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 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 $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}\, n_{.k}}{n}\right)^2}{\frac{n_{j.}\, n_{.k}}{n}}$ 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 (Hi – vysoké; Mi – střední; Li – nízké) 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() a pomocí operátoru [,] vybereme z datové tabulky pouze sloupce týkající se mužů. Vybrané sloupce přeuspořádáme tak, aby pořadí zakončení tří hlavních dlaňových linií bylo od nejnižšího po nejvyšší, a výslednou tabulku vložíme do proměnné data.M. Pro nalezení modu znaku X je třeba nejprve stanovit absolutní četnosti jednotlivých variant znaku X prostřednictvím řádkových součtů v tabulce data.M. Řádkové součty vypočítáme příkazem apply(), jehož vstupními argumenty budou datová tabulka data.M, argument MARGIN = 1 specifikující řádkovou operaci a argument FUN = sum určující, že hodnoty se po řádcích mají sečíst. Modem znaku X potom bude varianta s nejvyšší četností.

```
1 data <- read.delim("22-multinom-palmar-lines.txt", sep = "\t", dec = ".", row.names = 1)
2 data.M <- data[, 3:1]
3 apply(data.M, MARGIN = 1, FUN = sum)
```

```
LiH  MH DaH                                                                    4
 16  42  42                                                                    5
```

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

Analogicky stanovíme modus znaku Y. Nejprve vypočítáme absolutní četnosti jednotlivých variant znaku Y prostřednictvím sloupcových součtů v tabulce data.M. Sloupcové součty vypočítáme příkazem apply() s argumenty MARGIN = 2 a FUN = sum. Modem znaku Y bude varianta s nejvyšší četností.

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

```
Lo Mi Hi                                                                       7
23 33 44                                                                       8
```

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

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

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

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

Příklad 3.2. Neřešený příklad
Načtěte datový soubor 20-more-samples-probabilities-pubis.txt obsahující údaje o původu žen a změně kostního reliéfu na vnitřní straně stydké kosti v blízkosti stydké spony u těchto žen. Za předpokladu, že znak X popisuje původ žen a znak Y popisuje změnu 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 četností 110), modus znaku Y: nepřítomný kostní reliéf (s četností 102); (b) $V = 0,1517$, slabý stupeň závislosti. ♣

3.3 Číselné charakteristiky ordinálních znaků

3.3.1 Charakteristika polohy

α-kvantil, 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 3.3.4)).

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 stejné hodnoty 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 vztahu mezi dvěma diskrétními náhodnými veličinami, mezi dvěma spojitými náhodnými veličinami i mezi diskrétní a spojitou náhodnou veličinou.

Tabulka 3.2: Stupnice míry závislosti pro Spearmanův koeficient pořadové korelace r_S (resp. pro Pearsonův koeficient korelace r_{12} (viz sekce 3.3.4))

$\lvert r_S \rvert$, **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, kteří se zúčastnili sociologického výzkumu v rámci vzdělávacího programu Jeden svět na školách v roce 2014. Kompletní datový soubor viz na stránkách datového archivu Sociologického ústavu Akademie věd ČR (archiv.soc.cas.cz). Za předpokladu, že znak X popisuje odpověď na otázku *Jak často nakupujete v obchodních centrech?* (a.shop; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy) a znak Y popisuje odpověď na otázku *Jak často chodíte do kina?* (a.cinema; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy), (a) vytvořte tabulku základních číselných charakteristik pro znak X, resp. pro znak Y; (b) vypočítejte Spearmanův koeficient pořadové korelace r_S a nakreslete dvourozměrný tečkový diagram. Všechny vypočítané hodnoty řádně interpretujte.
Poznámka: Tabulka základních číselných charakteristik bude obsahovat: rozsah náhodného výběru, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu a interkvartilové rozpětí.

