3 Číselné charakteristiky datového souboru

3.1 Typy znaků

- **Nominální znak:** umožňuje obsahovou interpretaci pouze u relace rovnosti. O dvou variantách můžeme konstatovat jen to, zda jsou stejné nebo různé.
- **Ordinální znak:** vedle relace rovnosti lze obsahově interpretovat také relaci uspořádání. Varianty znaku můžeme tedy uspořádat podle velikosti.
- **Intervalový znak:** kromě relací rovnosti a uspořádání umožňuje obsahově interpretovat operaci rozdílu. Stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v intenzitě zkoumané vlastnosti. Charakteristická vlastnost: počátek měřicí stupnice byl stanoven konvencí.
- **Poměrový znak:** kromě relací rovnosti a uspořádání a operace rozdílu lze obsahově interpretovat také operaci podílu. Stejný poměr mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný podíl v intenzitě zkoumané vlastnosti. Charakteristická vlastnost: počátek měřicí stupnice je přirozený.
- **Alternativní znak:** stojí mimo uvedenou stupnici. Nabývá jen dvou hodnot, např. 0 a 1. Přitom 0 znamená nepřítomnost nějaké vlastnosti, 1 znamená přítomnost této vlastnosti. Může být ztotožněn s kterýmkoliv jiným typem znaku.

Upozornění: Číselné charakteristiky, které jsou určeny pro nižší typ znaku, mohou být použity pro vyšší typ znaku, ale naopak to není přípustné.

3.2 Číselné charakteristiky nominálních znaků

3.2.1 Charakteristika polohy

Modus – nejčetnější varianta, resp. střed nejčetnějšího třídicího intervalu.

3.2.2 Charakteristika těsnosti závislosti dvou znaků

Cramérův koeficient V. Počítá se na základě znalosti simultánních absolutních četností n_{jk} zapsaných v kontingenční tabulce.

$$V = \sqrt{\frac{K}{n(m-1)}},$$

kde $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.} n_{.k}}{n}\right)^2}{\frac{n_{j.} n_{.k}}{n}}$, n je rozsah datového souboru a $m = \min\{r, s\}$. Číslo $\frac{n_{j.} n_{.k}}{n}$ se nazývá teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})^T$. Cramérův koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi znaky, čím blíže je 0, tím je tato závislost volnější. Stupně závislosti podle hodnoty Cramérova koeficientu jsou uvedeny v tabulce 3.1.

Tabulka 3.1: Stupnice míry závislosti pro Cramérův koeficient

Cramérův koeficient V	**Interpretace**
$\langle 0,0; 0,1)$	zanedbatelný stupeň závislosti
$\langle 0,1; 0,3)$	slabý stupeň závislosti
$\langle 0,3; 0,7)$	střední stupeň závislosti
$\langle 0,7; 1,0\rangle$	silný stupeň závislosti

Příklad 3.1. Řešený příklad
Načtěte datový soubor 22-multinom-palmar-lines.txt obsahující údaje o zakončení tří hlavních dlaňových linií (Lo –

nízké; Mi – střední; Hi – vysoké) a údaje o odstínu barvy vlasů (LiH – světlý; MH – střední; DaH – tmavý) u mužů a žen. Za předpokladu, že znak X popisuje odstín barvy vlasů a znak Y popisuje zakončení tří hlavních dlaňových linií u mužů, vypočítejte (a) modus znaku X, resp. znaku Y; (b) Cramérův koeficient V. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkladu 3.1
Datový soubor načteme příkazem read.delim().

```
data <- read.delim("22-multinom-palmar-lines.txt", sep = "\t", row.names = 1)
```

```
    Lo.m Mi.m Hi.m Lo.f Mi.f Hi.f
LiH    4    6    6    6    6    4
MH     7   15   20   10   10   18
DaH   12   12   18   12   22   12
```

Pomocí operátoru [ ] vybereme z datové tabulky pouze sloupce týkající se mužů. Výslednou tabulku vložíme do proměnné data.M. Pro nalezení modu znaku X je třeba nejprve vypočítat absolutní četnosti jednotlivých variant znaku X, které odpovídají řádkovým součtům v tabulce data.M. Řádkové součty vypočítáme příkazem apply() s argumenty MARGIN = 1 a FUN = sum. Modem znaku X bude varianta s největší absolutní četností.

```
data.M <- data[, 1:3]
apply(data.M, MARGIN = 1, FUN = sum)
```

```
LiH  MH DaH
 16  42  42
```

Nejčetnějšími variantami (mody) znaku X jsou střední odstín barvy vlasů (MH) a tmavý odstín barvy vlasů (DaH), obě s absolutní četností 42.

Analogicky zjistíme hodnotu modu znaku Y. Nejprve vypočítáme absolutní četnosti jednotlivých variant znaku Y, které odpovídají sloupcovým součtům v tabulce data.M. Sloupcové součty vypočítáme příkazem apply() s argumenty MARGIN = 2 a FUN = sum. Modem znaku Y bude varianta s největší četností.

```
apply(data.M, MARGIN = 2, FUN = sum)
```

```
Lo.m Mi.m Hi.m
  23   33   44
```

Nejčetnější variantou (modem) znaku Y je vysoké zakončení tří hlavních dlaňových linií (Hi) s absolutní četností 44.

Cramérův koeficient vypočítáme příkazem cramersV() implementovaným v knihovně lsr.

```
V <- lsr::cramersV(data.M) # 0,1014841
```

Mezi odstínem barvy vlasů a zakončením tří hlavních dlaňových linií u mužů existuje slabý stupeň závislosti ($V = 0,1015$). ★

Příklad 3.2. Neřešený příklad
Načtěte datový soubor 20-more-samples-probabilities-pubis.txt obsahující údaje o původu žen (european – evropský; african – africký; inuits – inuitský) a o míře změn kostního reliéfu na vnitřní straně stydké kosti v blízkosti stydké spony (absence – nepřítomnost změn; trace.to small – stopové až malé změny; moderate.to.large – střední až výrazné změny). Za předpokladu, že znak X popisuje původ žen a znak Y popisuje míru změny kostního reliéfu u těchto žen, vypočítejte (a) modus znaku X, resp. znaku Y; (b) Cramérův koeficient V. Všechny vypočítané hodnoty řádně interpretujte.
Výsledky: (a) modus znaku X: africký původ (s absolutní četností 110), modus znaku Y: nepřítomnost změn kostního reliéfu (s absolutní četností 102); (b) $V = 0,1517$, slabý stupeň závislosti. ★

3.3 Číselné charakteristiky ordinálních znaků

3.3.1 Charakteristika polohy

α-kvantil (značíme x_α), kde $\alpha \in (0;1)$. Počítá se na základě uspořádaného datového souboru rozsahu n takto:

$$n\alpha = \begin{cases} \text{celé číslo } c \to x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}, \\ \text{necelé číslo } \to \text{ zaokrouhlíme nahoru na nejbližší celé číslo } c \to x_\alpha = x_{(c)}. \end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}, \ldots, x_{0,9}$ – decily, $x_{0,01}, \ldots, x_{0,99}$ – percentily.

3.3.2 Charakteristika variability

Interkvartilové rozpětí IQR (též mezikvartilové rozpětí nebo kvartilová odchylka): $IQR = x_{0,75} - x_{0,25}$ (značí se též q, viz definice krabicového diagramu v sekci 3.4.6).

