Deskriptivna statistika

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

2024-12-03

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

- još malo o **izgledima distribucije** podataka
- mjere centralne tendencije: aritmetička sredina, medijan
- standardna pogreška (SE - *standard error*)
- mjere raspršenosti: standardna devijacija, interkvartilni raspon, raspon
- primjer samostalnog zadatka

Otvrorimo skriptu i učitajmo pakete

```
# Instalacija i učitavanje potrebnih paketa
# install.packages("") # nadopuni za nove pakete!

library(dplyr) # manipulacija tablicama
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2) # crtanje grafova
library(patchwork) # spajanje više grafova u jedan plot
library(data.table) # statistika po postajama
```



```
## Warning: package 'data.table' was built under R version 4.4.2
```

Provjera i postavljanje radnog direktorija

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

```
## [1] "C:/Users/Hrvoje/Documents/APUBI/06_Deskriptivna_statistika"
```

```
setwd("C:/Users/Hrvoje/Documents/APUBI/06_Deskriptivna_statistika")
```

Dataset rakovi

- Novi set podataka “rakovi.csv” sastoji se od mjerenja duljine, mase i broja patogena rakova na dvije postaje: istočnoj (“istok”) i zapadnoj (“zapad”).
- U ovoj vježbi cilj nam je opisati distribucije i odrediti deskriptivnu statistiku za svaku numeričku varijablu u cijelom setu podataka i između dvije istraživanje postaje.

Učitavanje seta podataka o rakovima u na postaji istok i zapad

```r
# Učitavanje seta podataka o rakovima u na postaji istok i zapad
rakovi <- read.csv("rakovi.csv", header = TRUE)

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

```
## 'data.frame':    200 obs. of  5 variables:
##  $ duljina    : num  21.4 19.4 20.4 20.6 20.4 ...
##  $ masa       : num  270.6 250.2 89.9 374 106.3 ...
##  $ temperatura: num  15.2 20.1 21.3 19.2 23.8 ...
##  $ patogeni   : num  7.39 9.02 0.2 3.11 12.39 ...
##  $ postaja    : chr  "zapad" "zapad" "zapad" "zapad" ...
```

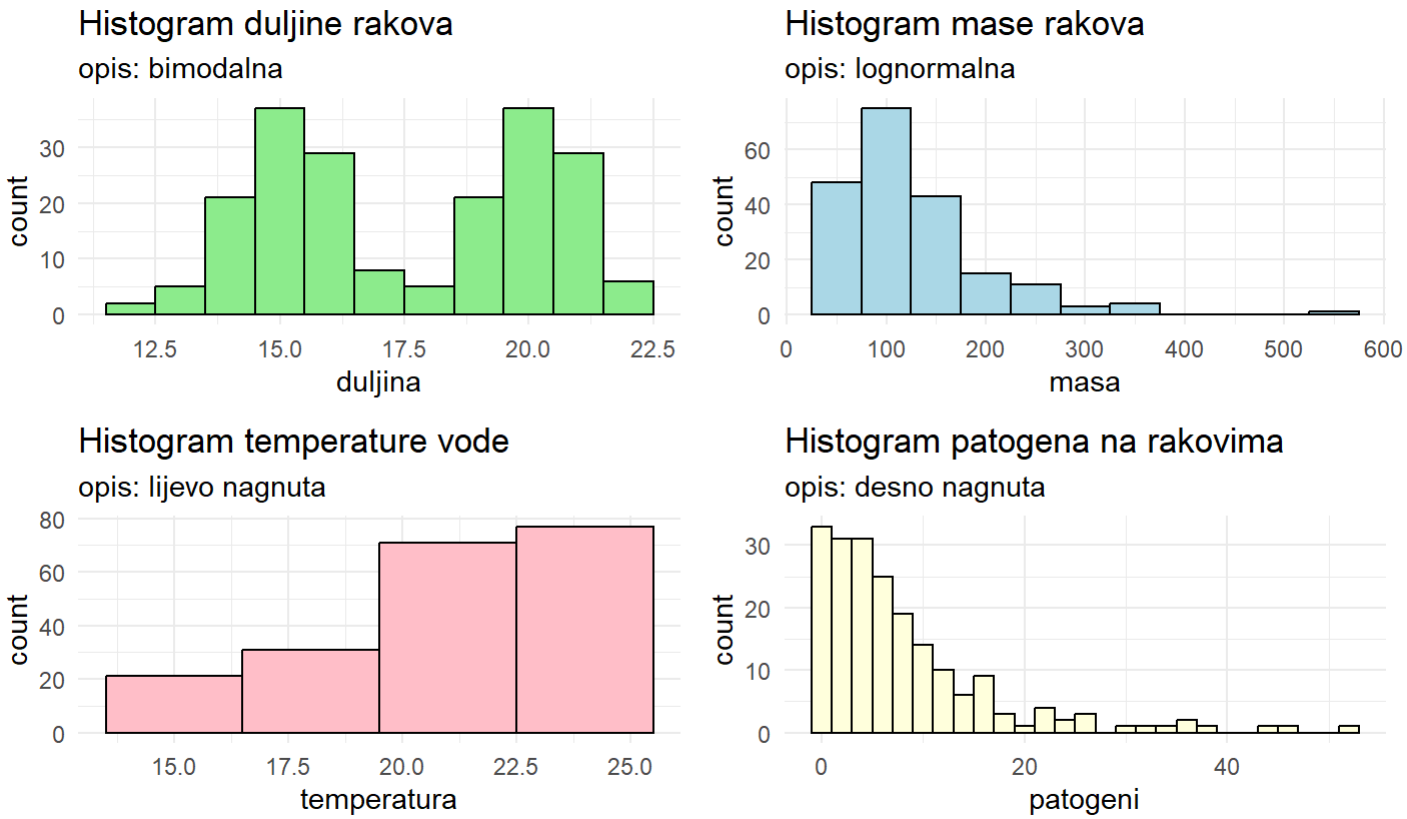
```r
# Pregledajte set podataka!
head(rakovi)
```

