

3 Základní číselné charakteristiky

V předchozí kapitole jsme se seznámili se základními metodami umožňující prvotní náhled na datový soubor, přičemž jsme se v závislosti na typu sledovaného znaku, který byl buď kategoriální nebo spojitý zabývali různými metodami číselné a grafické vizualizace. Metody představené v kapitole ?? mají jednu společnou vlastnost. Vždy nám poskytují široké množství informací o sledovaném znaku, což nám umožňuje utvořit si globální a ucelený pohled na tento znak. Nevýhodou však může být právě přemíra informací, která se jednak hůře interpretuje a jednak neumožňuje snadné porovnávání znaků z různých datových souborů.

Výše uvedené nedostatky vedly k potřebě zavedení pojmů, které elegantně a jednoduše vystihují základní charakteristické rysy sledovaného znaku. Tyto pojmy se nazývají *číselné charakteristiky* a jejich výhodou je, že sledované vlastnosti znaku dokáží vystihnout pomocí jednoho čísla. Podle vlastnosti, kterou popisují rozlišujeme celkem čtyři základní typy číselných charakteristik: (1) charakteristiky polohy; (2) charakteristiky variability; (3) charakteristiky symetrie; (4) charakteristiky závislosti.

Podoně jako jsme si v sekci ?? představili odlišné výpočetní a grafické metody pro znaky diskrétního typu a pro znaky spojitého typu, tak i zde používáme různé číselné charakteristiky pro různé typy znaků. Celkem rozlišujeme tři základní typy znaků: (a) nominální znaky; (b) ordinální znaky; (c) intervalové znaky. Přehled číselných charakteristik používaných v závislosti na typu znaku a vlastnosti, kterou popisují, je uveden v tabulce 1.

Tabulka 1: Přehled číselných charakteristik v závislosti na typu znaku a popisované vlastnosti

	Poloha	Variabilita	Symetrie	Závislost
Nominální	modus	–	–	Cramérův koeficient
Ordinální	medián	mezikvartilové rozpětí	–	Spearmanův korel. koeficient
Intervalová	aritmetický průměr	rozptyl směrodatná odchylka	koeficient šikmosti koeficient špičatosti	Pearsonův korel. koeficient

3.1 Číselné charakteristiky pro nominální znaky

Příklad 3.1. Charakteristika polohy nominálního znaku

Navážme na práci s datasetem `17-anova-newborns.txt`. V rámci sekce ?? jsme jako mezivýstup příkladu ?? získali kontingenční tabulku simultánních absolutních četností znaků $X = \text{vzdělání matky}$ a $Y = \text{porodní hmotnost novorozence}$ (viz tabulka 2). Znaky X a Y jsou typickým příkladem znaků nominálního typu. Najděte modus pro znak *vzdělání matky* i pro znak *porodní hmotnost novorozence*.

Tabulka 2: Simultánní absolutní četnosti znaků *vzdělání matky* a *porodní hmotnost novorozence*

	nízká	norma	vysoká
ZS	75	264	8
SS	79	325	20
SSm	73	341	11
VS	13	63	4

Řešení příkladu 3.1

Kontingenční tabulku simultánních absolutních četností znaků X a Y bychom mohli získat provedením posloupností kroků uvedených příkladech ??–??, tj. načtením datového souboru `17-anova-newborns.txt`, odstraněním neznámých hodnot, vyselektováním údajů o novorozencích s maximálně dvěma staršími sourozenci, kategorizací spojitě proměnné `weight.C` a vytvořením tabulky simultánních absolutních četností pro znaky X a Y . My však využijeme znalosti tabulky 2 a kontingenční tabulku simultánních absolutních četností vytvoříme pomocí příkazu `data.frame()`.

```
1 (data <- data.frame(nizka = c( 75,  79,  73, 13),
2                       norma = c(264, 325, 341, 63),
```

```
3     vysoka = c( 8, 20, 11, 4),
4     row.names = c('ZS', 'SS', 'SSm', 'VS'))
```

	nizka	norma	vysoka
ZS	75	264	8
SS	79	325	20
SSm	73	341	11
VS	13	63	4

5
6
7
8
9

Zaměříme se nejprve na znak X . Číselná charakteristika *modus* je definována jako nejčetnější varianta sledovaného znaku. K získání modu znaku $X = \text{vzdělání matky}$ musíme zjistit četnost výskytu jednotlivých variant tohoto znaku bez ohledu na porodní hmotnost novorozence. Jinými slovy potřebujeme vypočítat vektor absolutních marginálních četností $n_{j.}$ pro varianty znaku X . Analogicky jako v příkladu ?? použijeme funkci `apply()` se specifikací argumentů `MARGIN = 1` a `FUN = sum`.

```
10 (n.j. <- apply(data, MARGIN = 1, FUN = sum))
```

ZS	SS	SSm	VS
347	424	425	80

11
12

Interpretace výsledků: Nejčetnější variantou znaku *vzdělání matky* je středoškolské vzdělání s maturitou ($n_{SSm} = 425$). Nejvíce novorozenců v datovém souboru s maximálně dvěma staršími sourozenci se narodilo matkám s dokončeným středoškolským vzděláním s maturitou.

