Popisná statistika

Komentované řešení pomocí programu R

Ústav matematiky
Fakulta chemicko inženýrská
Vysoká škola chemicko-technologická v Praze

Načtení vstupních dat

Máme k dispozici data o počtech bodů z 1. a 2. zápočtového testu z Matematiky I v zimním semestru 2015/2016 a to za všechny 762 studenty, kteří psaly oba testy.

Vstupní data se nacházejí v souboru "data-popisna_statistika.csv".

Předpokládejme, že data jsou uložena na disku F ve složce Aplikovana_statistika.

Načtení vstupních dat do pracovního objektu DATA a vypsání na obrazovku provedeme příkazem

```
DATA<-read.csv("f:\data-popisna_statistika.csv", header=FALSE)
DATA
```

```
  Student.ka. X1..PP X2.PP
1           1     43    65
2           2     37    43
3           3     27     7
4           4     33    70
5           5     41    19
...
```

Jak vidíme, v R se nám automaticky očíslovaly řádky, nebudeme proto první sloupec vůbec potřebovat. Tento sloupec můžeme jednoduše odstranit příkazem

```
DATA<-DATA[,-1]
```

Poznámka: Napíšeme-li "DATA<-DATA..." znamená to, že přímo měníme objekt DATA. Samozřejmě bychom mohli vytvořit z objektu DATA úplně jiný objekt.

Obdobně se dá odstranit i nějaký sloupec. Například budeme-li chtít odstranit $i-$tý řádek v objektu DATA, použijeme příkaz

```
DATA<-DATA[-i,].
```

Povšimněme si názvu jednotlivých datových sloupců. Pro přehlednost je můžeme přejmenovat

```
names(DATA)=c("bodyPP1","bodyPP2")
DATA
```

```
    bodyPP1 bodyPP2
1        43      65
2        37      43
3        27       7
4        33      70
5        41      19
6        55      34
7        82      50
8        48      52
9        19      11
...
```

Kvantily a výběrový průměr

Podívejme se na kvartily a horní a dolní decil

```
quantile(DATA$bodyPP1,c(0.1,0.25,0.5,0.75,0.9))
```

```
10%  25%  50%  75%  90%
16.0 30.0 45.0 65.0 78.9
```

```
quantile(DATA$bodyPP2,c(0.1,0.25,0.5,0.75,0.9))
```

```
10%  25%  50%  75%  90%
8.0  21.0 39.0 56.0 71.0
```

Obecně, chceme-li určit **u**% kvantil bodů z první písemky, použijeme příkaz

```
quantile(DATA$bodyPP1,u)
```

Chceme-li jich vypsat více najednou, zapíšeme je do vektoru „c()“, jak je to provedeno na začátku tohoto slidu.

Ke spočtení výběrového průměru použijeme příkaz mean. Ten lze použít jak přímo na množinu DATA, tak i na její jednotlivé soupce.

```
mean(DATA)
```

```
bodyPP1  bodyPP2
46.57087 39.45013
```

Rozptyl, směrodatná odchylka a variační koeficient

- Rozptyl

```
var(DATA$bodyPP1)
[1] 539.7092
var(DATA$bodyPP2
[1] 553.6618
```

  Kdybychom použili příkaz var přímo na celý objekt DATA, obdrželi bychom kovarianční matici.

- Směrodatná odchylka
  Příkaz lze opět použít rovnou na celý objekt DATA (ale i zvlášť)

```
sd(DATA)
 bodyPP1  bodyPP2
23.23164 23.53002
```

  Poznámka: Směrodatnou odchylku by šlo samozřejmě spočítat i „ručně“ jako druhou odmocninu z rozptylu:

```
sqrt(var(DATA$bodyPP1))
[1] 23.23164
```

- Variační koeficient

  Je dán vztahem

$$V = \frac{\sigma}{\bar{x}}.$$

  Spočítáme jej z definice

```
var_koeficient<-sd(DATA)/mean(DATA)
  bodyPP1   bodyPP2
0.4988450 0.5964497
```