```
##   duljina   masa temperatura patogeni postaja
## 1   21.37 270.56       15.23     7.39   zapad
## 2   19.44 250.23       20.13     9.02   zapad
## 3   20.36  89.87       21.31     0.20   zapad
## 4   20.63 373.99       19.19     3.11   zapad
## 5   20.40 106.34       23.79    12.39   zapad
## 6   19.89 156.45       16.08     7.79   zapad
```

Izrada histograma

```
# Histogrami - pregled distribucije kontinuiranih numeričkih varijabli
# Prvo cemo kreirati objekte, a onda ih ispisati sve na jednom plotu

histogram_duljina <- ggplot(rakovi, aes(x = duljina)) +
  geom_histogram(binwidth = 1, color = "black", fill = "lightgreen") +
  labs(title = "Histogram duljine rakova", subtitle = "opis: bimodalna") +
  theme_minimal()

histogram_masa <- ggplot(rakovi, aes(x = masa)) +
  geom_histogram(binwidth = 50, color = "black", fill = "lightblue") +
  labs(title = "Histogram mase rakova", subtitle = "opis: lognormalna") +
  theme_minimal()

histogram_temperatura <- ggplot(rakovi, aes(x = temperatura)) +
  geom_histogram(binwidth = 3, color = "black", fill = "pink") +
  labs(title = "Histogram temperature vode", subtitle = "opis: lijevo nagnuta") +
  theme_minimal()

histogram_patogeni <- ggplot(rakovi, aes(x = patogeni)) +
  geom_histogram(binwidth = 2, color = "black", fill = "lightyellow") +
  labs(title = "Histogram patogena na rakovima", subtitle = "opis: desno nagnuta") +
  theme_minimal()
```

```
# Zadatak: Opisati izgled distrubucije - simetričnost, nagnutost - u subtitle!
# Spajanje 4 grafa u 1 pomoću paketa patchwork
(histogram_duljina + histogram_masa) / (histogram_temperatura + histogram_patogeni)
```



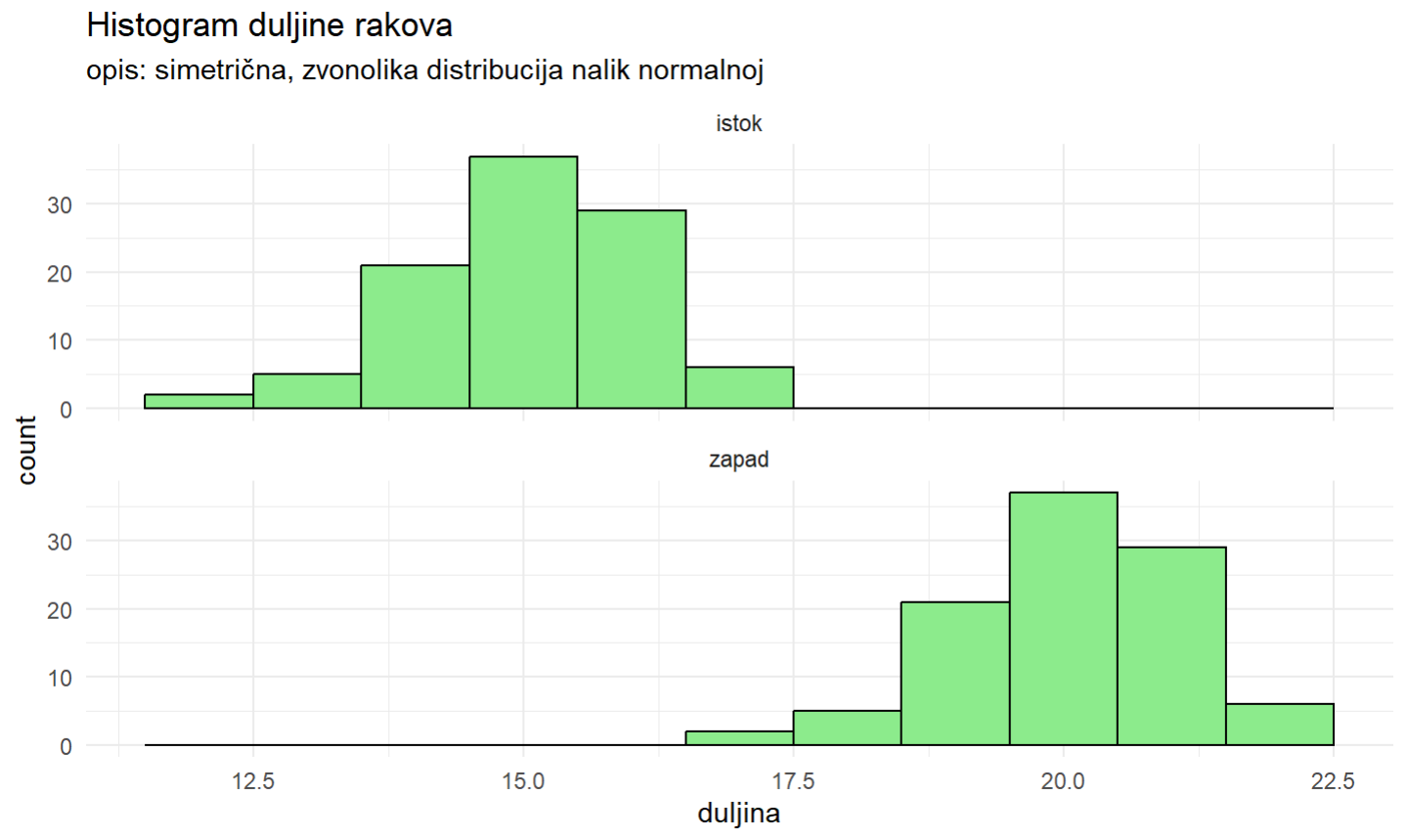
Eksport slike kao JPG