3.3.3 Charakteristika těsnosti pořadové závislosti dvou znaků

Spearmanův koeficient pořadové korelace r_S. Vyžaduje zavedení pojmu pořadí čísla v posloupnosti čísel $x_1, \ldots, x_n$:

a) jsou-li čísla navzájem různá, pak pořadím r_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i,

b) vyskytují-li se mezi danými čísly shodná čísla (tzv. *shody*, angl. *ties*), pak všem shodným číslům přiřadíme průměrné pořadí.

Ve dvourozměrném datovém souboru o rozsahu n označíme r_i pořadí hodnoty x_i a q_i pořadí hodnoty y_i, $i = 1, \ldots, n$. Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (r_i - q_i)^2$. Spearmanův koeficient pořadové korelace používáme pro kvantifikaci monotónního vztahu dvou znaků. Koeficient nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi znaky X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi znaky X a Y. Je-li $r_S = 1$, resp. $r_S = -1$, pak ve dvourozměrném tečkovém diagramu leží dvojice $(x_i, y_i)^T$ na nějaké rostoucí, resp. klesající křivce. Stupně závislosti podle absolutní hodnoty Spearmanova koeficientu pořadové korelace jsou uvedeny v tabulce 3.2. V souvislosti se Spearmanovým koeficientem pořadové korelace hovoříme o pořadové závislosti.

Poznámka: Spearmanův koeficient pořadové korelace lze použít na kvantifikaci monotónního vztahu mezi dvěma ordinálními znaky, mezi ordinálním a intervalovým znakem, nebo mezi dvěma intervalovými či poměrovými znaky.

Tabulka 3.2: Stupnice míry závislosti pro Spearmanův koeficient pořadové korelace r_S (resp. pro Pearsonův koeficient korelace r_{12} (viz sekce 3.4.5))

$\lvert r_S \rvert$, **resp.** $\lvert r_{12} \rvert$	**Interpretace**
$0,0$	pořadová (resp. lineární) nezávislost
$(0,0;0,1\rangle$	velmi nízký stupeň závislosti
$(0,1;0,3\rangle$	nízký stupeň závislosti
$(0,3;0,5\rangle$	mírný stupeň závislosti
$(0,5;0,7\rangle$	význačný stupeň závislosti
$(0,7;0,9\rangle$	vysoký stupeň závislosti
$(0,9;1,0)$	velmi vysoký stupeň závislosti
$1,0$	úplná pořadová (resp. lineární) závislost

3.3.4 Grafické znázornění ordinálních dat

Vztah mezi znaky X a Y vizualizujeme pomocí dvourozměrného tečkového diagramu.

Příklad 3.3. Řešený příklad
Načtěte datový soubor 28-one-world-2014.csv obsahující odpovědi respondentů (studentů středních škol) na otázku *Jak často nakupujete v obchodních centrech?* (a.shop; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy) a na otázku *Jak často chodíte do kina?* (a.cinema; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy). Za předpokladu, že znak X popisuje odpověď na otázku *Jak často nakupujete v obchodních centrech?* a znak Y popisuje odpověď na otázku *Jak často chodíte do kina?*, (a) vytvořte tabulku základních číselných charakteristik pro znak X, resp. pro znak Y; (b) vytvořte kontingenční tabulku simultánních absolutních četností pro znaky X a Y a nakreslete dvourozměrný tečkový diagram; (c) vypočítejte Spearmanův koeficient pořadové korelace r_S. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkladu 3.3
Datový soubor načteme příkazem read.delim() s argumentem sep = ";". Pomocí funkce head() si vypíšeme prvních pět řádků a prvních dvanáct sloupců tabulky.

```
data <- read.delim("28-one-world-2014.csv", sep = ";")
head(data, n = c(5, 12))
```

```
  id sex age edu.M edu.F a.course a.school a.parttime a.friends a.shop a.cinema a.concert
1  1   m  19     3     4        5        2          3         1      3        4         4
2  2   f  18     3     2        3        2          4         1      2        4         3
3  3   m  17     2     3        5        4          3         2      4        5         5
4  4   m  16     3     1        2        3         NA         2      3        4         4
5  5   m  16     3     3        5        3          5         1      4        3         5
```

Z načtené tabulky vybereme pouze sloupce a.shop a a.cinema a příkazem na.omit() z tohoto výběru odstraníme řádky s chybějícími hodnotami. V dalším kroku si z výběru separujeme hodnoty proměnné a.shop a hodnoty proměnné a.cinema.

```
data.SC <- na.omit(data[, c("a.shop", "a.cinema")])
a.shop <- data.SC$a.shop
a.cinema <- data.SC$a.cinema
```

Nyní se zaměříme na vytvoření tabulky základních číselných charakteristik pro znak X. Tabulka bude obsahovat rozsah datového souboru, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu a interkvartilové rozpětí. Rozsah datového souboru zjistíme příkazem length(), minimální a maximální naměřenou hodnotu příkazy min() a max(). Dolní kvartil, medián a horní kvartil vypočítáme najednou pomocí funkce quantile() s argumentem probs = c(0.25, 0.50, 0.75) a s argumentem type = 2, který určuje, že hodnoty všech tří kvantilů se vypočítají způsobem popsaným v sekci 3.3.1. Interkvartilové rozpětí vypočítáme příkazem IQR() opět s argumentem type = 2. Všechny číselné charakteristiky vložíme do souhrnné tabulky, kterou vytvoříme příkazem data.frame(). Argumentem row.names() specifikujeme název řádku tabulky.

```
n <- length(a.shop) # 1090
min.S <- min(a.shop) # 1
q.S <- quantile(a.shop, probs = c(0.25, 0.50, 0.75), type = 2) # 2; 3; 3
max.S <- max(a.shop) # 5
iqr.S <- IQR(a.shop, type = 2) # 1
tab.S <- data.frame(n, min = min.S, x0.25 = q.S[1], x0.50 = q.S[2], x0.75 = q.S[3], max
    = max.S, IQR = iqr.S, row.names = "znak X")
```

```
          n min x0.25 x0.50 x0.75 max IQR
znak X 1090   1     2     3     3   5   1
```

Základní číselné charakteristiky pro znak X byly počítány na základě 1090 získaných odpovědí. Na otázku *Jak často nakupujete v obchodních centrech?* bylo možné odpovědět jednou z pěti možností od odpovědi *velmi často* (min = 1) po odpověď *nikdy* (max = 5). 25 % respondentů nakupuje v obchodních centrech celkem často nebo velmi často. 50 % respondentů nakupuje v obchodních centrech občas nebo častěji. 75 % respondentů nakupuje v obchodních centrech občas nebo častěji. 50 % prostředních hodnot odpovědí v uspořádaném datovém souboru se

nachází v intervalu o délce 1.

Analogickým způsobem vypočítáme tabulku základních číselných charakteristik pro znak Y.

```
min.C <- min(a.cinema) # 1
q.C <- quantile(a.cinema, probs = c(0.25, 0.50, 0.75), type = 2) # 3; 3; 4
max.C <- max(a.cinema) # 5
iqr.C <- IQR(a.cinema, type = 2) # 1
tab.C <- data.frame(n, min = min.C, y0.25 = q.C[1], y0.50 = q.C[2], y0.75 = q.C[3], max
    = max.C, IQR = iqr.C, row.names = "znak Y")
```

```
          n min y0.25 y0.50 y0.75 max IQR
znak Y 1090   1     3     3     4   5   1
```

Základní číselné charakteristiky pro znak Y byly počítány na základě 1090 získaných odpovědí. Na otázku *Jak často chodíte do kina?* bylo možné odpovědět jednou z pěti možností od odpovědi *velmi často* (min = 1) po odpověď *nikdy* (max = 5). 25 % respondentů chodí do kina občas nebo častěji. 50 % respondentů chodí do kina občas nebo častěji. 75 % respondentů chodí do kina výjimečně nebo častěji. 50 % prostředních hodnot odpovědí v uspořádaném souboru se nachází v intervalu o délce 1.

