

Deskriptivna analiza

Uvodne napomene

Pakete neophodne za izvršavanje koda u skriptama **instaliramo uslovno** na sledeći način

```
if (!require(ggpubr)) install.packages("ggpubr")
```

Korišćene skripte možemo renderovati u html direktno korišćenjem **render** funkcije iz paketa **rmarkdown**

```
rmarkdown::render(input=here::here("scripts","deskriptivna_analiza.r"),  
output_dir=here::here("reports"))
```

ili ih pretvoriti u **rmd** format korišćenjem **spin** funkcije iz paketa **knitr**

```
knitr::spin(here::here("scripts","deskriptivna_analiza.r"),knit = FALSE)
```

Deskriptivna statistika

Description	R function
Mean	mean()
Standard deviation	sd()
Variance	var()
Minimum	min()
Maximum	maximum()
Median	median()
Range of values (minimum and maximum)	range()
Sample quantiles	quantile()
Generic function	summary()
Interquartile range	IQR()

Analiziraćemo skup podataka **iris**, koji opisuje vrste perunike pomoću atributa dužina i širina latice i cašice

```
#Vrste perunike, atributi dužina i širina latice i cašice
my_data <- iris

head(my_data,10)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

```
# Compute the mean value
mean(my_data$Sepal.Length)

[1] 5.843333

# Compute the median value
median(my_data$Sepal.Length)

[1] 5.8

# Compute the mode
# The function mfv(), for most frequent value, in package modeest
# can be used to find the statistical mode of a numeric vector.
library(modeest)
mfv(my_data$Sepal.Length)

[1] 5

# Compute the minimum value
min(my_data$Sepal.Length)

[1] 4.3

# Compute the maximum value
max(my_data$Sepal.Length)
```

```

[1] 7.9

# Range
range(my_data$Sepal.Length)

[1] 4.3 7.9

quantile(my_data$Sepal.Length, probs = seq(0, 1, 0.25))

  0%   25%   50%   75%  100%
4.3  5.1  5.8  6.4  7.9

# deciles
quantile(my_data$Sepal.Length, seq(0, 1, 0.1))

  0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
4.30 4.80 5.00 5.27 5.60 5.80 6.10 6.30 6.52 6.90 7.90

IQR(my_data$Sepal.Length)

[1] 1.3

# Compute the variance
var(my_data$Sepal.Length)

[1] 0.6856935

# Compute the standard deviation =
# square root of the variance
sd(my_data$Sepal.Length)

[1] 0.8280661

# Compute the median
median(my_data$Sepal.Length)

[1] 5.8

# Compute the median absolute deviation
mad(my_data$Sepal.Length)

[1] 1.03782

summary(my_data$Sepal.Length)

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.300  5.100   5.800   5.843  6.400   7.900

summary(my_data, digits = 1)

```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4	Min. :2	Min. :1	Min. :0.1	setosa :50
1st Qu.:5	1st Qu.:3	1st Qu.:2	1st Qu.:0.3	versicolor:50
Median :6	Median :3	Median :4	Median :1.3	virginica :50
Mean :6	Mean :3	Mean :4	Mean :1.2	
3rd Qu.:6	3rd Qu.:3	3rd Qu.:5	3rd Qu.:1.8	
Max. :8	Max. :4	Max. :7	Max. :2.5	

```
# It's also possible to use the function sapply()
# to apply a particular function over a list or vector.
# Compute the mean of each column without the fifth column
sapply(my_data[, -5], mean)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.843333	3.057333	3.758000	1.199333

```
x4 <- sample(1:999, 33, replace=F)
df123 <- as.data.frame(matrix(x4, nrow = 3, byrow = T))
df123
```

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
200	75	770	602	208	365	962	614	594	378	523
339	825	466	748	742	333	419	478	373	386	606
392	186	661	786	173	847	595	83	49	732	808

```
# row-wise operacije pomoću funkcije pmap_dbl
library(dplyr)
library(purrr)
df123 %>% pmap_dbl(~median(c(...))) %>% head()
```

```
[1] 523 466 595
```

```
df123 %>%
  rowwise() %>%
  mutate(medijana=median(c(V1,V2,V3,V4,V5,V6,V7,V8,V9,V10,V11)))
```

```
Source: local data frame [3 x 12]
Groups: <by row>
```

```
# A tibble: 3 x 12
   V1    V2    V3    V4    V5    V6    V7    V8    V9    V10   V11 me
  <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
1   200    75   770   602   208   365   962   614   594   378   523
2   339   825   466   748   742   333   419   478   373   386   606
```

```

466
3 392 186 661 786 173 847 595 83 49 732 808
595

df345 <- df123 %>%
  mutate(median = select(., V1:V11) %>% pmap_dbl(~median(c(...))))
head(df345)

  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 median
1 200  75 770 602 208 365 962 614 594 378 523    523
2 339 825 466 748 742 333 419 478 373 386 606    466
3 392 186 661 786 173 847 595  83  49 732 808    595

# Compute descriptive statistics
library(pastecs)
res <- stat.desc(my_data[, -5])
round(res, 2)

              Sepal.Length Sepal.Width Petal.Length Petal.Width
nbr.val           150.00           150.00           150.00           150.00
nbr.null             0.00             0.00             0.00             0.00
nbr.na              0.00              0.00              0.00              0.00
min                 4.30              2.00              1.00              0.10
max                 7.90              4.40              6.90              2.50
range               3.60              2.40              5.90              2.40
sum                 876.50            458.60            563.70            179.90
median              5.80              3.00              4.35              1.30
mean                5.84              3.06              3.76              1.20
SE.mean             0.07              0.04              0.14              0.06
CI.mean.0.95        0.13              0.07              0.28              0.12
var                 0.69              0.19              3.12              0.58
std.dev             0.83              0.44              1.77              0.76
coef.var            0.14              0.14              0.47              0.64

library(skimr)
skim(iris)