```
# Spremite ovaj graf kao JPEG sliku!
histogrami <- (histogram_duljina + histogram_masa) / (histogram_temperatura + histogram_patogeni)

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

Usporedba distribucije varijabli po postajama istok i zapad

Napravit ćemo histograme kao i ranije, ali ćemo ih odvojiti po grupirajućoj varijabli “postaja” i usporediti jesu li distribucije ostale iste ili su se izmjenile.

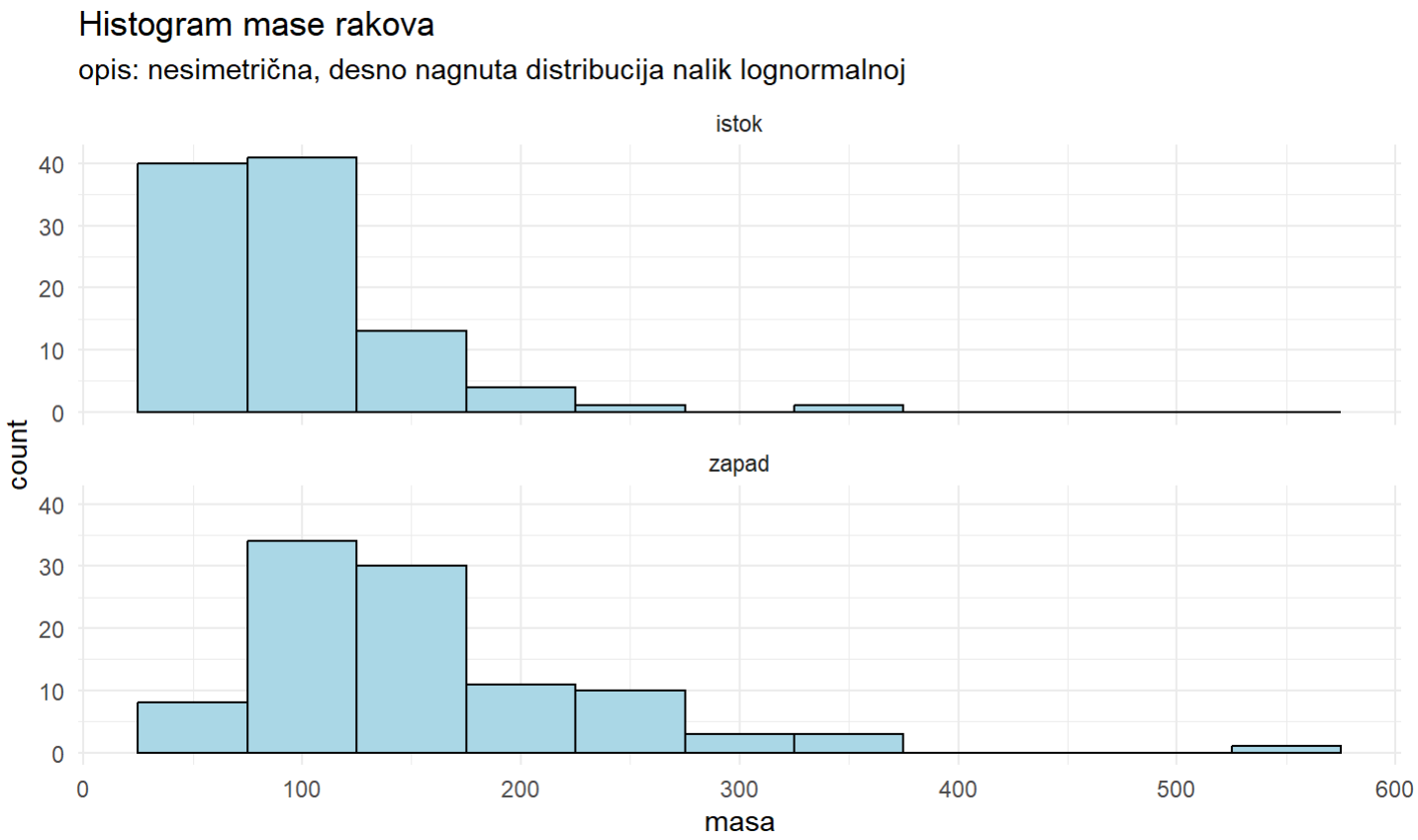
```r
# Crtamo isti histogram kao i ranije, ali koristimo facet_wrap za odvajanje po postajama
histogram_duljina_postaje <- ggplot(rakovi, aes(x = duljina)) +
  geom_histogram(binwidth = 1, color = "black", fill = "lightgreen") +
  labs(title = "Histogram duljine rakova", subtitle = "opis: simetrična, zvonolika distribucija nalik normalnoj") +
  facet_wrap(~ postaja, nrow = 2) + #odvaja histograme po grupirajucoj varijabli postaja
  theme_minimal()
print(histogram_duljina_postaje)
```