Kontingenční tabulku simultánních absolutních četností vypočítáme příkazem table(). Vstupními argumenty příkazu budou vektory a.shop a a.cinema. Výstupem příkazu je tabulka typu table. Tuto tabulku převedeme na tabulku typu data.frame pomocí funkce as.data.frame.matrix(). Argumentem row.names specifikujeme názvy řádků tabulky. Názvy sloupců doplníme samostatným příkazem names().

```
KT.abs <- table(a.shop, a.cinema)
KT.abs <- as.data.frame.matrix(KT.abs, row.names = c("velmi často", "celkem často",
    "občas", "výjimečně", "nikdy"))
names(KT.abs) <- c("velmi často", "celkem často", "občas", "výjimečně", "nikdy")
```

```
             velmi často celkem často občas výjimečně nikdy
velmi často           15           19    69         29     3
celkem často           8           37   140         77     6
občas                  4           27   156        231    17
výjimečně              0            7    60        130    30
nikdy                  0            1     4          9    11
```

Celkem 15 respondentů uvedlo, že v obchodních centrech nakupují velmi často a velmi často chodí i do kina, 17 respondentů uvedlo, že občas nakupují v obchodních centrech, ale do kina nechodí nikdy, apod.

Rozložení simultánních absolutních četností vizualizujeme tečkovým diagramem. Nejprve nastavíme okraje grafu tak, aby se pod osu x a vedle osy y vešly popisky variant znaku X a znaku Y. Příkazem par() s argumentem mar = c(6, 7, 2, 2) specifikujeme, že mezi dolním (resp. levým, horním, pravým) okrajem grafu a okrajem obrázku bude místo na šest (resp. sedm, dva, dva) řádků textu.

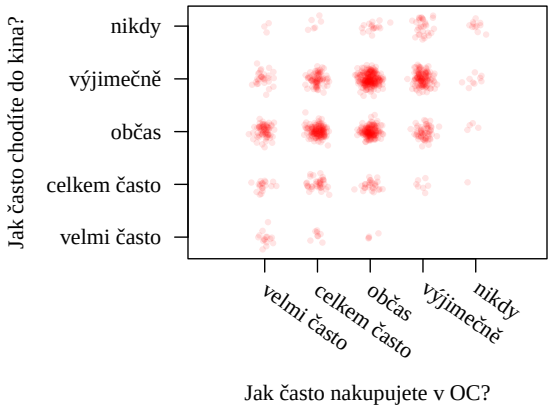
```
par(mar = c(6, 7, 2, 2))
```

Před vykreslením samotného tečkového diagramu je třeba vzít v potaz, že možných kombinací variant znaků X a Y je $5 \times 5 = 25$, zatímco rozsah datového souboru $n = 1090$. Odpovědi respondentů by tedy v klasickém tečkovém diagramu vykresleném příkazem plot() splývaly. Tečkový diagram proto vykreslíme pomocí funkce dotplot() implementované v  skriptu AS1-sbirka-funkce.R. Funkce dotplot() nejprve přičte ke každé zaznamenané odpovědi zanedbatelně malé pseudonáhodně vygenerované číslo, takže získané odpovědi respondentů nebudou ve výsledném tečkovém diagramu v překryvu. Vstupními argumenty funkce budou vektory a.shop a a.cinema. Dále ve funkci nastavíme velikost směrodatné odchylky pseudonáhodně vygenerovaných čísel (sd = 0.1), poměr stran osy x ku ose y jako jedna ku jedné (asp = T), potlačení vykreslení os x a y (axes = F), kruhový typ bodu s obrysem a výplní (pch = 19), poloviční velikost bodů (cex = 0.5), barvu bodů (col), popisek osy x (xlab) a popisek osy y (ylab). Barvu bodů přitom definujeme pomocí funkce rgb(red, green, blue, alpha). Všechny zmíněné argumenty nabývají hodnot z intervalu $\langle 0; 1 \rangle$, přičemž argumenty red, green a blue určují podíl červené, zelené a modré složky ve výsledné barvě (čím je hodnota větší, tím větší je podíl barevné složky) a argument alpha určuje průhlednost barvy (čím je hodnota větší, tím méně je výsledná barva průhledná).

```
source("AS1-sbirka-funkce.R")
dotplot(a.shop, a.cinema, sd = 0.1, asp = T, axes = F, pch = 19, cex = 0.5, col =
    rgb(1, 0, 0, 0.1), xlab = "", ylab = "")
```

Osy x a y dokreslíme do grafu zvlášť příkazem axis(), přičemž u osy x potlačíme vypsání popisků variant nastavením argumentu labels = NA. Popisky variant doplníme samostatně příkazem text(), kde pomocí argumentů x a y definujeme x-ové a y-ové souřadnice umístění popisků. V příkazu dále specifikujeme text popisků (labels), vykreslení popisků vně grafu, tj. v okrajové části obrázku (xpd = T), úhel otočení popisků o 35° ve směru hodinových ručiček (srt = -35) a zarovnání popisků do levého horního rohu (adj = 0). Nakonec pomocí funkce mtext() definujeme vypsání popisku osy x (side = 1) na pátém řádku od spodního okraje grafu (line = 5) a vypsání popisku osy y (side = 2) na šestém řádku od levého okraje grafu (line = 6). Výsledný graf je zobrazen na obrázku 3.1.

```
box(bty = "o")
axis(1, at = 1:5, labels = NA)
axis(2, at = 1:5, labels = c("velmi často", "celkem často", "občas", "výjimečně",
    "nikdy"), las = 1)
text(x = 1:5, y = 0.4, labels = c("velmi často", "celkem často", "občas", "výjimečně",
    "nikdy"), xpd = T, srt = -35, adj = 0)
mtext("Jak často nakupujete v OC?", side = 1, line = 5)
mtext("Jak často chodíte do kina?", side = 2, line = 6)
```



Obrázek 3.1: Dvourozměrný tečkový diagram pro odpovědi na otázku *Jak často nakupujete v obchodních centrech?* a na otázku *Jak často chodíte do kina?*

Nakonec vypočítáme Spearmanův koeficient pořadové korelace r_S příkazem cor() s argumentem method = "spearman".

```
rS <- cor(a.shop, a.cinema, method = "spearman") # 0,3817639
```

Mezi odpovědí na otázku *Jak často nakupujete v obchodních centrech?* a na otázku *Jak často chodíte do kina?* existuje mírný stupeň přímé pořadové závislosti ($r_S = 0,3818$).
★

Příklad 3.4. Neřešený příklad
Načtěte datový soubor 28-one-world-2014.csv obsahující odpovědi respondentů (studentů středních škol) na otázku *Jak často jste někde s kamarády?* (a.friends; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy) a na otázku *Jak často sledujete zprávy v médiích?* (c.news; 1 – pravidelně; 2 – občas; 3 – téměř nikdy). Za předpokladu, že znak X popisuje odpověď na otázku *Jak často jste někde s kamarády?* a znak Y popisuje odpověď na otázku *Jak často sledujete zprávy v médiích?*, (a) vytvořte tabulku základních číselných charakteristik pro znak X, resp. pro znak Y; (b) vytvořte kontingenční tabulku simultánních absolutních četností pro znaky X a Y a nakreslete dvourozměrný tečkový diagram; (c) vypočítejte Spearmanův koeficient pořadové korelace r_S. Všechny vypočítané

hodnoty řádně interpretujte.

