina_plivanja	starost_riba	broj_jaja
## 1	30.46	subadultna	65.76	2.95	6.46	129
## 2	28.13	subadultna	53.12	2.13	51.61	96
## 3	28.71	subadultna	35.74	2.81	8.52	87
## 4	37.86	adultna	73.11	2.18	3.27	105
## 5	36.78	adultna	73.11	2.41	10.35	96
## 6	28.66	subadultna	33.10	1.24	0.20	108
## 7	33.02	subadultna	47.96	2.32	1.61	109
## 8	26.06	subadultna	96.88	2.04	2.79	117
## 9	25.92	subadultna	65.09	1.74	26.00	83
## 10	33.64	subadultna	50.12	2.10	1.91	98
## 11	29.32	subadultna	117.74	1.75	15.24	96
## 12	29.02	subadultna	38.22	2.03	7.10	99
## 13	26.43	subadultna	94.10	2.98	5.67	96
## 14	23.72	juvenilna	62.16	2.31	16.80	83
## 15	23.85	juvenilna	39.36	1.86	29.57	119
## 16	26.80	subadultna	44.38	1.98	6.00	92
## 17	25.46	subadultna	66.93	2.49	5.29	105
## 18	36.05	adultna	88.67	2.65	16.90	101
## 19	34.30	subadultna	30.60	2.25	10.89	123

Faktori u R-u - organizacija kategoričkih varijabli

U R-u, faktori su strukture podataka koje se koriste za predstavljanje **kategoričkih podataka**, kao što su spol, razine obrazovanja ili bilo koja kvalitativna karakteristika s ograničenim skupom vrijednosti (koji se nazivaju razinama).

Faktori su posebna vrsta vektora koji pohranjuju ove kategoričke vrijednosti kao cjelobrojne kodove (*integers*), ali prikazuju razine kao nizove, što ih čini korisnim za statističku obradu i vizualizaciju podataka.

Ključne karakteristike faktora

- Razine (*levels*): Ovo su jedinstvene vrijednosti koje faktor može poprimiti. Na primjer, faktor "Spol" može imati razine "Muški" i "Ženski".
- Redoslijed: Faktori mogu biti neuređeni ili uređeni (*ordered*). Poredani faktori održavaju poredak ili hijerarhiju među razinama, korisni za redne podatke kao što su "Nisko", "Srednje", "Visoko".
- Interno pohranjeni kao cijeli brojevi (*integers*): Svaka razina se interno preslikava na cjelobrojnu vrijednost, što faktore čini memorijski učinkovitima.

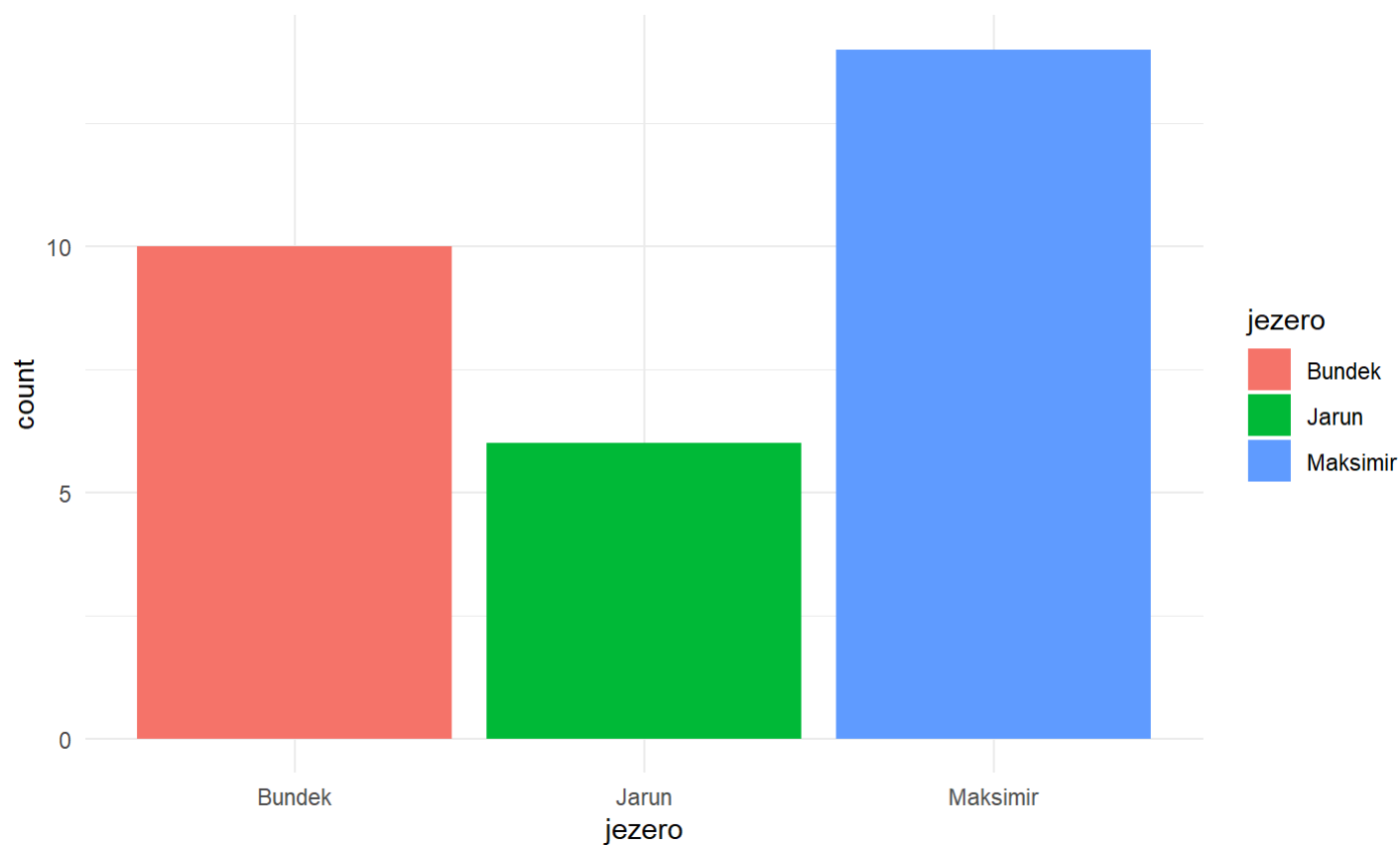
Koje kategoričke varijable nalazimo u setu podataka o ribama?