Variační, decilové a kvartilové rozpětí

- Variační rozpětí

```
var_rozpeti<-c(max(DATA$bodyPP1)-min(DATA$bodyPP1),max(DATA$bodyPP2)-min(DATA$bodyPP2))
var_rozpeti
[1] 100 100
```

- Decilové rozpětí

```
decilove_rozpeti<-c(quantile(DATA$bodyPP1,0.9)-quantile(DATA$bodyPP1,0.1),
       + quantile(DATA$bodyPP2,0.9)-quantile(DATA$bodyPP2,0.1))
decilove_rozpeti

      90%  90%
      62.9  63.0
```

Ono + na začátku druhého řádku posledního sloupce vypíše editor sám, pokud příkaz přeteče na další řádek.

- Kvartilové rozpětí

```
kvartilove_rozpeti<-c(quantile(DATA$bodyPP1,0.75)-quantile(DATA$bodyPP1,0.25),
       + quantile(DATA$bodyPP2,0.75)-quantile(DATA$bodyPP2,0.25))
kvartilove_rozpeti

       75% 75%
        35  35
```

Výběrová kovariance a korelční koeficient

Kovarianční matici spočteme příkazem

```
cov(DATA)
```

```
         bodyPP1  bodyPP2
bodyPP1 539.7092 371.1645
bodyPP2 371.1645 553.6618
```

Připomeňme si, že na hlavní diagonále matice jsou výběrové rozptyly jednotlivých bodů z paralelkových testů.

Pearsonův výběrový korelační koeficient (respektive korelační matici) spočteme příkazem "corr"

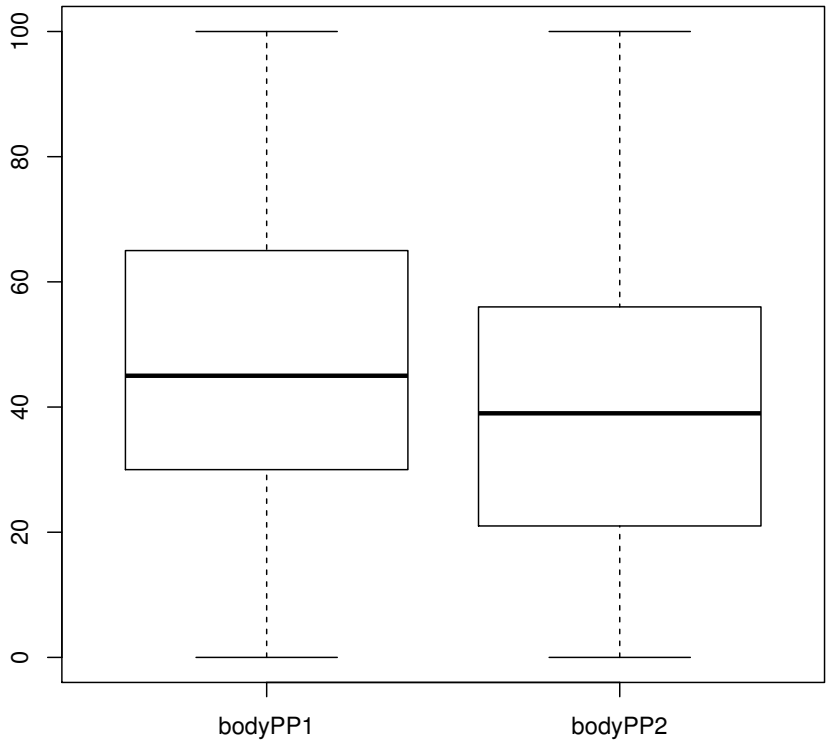
```
cor(DATA, method="pearson")
```

```
          bodyPP1   bodyPP2
bodyPP1 1.0000000 0.6789914
bodyPP2 0.6789914 1.0000000
```

Poznámka: Argument method značí typ korelačního koeficientu, který počítáme. V **R** jsou ve standardní nabídce jeste koeficienty "spearman"a "kendall".