Výsledky: (a) tabulka základních číselných charakteristik pro znak X viz tabulka 3.3 a pro znak Y viz tabulka 3.4; (b) kontingenční tabulka simultánních absolutních četností pro znaky X a Y viz tabulka 3.5, dvourozměrný tečkový diagram viz obrázek 3.2; (c) $r_S = -0,0178$, velmi nízký stupeň nepřímé pořadové závislosti.

Tabulka 3.3: Základní číselné charakteristiky pro odpovědi na otázku *Jak často jste někde s kamarády?*

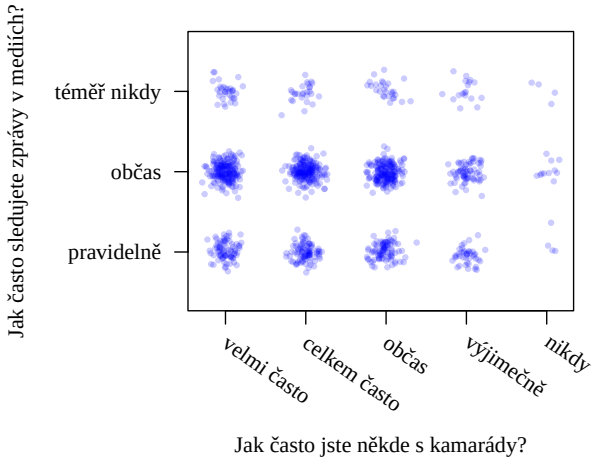
	n	min	$x_{0,25}$	$x_{0,50}$	$x_{0,75}$	max	IQR
znak X	1087	1	1	2	3	5	2

Tabulka 3.4: Základní číselné charakteristiky pro odpovědi na otázku *Jak často sledujete zprávy v médiích?*

	n	min	$y_{0,25}$	$y_{0,50}$	$y_{0,75}$	max	IQR
znak Y	1087	1	1	2	2	3	1

Tabulka 3.5: Kontingenční tabulka simultánních absolutních četností pro odpovědi na otázku *Jak často jste někde s kamarády?* (viz řádky) a na otázku *Jak často sledujete zprávy v médiích?* (viz sloupce)

	pravidelně	občas	téměř nikdy
velmi často	81	197	33
celkem často	93	203	27
občas	91	189	28
výjimečně	41	67	18
nikdy	4	11	4



Obrázek 3.2: Dvourozměrný tečkový diagram pro odpovědi na otázku *Jak často jste někde s kamarády?* a na otázku *Jak často sledujete zprávy v médiích?*

★

3.4 Číselné charakteristiky intervalových a poměrových znaků

3.4.1 Charakteristika polohy

Aritmetický průměr $m = \frac{1}{n}\sum_{i=1}^{n} x_i$. U poměrových znaků, které nabývají jen kladných hodnot, lze použít geometrický průměr $\sqrt[n]{x_1 \dots x_n}$. Pomocí aritmetického průměru se zavede i-tá centrovaná hodnota znaku X: $x_i - m$.

3.4.2 Charakteristika variability

Rozptyl $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m^2$, resp. směrodatná odchylka $s_n = \sqrt{s_n^2}$. U poměrových znaků lze jako charakteristiku variability použít koeficient variace $cv = \frac{s_n}{m}$. Je to bezrozměrné číslo, často se vyjadřuje v procentech a používá se při porovnání variability několika datových souborů. Pomocí aritmetického průměru a směrodatné odchylky se zavede i-tá standardizovaná hodnota znaku X: $\frac{x_i - m}{s_n}$.

Poznámka: Rozptyl můžeme také vypočítat pomocí vzorce $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$. Směrodatná odchylka $s_{n-1} = \sqrt{s_{n-1}^2}$ (viz kapitola 6). Výsledná hodnota rozptylu s_n^2 se bude od hodnoty rozptylu s_{n-1}^2 mírně lišit, neboť i vzorce se liší. Totéž platí pro směrodatné odchylky s_n a s_{n-1}.

3.4.3 Charakteristika nesymetrie

Koeficient šikmosti $g_1 = \frac{m_3}{\sqrt{m_2^3}}$, kde m_2 je druhý centrální moment a m_3 je třetí centrální moment. Koeficient špičatosti $g_2 = \frac{m_4}{m_2^2} - 3$, kde m_4 je čtvrtý centrální moment. Hodnotu p-tého centrálního momentu, $p \geq 2$, přitom vypočítáme pomocí vzorce $m_p = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^p$, kde m je aritmetický průměr a n je rozsah datového souboru. Ze vzorce je zjevné, že $m_2 = s_n^2$. Koeficient šikmosti, resp. špičatosti můžeme vypočítat v softwaru  příkazem skewness(), resp. kurtosis() s argumentem type = 1. Oba příkazy pochází z knihovny e1071.

Poznámka: Koeficient šikmosti můžeme také vypočítat pomocí vzorce $b_1 = \frac{m_3}{s_{n-1}^3} = g_1\sqrt{\left(\frac{n-1}{n}\right)^3}$, koeficient špičatosti podle vzorce $b_2 = \frac{m_4}{s_{n-1}^4} - 3 = (g_2 + 3)\left(1 - \frac{1}{n}\right)^2 - 3$, kde s_{n-1}^p je p-tá mocnina směrodatné odchylky s_{n-1}. Koeficient šikmosti b_1, resp. špičatosti b_2 můžeme vypočítat příkazem e1071::skewness(), resp. e1071::kurtosis() s argumentem type = 3. Výsledná hodnota koeficientu šikmosti b_1 se bude od hodnoty koeficientu šikmosti g_1 mírně lišit, neboť i vzorce se liší. Totéž platí pro koeficienty špičatosti b_2 a g_2.

3.4.4 Charakteristika společné variability dvou znaků

Kovariance $s_{n,12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2$, kde m_1 a m_2 jsou aritmetické průměry znaků X a Y. Je-li $s_{n,12} = 0$, pak řekneme, že znaky X a Y jsou nekorelované.

3.4.5 Charakteristika těsnosti lineární závislosti dvou znaků

Pearsonův koeficient korelace $r_{12} = \begin{cases} \frac{s_{n,12}}{s_{n,1}s_{n,2}} \text{ pro } s_{n,1}s_{n,2} > 0, \\ 0 \text{ jinak,} \end{cases}$

kde $s_{n,1}$, resp. $s_{n,2}$ je směrodatná odchylka znaku X, resp. znaku Y.

Pearsonův koeficient korelace používáme pro kvantifikaci lineárního vztahu dvou znaků. Koeficient nabývá hodnot mezi -1 a 1. Čím je bližší 1, tím je silnější přímá lineární závislost mezi znaky X a Y, čím je bližší -1, tím je silnější nepřímá lineární závislost mezi znaky X a Y. Je-li $r_{12} = 1$, resp. $r_{12} = -1$, pak ve dvourozměrném tečkovém diagramu leží dvojice $(x_i, y_i)^T$ na přímce s kladnou, resp. zápornou směrnicí. Stupně závislosti podle absolutní hodnoty Pearsonova koeficientu korelace jsou analogické jako u Spearmanova koeficientu pořadové korelace (viz tabulka 3.2), hovoříme však o lineární závislosti.