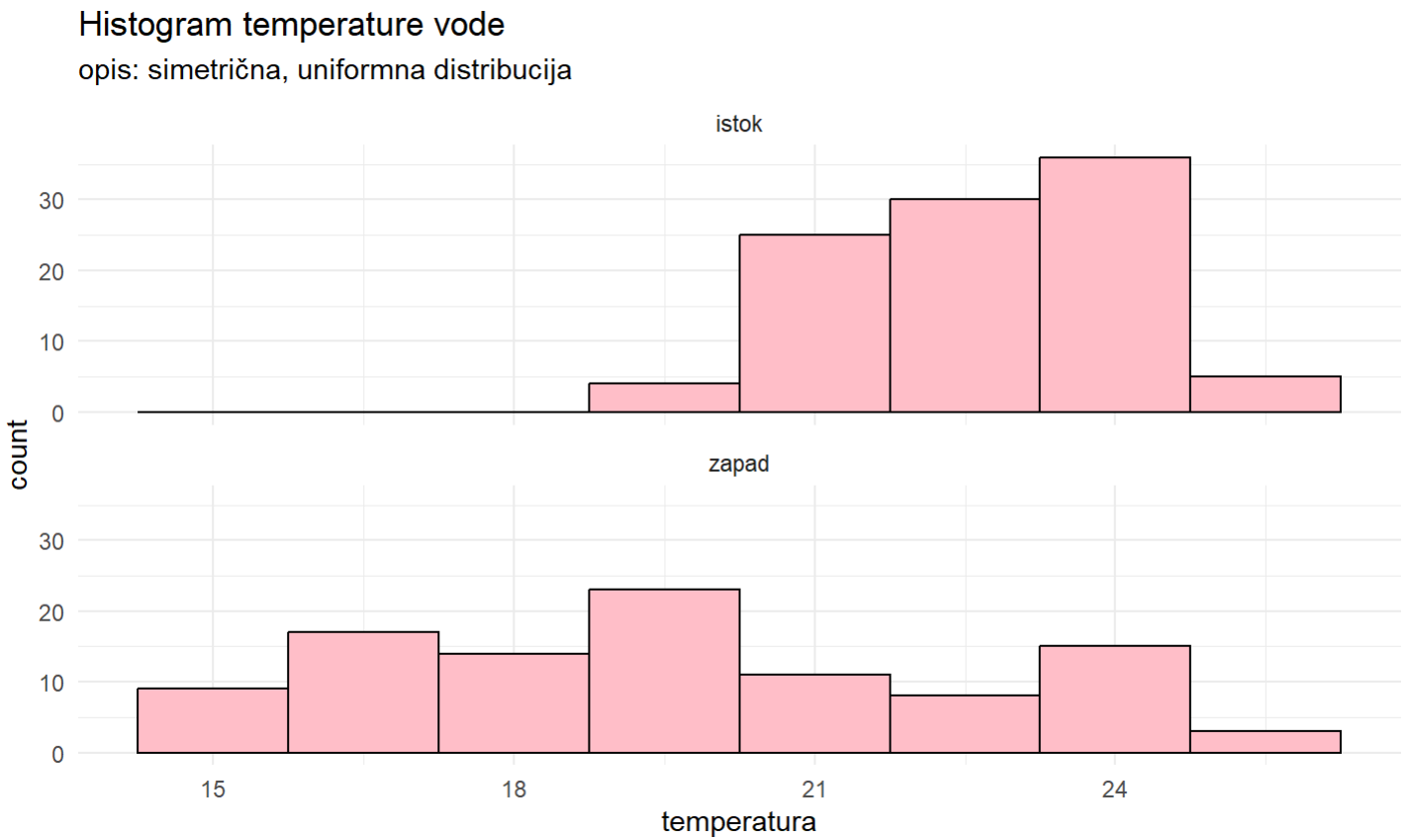
Kako sad izgleda distrubucija? Opišite u subtitle!

- Napravite histograme usporedbe po postajama i za varijable masa, temperatura i patogeni!
- Opišite nove izgled distribucije u subtitle!

