Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Aplikovaná statistika I

Téma 3: Základní číselné charakteristiky

Veronika Bendová

bendova.veroonika@gmail.com

Úvod a motivace

- minulá hodina → bodové/intervalové rozložení četností
  - seznámení s daty
  - výhody: široké množství informací, globální pohled na data
  - nevýhody: přemíra informací, horší interpretovatelnost, ztížené srovnávání dvou datasetů
  - Adam a Bára provádí nezávislý výzkum na stejné téma → dvě různé nemocnice → dva různé datasety (A) a (B) → dvě různé variační řady → porovnávání variačních řad (nepřehledné a neefektivní)
- → vznik *číselných charakteristik*
  - elegantní a jednoduché vystihnutí charakteristických rysů znaku zpravidla pomocí jednoho čísla
  - snadno spočítatelné i interpretovatelné
- různá data → různé charakteristiky
- typy dat
  - Nomiální,
  - Ordinální,
  - Intervalová,

- typy charakteristik
    - polohy
    - variability
    - závislosti
    - + nesymetrie (intervalové znaky)
- přehled číselných charakteristik podle typu znaku a sledované vlastnosti

	Poloha	**Variabilita**	**Symetrie**	**Závislost**
Nominální	modus	–	–	Cramérův koeficient
Ordinální	medián	interkvartilové rozpětí	–	Spearmanův koef. poř. korel.
Intervalový	aritmetický průměr	rozptyl směrodatná odchylka	koeficient šikmosti koeficient špičatosti	Pearsonův korel. koeficient

Nominální znaky

- varianty znaku jsou neporovnatelné
  - vzdělání: ZŠ, SŠ, SŠm, VŠ
  - barva očí: modrá, zelená, hnědá
  - pohlaví: m, f
- charakteristika polohy
  - *modus* . . . nejčetnější varianta znaku
- charakteristika závislosti
  - Cramérův koeficient r_C - těsnost závislosti u nominálních znaků
  - $r_C \in \langle 0; 1 \rangle$.
  - stunice míry závislosti podle Cramérova koeficientu

Cramérův koeficient	**Interpretace**
0.0 – 0.1	zanedbatelný stupeň závislosti
0.1 – 0.3	slabý stupeň závislosti
0.3 – 0.7	střední stupeň závislosti
0.7 – 1.0	silný stupeň závislosti

Příklad 3.1. Charakteristika polohy nominálního znaku
Navažme na práci s datasetem 17-anova-newborns-2.txt. V rámci cvičení 2 jsme jako mezivýstup příkladu 2.5 získali kontingenční tabulku simultánních absolutních četností znaků X = *vzdělání matky* a Y = *porodní hmotnost novorozence* (viz tabulka 3.1). Znaky X a Y jsou typickým příkladem znaků nominálního typu. Najděte modus pro znak *vzdělání matky* i pro znak *porodní hmotnost novorozence*.

Tabulka 3.1: Simultánní absolutní četnosti pro znaky *vzdělání matky* a *porodní hmotnost novorozence*

	nízká	norma	vysoká
ZS	97	312	8
SS	82	346	20
SSm	74	349	12
VS	13	64	4

Řešení příkladu 3.1

```
(data <- data.frame(nizka = c(97, 82, 74, 13),
                    norma  = c(312, 346, 349, 64),
                    vysoka = c(8, 20, 12, 4),
                    row.names = c('ZS', 'SS', 'SSm', 'VS')))
```

```
    nizka norma vysoka
ZS     97   312      8
SS     82   346     20
SSm    74   349     12
VS     13    64      4
```

Zaměřme se nejprve na znak $X =$ *vzdělání matky*. Číselná charakteristika *modus* je definována jako nejčetnější varianta sledovaného znaku.

```
(nj. <- apply(data, MARGIN = 1, FUN = sum))
```

```
 ZS  SS SSm  VS
417 448 435  81
```

Interpretace výsledků: Nejčetnější variantou znaku *vzdělání matky* je .... ($n =$ ....). Nejvíce novorozenců v datovém souboru se narodilo matkám s dokončeným ....