Koja varijabla je nominalna, a koja ordinalna?

Nominalna kategorička varijabla

```
### Nominalna kategorička varijabla - broj riba po jezerima
```

```
ggplot(ribe, aes(x = jezero)) + geom_bar(aes(fill = jezero)) + theme_minimal()
```



Abecedni redoslijed - zadani prikaz u R-u

```
# Koji je problem? R crta kategorije abecednim redom!  
# Provjerite tip varijable jezero!  
class(ribe$jezero)
```

```
## [1] "character"
```

```
str(ribe$jezero)
```

```
## chr [1:30] "Bundek" "Bundek" "Maksimir" "Maksimir" "Maksimir" "Bundek" ...
```

Narebda factor()

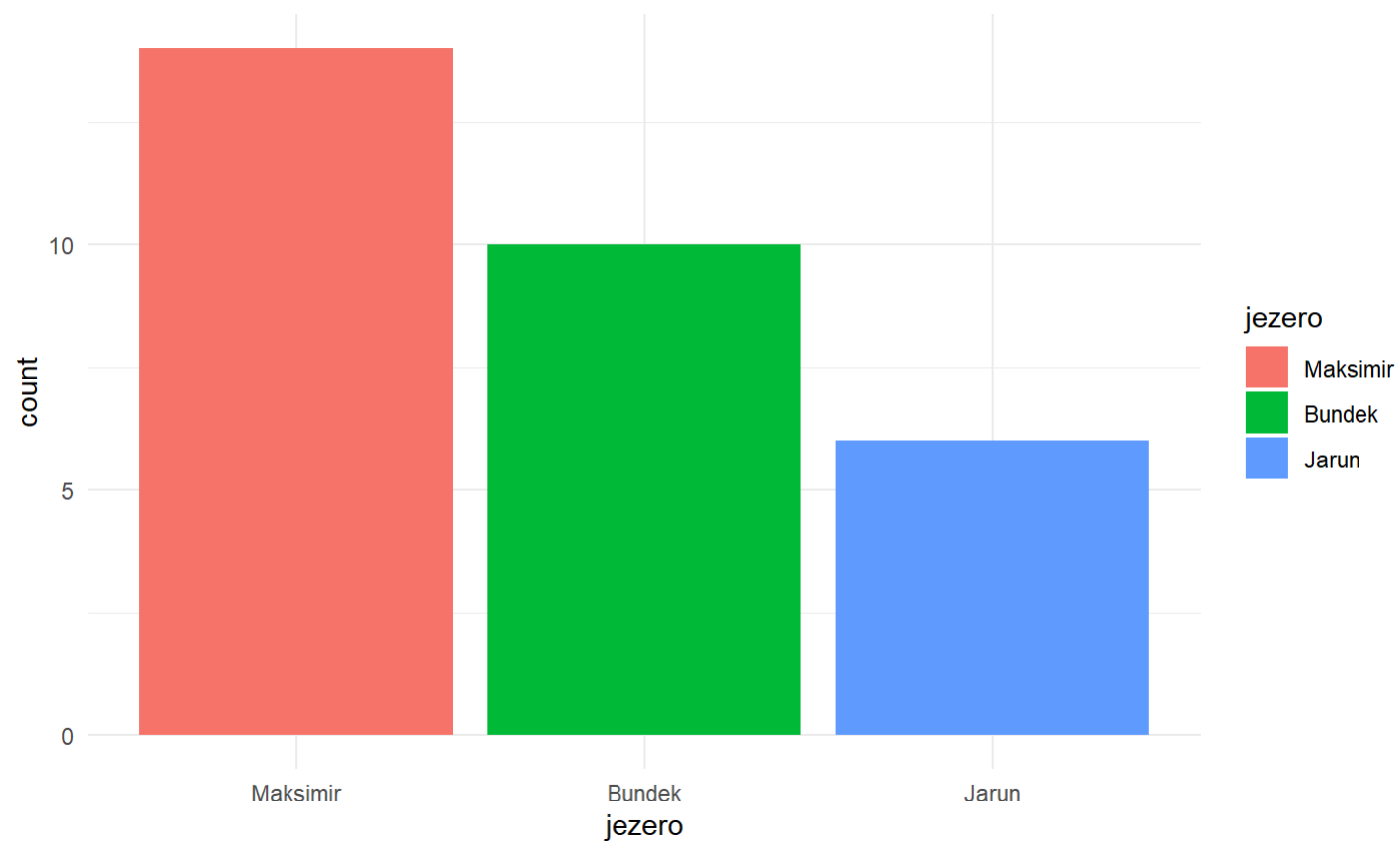
```
# Koristimo naredbu factor() da postavimo padajući poredak varijable jezero  
ribe$jezero <-  
  factor(ribe$jezero,  
        levels = names(sort(table(ribe$jezero), decreasing = TRUE)))  
  
# Ponovo provjerite tip varijable i nacrtajte barplot! Što se promjenilo?  
class(ribe$jezero)
```

```
## [1] "factor"
```

```
str(ribe$jezero)
```