Analogicky najdeme modus znaku $Y = \text{porodní hmotnost novorozence}$. K získání modu pro znak Y musíme zjistit četnost výskytu jednotlivých variant tohoto znaku bez ohledu na vzdělání matky. Jinými slovy potřebujeme vypočítat vektor absolutních marginálních četností $n_{.k}$ pro varianty znaku Y . Opět použijeme funkci `apply()` tentokrát se specifikací argumentu `MARGIN = 2`.

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

nizka	norma	vysoka
240	993	43

14
15

Interpretace výsledků: Nejvíce novorozenců v datovém souboru s maximálně dvěma staršími sourozenci mělo porodní hmotnost v normě ($n_{norma} = 993$).



Příklad 3.2. Charakteristika závislosti mezi dvěma nominálními znaky

Zaměříme se nyní na oba znaky $X = \text{vzdělání matky}$ a $Y = \text{porodní hmotnost novorozence}$ najednou. Určete míru závislosti mezi znaky X a Y .

Řešení příkladu 3.2

Protože X a Y jsou znaky nominálního typu, použijeme na určení míry závislosti mezi nimi *Cramérův koeficient*. Tento koeficient nabývá hodnoty z intervalu $\langle 0; 1 \rangle$, přičemž vyšší hodnota Cramérova koeficientu ukazuje na těsnější vztah mezi oběma znaky. Stupnice míry závislosti podle hodnoty Cramérova koeficientu je uvedena v tabulce 3.

Tabulka 3: Stupnice míry závislosti pro Cramérův koeficient

Cramérův koeficient r_C	Interpretace
$\langle 0.0; 0.1 \rangle$	Zanedbatelný stupeň závislosti
$\langle 0.1; 0.3 \rangle$	Slabý stupeň závislosti
$\langle 0.3; 0.7 \rangle$	Střední stupeň závislosti
$\langle 0.7; 1.0 \rangle$	Silný stupeň závislosti

Přesným postupem výpočtu Cramérova koeficientu se budeme zabývat v kapitole ???. Nyní stanovíme hodnotu Cramérova koeficientu pomocí funkce `cramersV()`, která je součástí knihovny `lsr`. Abychom mohli funkci `cramersV()` použít, musíme knihovnu `lsr` nainstalovat (RStudio → multifunkční okno → záložka Packages → ikona Install → knihovna: `lsr` → Install) a načíst. Celou knihovnu `lsr` je možné načíst příkazem `library(lsr)`. Pro nás je však zbytečné načítat celou knihovnu, proto pomocí operátoru `::` pouze zavoláme z knihovny `lsr` funkci `cramersV()`.

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

```
[1] 0.05502639
```

17

Interpretace výsledků: Cramérovův koeficient nabývá hodnoty 0.0550. Mezi vzděláním matky a porodní hmotností novorozence existuje zanedbatelný stupeň závislosti. ♣

Dataset 3: Zakončení palmárních linií

Vě vzorku, který tvořilo 200 studentů (100 mužů a 100 žen), byly standartní dermatoglyfickou metodikou snímané dermatoglyfy dlaně (Býmová, 1990; soubor `22-multinom-palmar-lines.txt`). Na otiscích byla hodnocena zakončení tří hlavních palmárních linií (D, C, a B). Případy byly podle vzoru zakončení (vyústění proximálních radiantů digitálních trirádií na standartně číslovaných polohách okraje dlaně) rozdělené do tří kategorií. Současně byla hodnocena barva vlasů podle standartní Fischer-Sallerové stupnice 30 odstínů (Martin a Saller, 1957–1966, s. 391), které byly rozděleny do tří skupin. K dispozici máme početnosti jedinců v jednotlivých kategoriích, zvláště pro muže a pro ženy.

Popis proměnných v datasetu 3:

- `sex` – pohlaví (m – muž, f – žena);
- `palmar.lines` – zakončení tří palmárních linií (Hi - vysoké (nejčastější vzorec 11 9 7), Mi - střední (nejčastější vzorec 9 7 5), Lo - nízké (nejčastější vzorec 7 5 5));
- `hair.C` - barva vlasů (LiH - světlé, MH - střední, DaH - tmavé).

Příklad 3.3. Charakteristika polohy nominálního znaku

Načtete datový soubor `22-multinom-palmar-lines.txt` a prohlédněte si jej. Z tabulky vyselektujte pouze údaje týkající se znaků $X = \text{barva vlasů}$ a $Y = \text{zakončení palmárních linií}$ u žen. Změňte záhlaví tabulky tak, aby názvy jednotlivých variant znaku $X = \text{barva vlasů}$ byly: světlé, střední a tmavé; a názvy jednotlivých variant znaku $Y = \text{zakončení palmárních linií}$ byly: vysoké, střední a nízké. Stanovte vhodnou charakteristiku polohy pro znak X i pro znak Y .