Analogicky nyní najdeme modus znaku $Y =$ *porodní hmotnost novorozence*.

```
(n.k <- apply(data, MARGIN = 2, FUN = sum))
```

```
 nizka  norma vysoka
   266   1071     44
```

Interpretace výsledků: Nejvíce novorozenců v datovém souboru mělo porodní hmotnost .... ($n =$ ....).

Příklad 3.2. Charakteristika závislosti mezi dvěma nominálními znaky
Zaměřte se nyní na oba znaky X = *vzdělání* a Y = *porodní hmotnost novorozence* najednou. Určete míru závislosti mezi znaky X a Y .

Řešení příkladu 3.2
Protože X a Y jsou znaky .... typu, použijeme na určení míry závislosti mezi nimi .... Stupnice míry závislosti podle hodnoty Cramérova koeficientu je uvedena v tabulce 3.2

Tabulka 3.2: Stunice míry závislosti podle Cramérova koeficientu

Cramérův koeficient	**Interpretace**
0.0 – 0.1	Zanedbatelný stupeň závislosti
0.1 – 0.3	Slabý stupeň závislosti
0.3 – 0.7	Střední stupeň závislosti
0.7 – 1.0	Silný stupeň závislosti

```
lsr::cramersV(data)
```

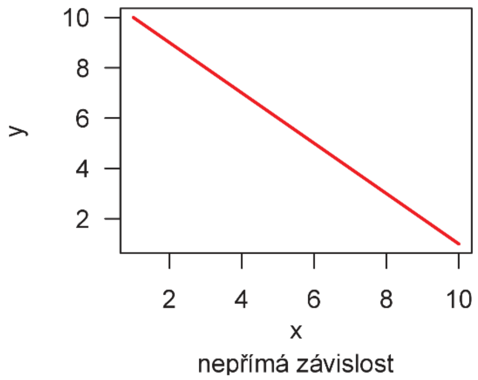
```
[1] 0.0645725
```

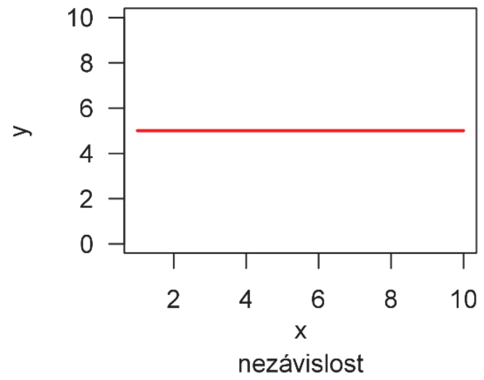
Interpretace výsledků: Hodnota Cramérova koeficientu vyšla .... Mezi vzděláním matky a porodní hmotností novorozence existuje .... stupeň ....

Ordinální znaky

- hodnoty znaku můžeme porovnávat, ale nemůžeme stanovit, jak velký je mezi nimi rozdíl
    - počet starších sourozenců, pořadí 10 pacientů podle závažnosti onemocnění, ...
- charakteristika polohy
    - α-kvantil x_α ... takové číslo, že $\alpha \times 100\%$ hodnot v datovém souboru je menších nebo rovných hodnotě x_α; $\alpha \in \langle 0; 1 \rangle$
        - medián $x_{0.5}$
        - dolní kvartil $x_{0.25}$
        - horní kvartil $x_{0.75}$
    - $n\alpha =$ celé číslo $c \to x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2}$
    - $n\alpha =$ necelé číslo $\to$ zaokrouhlíme nahoru na nejbližší vyšší celé číslo $c \to x_\alpha = x_{(c)}$
- charakteristika variability
    - (inter)kvartilové rozpětí IQR
    - $IQR = x_{0.75} - x_{0.25}$
    - v intervalu leží 50 % dat.
- charakteristika závislosti
    - dva znaky; aspoň jeden znak je ordinální: X – počet starších sourozenců (ord.), Y – porodní hmotnost novorozence (int.)
    - Spearmanův koeficient pořadové korelace r_S

- určuje míru **pořadové** závislosti mezi znaky X a Y
- $r_S \in \langle -1; 1 \rangle$
- $r_S > 0 \ldots$ přímá závislost
- $r_S < 0 \ldots$ nepřímá závislost
- $r_S = 0 \ldots$ nezávislost