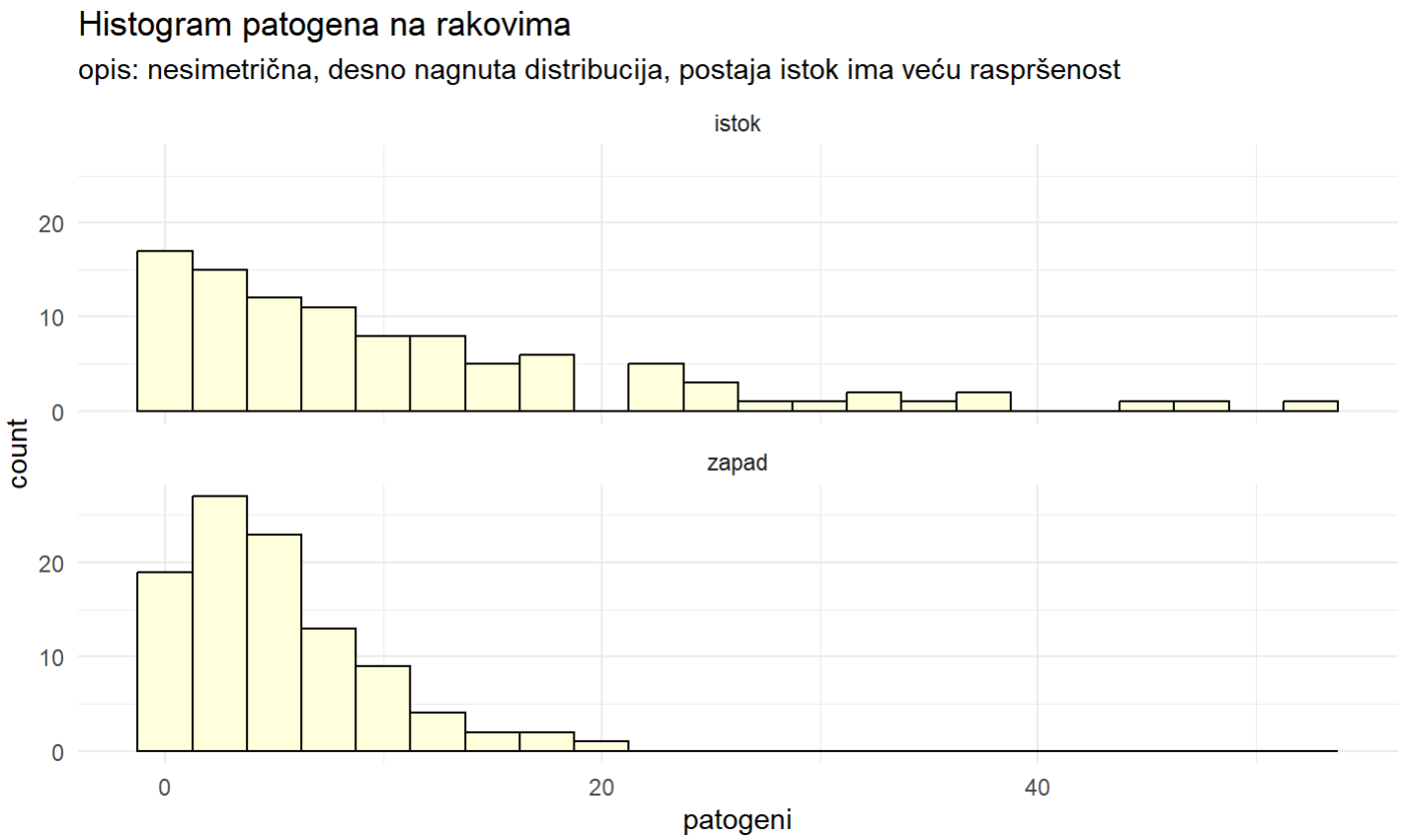
```
histogram_masa_postaje <- ggplot(rakovi, aes(x = masa)) +
  geom_histogram(binwidth = 50, color = "black", fill = "lightblue") +
  labs(title = "Histogram mase rakova", subtitle = "opis: nesimetrična, desno nagnuta distribucija nalik lognormalnoj") +
  facet_wrap(~ postaja, nrow = 2) +
  theme_minimal()
print(histogram_masa_postaje)
```



```r
histogram_temperatura_postaje <- ggplot(rakovi, aes(x = temperatura)) +
  geom_histogram(binwidth = 1.5, color = "black", fill = "pink") +
  labs(title = "Histogram temperature vode", subtitle = "opis: simetrična, uniformna distribucija") +
  facet_wrap(~ postaja, nrow = 2) +
  theme_minimal()
print(histogram_temperatura_postaje)
```



```
histogram_patogeni_postaje <- ggplot(rakovi, aes(x = patogeni)) +
  geom_histogram(binwidth = 2.5, color = "black", fill = "lightyellow") +
  labs(title = "Histogram patogena na rakovima", subtitle = "opis: nesimetrična, desno nagnuta distribucija, postaja istok ima ve
  facet_wrap(~ postaja, nrow = 2) +
  theme_minimal()
print(histogram_patogeni_postaje)
```



Deskriptina statistika

Mjere centralne tendencije

- Odgovaraju na pitanje: “Koja je tipična vrijednost u populaciji?”
- aritmetička sredina (“ravnotežna točka seta podataka”)
- medijan (vrijednost na sredini uzorka)
- mod (najčešća vrijednost - za kategoričke varijable)