Řešení příkladu 3.3

Datový soubor načteme příkazem `read.delim()`.

```
18 (data <- read.delim('22-multinom-palmar-lines.txt'))
```

	m	Hi	Mi	Lo	X	f	Hi.1	Mi.1	Lo.1
1	LiH	6	6	4	NA	LiH	4	6	6
2	MH	20	15	7	NA	MH	18	10	10
3	DaH	18	12	12	NA	DaH	12	22	12

19
20
21
22

Načtená datová tabulka obsahuje celkem 9 sloupců, z nichž první čtyři sloupce tvoří tabulku simultánních absolutních četností výskytu dvojic variant znaků X a Y pro muže, pátý sloupec obsahuje NA hodnoty sloužící jako oddělovače tabulky s údaji pro muže od tabulky s údaji pro ženy a poslední čtyři sloupce tvoří tabulku simultánních absolutních četností výskytu dvojic variant znaků X a Y pro ženy.

Pomocí logického operátoru `[]` vybereme z tabulky `data` pouze simultánní absolutní četnosti znaků *barva vlasů* a *zakončení palmárních linií* u žen a vložíme je do proměnné `data.f`. Příkazem `row.names()` doplníme do tabulky `data.f` názvy řádků příslušející jednotlivým variantám znaku $X = \text{barva vlasů}$. Příkazem `names()` doplníme do tabulky názvy sloupců příslušející variantám znaku $Y = \text{zakončení palmárních linií}$.

```

23 data.f <- data[, 7:9]
24 row.names(data.f) <- c('svetle', 'stredni', 'tmave')
25 names(data.f) <- c('vysoke', 'stredni', 'nizke')
26 data.f

```

	vysoke	stredni	nizke
svetle	4	6	6
stredni	18	10	10
tmave	12	22	12

27
28
29
30

Znaky X a Y jsou nominálního typu, proto jako vhodnou charakteristiku polohy zvolíme v obou případech *modus*. K získání modu znaku X musíme zjistit četnost výskytu jednotlivých variant tohoto znaku bez ohledu na typ zakončení palmárních linií. Pomocí funkce `apply()` se specifikací argumentů `MARGIN = 1` a `FUN = sum` najdeme marginální vektor absolutních četností jednotlivých variant znaku X .

```

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

```

svetle	stredni	tmave
16	38	46

32
33

Interpretace výsledků: Nejčetnější variantou znaku *barva vlasů* u žen v datovém souboru je tmavá barva ($n_{\text{tmavé}} = 46$). Nejvíce žen v datovém souboru mělo tmavé vlasy.

Analogicky najdeme modus znaku $Y = \text{zakončení palmárních linií}$. K získání modu znaku Y u žen musíme zjistit četnost výskytu jednotlivých variant tohoto znaku bez ohledu na barvu vlasů žen. Funkci `apply()` nyní použijeme se specifikací argumentu `MARGIN = 2`.

```

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

```

vysoke	stredni	nizke
34	38	28

35
36

Interpretace výsledků: Nejvíce žen v datovém souboru mělo střední zakončení palmárních linií ($n_{\text{střední}} = 38$). ♣

Příklad 3.4. Charakteristika závislosti mezi dvěma nominálními znaky

Zaměříme se nyní na oba znaky $X = \text{barva vlasů}$ a $Y = \text{zakončení palmárních linií}$ u žen najednou. Určete míru závislosti mezi znaky X a Y .

Řešení příkladu 3.4

Protože X a Y jsou znaky nominálního typu, použijeme na určení míry závislosti mezi nimi *Cramérův koeficient*. Hodnotu Cramérova koeficientu stanovíme pomocí funkce `cramersV()` z knihovny `lsr`.

```

37 lsr::cramersV(data.f)

```

```
[1] 0.1785374
```

38

Interpretace výsledků: Cramérův koeficient nabývá hodnoty 0.1785. Mezi barvou vlasů a zakončením palmárních linií u žen existuje slabý stupeň závislosti. ♣

3.2 Číselné charakteristiky pro ordinální znaky

Příklad 3.5. Základní číselné charakteristiky pro ordinální znak

Načtěte datový soubor `17-anova-newborns.txt`, ze souboru odstraňte neznámé hodnoty a zjistěte dimenzi datové tabulky. Zaměřte se tentokrát na všechny novorozence v datovém souboru a vytvořte tabulku vhodných základních číselných charakteristik pro znak $X = \text{počet starších sourozenců}$.