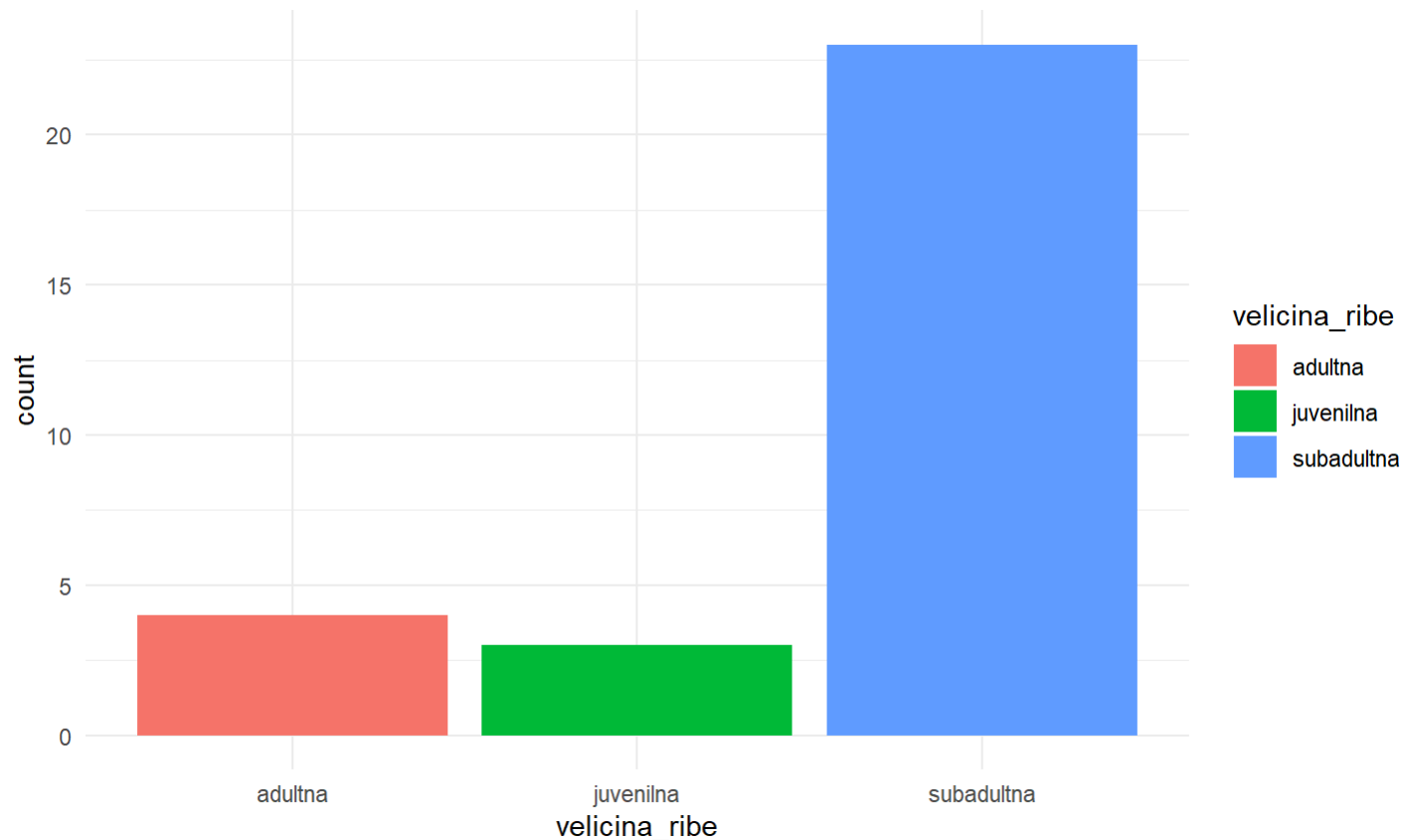
```
## Factor w/ 3 levels "Maksimir","Bundek",...: 2 2 1 1 1 2 1 3 1 3 ...
```

```
ggplot(ribe, aes(x = jezero)) + geom_bar(aes(fill = jezero)) + theme_minimal()
```



Ordinalna kategorička varijabla

```
### Ordinalna kategorička varijabla - broj riba po veličinskoj kategoriji  
graf_velicina <- ggplot(ribe, aes(x = velicina_ribe)) +  
  geom_bar(aes(fill = velicina_ribe)) + theme_minimal()  
print(graf_velicina)
```




```
# Provjerite tip varijable velicina_riba!  
class(riba$velicina_riba)
```