Mjere raspršenosti

- Koliko su varijabline vrijednosti u populaciji?
- standardna devijacija (prikazuje se uz aritmetičku sredinu)
- interkvartilni raspon (Q1-Q3)
- raspon min-max

Deskriptina statistika u R-u

```
# Naredba summary() - pregled kvartila, medijana i aritmetičke sredine
summary(rakovi) # cijeli dataset
```

```
##     duljina           masa         temperatura       patogeni
##  Min.   :12.01   Min.   : 32.71   Min.   :15.01   Min.   : 0.000
##  1st Qu.:15.09   1st Qu.: 75.59   1st Qu.:19.43   1st Qu.: 1.930
##  Median :17.16   Median :111.91   Median :21.64   Median : 5.160
##  Mean   :17.53   Mean   :126.86   Mean   :21.12   Mean   : 8.077
##  3rd Qu.:20.09   3rd Qu.:157.18   3rd Qu.:23.45   3rd Qu.:10.405
##  Max.   :22.29   Max.   :573.03   Max.   :24.96   Max.   :52.380
##    postaja
##  Length:200
##  Class :character
##  Mode  :character
##
##
##
```

```
summary(rakovi$duljina) # samo jedna varijabla
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.01   15.09   17.16   17.53   20.09   22.29
```

Simetrične i normalne distribucije

Za podatke koje su normalno i simetrično distribuirani koristimo:

- **aritmetičku sredinu** - kao mjeru centralne tendencije
- **standardnu devijaciju** - kao mjeru raspršenosti podataka u našem uzorku
- **standardnu pogrešku** - kao mjeru koliko precizno srednja vrijednost našeg uzorka dobro procjenjuje srednju vrijednost prave populacije

Ovakve distribucije mogu dalje koristiti metode **parametrijske statistike** (npr. t-test, ANOVA, …)

Simetrične i normalne distribucije

```
# Simetrične i normalne distribucije
# Aritmetička sredina za duljinu
mean(rakovi$duljina)
```

```
## [1] 17.5326
```

```
# SD - Standardna devijacija za duljinu
sd(rakovi$duljina)
```

```
## [1] 2.71285
```

```
# SE - Standardna pogreška za duljinu (Standard Error)
sd(rakovi$duljina) / sqrt(length(rakovi$duljina))
```

```
## [1] 0.1918275
```

Asimetrične i ne-normalne distribucije

Nemaju normalnu raspodjelu podataka i za njih se **ne preporuča** opisivati ih aritmetičkom sredinom i standardnom devicijom! Umjesto toga za njih prikazujemo:

- **medijan** - kao mjeru centralne tendencije
- **interkvartilni raspon** ili **raspon min-max** - kao mjeru raspršenosti uzorka

Distrubucije za koje ne očekujemo normalnu distribuciju podataka koristimo dalje metode **neparametrijske statistike** (npr. Mann-Whitney U test, Kruskal-Wallis test, Wilcoxonov test, ..)

Asimetrične i ne-normalne distribucije

```
# Asimetrične distribucije
# Medijan za masu
median(rakovi$masa)
```

```
## [1] 111.905
```

```
# Interkvartilni range (IQR) za masu
IQR(rakovi$masa)
```

```
## [1] 81.5925
```

```
# Raspon (Range) za temperaturu
range(rakovi$masa)
```

```
## [1]  32.71 573.03
```

Deskriptivna statistika po grupirajućoj varijabli “postaja”

```
# Deskriptivna statistika po grupirajućoj varijabli "postaja"

# Paket "data.table" - daje lijep tablični prikaz deskritipne statistike
# setDT() naredba preoblikuje dataset da ga mozemo koristiti s paketom data.table
setDT(rakovi)

# Kategoricku grupirajucu varijablu "postaja" moramo pretvoriti u faktor naredbom as.factor()
rakovi$postaja <- as.factor(rakovi$postaja)
str(rakovi$postaja)
```

```
##  Factor w/ 2 levels "istok","zapad": 2 2 2 2 2 2 2 2 2 2 ...
```

```
# Summary - pregled kvartila, medijana i aritmetičke sredine
# za varijablu "duljina"
rakovi[, as.list(summary(duljina)), by = list(postaja)]
```

```
##    postaja  Min. 1st Qu. Median    Mean 3rd Qu.  Max.
##     <fctr> <num>   <num>  <num>   <num>   <num> <num>
## 1:   zapad 17.01 19.3825  20.09 20.0326  20.665 22.29
## 2:   istok 12.01 14.3825  15.09 15.0326  15.665 17.29
```

```
# Izmjenite gornju naredbu da umjesto summary vrijednosti standarne devijacije
```

```r
# Sličnu tablicu možemo generirati pomoću "dplyr" objekta
# Podsjetimo se pipe operatora i vježbe "Ras s podacima"!
summary_duljina_postaje <- rakovi %>%
  group_by(postaja) %>%
  summarise(
    minimum = min(duljina),
    Q1 = quantile(duljina, 0.25),
    medijan = median(duljina),
    prosjek = mean(duljina),
    Q3 = quantile(duljina, 0.75),
    maximun = max(duljina),
    stdev = sd(duljina)  # Dodajemo SD
  )
```