Řešení příkladu 3.3
Datový soubor načteme příkazem read.delim() a pomocí operátoru [,] z něj vybereme sloupce a.shop a a.cinema. Dále se zaměříme na vytvoření tabulky základních číselných charakteristik pro znak X. Minimální, resp. maximální hodnotu stanovíme příkazem min(), resp. max(). Dolní kvartil, medián a horní kvartil vypočítáme najednou pomocí příkazu 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().

```
10 data <- read.delim('28-one-world-2014.csv', sep = ';')
11 data.S <- na.omit(data[, c('a.shop', 'a.cinema')])
12 a.shop <- data.S$a.shop
13 a.cinema <- data.S$a.cinema
14 n <- length(a.shop) # 1090
15 min.SC <- min(a.shop) # 1
16 q.SC <- quantile(a.shop, probs = c(0.25, 0.50, 0.75), type = 2) # 2; 3; 3
17 max.SC <- max(a.shop) # 5
18 iqr.SC <- IQR(a.shop, type = 2) # 1
19 tab.SC <- data.frame(n, min.SC, t(q.SC), max.SC, iqr.SC, row.names = "znak X")
```

```
20           n min.SC X25. X50. X75. max.SC iqr.SC
21 znak X 1090      1    2    3    3      5      1
```

Základní číselné charakteristiky 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 % nejčastějších odpovědí na zadanou otázku se realizuje v intervalu o šířce 1.

Analogickým způsobem vypočítáme tabulku základních číselných charakteristik pro znak Y.

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

```
27           n min.C X25. X50. X75. max.C iqr.C
28 znak Y 1090     1    3    3    4     5     1
```

Základní číselné charakteristiky 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 % nejčastějších odpovědí na zadanou otázku se realizuje v intervalu o šířce 1.

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

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

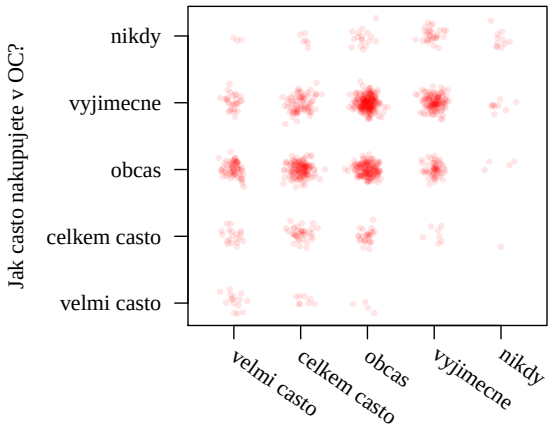
Mezi odpovědí na otázku *Jak často nakupujete v obchodních centrech?* a na otázku *Jak často chodíte do kina?* existuje mírný stupeň přímé pořadové závislosti ($r_S = 0,3818$).