Známe-li absolutní četnosti n_j, n_k, resp. relativní četnosti p_j, p_k variant $x_{[j]}$, $y_{[k]}$, resp. třídicích intervalů (se středy $x_{[j]}$, $y_{[k]}$) a simultánní absolutní četnosti n_{jk}, resp. simultánní relativní četnosti p_{jk}, pak počítáme vážený aritmetický průměr $m_w = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} = \sum_{j=1}^{r} p_j x_{[j]}$, vážený rozptyl $s_w^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - m_w)^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^2 - m_w^2 = \sum_{j=1}^{r} p_j x_{[j]}^2 - m_w^2$, váženou směrodatnou odchylku $s_w = \sqrt{s_w^2}$ a váženou kovarianci $s_{w,12} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk}(x_{[j]} - m_{w,1})(y_{[k]} - m_{w,2}) = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk} x_{[j]} y_{[k]} - m_{w,1}m_{w,2} = \sum_{j=1}^{r}\sum_{k=1}^{s} p_{jk} x_{[j]} y_{[k]} - m_{w,1}m_{w,2}$.

3.4.6 Grafické znázornění intervalových dat

Krabicový diagram (boxplot). Při jeho konstrukci potřebujeme znát medián, dolní kvartil, horní kvartil, minimum, maximum, vnitřní hradby a vnější hradby. Dolní vnitřní hradba $= x_{0,25} - 1,5q$, horní vnitřní hradba $= x_{0,75} + 1,5q$, dolní vnější hradba $= x_{0,25} - 3q$, horní vnější hradba $= x_{0,75} + 3q$. Dolní, resp. horní hrana krabičky je ve výši dolního, resp. horního kvartilu, zesílená vodorovná čára uvnitř krabičky odpovídá mediánu. Dolní, resp. horní svislá úsečka vycházející z dolní, resp. horní hrany krabičky končí ve výši max{minimum, dolní vnitřní hradba}, resp. min{maximum, horní vnitřní hradba}. Hodnoty ležící mezi vnitřními a vnějšími hradbami se nazývají odlehlé, hodnoty ležící za vnějšími hradbami se nazývají extrémní.

Vztah mezi znaky X a Y vizualizujeme pomocí dvourozměrného tečkového diagramu.

Příklad 3.5. Řešený příklad
Načtěte datový soubor 31-goldman-alaska-hfrt.csv obsahující údaje o délce stehenní kosti z levé strany v mm (femur.L) a délce pažní kosti z levé strany v mm (humerus.L) u mužů a žen tří aljašských populací. Za předpokladu, že znak X popisuje délku stehenní kosti z levé strany u žen a znak Y popisuje délku pažní kosti z levé strany u žen z kmene Tigara (a) vytvořte tabulku základních číselných charakteristik pro znak X, resp. pro znak Y; (b) nakreslete krabicový diagram pro znak X, resp. pro znak Y; (c) vypočítejte kovarianci $s_{n,12}$ a Pearsonův koeficient korelace r_{12} a nakreslete dvourozměrný tečkový diagram. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkladu 3.5
Datový soubor načteme a vypíšeme prvních pět řádků a deset sloupců tabulky.

```
data <- read.delim("31-goldman-alaska-hfrt.csv", sep = ";", dec = ".")
head(data, n = c(5, 10))
```

```
  loc pop sex humerus.L humerus.R humerus.HDL humerus.HDR humerus.ADL humerus.ADR femur.L
1 ala tig   m     308,5        NA       47,55          NA       22,00          NA     443
2 ala ipi   m     311,0       310       44,44       44,11       22,12       22,68     415
3 ala ipi   m     289,0       298       42,94       44,41       20,36       22,09     398
4 ala ipi   f     295,0       302       42,51       42,06       19,35       19,97     395
5 ala ipi   f     270,5       281       39,74       39,84       19,42       19,38      NA
```

Z načtené tabulky vybereme pouze řádky týkající se žen z kmene Tigara a sloupce femur.L a humerus.L. Z výběru odstraníme řádky s chybějícími hodnotami a separujeme naměřené délky stehenních kostí a naměřené délky pažních kostí.

```
data.F <- na.omit(data[data$sex == "f" & data$pop == "tig", c("femur.L", "humerus.L")])
femur.LF <- data.F$femur.L
humerus.LF <- data.F$humerus.L
```

Nyní se zaměříme na vytvoření tabulky základních číselných charakteristik pro znak X. Tabulka bude obsahovat rozsah datového souboru, aritmetický průměr, směrodatnou odchylku, koeficient variace, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu, interkvartilové rozpětí, koeficient šikmosti a koeficient špičatosti. Rozsah datového souboru zjistíme příkazem length(). Aritmetický průměr vypočítáme příkazem mean(). Směrodatnou odchylku vypočítáme přepisem vzorce $s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2}$, přičemž operaci odmocnění provedeme příkazem sqrt() a operaci součtu příkazem sum(). Koeficient variace získáme jako podíl směrodatné odchylky a aritmetického průměru. Minimální a maximální naměřenou hodnotu nalezneme pomocí funkcí min() a max(). Dolní kvartil, medián a horní kvartil vypočítáme najednou příkazem quantile() s argumenty probs = c(0.25, 0.50, 0.75) a type = 2. Interkvartilové rozpětí vypočítáme příkazem IQR(). Hodnotu koeficientu šikmosti, resp. špičatosti zjistíme příkazem skewness(), resp. kurtosis() z knihovny e1071 s argumentem type = 1 (viz sekce 3.4.3). Všechny číselné charakteristiky vložíme do souhrnné tabulky.

```
n <- length(femur.LF) # 23
m.f <- mean(femur.LF) # 392,8696
sn.f <- sqrt(1 / n * sum((femur.LF - m.f) ^ 2)) # 15,67553
cv.f <- sn.f / m.f # 0,03990008
min.f <- min(femur.LF) # 365
q.f <- quantile(femur.LF, probs = c(0.25, 0.50, 0.75), type = 2) # 385; 391; 403
max.f <- max(femur.LF) # 427
```

```
77 iqr.f <- IQR(femur.LF, type = 2) # 18
78 g1.f <- e1071::skewness(femur.LF, type = 1) # 0,4756138
79 g2.f <- e1071::kurtosis(femur.LF, type = 1) # -0,1459509
80 tab.f <- data.frame(n, m = m.f, sn = sn.f, cv = cv.f, min = min.f, x0.25 = q.f[1],
       x0.50 = q.f[2], x0.75 = q.f[3], max = max.f, IQR = iqr.f, g1 = g1.f, g2 = g2.f,
       row.names = "znak X")
```

```
81         n      m    sn   cv min x0.25 x0.50 x0.75 max IQR   g1    g2
82 znak X 23 392,87 15,68 0,04 365   385   391   403 427  18 0,48 -0,15
```

Základní číselné charakteristiky byly počítány na základě 23 naměřených hodnot délky stehenní kosti (v mm) z levé strany u žen z kmene Tigara. Průměrná délka stehenní kosti z levé strany je 392,9 mm se směrodatnou odchylkou 15,7 mm. Směrodatná odchylka tvoří 4,0 % aritmetického průměru. Naměřené hodnoty se pohybují v rozmezí 365,0 až 427,0 mm. 25 % naměřených hodnot je menších nebo rovných 385,0 mm, 50 % naměřených hodnot je menších nebo rovných 391,0 mm a 75 % naměřených hodnot je menších nebo rovných 403,0 mm. 50 % prostředních hodnot v uspořádaném datovém souboru leží v intervalu o šířce 18,0 mm. Hodnota koeficientu šikmosti ukazuje na kladně zešikmené rozložení dat s prodlouženým pravým koncem ($g_1 = 0,48$). Hodnota koeficientu špičatosti ukazuje na mírně zploštělé rozložení dat ($g_2 = -0,15$).