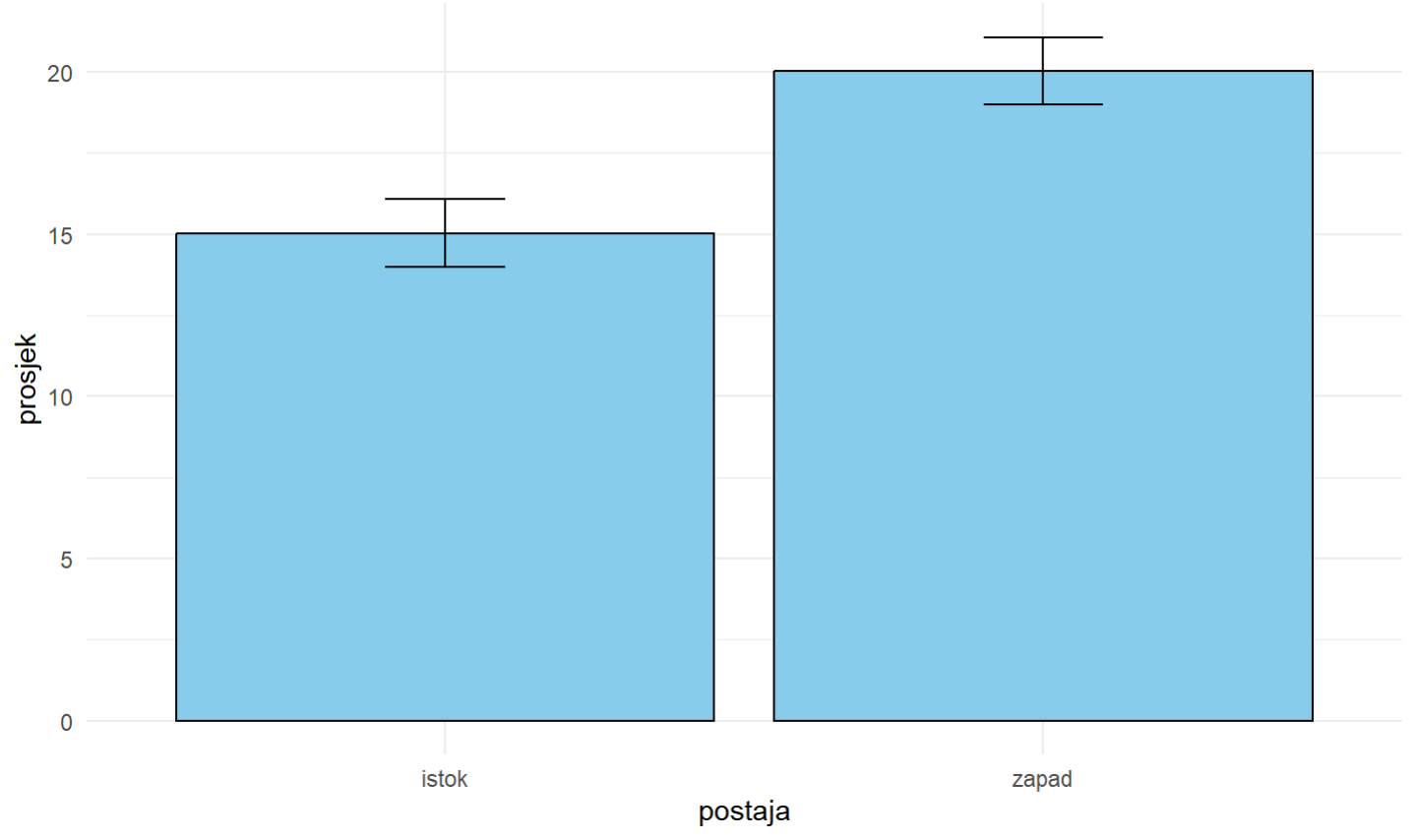
```r
print(summary_duljina_postaje)
```

```
## # A tibble: 2 × 8
##   postaja minimum    Q1 medijan prosjek    Q3 maximun stdev
##   <fct>     <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>
## 1 istok      12.0  14.4    15.1    15.0  15.7    17.3  1.04
## 2 zapad      17.0  19.4    20.1    20.0  20.7    22.3  1.04
```

```r
# Izvoz u tablicu
write.csv(summary_duljina_postaje, #naziv R objekta
          "summary_duljina_postaje.csv", #naziv nove CSV datoteke
          row.names = FALSE, quote = FALSE) #postavke
```