Nakonec vykreslíme tečkový diagram. Nejprve si rozvrhneme okno obrázku tak, aby se na levou stranu od grafu vešly popisky variant znaku X. K tomu využijeme příkaz par() s argumentem mar = c(2, 7, 2, 2), kterým specifikujeme, že mezi dolním, horním a pravým okrajem grafu a okrajem obrázku bude místo na šest řádků textu a mezi levým okrajem grafu a okrajem obrázku bude místo na sedm řádků textu. Jelikož možných kombinací variant znaků X a Y je pouze $5 \times 5 = 25$, zatímco respondentů s kompletním dotazníkem $n = 1090$, nebylo by použití klasického tečkového diagramu vykresleného příkazem plot() příliš praktické, neboť shodné dvojice odpovědí jednotlivých respondentů by se v takovém grafu překrývaly. Tečkový diagram proto vykreslíme pomocí funkce dotplot(), implementované v R-skriptu AS-sbirka-funkce.R, který je součástí této publikace. Funkce dotplot() před vykreslením tečkového diagramu přičte ke každé zaznamenané odpovědi zanedbatelně malé pseudonáhodně vygenerované číslo, tudíž hodnoty, které se původně překrývaly, se již překrývat nebudou. Povinnými vstupními argumenty funkce jsou vektory a.shop a a.cinema. Z povinně volitelných argumentů zmiňme argument sd stanovující, jak velká čísla se k původním odpovědím přičtou, a argument asp = T zajišťující, že poměr stran osy x ku ose y bude jedna ku jedné. Barvu bodů specifikujeme pomocí funkce rgb() se syntaxí rgb(red, green, blue, alpha). Argumenty red, green a blue nabývají hodnoty z intervalu $\langle 0; 1\rangle$ a určují podíl červené, zelené a modré složky ve výsledné barvě (čím je hodnota vyšší, tím vyšší je podíl barevné složky). Argument alpha rovněž nabývá hodnoty z intervalu $\langle 0; 1\rangle$ a určuje průhlednost barvy (čím je hodnota vyšší, tím méně je výsledná barva průhledná). Protože popisky variant proměnné a.shop jsou celkem dlouhé a hrozí, že by se nám k jednotlivým variantám zobrazeným na ose x nevešly, otočíme je o vhodný úhel. Nejprve v příkazu dotplot() argumentem axes = F potlačíme vykreslení os x a y. Obě osy následně vykreslíme zvlášť příkazem axis(), přičemž u osy x potlačíme vypsání popisků variant nastavením argumentu label = NA. Popisky variant doplníme zvlášť příkazem text(), kde argumenty x, resp y definujeme x-ové, resp. y-ové souřadnice umístění popisků, argumentem labels specifikujeme text popisků, argumentem xpd nastavíme vykreslení popisků vně grafu (tj. v okrajové části obrázku), argumentem srt stanovíme úhel otočení popisků (zde konkrétně otáčíme o 35° ve směru hodinových ručiček) a argumentem adj = 0 zvolíme zarovnání popisků do levého horního rohu. Popisek osy x, resp. osy y definujeme samostatně příkazem mtext(). Výsledný graf je zobrazen na obrázku 3.1.

```
30 source("AS-sbirka-funkce.R")
31 names <- c('velmi casto', 'celkem casto', 'obcas', 'vyjimecne', 'nikdy')
32 par(mar = c(6, 7, 2, 2))
33 dotplot(a.shop, a.cinema, sd = 0.1, pch = 19, col = rgb(1, 0, 0, 0.1),
34         xlab = "", ylab = "", axes = F, cex = 0.5, asp = T)
35 box(bty = "o")
36 axis(1, at = 1:5, labels = NA)
37 axis(2, at = 1:5, labels = names, las = 1)
38 text(x = seq(1, 5, by = 1), y = 0.4, labels = names, xpd = T, srt = -35, adj = 0)
39 mtext("Jak casto chodite do kina?", side = 1, line = 5)
40 mtext("Jak casto nakupujete v OC?", side = 2, line = 6)
```

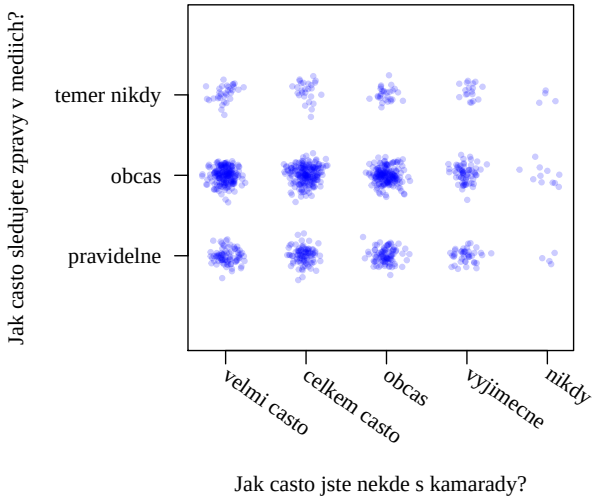
♣

Příklad 3.4. Neřešený příklad
Načtěte datový soubor 28-one-world-2014.csv obsahující odpovědi respondentů, studentů středních škol, kteří se zúčastnili sociologického výzkumu v rámci vzdělávacího programu Jeden svět na školách v roce 2014. Kompletní datový soubor viz na stránkách datového archivu Sociologického ústavu Akademie věd ČR (archiv.soc.cas.cz). Za předpokladu, že znak X popisuje odpověď na otázku *Jak často jste někde s kamarády?* (a.friends; 1 – velmi často; 2 – celkem často; 3 – občas; 4 – výjimečně; 5 – nikdy) a znak Y popisuje odpověď na otázku *Jak často sledujete zprávy v médiích?* (c.news; 1 – pravidelně; 2 – občas; 3 – téměř nikdy), (a) vytvořte tabulku základních číselných