Analogickým způsobem vytvoříme tabulku základních číselných charakteristik pro znak Y.

```
83 m.h <- mean(humerus.LF) # 275,7826
84 sn.h <- sqrt(1 / n * sum((humerus.LF - m.h) ^ 2)) # 9,500373
85 cv.h <- sn.h / m.h # 0,03444878
86 min.h <- min(humerus.LF) # 249,5
87 q.h <- quantile(humerus.LF, probs = c(0.25, 0.50, 0.75), type = 2) # 269; 275,5; 283
88 iqr.h <- IQR(humerus.LF, type = 2) # 14
89 max.h <- max(humerus.LF) # 297
90 g1.h <- e1071::skewness(humerus.LF, type = 1) # -0,3864936
91 g2.h <- e1071::kurtosis(humerus.LF, type = 1) # 1,108222
92 tab.h <- data.frame(n, m = m.h, sn = sn.h, cv = cv.h, min = min.h, y0.25 = q.h[1],
       y0.50 = q.h[2], y0.75 = q.h[3], max = max.h, IQR = iqr.h, g1 = g1.h, g2 = g2.h,
       row.names = "znak Y")
```

```
93         n      m  sn   cv   min y0.25 y0.50 y0.75 max IQR    g1   g2
94 znak Y 23 275,78 9,5 0,03 249,5   269 275,5   283 297  14 -0,39 1,11
```

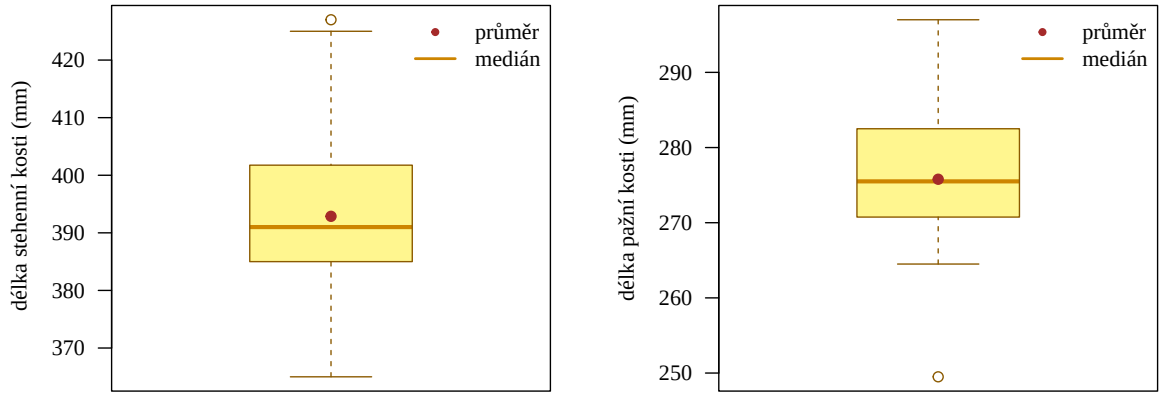
Základní číselné charakteristiky byly počítány na základě 23 naměřených hodnot délky pažní kosti (v mm) z levé strany u žen z kmene Tigara. Průměrná délka pažní kosti z levé strany je 275,8 mm se směrodatnou odchylkou 9,5 mm. Směrodatná odchylka tvoří 3,4 % aritmetického průměru. Naměřené hodnoty se pohybují v rozmezí 249,5 až 297,0 mm. 25 % naměřených hodnot je menších nebo rovných 269,0 mm, 50 % naměřených hodnot je menších nebo rovných 275,5 mm a 75 % naměřených hodnot je menších nebo rovných 283,0 mm. 50 % prostředních hodnot v uspořádaném datovém souboru leží v intervalu o šířce 14,0 mm. Hodnota koeficientu šikmosti ukazuje na záporně zešikmené rozložení dat s prodlouženým levým koncem ($g_1 = -0,39$). Hodnota koeficientu špičatosti ukazuje na strmé rozložení dat ($g_2 = 1,11$).

Krabicový diagram pro znak X vykreslíme příkazem boxplot() s argumentem type = 2. Tento argument zajistí, že se hodnoty všech tří kvantilů vystupujících v krabicovém diagramu vypočítají způsobem popsaným v sekci 3.4.6. V příkazu boxplot() dále nastavíme barvu výplně krabičky (col = "khaki1"), barvu obrysu krabičky (border = "orange4"), barvu mediánu (medcol = "orange3"), popisek osy y (ylab) a otočení měřítka osy y o 90° (las = 1). Do krabicového diagramu dále příkazem points() dokreslíme hnědý bod reprezentující hodnotu aritmetického průměru. Nakonec do grafu doplníme legendu informující o tom, že plný bod hnědé barvy značí aritmetický průměr a silná oranžová úsečka značí medián. Legendu vytvoříme příkazem legend(). V příkazu nastavíme vykreslení legendy v pravém horním rohu (x = "topright"), zobrazení první položky jako plného bodu (pch = c(19, NA)), zobrazení druhé položky jako silné úsečky (lwd = c(NA, 2)), barvu každé položky (col = c("brown", "orange3")), popisek každé položky (legend = c("průměr", "medián")) a potlačení vykreslení rámečku okolo legendy (bty = "n"). Výsledný krabicový diagram je zobrazen na obrázku 3.3 vlevo.

```
95  boxplot(femur.LF, type = 2, col = "khaki1", border = "orange4", medcol = "orange3",
        ylab = "délka stehenní kosti (mm)", las = 1)
96  points(m.f, pch = 19, col = "brown")
97  legend(x = "topright", pch = c(19, NA), lwd = c(NA, 2), col = c("brown", "orange3"),
        legend = c("průměr", "medián"), bty = "n")
```

Analogickým způsobem vykreslíme krabicový diagram pro znak Y. Graf je zobrazen na obrázku 3.3 vpravo. V obou diagramech vidíme odlehlé hodnoty.

```
98  boxplot(humerus.LF, type = 2, col = "khaki1", border = "orange4", medcol = "orange3",
        ylab = "délka pažní kosti (mm)", las = 1)
99  points(m.h, pch = 19, col = "brown")
100 legend(x = "topright", pch = c(19, NA), lwd = c(NA, 2), col = c("brown", "orange3"),
        legend = c("průměr", "medián"), bty = "n")
```



Obrázek 3.3: Krabicový diagram pro délku stehenní kosti (vlevo), resp. pro délku pažní kosti (vpravo) z levé strany u žen z kmene Tigara

Hodnotu kovariance mezi znaky X a Y vypočítáme dosazením do vzorce $s_{n,12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2)$, hodnotu Pearsonova koeficientu korelace získáme příkazem cor() s argumentem method = "pearson".

```
101 sn12 <- 1 / n * sum((femur.LF - m.f) * (humerus.LF - m.h)) # 134,5151
102 r12 <- cor(femur.LF, humerus.LF, method = "pearson") # 0,9032507
```

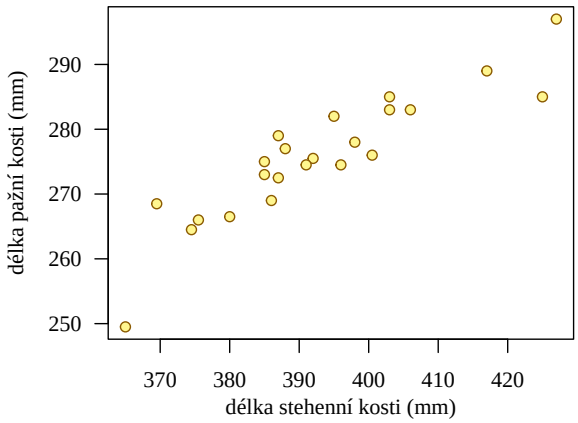
Kovariance mezi znaky X a Y nabývá hodnoty $134,5\,\mathrm{mm}^2$. Mezi délkou stehenní kosti a délkou pažní kosti z levé strany u žen z kmene Tigara existuje velmi vysoký stupeň přímé lineární závislosti ($r_{12} = 0,9033$).