Řešení příkladu 3.5

Nejprve načteme datový soubor (`read.delim()`), odstraníme neznámé hodnoty (`na.omit()`) a vypíšeme dimenzi datové tabulky (`dim()`).

```
39 data <- read.delim('17-anova-newborns.txt')
40 data <- na.omit(data)
41 dim(data)
```

42

```
[1] 1382 4
```

Po odstranění neznámých hodnot obsahuje datová tabulka údaje o 1382 novorozencích, přičemž u každého novorozence máme záznamy o čtyřech znacích.

Znak $X = \text{počet starších sourozenců}$ novorozence je příkladem ordinálních dat. Ordinální data můžeme navzájem porovnávat, (nula starších sourozenců je méně než jeden starší sourozenec a to je méně než dva starší sourozenci), ale uvědomujeme si, že rozestupy mezi sousedními variantami nejsou stejné (rozdíl prvorodičkou a druhorodičkou je propastnější než rozdíl mezi druhorodičkou a třetírodičkou). V tabulce základních charakteristik budou obsaženy následující charakteristiky: minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota a mezikvartilové rozpětí.

Výpočet α -kvantilu x_α

Předpokládejme, že α je libovolná hodnota z intervalu $(0; 1)$. Pojmeme α -kvantil, nebo také $\alpha \times 100\%$ kvantil, značíme takové číslo x_α , pro které $\alpha \times 100\%$ hodnot z datového souboru leží nalevo od x_α a $(1 - \alpha) \times 100\%$ hodnot leží napravo od x_α . Výpočet α -kvantilu je tedy úzce spjatý s počtem objektů v datovém souboru n . Při výpočtu α -kvantilu mohou nastat dvě situace:

1. $n \times \alpha = c$, kde c je celé číslo. V takovém případě dopočítáme hodnotu kvantilu x_α jako aritmetický průměr c -tého a $(c + 1)$ -tého čísla v posloupnosti **seřazených** naměřených hodnot, tj.

$$x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}, \quad (1)$$

2. $n \times \alpha = c$, kde c není celé číslo. V takovém případě zaokrouhlíme c na nejbližší vyšší celé číslo a hodnota kvantilu x_α je rovna c -tému číslu v posloupnosti **seřazených** naměřených hodnot, tj.

$$x_\alpha = x_{(c)}. \quad (2)$$

Nejprve se zaměříme na výpočet dolního kvantilu znaku $X = \text{počet starších sourozenců}$. Koeficient α je v tomto případě rovný 0.25, počet novorozenců $n = 1382$. Součin $c = n \times \alpha = 1382 \times 0.25 = 345.5$ není celé číslo, proto jej zaokrouhlíme na nejbližší vyšší celé číslo, tj. 346. Dolní kvartil $x_{0.25}$ bude potom odpovídat 346. hodnotě v posloupnosti seřazených naměřených hodnot. Hodnoty ve vektoru `prch` seřadíme vzestupně pomocí příkazu `sort()`. V pořadí 346. hodnotu ze seřazeného vektoru `prch` získáme pomocí operátoru `[]`.

```
43 prch <- sort(data$prch)
44 prch[346]
```

45

```
[1] 0
```

$$x_{0.25} = x_{(346)} = 0. \quad (3)$$

V případě výpočtu mediánu $x_{0.50}$ je $\alpha = 0.50$ a počet novorozenců $n = 1382$. Součin $c = n \times \alpha = 1382 \times 0.50 = 691$ je celé číslo, proto hodnotu mediánu stanovíme jako průměr hodnot na 691. a 692. pozici v seřazeném vektoru `prch`.

```
46 prch[691]
```

47

```
[1] 1
```

```
48 prch[692]
```

49

```
[1] 1
```

$$x_{0.50} = \frac{x_{(691)} + x_{(692)}}{2} = \frac{1 + 1}{2} = 1.$$

Při výpočtu horního kvartilu $x_{0.75}$ je $\alpha = 0.75$ a počet novorozenců $n = 1382$. Součin $c = n \times \alpha = 1382 \times 0.75 = 1036.5$ není celé číslo, proto jej zaokrouhlíme na nejbližší vyšší celé číslo, tj. 1037, a horní kvartil bude odpovídat 1037. hodnotě v posloupnosti seřazených naměřených hodnot.

```
50 prch [1037]
```


```
[1] 1
```

51

$$x_{0.75} = x_{(1037)} = 1.$$

Mezikvartilové rozpětí spočítáme odečtením dolního kvartilu od horního kvartilu, tj.