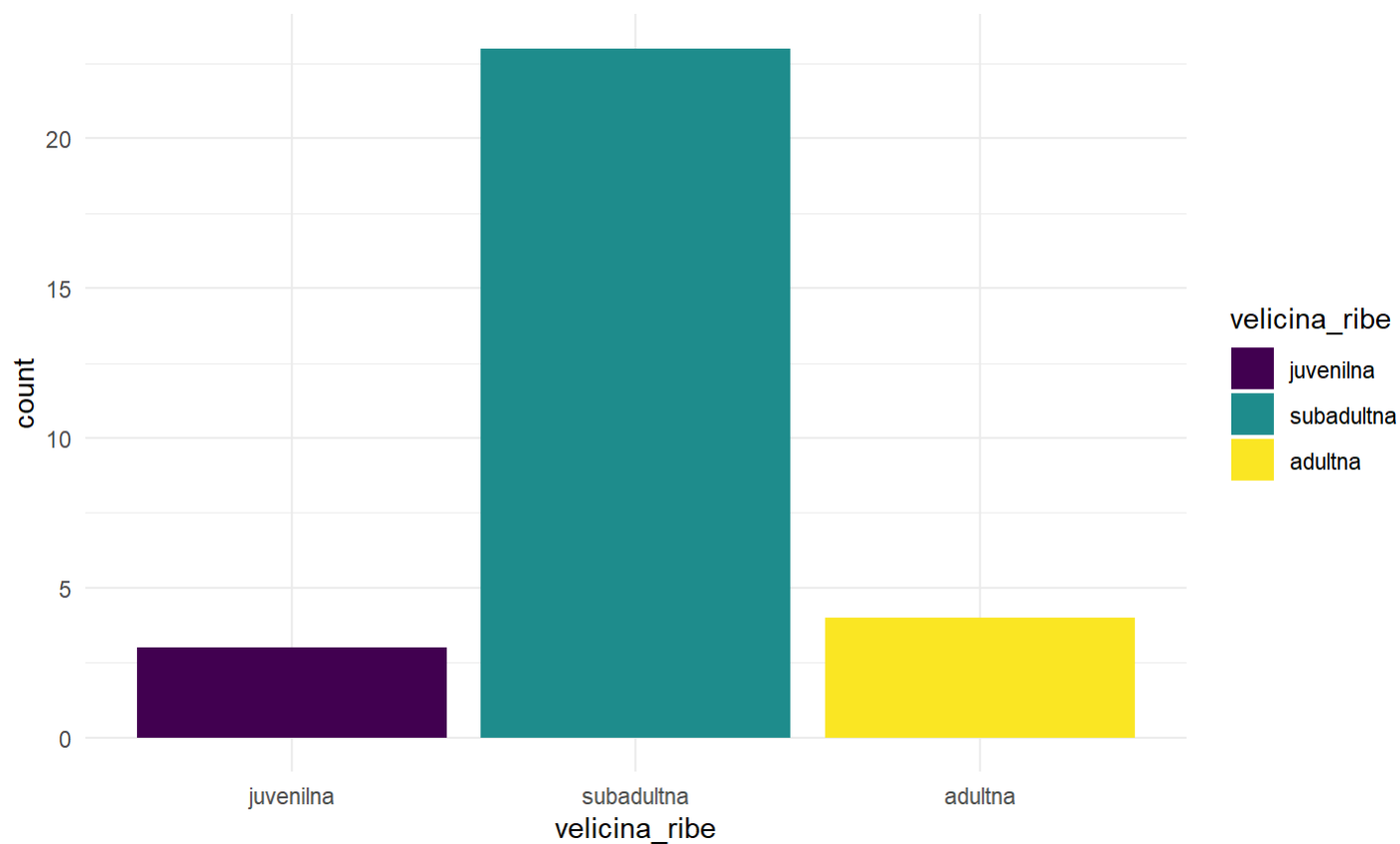
```
## [1] "character"
```

```
str(riba$velicina_riba)
```

```
## chr [1:30] "subadultna" "subadultna" "subadultna" "adultna" "adultna" ...
```

```
# Koristimo naredbu factor() da ručno uredimo poredak ove ordinalne kategoričke varijable  
riba$velicina_riba <-  
  factor(riba$velicina_riba,  
        levels = c("juvenilna", "subadultna", "adultna"), ordered = TRUE)
```

```
# Ponovo provjerite tip varijable i nacrtajte barplot! Što se promjenilo?  
graf_velicina <- ggplot(ribe, aes(x = velicina_ribe)) +  
  geom_bar(aes(fill = velicina_ribe)) + theme_minimal()  
print(graf_velicina)
```



Eksport grafa iz ggplot-a

Eksportiranje ggplot grafa kao slike ili PDF dokumenta

```
ggsave(filename = "graf_velicina.jpg", # naziv JPG slike
        plot = graf_velicina, # koji objekt želimo eksportirati
        width = 8, height = 6, # dimenzije u inčima
        dpi = 300) # dots per inch
```

```
ggsave(filename = "graf_velicina.pdf", # naziv JPG slike
        width = 8, height = 6, # dimenzije u inčima
        device = cairo_pdf) # naziv metode eksporta za PDF
```

Je li ova veličina grafa dobra za A4 dokument?

Izmjenite dimenzije tako da se font i podaci jasno vide!

Ispitivanje normalnosti distribucije

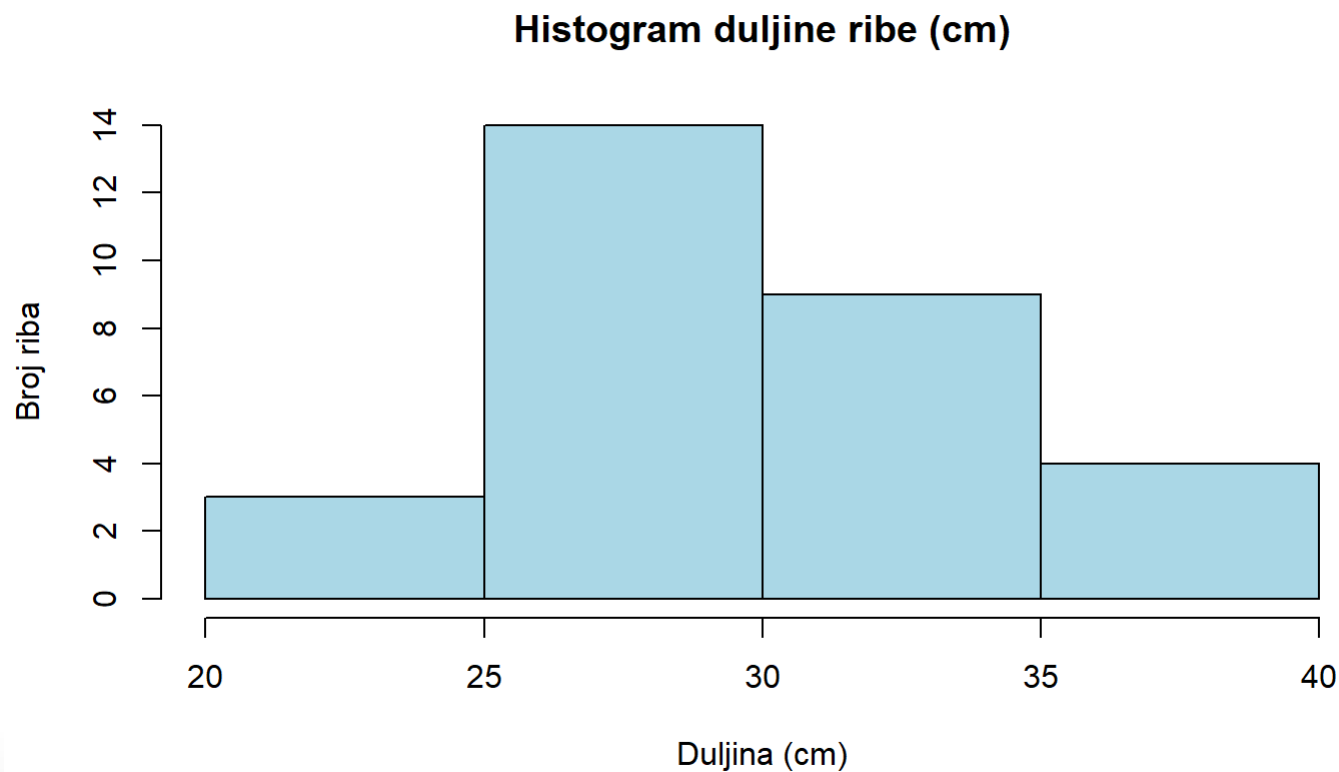
1. Histogram
2. Q-Q plot
3. Shapiro-Wilk test
4. Cullen-Fraser plot