Dvourozměrný tečkový diagram vykreslíme příkazem plot(). Graf je zobrazený na obrázku 3.4.

```
103 plot(femur.LF, humerus.LF, pch = 21, col = "orange4", bg = "khaki1", las = 1, xlab =
        "délka stehenní kosti (mm)", ylab = "délka pažní kosti (mm)")
```

Z dvourozměrného tečkového diagramu je patrný rostoucí lineární trend. Tečkový diagram tedy podporuje náš závěr o přímé lineární závislosti mezi délkou stehenní kosti a délkou pažní kosti z levé strany u žen z kmene Tigara. ★

Příklad 3.6. Řešený příklad
Načtěte datový soubor 31-goldman-alaska-hfrt.csv obsahující údaje o délce stehenní kosti z levé strany v mm (femur.L) a délce pažní kosti z levé strany v mm (humerus.L) u mužů a žen tří aljašských populací. Za předpokladu, že znak X popisuje délku stehenní kosti z levé strany a znak Y popisuje délku pažní kosti z levé strany u žen z kmene Tigara, (a) vytvořte tabulku vážených číselných charakteristik pro znak X, resp. pro znak Y; (b) vypočítejte váženou kovarianci $s_{w,12}$. Všechny vypočítané hodnoty řádně interpretujte.



Obrázek 3.4: Dvourozměrný tečkový diagram pro délku stehenní kosti a délku pažní kosti z levé strany u žen z kmene Tigara

Řešení příkladu 3.6
Úvod příkladu je totožný jako u příkladu 3.5.

```
104 data <- read.delim("31-goldman-alaska-hfrt.csv", sep = ";", dec = ".")
105 data.F <- na.omit(data[data$sex == "f" & data$pop == "tig", c("femur.L", "humerus.L")])
106 femur.LF <- data.F$femur.L
107 humerus.LF <- data.F$humerus.L
```

Po separování naměřených hodnot délky stehenní kosti a délky pažní kosti z výběru žen z kmene Tigara se zaměříme na vytvoření tabulky vážených číselných charakteristik pro znak X. Tabulka bude obsahovat vážený průměr, vážený rozptyl a váženou směrodatnou odchylku. Při výpočtu vážených číselných charakteristik pro intervalová nebo poměrová data je třeba naměřené hodnoty roztřídit do třídicích intervalů. Příkazem length() zjistíme nejprve rozsah datového souboru a příkazem range() rozsah naměřených hodnot znaku X. Pomocí Sturgesova pravidla (viz kapitola 2) vypočítáme optimální počet třídicích intervalů r.

```
108 n <- length(femur.LF) # 23
109 range(femur.LF) # 365; 427
110 r <- round(1 + 3.3 * log10(n)) # 5
111 # femur.LF: 427 - 364 = 63 -> 65 / 5 = 13 -> seq(363, 428, by = 13)
112 b.femur.LF <- seq(from = 363, to = 428, by = 13)
```

Optimální počet třídicích intervalů $r = 5$. Naměřené hodnoty délky stehenní kosti se pohybují v rozmezí 365 až 427 mm. Vzdálenost mezi minimální naměřenou hodnotou sníženou o 1 a maximální naměřenou hodnotou je $427 - 364 = 63$ mm. Nejbližší vyšší celé číslo dělitelné beze zbytku počtem třídicích intervalů, tj. pěti, je 65. Optimální délka jednoho třídicího intervalu $d = \frac{65}{5} = 13$ mm. Hranice třídicích intervalů stanovíme jako posloupnost 363, 376, ..., 428 mm.

Nyní zjistíme středy třídicích intervalů $x_{[j]}$, $j = 1, \ldots, 5$, pomocí funkce hist() s argumentem plot = F a jejího výstupu mids (viz kapitola 2). Dále roztřídíme naměřené hodnoty délky stehenní kosti do příslušných třídicích intervalů příkazem cut(), a následně vypočítáme četnostní zastoupení n_j naměřených hodnot v každém třídicím intervalu příkazem table().

```
113 xj <- hist(femur.LF, breaks = b.femur.LF, plot = F)$mids
114 femur.LF.c <- cut(femur.LF, breaks = b.femur.LF)
115 nj <- table(femur.LF.c)
```

Vážený průměr vypočítáme dosazením do vzorce $m_w = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}$. Vážený rozptyl vypočítáme například dosazením do vzorce $s_w^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - m_w)^2$ a váženou směrodatnou odchylku získáme odmocněním váženého rozptylu. Vypočítané vážené číselné charakteristiky vložíme do souhrnné tabulky.

```
m.wf <- 1 / n * sum(nj * xj) # 392,1087
s2.wf <- 1 / n * sum(nj * (xj - m.wf) ^ 2) # 267,7164
s.wf <- sqrt(s2.wf) # 16,36204
tab.wf <- data.frame(m.w = m.wf, s2.w = s2.wf, s.w = s.wf, row.names = "znak X")
```

```
          m.w   s2.w   s.w
znak X 392,11 267,72 16,36
```

Vážený průměr délky stehenní kosti z levé strany u žen z kmene Tigara nabývá hodnoty $392,1$ mm s váženým rozptylem $267,7\,\text{mm}^2$ (resp. s váženou směrodatnou odchylkou $16,4$ mm). Pro porovnání si připomeňme, že aritmetický průměr nabýval hodnoty $392,9$ mm se směrodatnou odchylkou $15,7$ mm (viz příklad 3.5).

Nyní se zaměříme na vytvoření tabulky vážených číselných charakteristik pro znak Y. V souladu se Sturgesovým pravidlem rozdělíme naměřené hodnoty délky pažní kosti do pěti ekvidistantních třídicích intervalů o optimální šířce $h = 10$ mm. Hranice třídicích intervalů stanovíme jako posloupnost 248, 258, ..., 298 mm. Dále vypočítáme středy třídicích intervalů $y_{[k]}$, $k = 1, \ldots, 5$. Výpočet vážených číselných charakteristik pro znak Y je analogický jako u znaku X.

```
range(humerus.LF) # 249,5; 297,0
# humerus.LF: 297 - 249 = 48 -> 50 / 5 = 10 -> seq(248, 298, by = 10)
b.humerus.LF <- seq(from = 248, to = 298, by = 10)
yk <- hist(humerus.LF, breaks = b.humerus.LF, plot = F)$mids
humerus.LF.c <- cut(humerus.LF, breaks = b.humerus.LF)
nk <- table(humerus.LF.c)
m.wh <- 1 / n * sum(nk * yk) # 275,1739
s2.wh <- 1 / n * sum(nk * (yk - m.wh) ^ 2) # 86,57845
s.wh <- sqrt(s2.wh) # 9,304754
tab.wh <- data.frame(m.w = m.wh, s2.w = s2.wh, s.w = s.wh, row.names = "znak Y")
```

```
          m.w  s2.w s.w
znak Y 275,17 86,58 9,3
```

Vážený průměr délky pažní kosti z levé strany u žen z kmene Tigara nabývá hodnoty $275,2$ mm s váženým rozptylem $86,6\,\text{mm}^2$ (resp. s váženou směrodatnou odchylkou $9,3$ mm). Pro porovnání si připomeňme, že aritmetický průměr nabýval hodnoty $275,8$ mm se směrodatnou odchylkou $9,5$ mm (viz příklad 3.5).