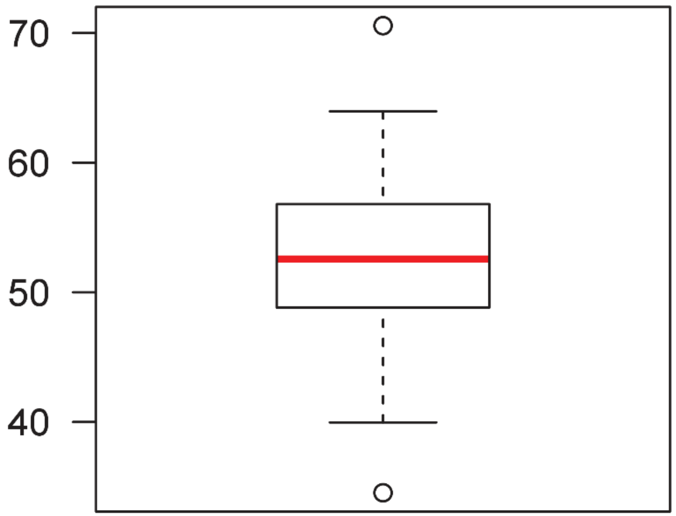




- stupnice míry závislosti podle Spearmanova koeficientu pořadové korelace

$\lvert r_S \rvert$	**Interpretace**
0.0	**pořadová** nezávislost
0.0 – 0.1	velmi nízký stupeň závislosti
0.1 – 0.3	nízký stupeň závislosti
0.3 – 0.5	mírný stupeň závislosti
0.5 – 0.7	význačný stupeň závislosti
0.7 – 0.9	vysoký stupeň závislosti
0.9 – 1.0	velmi vysoký stupeň závislosti
1.0	úplná **pořadová** závislost

- grafická viualizace ordinálních dat
    - krabicový diagram



Příklad 3.3. Základní číselné charakteristiky pro ordinální znak
Načtěte datový soubor 17-anova-newborns-2.txt a odstraňte neznámé hodnoty. Zaměřte se pouze na novorozence ženského pohlaví s vysokou porodní hmotností. Zjistěte dimenzi datové tabulky obsahující údaje o těchto novorozencích a vytvořte tabulku základních číselných charakteristik pro znak X = *počet starších sourozenců*.

Řešení příkladu 3.3

```
data <- read.delim('17-anova-newborns-2.txt', sep = '\t')
head(data, n = 4)
```

```
  edu.M prch.N sex.C weight.C weight.K
1     2      0     m     3470        2
2     2      0     m     3240        2
3     2      0     f     2980        2
4     1      0     m     3280        2
```

```
data <- na.omit(data)
data.FH <- data[data$sex == 'f' & data$weight.C > 4200, ]
dim(data.FH) # 14 x 5
```

Po odstranění neznámých hodnot obsahuje datová tabulka údaje o .... novorozencích ženského pohlaví s vysokou porodní hmotností, přičemž u každého z těchto novorozenců máme záznamy o .... znacích.

Znak X = *počet starších sourozenců novorozence* je příkladem .... dat. V tabulce základních charakteristik budou obsaženy následující charakteristiky: minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota a interkvartilové rozpětí.

1. minimální hodnota x_{min} = ...

```
prch <- data.FH$prch.N  # 0 0 2 0 1 1 0 1 1 2 2 3 1 2
prch.sort <- sort(prch) # 0 0 0 0 1 1 1 1 1 2 2 2 2 3
min <- min(prch) # 0
```

2. dolní kvartil $x_{0.25}$
   - n = ..., α = ...
   - $\alpha \times n$ = ... $\rightarrow$ je / není celé číslo
   - $x_{0.25}$ =

```
n <- length(prch) # 14
alpha <- 0.25
alpha * n # 3.5 ... neni cele cislo -> zaokrouhleni
ceiling(alpha * n) # 4
prch.sort[4] # 0
x0.25 <- quantile(prch, probs = 0.25, type = 2) # 0
```