$$IQR = x_{0.75} - x_{0.50} = 1 - 0 = 1.$$

Všechny výše zmíněné charakteristiky můžeme vypočítat také pomocí funkcí implementovaných v softwaru . Hodnoty kvantilů stanovíme příkazem `quantile()`. Prvním argumentem příkazu bude vektor seřazených nebo neseřazených údajů o počtu starších sourozenců (`prch`). Druhým argumentem `probs` specifikujeme hodnotu α (0.25, 0.50, resp. 0.75). Nakonec specifikací argumentu `type = 2` vybereme z devíti možných metod výpočtu, které funkce `quantile()` poskytuje, metodu odpovídající ručnímu výpočtu. Interkvartilové rozpětí vypočítáme pomocí funkce `IQR()` opět se specifikací argumentu `type = 2`. Nakonec stanovíme minimální, resp. maximální počet starších sourozenců u novorozenců v datovém souboru pomocí příkazu `min()`, resp. `max()` a všechny hodnoty vložíme do tabulky příkazem `data.frame()`.

```
52 x0.25 <- quantile(prch, probs = 0.25, type = 2)
53 x0.50 <- quantile(prch, probs = 0.50, type = 2)
54 x0.75 <- quantile(prch, probs = 0.75, type = 2)
55 IQR <- IQR(prch, type = 2)
56 min <- min(prch)
57 max <- max(prch)
58 (Tab <- data.frame(min = min, dolni.kv = x0.25, median = x0.50,
59                   horni.kv = x0.75, max = max, IQR = IQR,
60                   row.names = 'pocet st. sourozencu'))
```

```
min dolni.kv median horni.kv max IQR
pocet st. sourozencu 0 0 1 1 9 1
```

61
62

Interpretace výsledků: Počet starších sourozenců u novorozenců v datovém souboru se pohybuje v rozmezí 0–9. Dolní kvartil počtu starších sourozenců nabývá hodnoty 0, tj. 25% novorozenců v datovém souboru nemá více než nula starších sourozenců. Medián počtu starších sourozenců nabývá hodnoty 1, tj. 50% novorozenců v datovém souboru má jednoho staršího sourozence nebo méně. Horní kvartil počtu starších sourozenců nabývá hodnoty 1, tj. 75% novorozenců v datovém souboru má jednoho staršího sourozence nebo méně. Hodnota mezikvartilového rozpětí je rovna jedné. ♣

Příklad 3.6. Krabicový diagram

Sestrojte krabicový diagram pro znak $X = \text{počet starších sourozenců novorozence}$. Zaměřte se na vzhled vykresleného grafu a zamyslete se nad polohou mediánu, dolního kvartilu, horního kvartilu a mezikvartilového rozpětí v krabicovém diagramu.

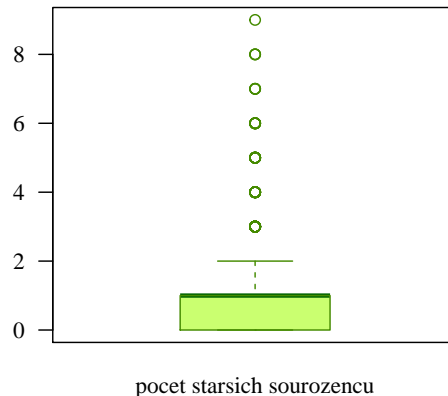
Řešení příkladu 3.6

Krabicový diagram vykreslíme příkazem `boxplot()`. Prvním argumentem bude vektor počtu starších sourozenců `prch`, argumentem `type = 2` vybereme k výpočtu kvantilů zobrazených v grafu metodu analogickou ručnímu výpočtu. Dále nastavíme barvu výplně grafu (`col`), barvu ohrazení grafu (`border`), barvu mediánu (`medcol`) v zelených odstínech a vodorovné vykreslení popisků u měřítka osy y (`las`). Argumentem `xlab = ''` zamezíme vypsání popisku osy x . Ten doplníme do grafu samostatně pomocí příkazu `mtext()`. Prvním argumentem tohoto příkazu bude text popisku. Argumentem `side = 1` specifikujeme umístění popisku pod dolní stranu grafu a argumentem `line` zvolíme umístění popisku do výšky 1.5.

```

63 boxplot(prch, type = 2, col = 'darkolivegreen1',
64         border = 'chartreuse4', medcol = 'darkgreen',
65         las = 1, xlab = '')
66 mtext('pocet starsich sourozencu', side = 1, line = 1.5)

```



Příklad 3.7. Charakteristika závislosti mezi ordinálními znaky

Zaměříme se nyní na znaky $X = \text{počet starších sourozenců}$ a $Y = \text{porodní hmotnost novorozence}$ najednou. Určete míru závislosti mezi znaky X a Y .

Řešení příkladu 3.7

Znak X je ordinálního typu, zatímco znak Y je typickým případem znaku intervalového typu. Vzhledem k tomu, že znaky intervalového typu jsou bohatší na informace než znaky ordinálního typu, můžeme k nim bez jakékoli újmy přistupovat jako k ordinálním znakům. Konkrétně tedy na znak Y budeme v tomto případě nahlížet jako na ordinální znak.