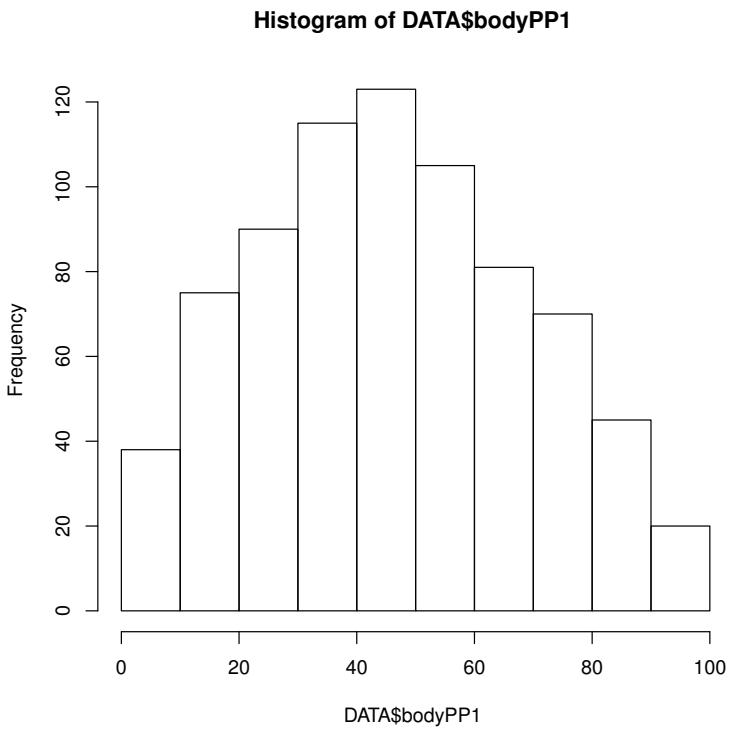
Grafické výstupy - Boxplot

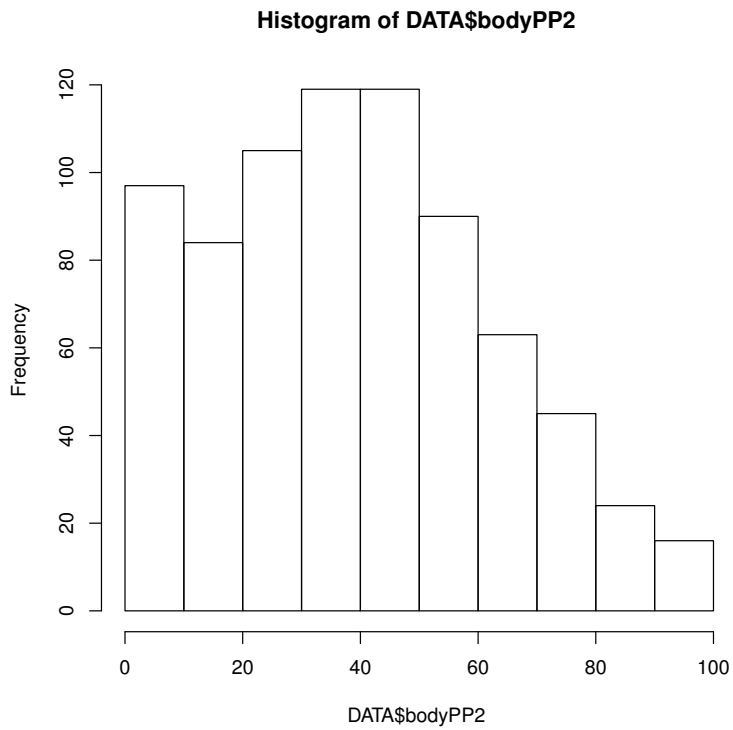
```
boxplot(DATA)
```



Grafické výstupy - Histogramy

```
hist(DATA$body)
hist(DATA$dochazka)
```





Grafické výstupy - Bodový graf

```
plot(DATA,col="red")
```