3. medián $x_{0.50}$
   - n = ..., α = ...
   - $\alpha \times n$ = ... $\rightarrow$ je / není celé číslo
   - $x_{0.50}$ =

```
alpha <- 0.5
alpha * n # 7
(prch.sort[7] + prch[8]) / 2 # 1
x0.5 <- median(prch, type = 2) # 1
x0.5 <- quantile(prch, probs = 0.50, type = 2) # 1
```

4. horní kvartil $x_{0.75}$

   - $n =$ ...., $\alpha =$ ....
   - $\alpha \times n =$ .... $\rightarrow$ je / není celé číslo
   - $x_{0.75} =$

```
alpha <- 0.75
alpha * n # 10.5
ceiling(alpha * n) # 11
prch.sort[11] # 2
x0.75 <- quantile(prch, probs = 0.75, type = 2) # 2
```

5. maximální hodnota $x_{max} =$ ....

```
max <- max(prch) # 3
```

6. interkvartilové rozpětí $IQR = x_{0.75} - x_{0.25} =$ ....

```
IQR <- x0.75 - x0.25 # 2
(tab <- data.frame(min = min, dolni.kv = x0.25, median = x0.5,
                   horni.kv = x0.75, max = max, IQR = IQR))
```

```
    min dolni.kv median horni.kv max IQR
25%   0        0      1        2   3   2
```

Interpretace výsledků: Počet starších sourozenců u novorozenců ženského pohlaví s vysokou porodní hmotností se pohyboval v rozmezí .... – .... Dolní kvartil počtu starších sourozenců nabývá hodnoty ...., tj. .... % novorozenců ženského pohlaví s vysokou porodní hmotností má .... starších sourozenců. Medián počtu starších sourozenců nabývá hodnoty ...., tj. .... % novorozenců ženského pohlaví s vysokou porodní hmotností má .... nebo .... starších sourozenců.

Horní kvartil počtu starších sourozenců nabývá hodnoty ...., tj. .... % novorozenců ženského pohlaví s vysokou porodní hmotností má ...., ... nebo .... starších sourozenců. Hodnota interkvartilového rozpětí je rovna ....

Příklad 3.4. Krabicový diagram

Sestrojte krabicový diagram pro znak X = *počet starších sourozenců* pro novorozence ženského pohlaví s vysokou porodní hmotností. Zaměřte se na vzhled krabicového diagramu a zamyslete se nad tím, kde je v krabicovém diagramu zobrazen medián, dolní kvartil, horní kvartil a mezikvartilové rozpětí.

Řešení příkladu 3.4

```
par(mar = c(2, 3, 1, 1))
boxplot(prch, type = 2, horizontal = F, las = 1, col = 'darkolivegreen1',
        border = 'chartreuse4', medcol = 'darkgreen', xlab = '')
mtext('pocet starsich sourozencu', side = 1, line = 1)
```

Příklad 3.5. Charakteristika závislosti mezi ordinálními znaky
Zaměřme se nyní na oba znaky X = *počet starších sourozenců* a Y = *porodní hmotnost novorozence* najednou. Určete míru závislosti mezi znaky X a Y u novorozenců ženského pohlaví.

Řešení příkladu 3.5
Znak X je .... typu, zatímco znak Y je .... typu $\rightarrow$ ke znaku Y budeme přistupovat jako ke znaku .... typu. Ke stanovení míry závislosti použijeme .... koeficient .... korelace.

Tabulka 3.3: Stupnice míry závislosti podle Spearmanova koeficientu pořadové korelace

$\lvert r_S \rvert$	**Interpretace**
0.0	pořadová nezávislost
0.0 – 0.1	velmi nízký stupeň závislosti
0.1 – 0.3	nízký stupeň závislosti
0.3 – 0.5	mírný stupeň závislosti
0.5 – 0.7	význačný stupeň závislosti
0.7 – 0.9	vysoký stupeň závislosti
0.9 – 1.0	velmi vysoký stupeň závislosti
1.0	úplná pořadová závislost

```
data.F <- data[data$sex == 'f', ]
prch.F <- data.F$prch.N
wei.F <- data.F$weight.C
cor(prch.F, wei.F, method = 'spearman')
```

