Testy nezávislosti kardinálních veličin

Komentované řešení pomocí programu R

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

Načtení vstupních dat

Vstupní data se nacházejí v souboru "data-nezavislost_korelace.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:\\Aplikovana_statistika\\data-nezavislost_korelace.csv", header=FALSE)
DATA
```

```
    V1    V2
1   33  70.4
2  135 100.0
3   79  92.6
4   42  85.2
5   27 100.0
6   14  88.9
...
```

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

```
names(DATA)=c("body","dochazka")
DATA
```

```
   body dochazka
1    33     70.4
2   135    100.0
3    79     92.6
4    42     85.2
5    27    100.0
6    14     88.9
...
```

Souhrnné charakteristiky souboru

Podívejme se nejprve na některé souhrnné charakteristiky našich dat pomocí příkazu summary. Ten nám vypíše minimum, maximum, všechny kvartily a medián

```
summary(DATA)
```

```
body           dochazka
 Min.   :  0.00   Min.   :  0.00
 1st Qu.: 22.00   1st Qu.: 67.30
 Median : 46.00   Median : 82.10
 Mean   : 54.65   Mean   : 71.46
 3rd Qu.: 81.00   3rd Qu.: 92.30
 Max.   :183.00   Max.   :100.00
```

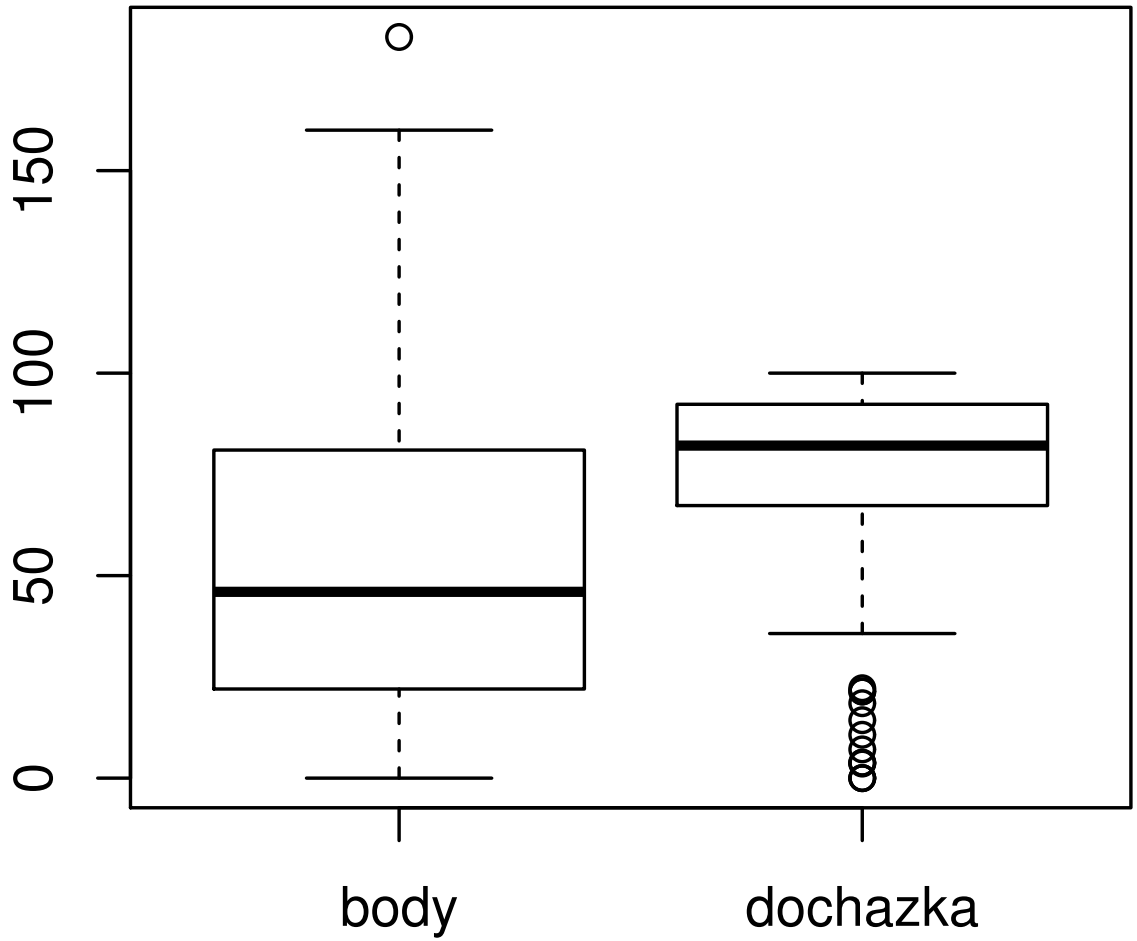
Součástí příkazu summary nejsou výběrové rozptyly, takže si necháme vypsat zvlášť výběrovou varianční matici příkazem var:

```
var(DATA)
```

```
41.40621 29.95059
```

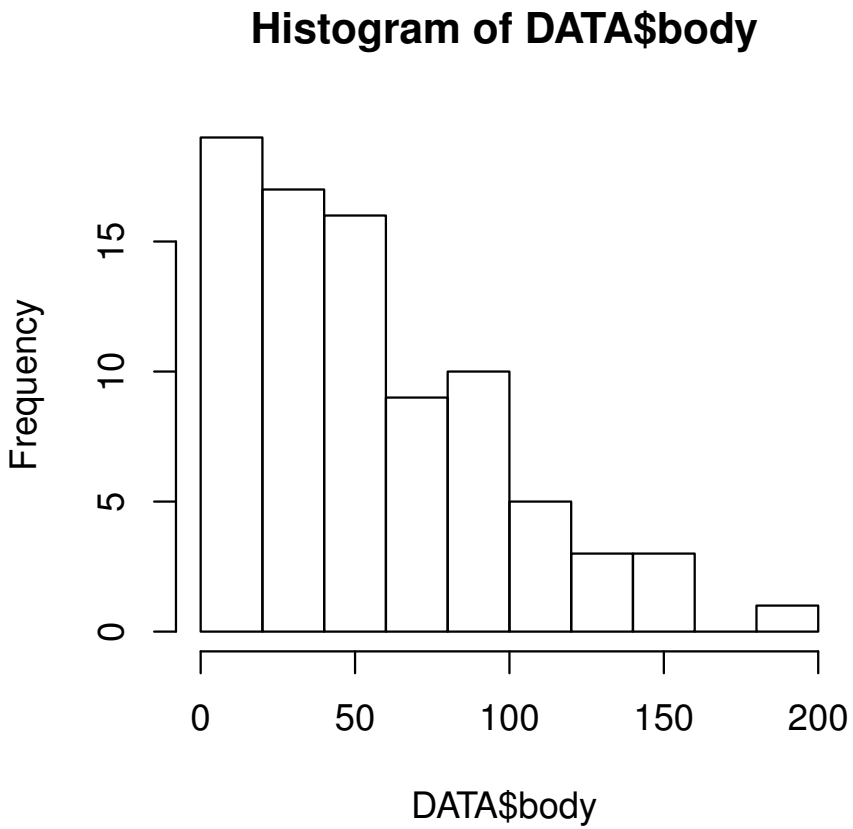
Grafické výstupy - Boxplot

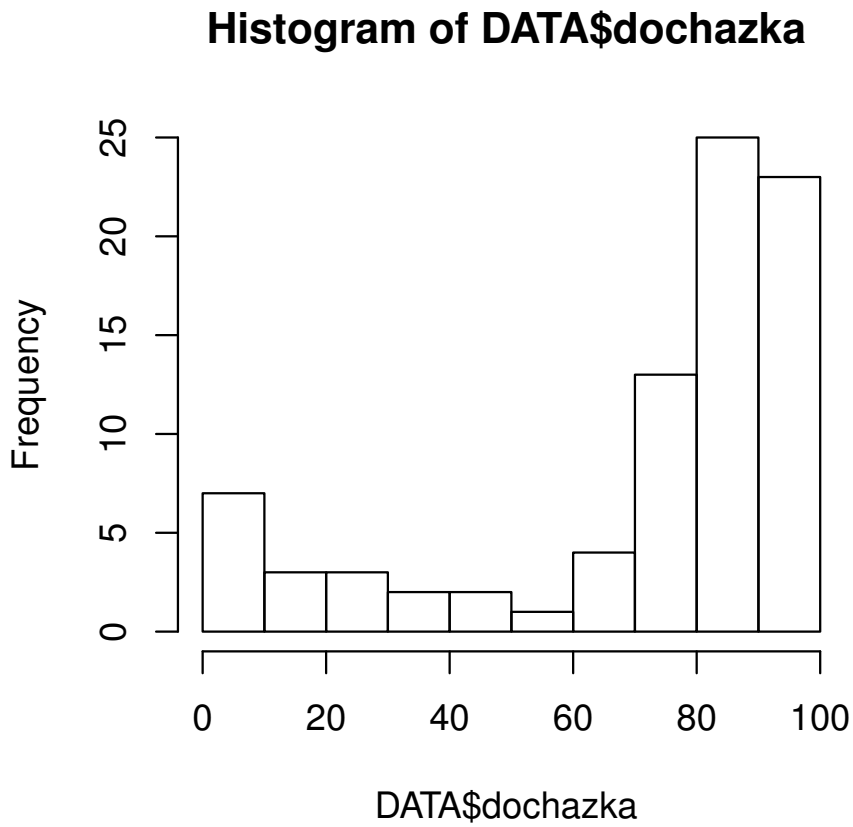
```
boxplot(DATA)
```



Grafické výstupy - Histogramy

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

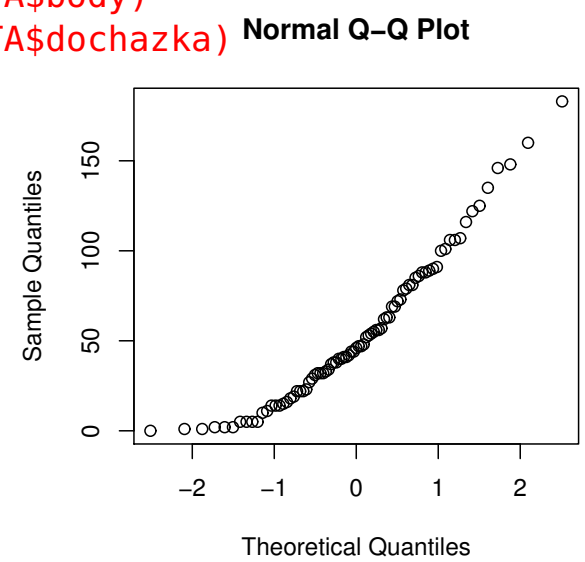


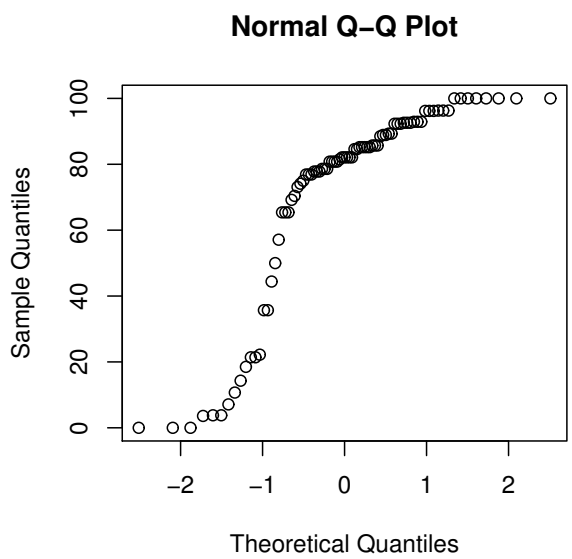


Grafické výstupy - Q-Q plot

Z histogramů je poměrně očividné, že ani počty bodů ani docházka v procentech nevypadají jako normálně rozdělené veličiny. To nám ostatně potvrzuje i normální Q-Q plot. Normální Q-Q plot je graf porovnávající kvantily výběru s teoretickými kvantily normalního rozdělení. Pokud by veličiny pocházely z normálního rozdělení, pak by grafem byla při bližně přímka $y = x$

```
qqnorm(DATA$body)
qqnorm(DATA$dochazka)
```





Test normality dat

Grafické výstupy ovšem nejsou exaktním důkazem, proto si ještě provedeme test normality dat. K tomuto účelu použijeme Shapiro-Wilkův test normality