1. Histogram

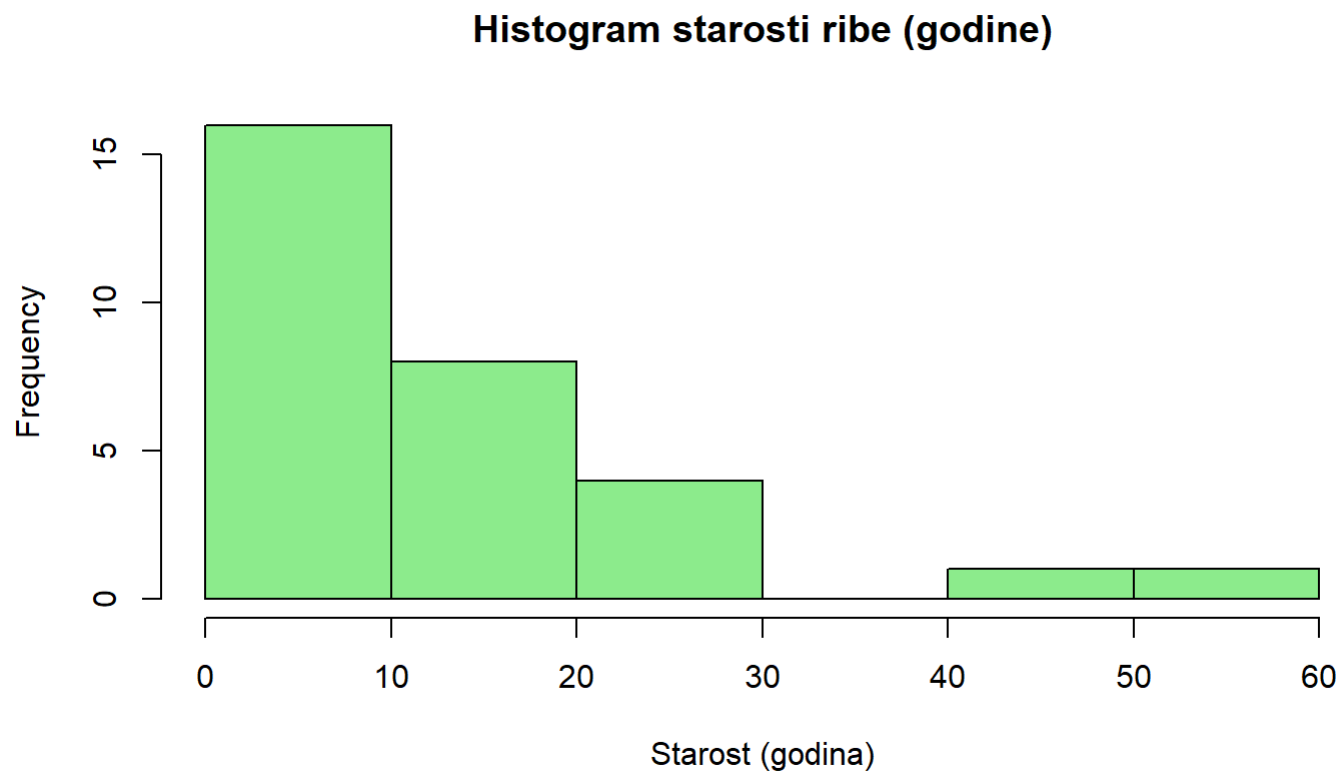
```
# Ispitivanje normalnosti distribucije za duljinu riba
```

```
# 1. Histogram duljine riba
```

```
hist(ribe$duljina_cm, main = "Histogram duljine ribe (cm)",  
     xlab = "Duljina (cm)", ylab = "Broj riba", col = "lightblue", border = "black")
```



```
# Histogram starosti riba  
hist(ribe$starost_riba, main = "Histogram starosti ribe (godine)",  
     xlab = "Starost (godina)", col = "lightgreen", border = "black")
```



Koju razliku primjećijete između dva histograma?