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

charakteristik pro znak X, resp. pro znak Y; (b) vypočítejte Spearmanův koeficient pořadové korelace r_S a nakreslete dvourozměrný tečkový diagram. Všechny vypočítané hodnoty řádně interpretujte.
Poznámka: Tabulka základních číselných charakteristik bude obsahovat: rozsah náhodného výběru, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu a interkvartilové rozpětí.
Výsledky: (a) tabulka základních číselných charakteristik pro znak X, resp. pro znak Y viz tabulka 3.3, resp. tabulka 3.4; (b) $r_S = -0,0178$, velmi nízký stupeň nepřímé pořadové závislosti, dvourozměrný tečkový diagram viz obrázek 3.2.



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

♣

Tabulka 3.3: Základní číselné charakteristiky znaku X

	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 znaku Y

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

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 $cv = \frac{s_n}{m}$. Pomocí 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 odhadneme pomocí vzorce $m_p = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^p$, kde m je aritmetický průměr a n je rozsah náhodného výběru. Koeficient šikmosti, resp. špičatosti můžeme vypočítat v softwaru R příkazem skewness(), resp. kurtosis() s argumentem type = 1. Oba příkazy pochází z knihovny e1071.

Poznámka: Koeficient šikmosti můžeme také vypočítat pomocí vzorce $b_1 = \frac{m_3}{s_n^3} = g_1\sqrt{\left(\frac{n-1}{n}\right)^3}$, koeficient špičatosti podle vzorce $b_2 = \frac{m_4}{s_n^4} - 3 = (g_2+3)\left(1 - \frac{1}{n}\right)^2 - 3$, kde s_n^p je p-tá mocnina směrodatné odchylky s_n. Koeficient šikmosti b_1, resp. špičatosti b_2 můžeme vypočítat příkazem e1071::skewness(), resp. e1071::kurtosis() s argumentem type = 3. Výsledná hodnota koeficientu šikmosti g_1 se bude od hodnoty koeficientu šikmosti b_1 mírně lišit, neboť i vzorce se liší. Totéž platí pro koeficienty špičatosti g_2 a b_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, m_2 jsou průměry znaků X, Y. Je-li $s_{n,12} = 0$, pak řekneme, že znaky X, 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 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.csv obsahující údaje o délce stehenní kosti z levé strany (femur.L) a délce pažní kosti z levé strany (humer.L) mužů a žen aljašské populace z kmene Tigara. 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 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.
Poznámka: Tabulka základních číselných charakteristik bude obsahovat: rozsah náhodného výběru, aritmetický průměr, směrodatnou odchylku, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu, interkvartilové rozpětí, koeficient šikmosti a koeficient špičatosti.