Ke stanovení míry závislosti mezi znaky X a Y použijeme *Spearmanův koeficient pořadové korelace* r_S . Tento koeficient nabývá hodnoty mezi -1 a 1 , tj. $r_S \in (-1; 1)$, přičemž kladné znaménko koeficientu určuje přímý směr *pořadové závislosti* a záporné znaménko určuje nepřímý směr *pořadové závislosti*. Stupnice těsnosti závislosti mezi dvěma znaky podle hodnoty Spearmanova koeficientu r_S je uvedena v tabulce 4. Detailněji se na výpočet Spearmanova koeficientu pořadové korelace zaměříme v kapitole ??.

Spearmanův koeficient pořadové korelace r_S vypočítáme pomocí funkce `cor()` se specifikací argumentu `method = 'spearman'`. První dva argumenty zadané do funkce budou vektory naměřených hodnot znaků X (`prch`) a Y (`wei`).

```

67 prch <- data$prch.N
68 wei  <- data$wei
69 (rS <- cor(prch, wei, method = 'spearman'))

```

```
[1] 0.04761724
```

70

Interpretace výsledků: Hodnota Spearmanova koeficientu pořadové korelace $r_S = 0.0476$. Mezi počtem starších sourozenců a porodní hmotností novorozence existuje velmi nízký stupeň přímé pořadové závislosti.



Příklad 3.8. Dvourozměrný tečkový diagram

Pro znaky $X = \text{počet starších sourozenců}$ a $Y = \text{porodní hmotnost novorozence}$ vykreslete dvourozměrný tečkový diagram. Pozastavte se nad vzhledem tečkového diagramu a jeho vztahem k hodnotě Spearmanova koeficientu pořadové korelace.

Tabulka 4: Stupnice míry závislosti pro Spearmanův a Pearsonův korelační koeficient

$ r_S $, resp. $ r_{12} $	Interpretace
0.0	Pořadová (resp. 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 pořadová (resp. lineární) závislost

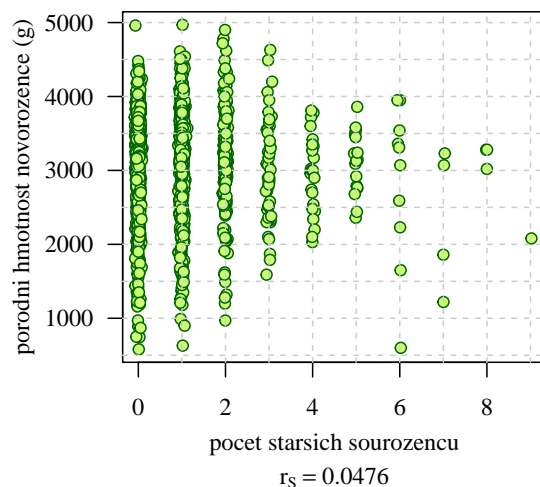
Řešení příkladu 3.8

Dvourozměrný tečkový diagram sestojíme příkazem `dotplot()`, který je součástí RSkriptu `Sbirka-AS-I-2018-funkce.R`. Ten načteme příkazem `source()`. Vykreslovaným bodům přisoudíme kulatý tvar `pch = 21` s tmavě zeleným obvodem (`col`) a světlezelenou výplní (`bg`). Pomocí příkazu `abline()` dokreslíme do grafu horizontální referenční čáry (specifikace argumentu `h`) a vertikální referenční čáry (specifikace argumentu `v`). Poznamenejme, že v příkazu `dotplot()` jsme volbou argumentu `main = ''` zakázali vypsání nadpisu a volbou argumentu `xlab = ''` zase vypsání popisku osy x . Popisek osy x doplníme do grafu samostatně příkazem `mtext()`. Pomocí stejné funkce v kombinaci s funkcí `bquote()` přidáme do grafu druhý popisek s hodnotu Spearmanova koeficientu pořadové korelace r_S zaokrouhlenou na čtyři desetinná místa. Funkce `bquote()` zadaná uvnitř příkazu `mtext()` umožňuje vytvoření specifického popisku. Zápis `r[S]` vysází písmeno r s indexem S , tj. r_S . Symbol `==` v příkazu `bquote()` odpovídá syntaxi symbolu $=$ a vyjádření `.(rS)` vyčíslí hodnotu uloženou v proměnné `rS`, tj. 0.0476.

```

71 source('Sbirka-AS-I-2018-funkce.R')
72 rS <- round(rS, digits = 4)
73
74 dotplot(prch, wei, main = '', xlab = '',
75         ylab = 'porodni hmotnost novorozence (g)', pch = 21,
76         bg = 'darkolivegreen1', col = 'darkgreen')
77
78 abline(h = seq(0, 5000, by = 500), col = 'grey80', lty = 2)
79 abline(v = seq(1, 10, by = 1), col = 'grey80', lty = 2)
80 mtext('pocet starsich sourozencu', side = 1, line = 2.2)
81 mtext(bquote(r[S] == .(rS)), side = 1, line = 3.5)

```

3.3 Číselné charakteristiky pro intervalové znaky

Příklad 3.9. 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 3.9


Načtení datového souboru provedeme příkazem `read.delim()`, odstranění NA hodnot příkazem `na.omit()`. Pomocí podmnožinového operátoru `[]` vybereme z tabulky `data` pouze údaje o největší šířce mozkovny (`skull.B`) pro muže. Naměřené hodnoty si příkazem `sort()` vzestupně seřadíme.

```
82 data <- read.delim('01-one-sample-mean-skull-mf.txt')
83 data <- na.omit(data)
84 skull.BM <- data[data$sex == 'm', 'skull.B']
85 skull.BM <- sort(skull.BM)
86 length(skull.BM)
```

```
[1] 216
```

87

Po odstranění neznámých hodnot obsahuje datová tabulka údaje o 216 skeletech mužského pohlaví.