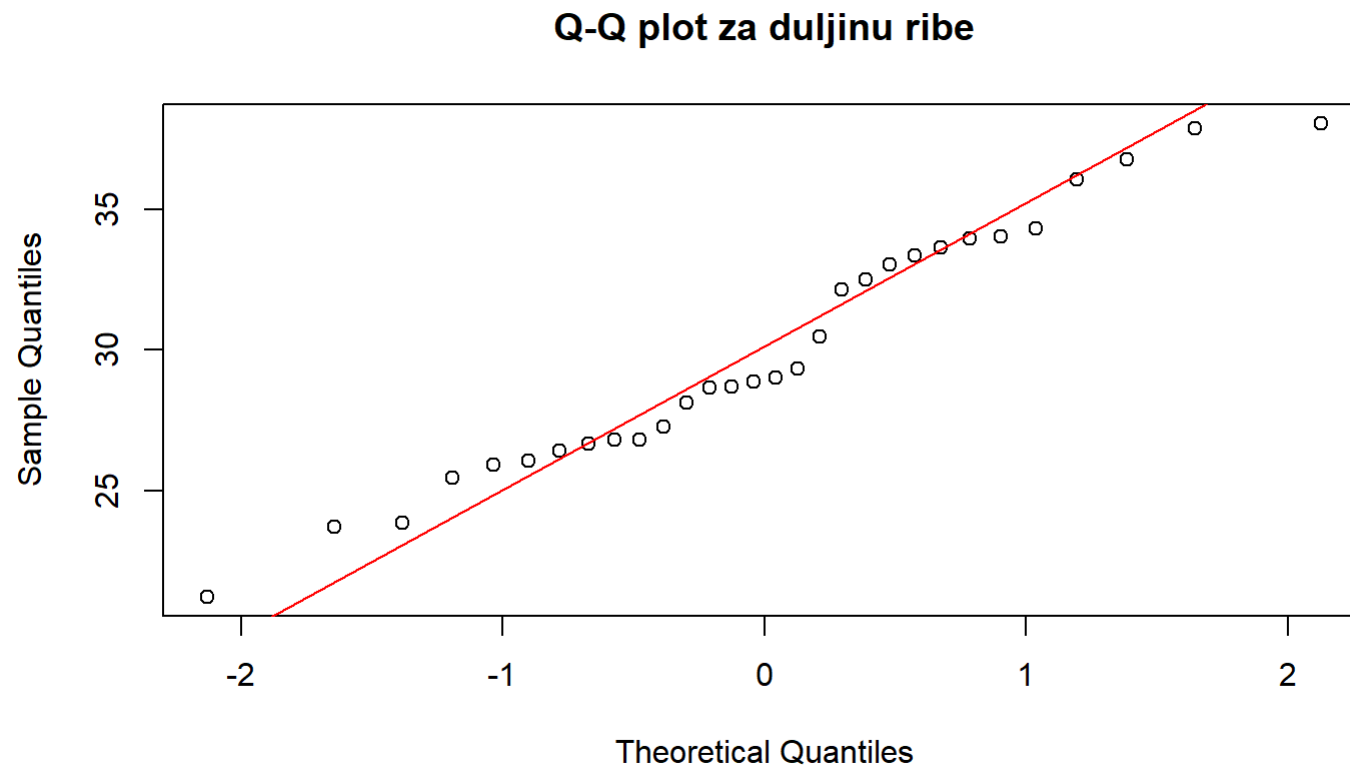
2. Q-Q plot

- **Q-Q plot (Quantile-Quantile plot)** je grafički alat koji pomaže u procjeni da li skup podataka slijedi određenu distribuciju, najčešće normalnu.
- Na grafu su naneseni kvantili podataka iz uzorka (na y-osi) u odnosu na kvantile teoretske distribucije (na x-osi) koju želimo provjeriti (u ovom slučaju, normalnu distribuciju).
- Ako podaci prate normalnu distribuciju, točke na Q-Q grafu bi trebale slijediti približno ravnu liniju.

2. Q-Q plot

```
qqnorm(ribe$duljina_cm, main = "Q-Q plot za duljinu ribe")
```

```
qqline(ribe$duljina_cm, col = "red")
```



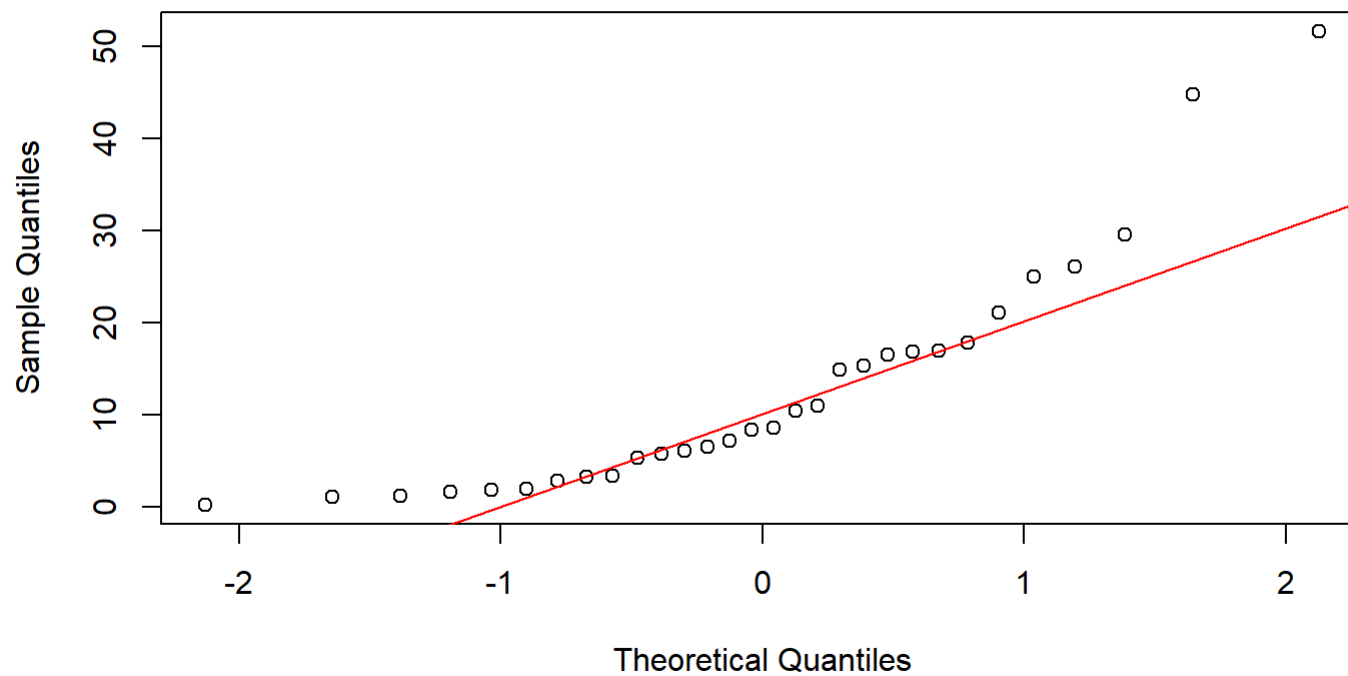
Interpretacija Q-Q plota za duljinu ribe