Řešení příkladu 3.5
Datový soubor načteme a pomocí operátoru [,] z něj vybereme řádky týkající se žen z kmene Tigara a sloupce femur.L a humer.L. Dále se zaměříme na vytvoření tabulky základních číselných charakteristik pro znak X. Nejprve příkazem length() stanovíme rozsah náhodného výběru n. 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(). Minimální, resp. maximální naměřenou hodnotu nalezneme pomocí příkazu min(), resp. 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 stanovíme příkazem IQR(). Hodnotu koeficientu šikmosti, resp. špičatosti vypočítáme příkazem skewness(), resp. kurtosis() z knihovny e1071 s argumentem type = 1 (viz sekce 3.4.3). Všechny číselné charakteristiky vložíme do souhrnné tabulky, kterou vytvoříme pomocí příkazu data.frame().

```
41 data <- read.delim("31-goldman-alaska.csv", sep = ";", dec = ".")
42 data.F <- na.omit(data[data$sex == "f" & data$pop == "Tigara", c("femur.L", "humer.L")])
43 femur.LF <- data.F$femur.L
44 humer.LF <- data.F$humer.L
45 n <- length(femur.LF) # 23
46 m.f <- mean(femur.LF) # 392,8696
47 sn.f <- sqrt(1 / n * sum((femur.LF - m.f) ^ 2)) # 15,67553
48 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
iqr.f <- IQR(femur.LF, type = 2) # 18
b1.f <- e1071::skewness(femur.LF, type = 1) # 0,4756138
b2.f <- e1071::kurtosis(femur.LF, type = 1) # -0,1459509
tab.f <- data.frame(n, m.f, sn.f, min.f, t(q.f), max.f, iqr.f, b1.f, b2.f,
                    row.names = "znak X")
```

```
       n    m.f  sn.f min.f X25. X50. X75. max.f iqr.f b1.f  b2.f
znak X 23 392,87 15,68   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. 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 $427,0$ mm. 50 % nejčastěji naměřených hodnot leží v intervalu o šířce 18,0 mm. Hodnota koeficientu šikmosti ukazuje na kladné zešikmení dat s prodlouženým pravým koncem ($g_1 = 0,48$). Hodnota koeficientu špičatosti ukazuje na mírně zploštělý charakter dat ($g_2 = -0,15$).

Analogickým způsobem vytvoříme tabulku základních číselných charakteristik pro znak Y.

```
m.h <- mean(humer.LF) # 275,7826
sn.h <- sqrt(1 / n * sum((humer.LF - m.h) ^ 2)) # 9,500373
min.h <- min(humer.LF) # 249,5
q.h <- quantile(humer.LF, probs = c(0.25, 0.50, 0.75), type = 2) # 269; 275,5; 283
iqr.h <- IQR(humer.LF, type = 2) # 14
max.h <- max(humer.LF) # 297
b1.h <- e1071::skewness(humer.LF, type = 3) # -0,3615635
b2.h <- e1071::kurtosis(humer.LF, type = 3) # 0,7587516
tab.h <- data.frame(n, m.h, sn.h, min.h, t(q.h), max.h, iqr.h, b1.h, b2.h,
                    row.names = "znak Y")
```

```
       n    m.h sn.h min.h X25.  X50. X75. max.h iqr.h  b1.h b2.h
znak Y 23 275,78  9,5 249,5  269 275,5  283   297    14 -0,36 0,76
```

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. 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 % nejčastěji naměřených hodnot leží v intervalu o šířce 14,0 mm. Hodnota koeficientu šikmosti ukazuje na záporné zešikmení dat s prodlouženým levým koncem ($g_1 = -0,36$). Hodnota koeficientu špičatosti ukazuje na strmý charakter dat ($g_2 = 0,76$).

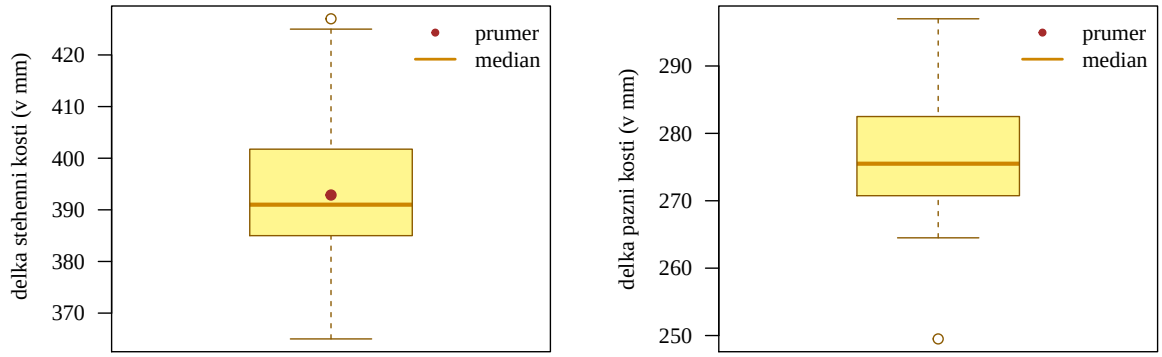
Krabicový diagram pro znak X vykreslíme příkazem boxplot() s argumentem type $= 2$. Tento argument zajistí, že se hodnoty všech tří kvantilů vystupujících v krabicovém diagramu vypočítají způsobem popsaným v sekci 3.3.1. Do krabicového diagramu dále příkazem points() dokreslíme bod reprezentující hodnotu aritmetického průměru. Aby bylo patrné, co vykreslený bod reprezentuje, doplníme do grafu legendu. Výsledný krabicový diagram je zobrazen na obrázku 3.3 vlevo.