```
shapiro.test(DATA$body)

        Shapiro-Wilk normality test

data:  DATA$body
W = 0.9353, p-value = 0.0004168

shapiro.test(DATA$dochazka)

        Shapiro-Wilk normality test

data:  DATA$dochazka
W = 0.7771, p-value = 7.008e-10
```

Jak je vidět, p-hodnota je velmi nízká, takže normalitu dat zamítáme i na velmi nízkých hladinách testu (hodnota W je hodnota použité testové statistiky). Námi použité testy budou tedy **pouze asymptotické**.

Výběrový korelční koeficient

Pearsonův výběrový korelační koeficient (respektive korelační matici) spočteme příkazem cor

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

```
              body  dochazka
body     1.0000000 0.2621148
dochazka 0.2621148 1.0000000
```

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

Test nulovosti korelačního koeficientu

Test nulovosti korelačního koeficientu oproti oboustranné alternativě - úloha **(A)**

Příkaz cor.test má čtyři různé argumenty - první dva jsou výběry u nichž provádíme test, třetí je typ alternativy. Zde může být oboustranná ("two.sided") nebo jednostranné ("greater", "less") a konečně poslední argument je typ metody (zde "pearson").

```
cor.test(DATA$body,DATA$dochazka,alternative="two.sided", method="pearson")

   Pearson's product-moment correlation

data:  DATA$body and DATA$dochazka
t = 2.4445, df = 81, p-value = 0.01668
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04920743 0.45223670
sample estimates:
      cor
0.2621148
```

Pojďme si podrobně probrat jednotlivé výstupy testu.

- $t = 2.4445$ . . . hodnota testové statisiky (toto číslo porovnáváme s kritickým oborem
- $df = 81$ počet stupňů volnosti
- p-value . . . numericky spočtená skutečná hladina testu
- 95 percent confidence interval . . . 95% konfidenční interval (pokud je 0 uvnitř tohoto intervalu, nemůžeme na hladině 5% zamítnout nulovou hypotézu)
- sample estimates: corr . . . hodnota výběrového korelačního koeficientu, na níž je test založen

Neboť p-hodnota je rovna $p = 0,01668$, je vidět v našem případě ulvou hypotézu zamítáme na hladinách větších než $1,668\%$.

Test nulovosti korelačního koeficientu

Test nulovosti korelačního koeficientu oproti alternativě, ze je větší než nula - úloha **(B)**

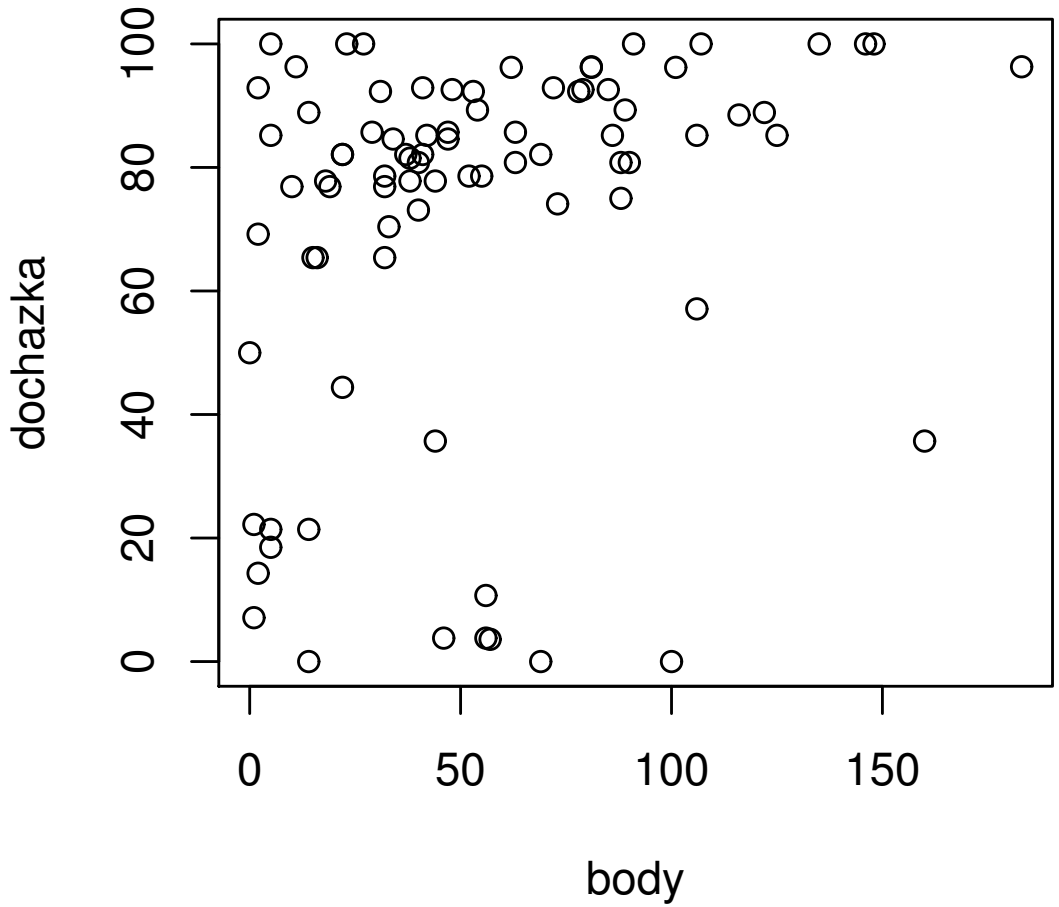
```
cor.test(DATA$body,DATA$dochazka, alternative="greater", method ="pearson")

