

Cvičenie 7 – Základné spracovanie dát, opisná štatistika

práca s pomocou R

Nižšie uvedené príkazy predstavujú iba nevyhnutné minimum potrebné na vykonanie toho, čo sa týka preberanej témy. V záujme pružnejšieho zvládnutia prostredia R sa výrazne odporúča samoštúdium dostupných základných úvodov.

Nastavenie pracovného adresára

Default nastavenie je spravidla `C:/Users/MenoPoužívateľa/Documents`

Overenie `getwd()`

Zmena `setwd("D:/.../.../MASTA")`

!!! pozor na syntax – úvodzovky, štandardná lomka / , nie \

Dáta

Dnes budeme pracovať najmä s vektorom dát.

– Dáta môžeme zadať ručne.

Názov vektora zvolte tak, aby ste ním neprekryli interné funkcie a príkazy v rámci R, situáciu si viete overiť príkazom `?názov`

Priradenie do premennej pomocou znaku `<-` (skratka `alt -`)

Vektor (nielen číselný) tvoríme príkazom `c(...)`, kde `c` je od slova *concatenate*.

`uvod <- c(165, 171, 194, 184, 164, 142, 198, 177, 166, 130, 159, 179, 181, 163)`

Vektor sa uloží do premennej `uvod`. Obsah premennej si vieme overiť jej názvom + `enter`.

– Dáta môžeme načítať zo súboru. Preferovaný formát je `.csv`.

Charakteristiky polohy

Výpis vektora udajee prezradil počet čísel 14. Ale overme si to a dajme do premennej n:

```
n=length(udajee)
n
```

Minimum a maximum:

```
min(udajee)
max(udajee)
```

Na jeden krok:

```
range(udajee)
```

Celkové rozpätie:

```
max(udajee)-min(udajee)
```

Ak si dáta usporiadame podľa veľkosti, uvidíme minimum a maximum hneď:

```
udaje <- sort(udajee)
udaje
130 142 159 163 164 165 166 171 177 179 181 184 194 198
```

Aritmetický priemer – zabudovaným príkazom mean a “od podlahy”:

```
mean(udaje)
sum(udaje)/n
169.5
```

Medián – systémovo:

```
median(udaje)
168.5
```

Minimum, maximum, medián a oba kvartily na jeden krok:

```
quantile(udaje, type=1)
0% 25% 50% 75% 100%
130 163 166 181 198
```

R pozná 9 rôznych algoritmov na výpočet kvantilov, našej štandardnej definícii zodpovedá typ 1. Treba to uviesť do parametrov.

Medzikvartilové rozpätie:

```
IQR(udaje, type=1)
18
```

Iné kvantily:

```
quantile(udaje, 0.15, type=1)
15%
159
```

Rozptyl

Disperzia (variancia):

```
var(udaje)
341.1923
```

Od podlahy:

```
u1=udaje-mean(udaje)
u2=u1*u1
sum(u2)/13
341.1923
```

Ak vektor udaje predstavuje celú *populáciu*, treba deliť počtom 14.

Smerodajná odchýlka je odmocninou z tejto hodnoty:

```
sd(udaje)
sqrt(341.1923)
```

Šikmosť a špicatosť

!! Je potrebné nainštalovať balík **psych** a načítať si ho.

```
skew(udaje)
-0.4584723
kurtosi(udaje)
-0.4752467
```

Vizualizácia údajov

Histogram – počet tried určí R podľa Sturgesovho pravidla (rovnomerné hranice tried)

```
hist(
  udaje,
  breaks = "Sturges",
  main = "Histogram",
  xlab = "vyska",
  ylab = "pocetnost"
)
```

Rozloženie údajov na 1 osi:

```
stripchart(udaje)
```

Krabicový diagram (boxplot):

```
boxplot(udaje)
boxplot(udaje, horizontal=T)
```

Koláčový diagram:

Použijeme iné dáta, v ktorých sa opakujú čísla.

```
pocety <- c(1,1,1,2,2,3,3,3,3,4,5,5,7,8,9,9)
tp <- table(pocety)
tp
      pocety
      1 2 3 4 5 7 8 9
      3 2 4 1 2 1 1 2
pie(tp)
```

Úlohy

1. Načítajte dané dáta zo súboru `cisla.csv` a “vytiahnite si” odtiaľ stĺpec BMI:

Načítame si knižnicu **readr**, najľahšie odkliknutím v záložke Packages.

Alebo príkazom

```
library(readr)
```

Ak ju tam nevidno, treba ju stiahnuť – na záložke Packages kliknete na Install a v zozname si vyberiete balík readr.

```
cisla <- read_csv("cisla.csv")  
bmi <- cisla$BMI
```

Pre dátový vektor vypočítajte všetky vyššie spomenuté charakteristiky.
Urobte vizualizáciu dát všetkými spomenutými spôsobmi.