```
boxplot(femur.LF, type = 2, col = 'khaki1', border = "orange4", medcol = "orange3",
        ylab = " delka stehenni kosti (v mm)", las = 1)
points(m.f, pch = 20, col = 'brown')
legend('topright', lwd = c(NA, 2), pch = c(20, NA), col = c('brown', 'orange3'),
       legend = c('prumer', 'median'), bty = 'n')
```

Analogickým způsobem vykreslíme krabicový diagram pro znak Y. Graf je zobrazen na obrázku 3.3 vpravo.

```
boxplot(humer.LF, type = 2, col = 'khaki1', border = "orange4", medcol = "orange3",
        ylab = "delka pazni kosti (v mm)", las = 1)
points(m.f, pch = 20, col = 'brown')
```

```
legend('topright', lwd = c(NA, 2), pch = c(20, NA), col = c('brown', 'orange3'),
       legend = c('prumer', 'median'), 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

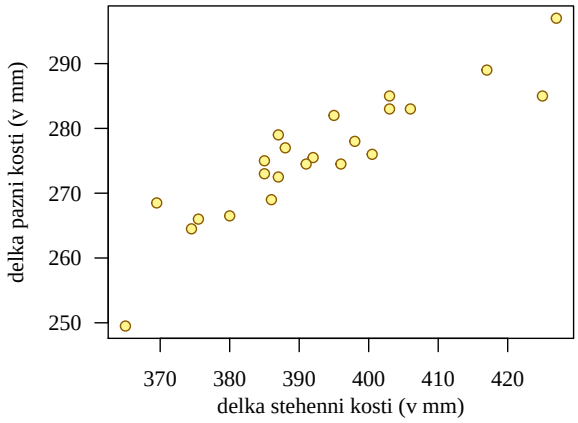
Hodnotu kovariance mezi znaky X a Y vypočítáme dosazením do vzorce $s_{n,12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2)$, hodnotu Pearsonova korelačního koeficientu získáme příkazem cor() s argumentem method = "pearson".

```
sn12 <- 1 / n * sum((femur.LF - m.f) * (humer.LF - m.h)) # 134,5151
r12 <- cor(femur.LF, humer.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 ($r_{12} = 0,90$).

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

```
plot(femur.LF, humer.LF, pch = 21, col = "orange4", bg = "khaki1",
     xlab = "delka stehenni kosti (v mm)", ylab = "delka pazni kosti (v mm)", las = 1)
```



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

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.csv obsahující údaje o délce stehenní kosti z levé strany (femur.L) a délce pažní kosti z levé strany (humer.L) mužů a žen aljašské populace z kmene Tigara. 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.
Poznámka: Tabulka vážených číselných charakteristik bude obsahovat: vážený průměr, vážený rozptyl a váženou směrodatnou odchylku.

Řešení příkladu 3.6
Datový soubor načteme a pomocí operátoru [,] z něj vybereme řádky týkající se žen z kmene Tigara a sloupce femur.L a humer.L. Dále se zaměříme na vytvoření tabulky vážených číselných charakteristik pro znak X. Příkazem length() určíme rozsah náhodného výběru a pomocí Sturgesova pravidla (viz kapitola 2) stanovíme optimální počet třídicích intervalů r. Dále příkazem range() zjistíme rozsah naměřených délek stehenních kostí.