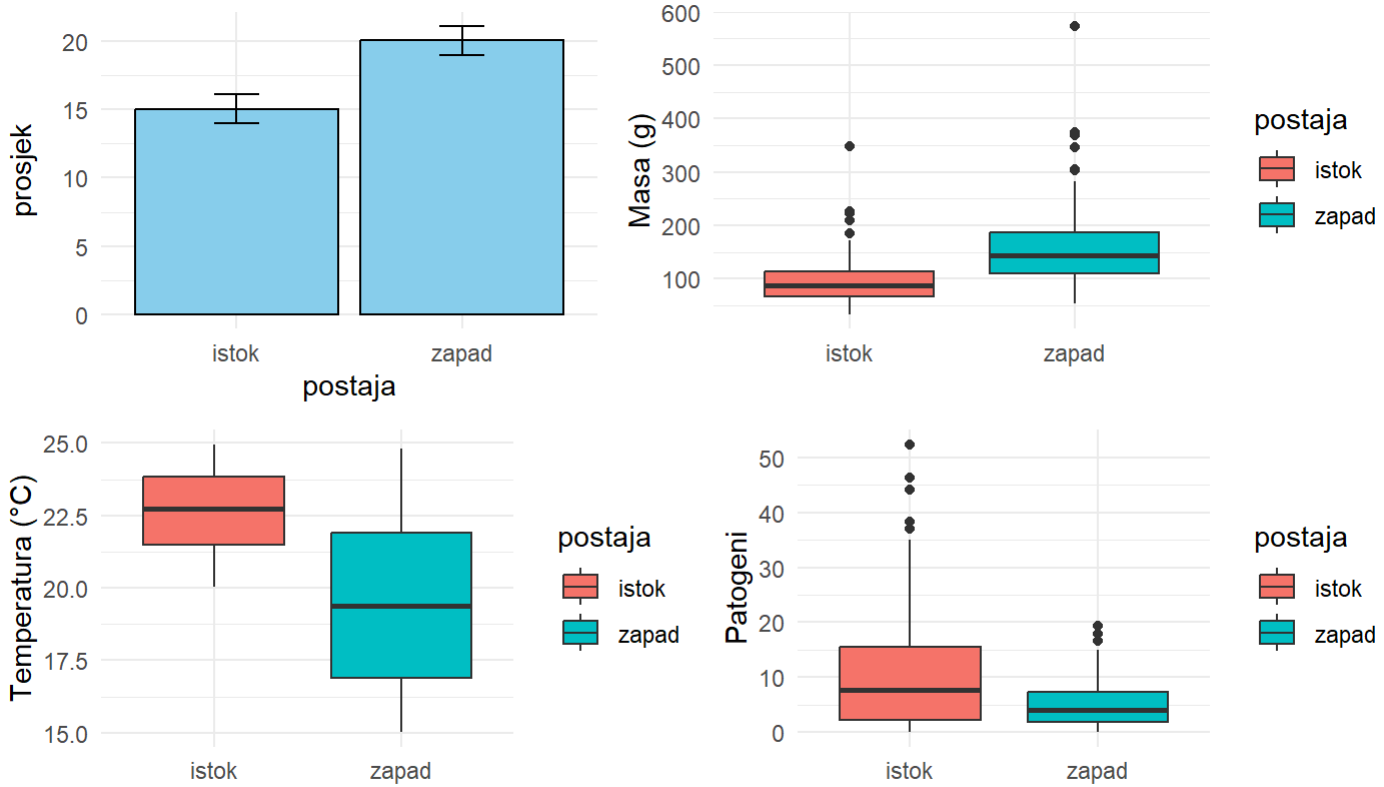
Grafički prikazi deskriptivne statistike

```
# Stupičasti dijagram: duljina (aritmetička sredina i standardna devijacija)
barplot_duljina <- ggplot(summary_duljina_postaje, aes(x = postaja, y = prosjek)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_errorbar(aes(ymin = prosjek - stdev, ymax = prosjek + stdev), width = 0.2) +
  theme_minimal()
print(barplot_duljina)
```



```r
# Boxplot usporedbe: masa, temperatura i patogeni između postaja zapad i istok
boxplot_masa <- ggplot(rakovi, aes(x = postaja, y = masa, fill = postaja)) +
  geom_boxplot() +
  theme_minimal() +
  labs(x = "", y = "Masa (g)")

boxplot_temperatura <- ggplot(rakovi, aes(x = postaja, y = temperatura, fill = postaja)) +
  geom_boxplot() +
  theme_minimal() +
  labs(x = "", y = "Temperatura (°C)")

boxplot_patogeni <- ggplot(rakovi, aes(x = postaja, y = patogeni, fill = postaja)) +
  geom_boxplot() +
  theme_minimal() +
  labs(x = "", y = "Patogeni")
```

```
# Prikaz svih plotova na jednom plotu
(barplot_duljina + boxplot_masa) / (boxplot_temperatura + boxplot_patogeni)
```



Zadatak

U setu podataka o pingvinima (“pingvini.xlsx”) napravite usporedbu mase pingvina po vrstama!

- Napravite novu mapu naziva “samostalni_zadatak” i postavite ju kao radni direktorij.
- Učitajte tablicu u radno okruženje.
- Napravite tablicu da sadrži samo podatke o vrsti i masi pingvina.
- Grafički prikažite broj pingvina svake vrste. Koristite odgovarjući graf za kategoričku varijable.
- Napravite histogram za masu pingvina svake vrste.
- Opišite kako izgleda distribucija podataka o masi s obzirom na simetričnost i nagnutost.
- Ispitajte je li distribucija normalna koristeći Q-Q plot i Shapiro-Wilk test.
- Izračunajte deskriptivnu statistiku (summary) za masu svake vrste pingvina.
- Izračunajte aritmetičku standardnu devijaciju i standardnu pogrešku.
- S obzirom izgled distribucije, je li bolje kosristiti aritmetičku sredinu ili medijan za prikaz centralne tendencije?
- Grafički usporedite mase između vrsta (npr. boxplot).
- Komentirajte usporedbu deskriptive statistike mase između vrsta.