        Pearson's product-moment correlation

data:  DATA$body and DATA$dochazka
t = 2.4445, df = 81, p-value = 0.008338
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.08427723 1.00000000
sample estimates:
      cor
0.2621148
```

Výstupy testu jsou stejné jako v případě testování pří oboustranné alternativě. Jak vidíme, p-hodnota testu je $p = 0,008338$, takže nulovou hypotézu zamítáme ve prospěch alternativy, že veličiny jsou kladně korelované, na každé hladině testu větší než $0,8338\%$. Pro všechny takové hladiny tedy lze statisticky významně tvrdit, že veličiny se pozitivně ovlivňují.

Typ závislosti

Pokud bychom si chtěli jednoduše bodově vykreslit závislost mezi body a docházkou, můžeme použít příkazu plot:

```
plot(DATA)
```



Spermanův test nezávislosti

Jak je vidět, data nevypadají, že by byla lineárně závislá. Přesto bychom přirozeně čekali, že by vysoká docházka mohla mít pozitivní vliv na výsledky testů. Neboli, že vztah mezi body a docházkou je (až na náhodný faktor) rostoucí funkce. Tímto monotónním typem závislosti se zabývá Spearmanův korelační koeficient. Můžeme tedy použít k testování tento koeficient a to jednoduše tím, že v argumentu method nahradíme "pearson" slovem "spearman".

```
cor.test(DATA$body,DATA$dochazka,alternative="greater", method="spearman")

        Spearman's rank correlation rho

data:  DATA$body and DATA$dochazka
S = 62343.62, p-value = 0.0006845
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.3457074

Warning message:
In cor.test.default(DATA$body, DATA$dochazka, alternative = "greater",  :
  Cannot compute exact p-values with ties
```

Poznámka: rho je výběrový Spearmanův korelační koeficient. Warning message pouze říká, že pro naše data není spočtená přesná hodnota p, ale je pouze odhadnutá pomocí numerického algoritmu. Ježto je ale menší než jedno procento, je to pro nás i tak dostačující údaj, abychom zamítli nekorelovanost ve prospěch alternativy, že vztah mezi body a docházkou je rostoucí funkce.

Závěr

Na hladinách minimálně jedno procento a výše se nám podařilo prokázat souvislost mezi docházkou a počtem bodů získaných z testů z Matematiky I. Dokonce se nám statisticky významně na hladinách nejméně jedno procento potvrdila hypotéza o tom, že se docházka a počty bodů navzájem pozitivně ovlivňují. Všechny provedené testy jsou, vzhledem k předpokládané nenormalitě dat, pouze asymptotické.