```
data <- read.delim("31-goldman-alaska.csv", sep = ";", dec = ".")
data.F <- na.omit(data[data$sex == "f" & data$pop == "Tigara", c("femur.L", "humer.L")])
femur.LF <- data.F$femur.L
humer.LF <- data.F$humer.L
n <- length(humer.LF) # 23
r <- round(3.3 * log10(n) + 1) # 5
range(femur.LF) # 365-427
# femur.LF: 427 - 364 = 63 -> 65 / 5 = 13 -> seq(363, 428, by = 13)
b.femur.LF <- seq(363, 428, by = 13)
```

Na základě Sturgesova pravidla stanovíme 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 šířka jednoho třídicího intervalu $d = \frac{65}{5} = 13$ mm. Hranice třídicích intervalů stanovíme jako posloupnost 363, 376, ..., 428 mm. Celkem získáme šest hranic definujících pět třídicích intervalů o ekvidistantní šířce 13 mm.

Nyní vypočítáme středy třídicích intervalů $x_{[j]}$, $j = 1, \ldots, 5$, a to buď ručním výpočtem, nebo automaticky jako výstup mids funkce hist() s argumentem plot = F. Po stanovení středů roztřídíme naměřené hodnoty délky stehenní kosti do příslušných třídicích intervalů příkazem cut(), a následně příkazem table() vypočítáme četnostní zastoupení n_j naměřených hodnot v každém třídicím intervalu.

```
xj <- hist(femur.LF, breaks = b.femur.LF, plot = F)$mids
femur.LF.c <- cut(femur.LF, breaks = b.femur.LF)
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 charakteristiky vložíme do souhrnné tabulky, kterou vytvoříme pomocí příkazu data.frame().

```
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.wf, s2.wf, s.wf, row.names = "znak X")
```

```
        m.wf  s2.wf  s.wf
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 mm^2 (resp. s váženou směrodatnou odchylkou 16, 4 mm). Pro porovnání si připomeňme, že průměrná délka nabývá 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 10 mm. Hranice třídicích intervalů stanovíme jako posloupnost 248, 258, ..., 298 mm. Dále postupujeme analogicky jako při výpočtu vážených číselných charakteristik pro znak X.

```
range(humer.LF) # 249,5-297,0
# humer.LF: 297 - 249 = 48 -> 50 / 5 = 10 -> seq(248, 298, by = 10)
b.humer.LF <- seq(248, 298, by = 10)
yk <- hist(humer.LF, breaks = b.humer.LF, plot = F)$mids
humer.LF.c <- cut(humer.LF, breaks = b.humer.LF)
nk <- table(humer.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.wh, s2.wh, s.wh, row.names = "znak Y")
```

```
         m.wh s2.wh s.wh
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 mm^2 (resp. s váženou směrodatnou odchylkou 9, 3 mm). Pro porovnání si připomeňme, že průměrná délka nabývá hodnoty 275, 8 mm se směrodatnou odchylkou 9, 5 mm (viz příklad 3.5).

K výpočtu vážené kovariance potřebujeme nejprve stanovit kontingenční tabulku simultánních absolutních četností n_{jk}, jež popisují četnostní zastoupení v dvourozměrných třídicích intervalech stanovených pro znaky X a Y najednou ($j, k = 1, \ldots, 5$). Každý třídicí interval je reprezentován dvojicí středů $(x_{[j]}, y_{[k]})^T$. Kontingenční tabulku vytvoříme příkazem table(), jehož vstupními argumenty budou vektory femur.LF.c a humer.LF.c, které jsme vypočítali dříve příkazem cut(). Hodnotu vážené kovariance potom získáme například 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 jk, $j, k = 1, \ldots, 5$, vypočítáme elegantně pomocí operace zvané *maticové násobení*, tj. $x_{[j]}y_{[k]} = xy^T$, kde x je vektor středů třídicích intervalů znaku X a y^T je transpozice vektoru středů třídicích intervalů znaku Y. Transpozici vektoru y provedeme příkazem t(), operaci maticového násobení zapíšeme pomocí operátoru %*%.