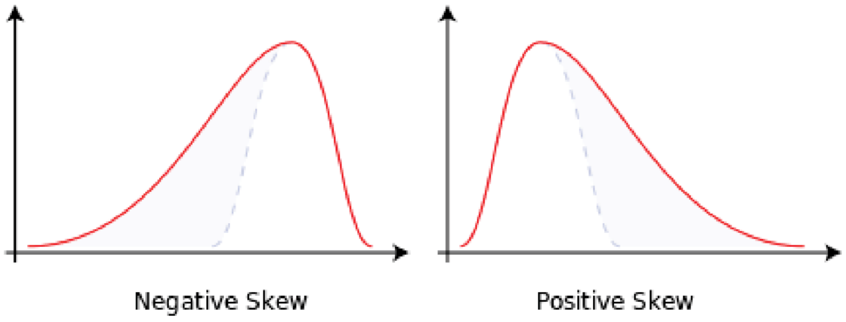
```
[1] 0.05918883
```

Interpretace výsledku: Hodnota Spearmanova koeficientu pořadové korelace vyšla .... Mezi počtem starších sourozenců a porodní hmotností novorozence ženského pohlaví existuje .... stupeň .... závislosti.

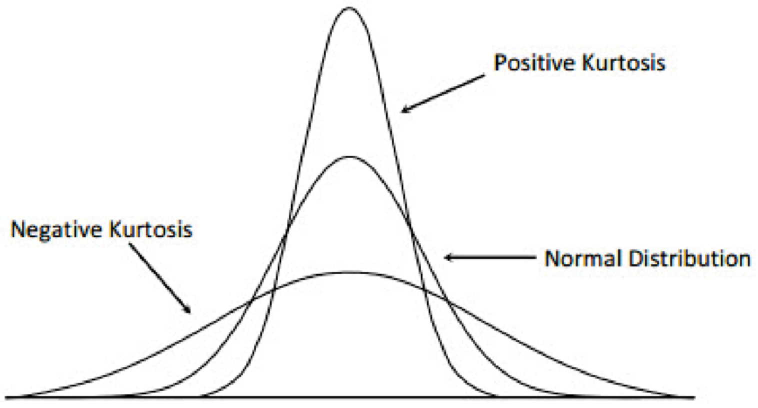
Intervalové znaky

- hodnoty znaků můžeme nejen vzájemně porovnat, ale můžeme též říci, o kolik se liší
  - porodní hmotnost novorozence (v g), největší šířka mozkovny (v mm), BMI (v kg/m^2)
- charakteristiky polohy
  - aritmetický průměr m
    - $m = \frac{1}{n} \sum_{i=1}^{n} x_i$
    - ovlivněn vybočujícími hodnotami $\rightarrow$ vhodný, máme-li symetrická data
  - α-kvantily x_α
    - medián $x_{0.5}$, dolní kvartil $x_{0.25}$, horní kvartil $x_{0.75}$, ...
    - nejsou ovlivněny vybočujícími hodnotami $\rightarrow$ vhodné, máme-li nesymetrická data
- charakteristiky variability
  - rozptyl s^2
    - $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2$
    - průměrná kvadratická odchylka hodnot od jejich aritmetického průměru.
    - $s^2 \geq 0$
    - ovlivněn vybočujícími hodnotami $\rightarrow$ vhodný, máme-li symetrická data
    - rozptyl s^2 $\rightarrow$ jednotky $^\wedge$ 2
  - směrodatná odchylka s
    - $s = \sqrt{s^2}$
    - převádí rozptyl do původních jednotek
  - interkvartilové rozpětí IQR

- charakteristiky nesymetrie
  - šikmost α_3
    - $\alpha_3 = 0 \rightarrow$ symetrické rozdělení dat
    - $\alpha_3 < 0 \rightarrow$ záporně zešikmené$\rightarrow$ prodloužený levý konec
    - $\alpha_3 > 0 \rightarrow$ kladně zešikmené $\rightarrow$ prodloužený pravý konec