Znak $X =$ největší šířka mozkovny pro skelety mužského pohlaví je příkladem intervalového typu dat. V tabulce základních číselných charakteristik budou obsaženy následující charakteristiky: aritmetický průměr, rozptyl, směrodatná odchylka, koeficient variace, minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota, mezikvartilové rozpětí, koeficient šikmosti a koeficient špičatosti. Nejprve se podíváme na ruční výpočet každé z těchto číselných charakteristik a následně provedeme kontrolu výsledků pomocí softwaru .

Aritmetický průměr m vypočítáme pomocí vzorce

$$m = \frac{1}{n} \sum_{i=1}^n x_i, \quad (4)$$

kde x_i , $i = 1, \dots, n$, je i -tá naměřená hodnota a $n = 216$.

$$m = \frac{1}{216}(124 + 127 + \dots + 149 + 149) = \frac{29\,632}{216} = 137.1852.$$

Rozptyl s^2 vypočítáme pomocí vzorce

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2, \quad (5)$$

kde x_i , $i = 1, \dots, n$, je i -tá naměřená hodnota, $n = 216$ a m je aritmetický průměr.

$$\begin{aligned} s^2 &= \frac{1}{216} ((124 - 137.1852)^2 + (127 - 137.1852)^2 + \dots + (149 - 137.1852)^2 + (149 - 137.1852)^2) \\ &= \frac{1}{216} ((-13.1852)^2 + (-10.1852)^2 + \dots + 11.8148^2 + 11.8148^2) \\ &\doteq 23.1694. \end{aligned}$$

Směrodatnou odchylku s vypočítáme jako odmocninu z rozptylu, tj.

$$s = \sqrt{s^2} = \sqrt{23.1694} \doteq 4.8135.$$

Koeficient variace v je definovaný jako podíl směrodatné odchylky a aritmetického průměru vynásobený 100%, tj.

$$v = \frac{s}{m} \times 100\% = \frac{4.8135}{137.1852} \times 100\% = 0.035087 \times 100\% \doteq 3.5087\%.$$

Minimální naměřenou hodnotu nalezneme na první pozici v seřazeném vektoru skull.BM.

88 skull.BM [1]

[1] 124

89

$$x_{\min} = 124.$$

V případě výpočtu dolního kvartilu postupujeme analogicky jako v příkladu 3.5. Koeficient $\alpha = 0.25$, počet naměřených hodnot $n = 216$. Součin $c = n \times \alpha = 216 \times 0.25 = 54$ je celé číslo, tedy hodnotu dolního kvartilu $x_{0.25}$ stanovíme jako průměr 54. a 55. hodnoty v posloupnosti seřazených naměřených hodnot.

90 skull.BM [54]

[1] 134

91

92 skull.BM [55]

[1] 134

93

$$x_{0.25} = \frac{x_{(54)} + x_{(55)}}{2} = \frac{134 + 134}{2} = 134.$$

Pro výpočet mediánu $x_{0.50}$ je $\alpha = 0.50$ a počet naměřených hodnot $n = 216$. Součin $c = n \times \alpha = 216 \times 0.50 = 108$ je celé číslo, proto hodnotu mediánu $x_{0.50}$ stanovíme jako průměr hodnot na 108. a 109. pozici v posloupnosti seřazených naměřených hodnot.

94 skull.BM [108]

[1] 137

95

96 skull.BM [109]

[1] 137

97

$$x_{0.50} = \frac{x_{(108)} + x_{(109)}}{2} = \frac{137 + 137}{2} = 137.$$

V případě výpočtu horního kvartilu je $\alpha = 0.75$ a počet naměřených hodnot $n = 216$. Součin $n \times \alpha = 216 \times 0.75 = 162$, je celé číslo, tedy hodnota horního kvartilu bude rovná průměru 162. a 163. hodnoty v posloupnosti seřazených naměřených hodnot.

98 skull.BM [162]

[1] 140

99

100 skull.BM [163]

[1] 140

101

$$x_{0.75} = \frac{x_{(162)} + x_{(163)}}{2} = \frac{140 + 140}{2} = 140.$$

Maximální naměřenou hodnotu nalezneme na poslední pozici v seřazeném vektoru skull.BM.