(duljina_cm)

- Ako je `duljina_cm` varijabla koja bi trebala biti normalno distribuirana, tada će točke na grafu pratiti ravnu liniju (naznačenu crvenom bojom) kroz čitav raspon podataka.
- Odstupanja od linije (posebno značajna odstupanja na početku ili kraju skale) ukazuju na to da podaci odstupaju od normalne distribucije.
- Na grafu možemo očekivati da će točke većinom pratiti liniju, uz manja odstupanja zbog slučajnih varijacija u uzorku.

```
qqnorm(ribe$starost_riba, main = "Q-Q plot za starost riba")  
qqline(ribe$starost_riba, col = "red")
```

Q-Q plot za starost riba



Koju razliku primjećujete u izgledima Q-Q plota?

Interpretacija Q-Q plota za starost ribe

(starost_riba)

- Varijabla `starost_riba` je eksponencijano distribuirana, što znači da njezini podaci neće imati normalnu distribuciju. Umjesto toga, podaci su pozitivno asimetrični – koncentrirani su bliže nuli i imaju dugu desnu stranu raspodjele.
- Na Q-Q plotu, eksponencijalno distribuirani podaci obično pokazuju značajna odstupanja od crvene linije (posebno u desnom kraju raspona), jer kvantili eksponencijalne distribucije ne odgovaraju kvantilima normalne distribucije.

3. Shapiro-Wilk test

Shapiro-Wilk test koristi se za procjenu normalnosti distribucije podataka. Ovaj test daje p-vrijednost, na temelju koje možemo donijeti zaključak o tome jesu li podaci iz uzorka normalno distribuirani.

Tumačenje rezultata Shapiro-Wilk testa

Hipoteze testa:

- Nulta hipoteza (H_0): Podaci su normalno distribuirani.
- Alternativna hipoteza (H_1): Podaci nisu normalno distribuirani.

p-vrijednost:

- Ako je p-vrijednost > 0.05 : Nema dovoljno dokaza da odbacimo nultu hipotezu. To znači da podaci mogu biti normalno distribuirani (prihvatamo pretpostavku normalnosti).
- Ako je p-vrijednost ≤ 0.05 : Postoje značajni dokazi protiv nulte hipoteze, što znači da podaci vjerojatno nisu normalno distribuirani (odbijamo pretpostavku normalnosti).

```
# 3. Shapiro-Wilk test
```

```
shapiro.test(ribe$duljina_cm)
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: ribe$duljina_cm
```

```
## W = 0.96505, p-value = 0.414
```

```
shapiro.test(ribe$starost_riba)
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: ribe$starost_riba
```

```
## W = 0.83155, p-value = 0.0002609
```

```
# Za koju varijablu je Shapiro-Wilk test značajan? Što to znači?
```

Shapiro-Wilk test je posebno koristan za manje uzorke, jer daje dobar uvid u normalnost podataka.

U ovom slučaju:

- duljina riba je normalno distribuirana varijabla jer je p-vrijednost = 0.414 što je veće od 0.05
- starost ribe nije normalno distribuirana varijabla jer je p-vrijednost = 0.0002609, što je manje od 0.05

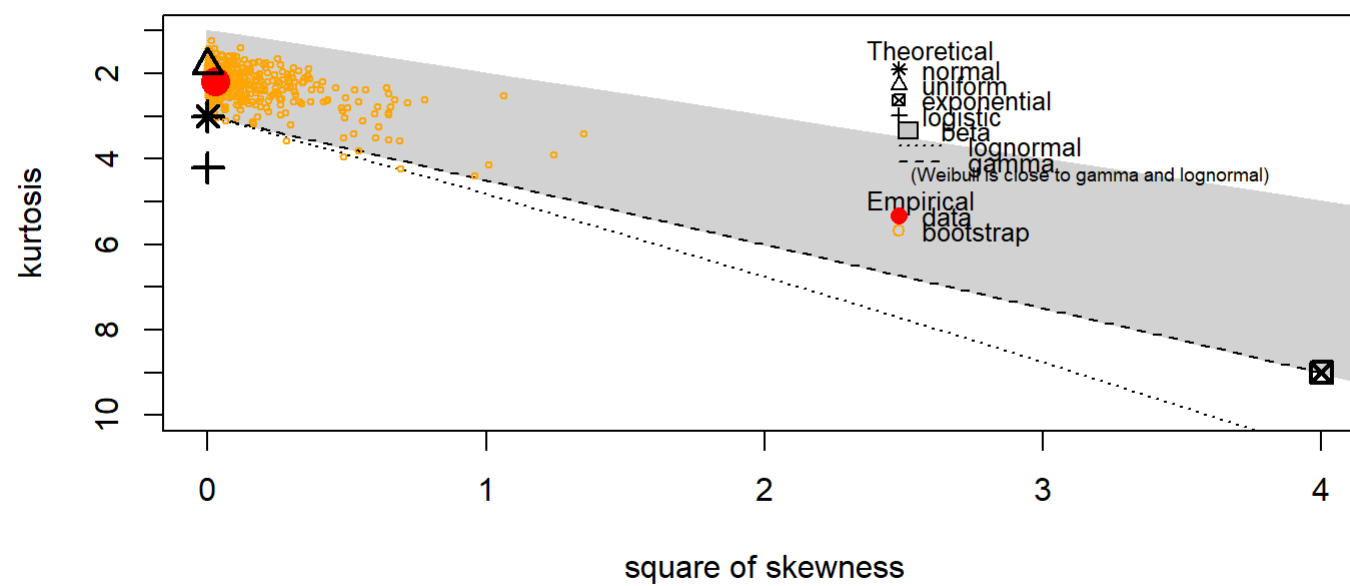
Napomena: Za veće uzorke, Shapiro-Wilk test može biti previše osjetljiv i pokazivati značajna odstupanja čak i kod blagih odstupanja od normalnosti.