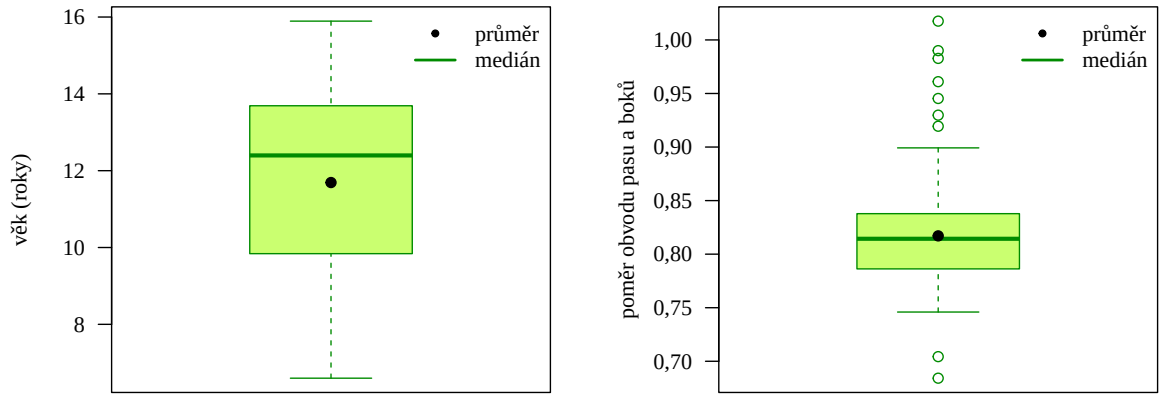
K výpočtu vážené kovariance potřebujeme nejprve znát kontingenční tabulku simultánních absolutních četností n_{jk}, jež popisují četnostní zastoupení naměřených hodnot v dvourozměrných třídicích intervalech pro znaky X a Y. Kontingenční tabulku vytvoříme příkazem table(), jehož vstupními argumenty budou vektory femur.LF.c a humerus.LF.c. Hodnotu vážené kovariance potom vypočítáme dosazením do vzorce $s_{w,12} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk}x_{[j]}y_{[k]} - m_{w,1}m_{w,2}$. Součiny $x_{[j]}y_{[k]}$ pro všechny kombinace indexů $j, k = 1, \ldots, 5$, vypočítáme maticovým vynásobením vektoru středů třídicích intervalů xj s transponovaným vektorem středů třídicích intervalů yk. Maticové násobení provedeme pomocí operátoru %*%, transpozici vektoru yk pomocí funkce t().

```
njk <- table(femur.LF.c, humerus.LF.c)
s12.w <- 1 / n * sum(njk * (xj %*% t(yk))) - m.wf * m.wh # 126,0681
```

Vážená kovariance mezi délkou stehenní kosti a délkou pažní kosti z levé strany u žen z kmene Tigara nabývá hodnoty $126,1\,\text{mm}^2$. Pro porovnání si připomeňme, že kovariance $s_{n,12}$ nabývala hodnoty $134,5\,\text{mm}^2$ (viz příklad 3.5). ★

Příklad 3.7. Neřešený příklad
Načtěte datový soubor 33-two-samples-whr-mf.csv obsahující údaje o věku v letech (age) a poměru obvodu pasu a boků (bez jednotky; WHR) u dětí ve věku do 16 let. Za předpokladu, že znak X popisuje věk a znak Y popisuje poměr obvodu pasu a boků u chlapců, (a) vytvořte tabulku základních číselných charakteristik pro znak X, resp. pro znak Y; (b) vykreslete krabicový diagram pro znak X, resp. pro znak Y; (c) vypočítejte kovarianci $s_{n,12}$ a Pearsonův koeficient korelace r_{12} a nakreslete dvourozměrný tečkový diagram. Všechny vypočítané hodnoty řádně interpretujte.
Výsledky: (a) tabulka základních číselných charakteristik pro znak X viz tabulka 3.6 a pro znak Y viz tabulka

3.7; (b) krabicový diagram pro znak X, resp. pro znak Y viz obrázek 3.5 vlevo, resp. vpravo; (c) $s_{n,12} = -0,0502$, $r_{12} = -0,4252$, mírný stupeň nepřímé lineární závislosti, dvourozměrný tečkový diagram viz obrázek 3.6.



Obrázek 3.5: Krabicový diagram pro věk (vlevo), resp. pro poměr obvodu pasu a boků (vpravo) u chlapců

Tabulka 3.6: Základní číselné charakteristiky pro věk u chlapců

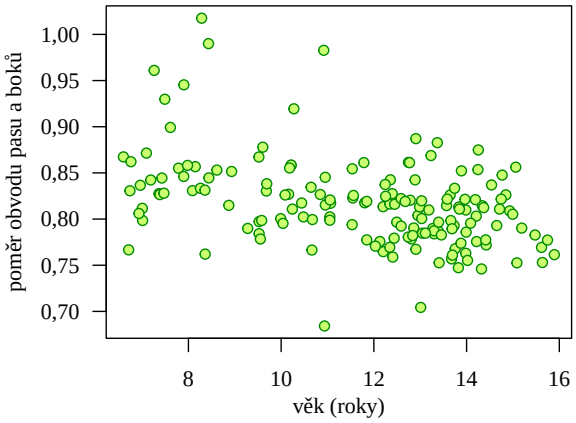
	n	m	s_n	cv	min	$x_{0,25}$	$x_{0,50}$	$x_{0,75}$	max	IQR	g_1	g_2
znak X	163	11,69	2,50	0,21	6,60	9,69	12,40	13,69	15,89	4,01	-0,49	-0,89

Tabulka 3.7: Základní číselné charakteristiky pro poměr obvodu pasu a boků u chlapců

	n	m	s_n	cv	min	$y_{0,25}$	$y_{0,50}$	$y_{0,75}$	max	IQR	g_1	g_2
znak Y	163	0,82	0,05	0,06	0,68	0,79	0,81	0,84	1,02	0,05	1,16	3,40

★

Příklad 3.8. Neřešený příklad
Načtěte datový soubor 33-two-samples-whr-mf.csv obsahující údaje o věku v letech (age) a poměru obvodu pasu a boků (bez jednotky; WHR) u dětí ve věku do 16 let. Za předpokladu, že znak X popisuje věk a znak Y popisuje poměr obvodu pasu a boků u chlapců, (a) vytvořte tabulku vážených číselných charakteristik pro znak X, resp. pro znak Y; (b) vypočítejte váženou kovarianci $s_{w,12}$. Všechny vypočítané hodnoty řádně interpretujte.
Výsledky: (a) $r = 8$, zvolené hranice třídicích intervalů pro znak X: $6,4$; $7,6$; $8,8$; $10,0$; $11,2$; $12,4$; $13,6$; $14,8$; $16,0$, zvolené hranice třídicích intervalů pro znak Y: $0,65$; $0,70$; $0,75$; $0,80$; $0,85$; $0,90$; $0,95$; $1,00$; $1,05$; (b) $m_{w,1} = 11,6896$, $s^2_{w,1} = 6,3750$, $s_{w,1} = 2,5249$, $m_{w,2} = 0,8179$, $s^2_{w,2} = 0,0024$, $s_{w,2} = 0,0489$; (c) $s_{w,12} = -0,0468$.
★



Obrázek 3.6: Dvourozměrný tečkový diagram pro věk a poměr obvodu pasu a boků u chlapců