  - špičatost α_4
    - $\alpha_4 = 0 \rightarrow$ normální rozdělení dat
    - $\alpha_4 > 0 \rightarrow$ strmé rozdělení dat
    - $\alpha_4 < 0 \rightarrow$ ploché rozdělení dat (Říp)



- charakteristika těsnosti závislosti
  - dva znaky X a Y, oba intervalového typu
  - Pearsonův korelační koeficient r_{12}
    - určuje míru **lineární** závislosti mezi znaky X a Y
    - $r_{12} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - m_1}{s1} \frac{y_i - m_2}{s_2}$
    - $r_{12} \in \langle -1 \,;\, 1 \rangle$
    - $r_{12} > 0 \ldots$ přímá závislost
    - $r_{12} < 0 \ldots$ nepřímá závislost
    - $r_{12} = 0 \ldots$ nezávislost
  - stupnice míry závislosti podle Pearsonova korelačního koeficientu

$\lvert r_{12} \rvert$	**Interpretace**
0.0	**lineární** nezávislost
0.0 – 0.1	velmi nízký stupeň závislosti
0.1 – 0.3	nízký stupeň závislosti
0.3 – 0.5	mírný stupeň závislosti
0.5 – 0.7	význačný stupeň závislosti
0.7 – 0.9	vysoký stupeň závislosti
0.9 – 1.0	velmi vysoký stupeň závislosti
1.0	úplná **lineární** závislost

Příklad 3.6. Základní číselné charakteristiky pro intervalový znak
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Zaměřte se pouze na znak X = *největší šířka mozkovny* pro skelety mužského pohlaví. Vytvořte tabulku základních číselných charakteristik pro znak X.

Řešení příkladu 3.6

```
data <- read.delim('01-one-sample-mean-skull-mf.txt')
head(data, n = 3)
```

```
   id   pop sex skull.L skull.B
1 416 egant   m     188     145
2 417 egant   m     172     139
3 420 egant   m     176     138
```

```
data <- na.omit(data)
skull.BM <- data[data$sex == 'm', 'skull.B']
```

Po odstranění neznámých hodnot obsahuje datová tabulka údaje o .... skeletech mužského pohlaví.

Znak X = *největší šířka mozkovny* pro skelety mužského pohlaví je příkladem .... dat. V tabulce základních číselných charakteristik budou obsaženy následující charakteristiky: aritmetický průměr, směrodatná odchylka, minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota, interkvartilové rozpětí, koeficient šikmosti a koeficient špičatosti.

1. aritmetický průmer m

   - $m = \frac{1}{n}\sum_{i=1}^{n} x_i =$

```
n <- length(skull.BM) # 216
m <- 1 / n * sum(skull.BM) # 137.1852
m <- mean(skull.BM) # 137.1852
```

2. rozptyl s^2

   - $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 =$

```
s2 <- 1 / n * sum((skull.BM - m) ^ 2) # 23.16941
```

3. směrodatná odchylka s

   - $s = \sqrt{s^2} =$

```
s <- sqrt(s2) # 4.813461
```

4. minimální hodnota $x_{min} =$ ....

```
skull.BM.sort <- sort(skull.BM) # 124, 127, 127, ..., 148, 149, 149
min <- min(skull.BM) # 124
```

5. dolní kvartil $x_{0.25}$

   - $n =$ ...., $\alpha =$ ....
   - $\alpha \times n =$ .... $\rightarrow$ je / není celé číslo
   - $x_{0.25} =$

```
alpha <- 0.25
n * alpha # 54
(skull.BM.sort[54] + skull.BM.sort[55]) / 2 # 134
x0.25 <- quantile(skull.BM, probs = 0.25, type = 2) # 134
```

6. medián $x_{0.50}$

- $n =$ ...., $\alpha =$ ....
- $\alpha \times n =$ .... $\rightarrow$ je / není celé číslo
- $x_{0.50} =$

```
alpha <- 0.5
n * alpha # 108
(skull.BM.sort[108] + skull.BM.sort[109]) / 2 # 137
x0.50 <- quantile(skull.BM, probs = 0.50, type = 2) # 137
```