4. Cullen and Fray graf

- Cullen-Fray graf je vizualni alat za procjenu distribucije podataka i koristi se u funkciji `descdist()` iz R paketa `fitdistrplus`.
- Ovaj graf prikazuje **kurtosis** (koncentriranost) i **skewness** (nagnutost) podataka kako bi nam pomogao odrediti koja distribucija najbolje odgovara skupu podataka.
- Analizom gdje podaci “padaju” na ovom grafu, možemo usporediti njihovu poziciju s teoretskim distribucijama.

```
descdist(ribe$duljina_cm, boot = 500)
```

Cullen and Frey graph



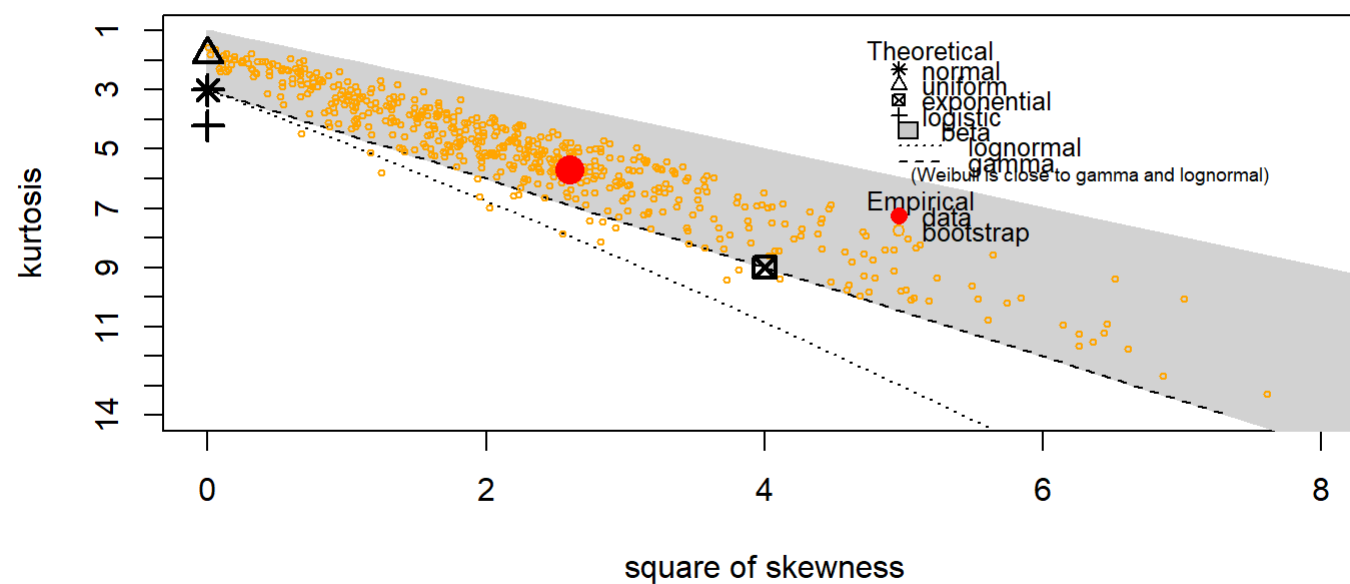
```
## summary statistics
## -----
## min: 21.24   max: 38.05
## median: 28.95
## mean: 29.96833
## estimated sd: 4.441661
## estimated skewness: 0.1611543
## estimated kurtosis: 2.190404
```


Interpretacija rezultata za varijablu `dujina_cm` (normalna distribucija)

- Ako su točke blizu područja normalne distribucije (skewness ≈ 0 i kurtosis ≈ 3), to potvrđuje da podaci `dujina_cm` odgovaraju normalnoj distribuciji.
- Ako postoji mala varijacija, očekujemo da će se točke nalaziti blizu normalne distribucije, s manjim odstupanjima zbog slučajnih varijacija.

```
descdist(ribe$starost_riba, boot = 500)
```

Cullen and Frey graph



```
## summary statistics
## -----
## min: 0.2    max: 51.61
## median: 8.435
## mean: 12.721
## estimated sd: 12.63815
## estimated skewness: 1.612995
## estimated kurtosis: 5.680622
```

Interpretacija rezultata za varijablu **starost_riba** (eksponencijalna distribucija)

- Visoka vrijednost za skewness i umjereno visoka za kurtosis ukazuju na distribuciju s jednim dominantnim repom, što je karakteristično za eksponencijalnu distribuciju..
- Ako su točke blizu područja eksponencijalne distribucije, to potvrđuje da podaci **starost_riba** imaju karakteristike te distribucije, što je u skladu s očekivanom dugom desnom stranom i koncentracijom podataka bliže nuli.

Eksportiranje grafova u base R-u

```
# Eksportiranje grafova u base R-u
# Otvorite uređaj za PNG format
png("graf.png", width = 2000, height = 1500, res = 300) # 'res' je parametar za dpi

# Nacrtajte graf
hist(ribe$duljina_cm, main = "Histogram duljine riba", xlab = "Duljina (cm)",
     col = "skyblue", border = "black")

# Zatvorite uređaj za izvođenje i spremite sliku
dev.off()
```

```
## png
## 2
```

Zadaci

- 1.a Eksportirajte napravljenje grafove, zalijepite ih u Word dokument.
- 1.b Objasnite svaki od korištenih grafova i testova.
- 2.a Napravite ispitivanje normalnosti za varijablu “brzina_plivanja”.
- 2.b Je li varijabla normalno distribuirana? Napravite dokument izvještaja.
- 3.a Napravite ispitivanje normalnosti za varijablu “broj_jaja”.
- 3.b Je li varijabla normalno distribuirana? Napravite dokument izvještaja.