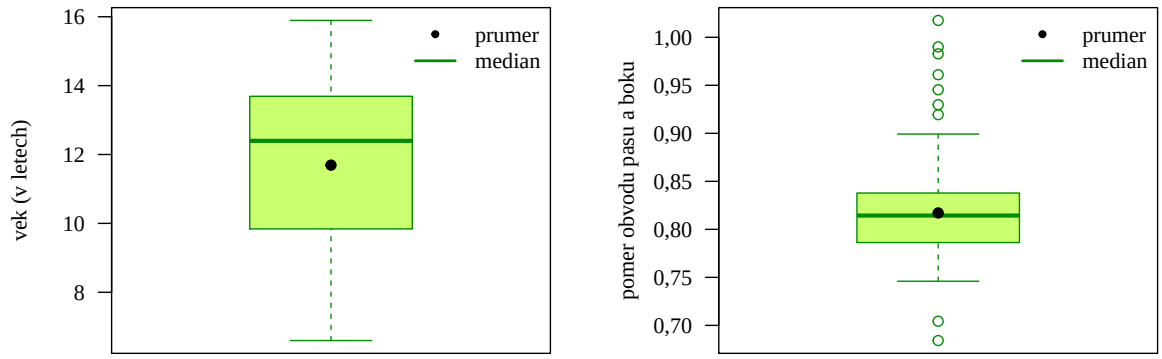
```
njk <- table(femur.LF.c, humer.LF.c)
s12.w <- 1 / n * sum(njk * (xj %*% t(yk))) - m.wf * m.wh # 126,0681
```

Vážená kovariance mezi délkou stehenní kosti a délkou pažní kosti z levé strany u žen z kmene Tigara nabývá hodnoty 126, 1 mm^2. Pro porovnání si připomeňme, že kovariance nabývá hodnoty 134, 5 mm^2 (viz příklad 3.5).

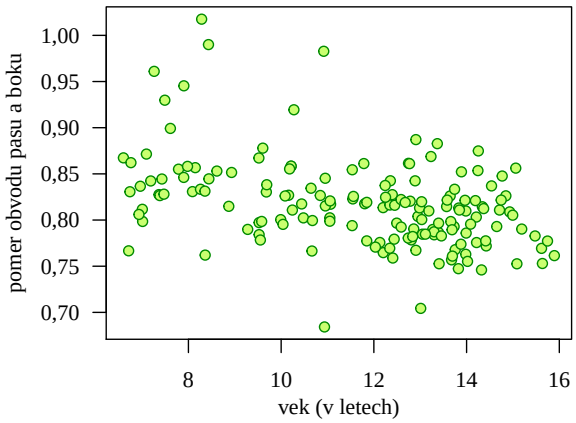
♣

Příklad 3.7. Neřešený příklad
Načtěte datový soubor 32-two-samples-whr-mf.csv obsahující údaje o věku (age) a poměru obvodu pasu a boků (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.
Poznámka: Tabulka základních číselných charakteristik bude obsahovat: rozsah náhodného výběru, aritmetický průměr, směrodatnou odchylku, minimální naměřenou hodnotu, dolní kvartil, medián, horní kvartil, maximální naměřenou hodnotu, interkvartilové rozpětí, koeficient šikmosti a koeficient špičatosti.
Výsledky: (a) tabulka základních číselných charakteristik pro znak X, resp. pro znak Y viz tabulka 3.5, resp. tabulka 3.6; (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.
♣

Příklad 3.8. Neřešený příklad
Načtěte datový soubor 32-two-samples-whr-mf.csv obsahující údaje o věku (age) a poměru obvodu pasu a boků



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



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

Tabulka 3.5: Základní číselné charakteristiky znaku X

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

Tabulka 3.6: Základní číselné charakteristiky znaku Y

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

(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.
Poznámka: Tabulka vážených číselných charakteristik bude obsahovat: vážený průměr, vážený rozptyl a váženou směrodatnou odchylku.
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) $s_{w,12} = -0,0468$. ♣