7. horní kvartil $x_{0.75}$

- $n =$ ...., $\alpha =$ ....
- $\alpha \times n =$ .... $\rightarrow$ je / není celé číslo
- $x_{0.75} =$

```
alpha <- 0.75
n * alpha # 162
(skull.BM.sort[162] + skull.BM.sort[163]) / 2 # 140
x0.75 <- quantile(skull.BM, probs = 0.75, type = 2) # 140
```

8. maximální hodnota $x_{max} =$ ....

```
max <- max(skull.BM) # 149
```

9. interkvartilové rozpětí $IQR = x_{0.75} - x_{0.25} =$ ....

```
IQR <- x0.75 - x0.25
```

6. koeficient šikmosti b_1

    - $b_1 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

```r
sikmost <- e1071::skewness(skull.BM, type = 3) # 0.08410943
```

7. koeficient špičatosti b_2

    - $b_2 = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

```r
spicatost <- e1071::kurtosis(skull.BM, type = 3) # -0.2956831
tab <- data.frame(m, s, min, dolni.kv = x0.25, median = x0.50,
                  horni.kv = x0.75, max, IQR, sikmost, spicatost,
                  row.names = c('muzi'))
(tab <- round(tab, digits = 2))
```

```
          m    s min dolni.kv median horni.kv max IQR sikmost spicatost
muzi 137.19 4.81 124      134    137      140 149   6    0.08      -0.3
```

Interpretace výsledků: Naměřené hodnoty největší šířky mozkovny pro skelety mužského pohlaví se pohybují v rozmezí .... – .... mm. Průměrná hodnota největší šířky mozkovny je .... mm se směrodatnou odchylkou .... mm. 25% naměřených hodnot je menších nebo rovných .... mm, 50% naměřených hodnot je menších nebo rovných .... mm, 75% naměřených hodnot je menších nebo rovných .... mm. Interkvartilové rozpětí naměřených hodnot je rovno .... mm. Hodnota koeficientu šikmosti, ..., ukazuje na .... zešikmená data (prodloužený .... konec). Hodnota koeficientu špičatosti, ..., ukazuje na .... charakter dat.

Příklad 3.7. Charakteristika závislosti pro znaky intervalového typu
Zaměřme se nyní na znaky X = *největší šířka mozkovny* a Y = *největší délka mozkovny* pro skelety mužského pohlaví najednou. Určete míru závislosti mezi znaky X a Y.

Řešení příkladu 3.7
Oba znaky X a Y jsou .... typu. Ke stanovení míry závislosti použijeme .... korelační koeficient.

Tabulka 3.4: Stupnice míry závislosti podle Pearsonova korelačního koeficientu

$\lvert r_{12} \rvert$	**Interpretace**
0.0	**lineární** nezávislost
0.0 – 0.1	velmi nízký stupeň závislosti
0.1 – 0.3	nízký stupeň závislosti
0.3 – 0.5	mírný stupeň závislosti
0.5 – 0.7	význačný stupeň závislosti
0.7 – 0.9	vysoký stupeň závislosti
0.9 – 1.0	velmi vysoký stupeň závislosti
1.0	úplná **lineární** závislost

```
skull.BM <- data[data$sex == 'm', 'skull.B']
skull.LM <- data[data$sex == 'm', 'skull.L']
cor(skull.BM, skull.LM, method = 'pearson')
```

```
[1] 0.168157
```

Interpretace výsledků: Pearsonův korelační koeficient nabývá hodnoty .... Mezi největší šířkou mozkovny a největší délkou mozkovny pro skelety mužského pohlaví existuje .... stupeň .... závislosti.

Příklad 3.8. Dvourozměrný tečkový diagram
Pro znaky X = *největší šířka mozkovny* a Y = *největší délka mozkovny* u mužů vykreslete dvourozměrný tečkový diagram.

Řešení příkladu 3.8

```
par(mar = c(3, 4, 2, 2))
plot(skull.BM, skull.LM, pch = 21, bg = 'mintcream', col = 'darkblue',
        xlab = '', ylab = 'nejvetsi delka mozkovny (mm)', main = '', las = 1)
mtext('nejvetsi sirka mozkovny (mm)', side = 1, line = 2.1)
```