```

Data summary

Name	iris
Number of rows	150
Number of columns	5

Column type frequency:

factor	1
--------	---

numeric 4

Group variables None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Species	0	1	FALSE	3	set: 50, ver: 50, vir: 50

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sepal.Length	0	1	5.84	0.83	4.3	5.1	5.80	6.4	7.9	
Sepal.Width	0	1	3.06	0.44	2.0	2.8	3.00	3.3	4.4	
Petal.Length	0	1	3.76	1.77	1.0	1.6	4.35	5.1	6.9	
Petal.Width	0	1	1.20	0.76	0.1	0.3	1.30	1.8	2.5	

```
skim(iris, Sepal.Length, Petal.Length)
```

Data summary

Name iris
Number of rows 150
Number of columns 5

Column type frequency:

numeric 2

Group variables None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sepal.Length	0	1	5.84	0.83	4.3	5.1	5.80	6.4	7.9	
Petal.Length	0	1	3.76	1.77	1.0	1.6	4.35	5.1	6.9	

```
iris %>%  
  dplyr::group_by(Species) %>%  
  skim()
```

Data summary

Name Piped data
Number of rows 150

Number of columns 5

Column type frequency:

numeric 4

Group variables Species

Variable type: numeric

skim_variable	Species	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sepal.Length	setosa	0	1	5.01	0.35	4.3	4.8	5.0	5.2	5.8	
Sepal.Length	versicolor	0	1	5.94	0.52	4.9	5.6	5.9	6.3	7.0	
Sepal.Length	virginica	0	1	6.59	0.64	4.9	6.23	6.5	6.9	7.9	
Sepal.Width	setosa	0	1	3.43	0.38	2.3	3.20	3.4	3.6	4.4	
Sepal.Width	versicolor	0	1	2.77	0.31	2.0	2.52	2.8	3.0	3.4	
Sepal.Width	virginica	0	1	2.97	0.32	2.2	2.8	3.0	3.18	3.8	
Petal.Length	setosa	0	1	1.46	0.17	1.0	1.40	1.50	1.58	1.9	
Petal.Length	versicolor	0	1	4.26	0.47	3.0	4.0	4.35	4.6	5.1	
Petal.Length	virginica	0	1	5.55	0.55	4.5	5.10	5.55	5.8	6.9	
Petal.Width	setosa	0	1	0.25	0.11	0.1	0.2	0.2	0.3	0.6	
Petal.Width	versicolor	0	1	1.33	0.20	1.0	1.20	1.30	1.50	1.8	
Petal.Width	virginica	0	1	2.03	0.27	1.4	1.80	2.0	2.30	2.5	

```
# funs <- list(  
#   iqr = IQR,  
#   mad = mad  
# )  
#  
# skim_with(numeric = funs, append = FALSE)
```

```
#
# skim(iris, Sepal.Length)
# skim_with_defaults()

my_skim <- skim_with(numeric = sfl(median, mad), append = FALSE)
my_skim(faithful)
```

Data summary

Name	faithful
Number of rows	272
Number of columns	2

Column type frequency:

numeric	2
---------	---

Group variables	None
-----------------	------

Variable type: numeric

skim_variable	n_missing	complete_rate	median	mad
eruptions	0	1	4	0.95
waiting	0	1	76	11.86

```
# Set multiple types of skimmers simultaneously.
my_skim <- skim_with(numeric = sfl(mean), character = sfl(length))

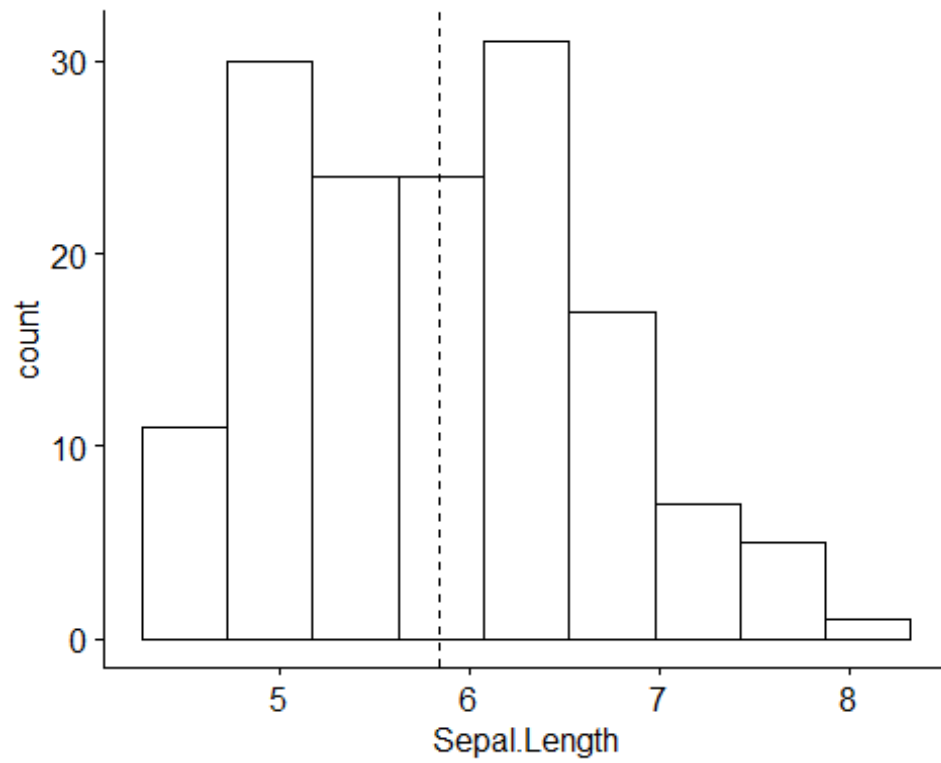
my_skimmers <- list(numeric = sfl(mean), character = sfl(length))
my_skim <- skim_with(!!!my_skimmers)
```

```
mean(my_data$Sepal.Length, na.rm = TRUE)
```

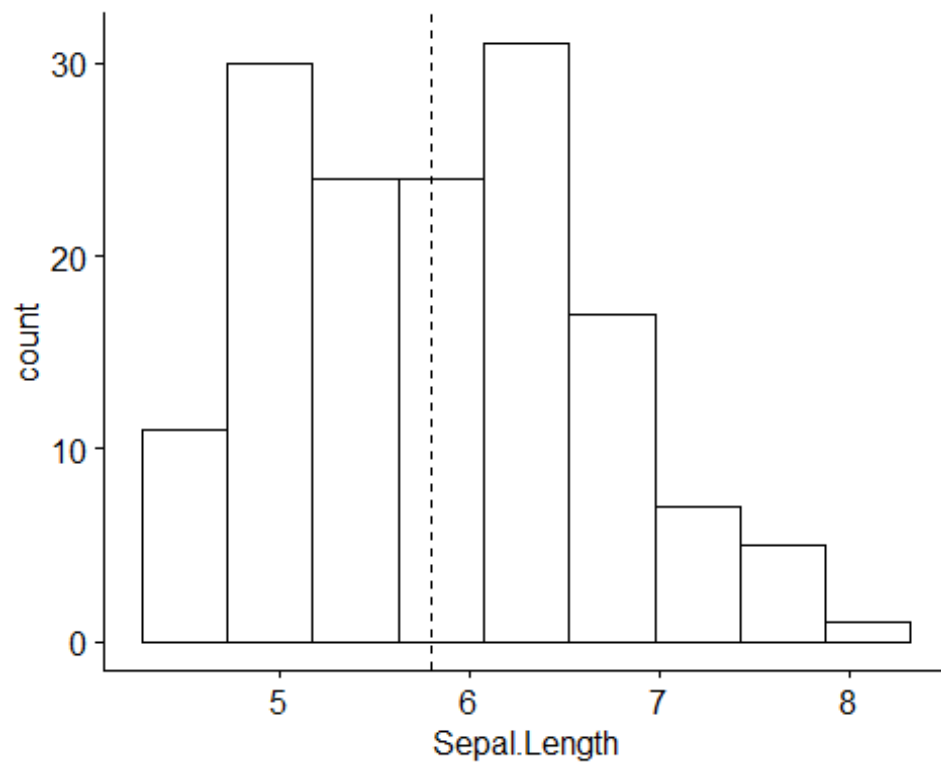
```
[1] 5.843333
```

```
library(ggpubr)
```

```
gghistogram(my_data, x = "Sepal.Length", bins = 9,
             add = "mean")
```



```
gghistogram(my_data, x = "Sepal.Length", bins = 9,  
            add = "median")
```



```
# Descriptive statistics by groups

library(dplyr)
group_by(my_data, Species) %>%
  summarise(
    count = n(),
    mean = mean(Sepal.Length, na.rm = TRUE),
    sd = sd(Sepal.Length, na.rm = TRUE)
  )

# A tibble: 3 x 4
  Species    count  mean    sd
  <fct>      <int> <dbl> <dbl>
1 setosa         50  5.01 0.352
2 versicolor    50  5.94 0.516
3 virginica     50  6.59 0.636
```