102 skull.BM [216]

[1] 149

103

$$x_{\max} = 149$$

Mezikvartilové rozpětí IQR získáme odečtením hodnoty dolního kvartilu od hodnoty horního kvartilu, tj.

$$IQR = x_{0.75} - x_{0.25} = 140 - 134 = 6.$$

Koeficient šikmosti b_1 vypočítáme pomocí vzorce

$$b_1 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - m)^3}{s^3}, \quad (6)$$

kde x_i , $i = 1, \dots, n$, je i -tá naměřená hodnota, $n = 216$, m je aritmetický průměr a s je směrodatná odchylka.


$$\begin{aligned} b_1 &= \frac{1}{216} \frac{(124 - 137.1852)^3 + (127 - 137.1852)^3 + \dots + (149 - 137.1852)^3 + (149 - 137.1852)^3}{23.2772^3} \\ &= \frac{1}{216} \frac{(-13.1852)^3 + (-10.1852)^3 + \dots + 11.8148^3 + 11.8148^3}{4.824642^3} \\ &= \frac{2040.299}{24257.67} = 0.0841094 \doteq 0.0841. \end{aligned}$$

Koeficient špičatosti b_2 vypočítáme pomocí vzorce

$$b_2 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - m)^4}{s^4} - 3, \quad (7)$$

kde x_i , $i = 1, \dots, n$, je i -tá naměřená hodnota, $n = 216$, m je aritmetický průměr a s je směrodatná odchylka.

$$\begin{aligned} b_2 &= \frac{1}{216} \frac{(124 - 137.1852)^4 + (127 - 137.1852)^4 + \dots + (149 - 137.1852)^4 + (149 - 137.1852)^4}{8.424642^4} - 3 \\ &= \frac{1}{216} \frac{(-13.1852)^4 + (-10.1852)^4 + \dots + 11.8148^4 + 11.8148^4}{4.824642^4} - 3 \\ &= \frac{316498.6}{117034.6} - 3 = -0.295683 \doteq -0.2957. \end{aligned}$$

Všechny výše zmíněné základní charakteristiky můžeme vypočítat pomocí funkcí zabudovaných v softwaru . Aritmetický průměr získáme příkazem `mean()`, rozptyl pomocí funkcí `mean()` a `sum()` a směrodatnou odchylku jako odmocninu z rozptylu pomocí příkazu `sqrt()`. Koeficient variace vypočítáme jako podíl směrodatné odchylky a

aritmetického průměru vynásobený stem. Minimální resp. maximální naměřenou hodnotu získáme příkazem `min()`, resp. `max()`. Hodnotu dolního kvartilu, mediánu a horního kvartilu vypočítáme funkcí `quantile()` s volbou ruční metody výpočtu (`type = 2`), kde specifikací argumentu `probs` stanovíme hodnotu koeficientu α (0.25, 0.50 a 0.75). Mezikvartilové rozpětí spočítáme příkazem `IQR()` opět se specifikací argumentu (`type = 2`). Koeficient šikmosti, resp. špičatosti získáme pomocí funkce `skewness()`, resp. `kurtosis()`, které jsou součástí balíčku `e1071`. Volbou argumentu `type = 3` vybereme ze tří dostupných metod výpočtu koeficientů metody analogické vzorcům 6 a 7. Poznamenejme, že balíček `e1071` není mezi defaultně nainstalovanými balíčky a je tedy potřeba jej doinstalovat.

Na závěr všechny hodnoty vložíme do jedné tabulky (`data.frame()`), kterou vypíšeme se zaokrouhlením na čtyři desetinná místa (`round()`).

```

104 m <- mean(skull.BM)
105 s2 <- 1 / 216 * sum((skull.BM - m)^2)
106 s <- sqrt(s2)
107 v <- s / m * 100
108
109 min <- min(skull.BM)
110 x0.25 <- quantile(skull.BM, probs = 0.25, type = 2)
111 x0.50 <- quantile(skull.BM, probs = 0.50, type = 2)
112 x0.75 <- quantile(skull.BM, probs = 0.75, type = 2)
113 IQR <- IQR(skull.BM, type = 2)
114 max <- max(skull.BM)
115
116 sikmost <- e1071::skewness(skull.BM, type = 3)
117 spicatost <- e1071::kurtosis(skull.BM, type = 3)
118
119 tab <- data.frame(m, var = s2, s, v, min, dolni.k = x0.25, median = x0.50,
120                 horni.k = x0.75, max, IQR, sikmost, spicatost,
121                 row.names = 'm-S')
122 round(tab, digits = 4)

```

	m	var	s	v	min	dolni.k	median	horni.k	max	IQR	sikmost	spicatost
m-S	137.1852	23.1694	4.8135	3.5087	124	134	137	140	149	6	0.0841	-0.2957

123
124

Interpretace výsledků: Naměřené hodnoty největší šířky mozkovny pro skelety mužského pohlaví se pohybují v rozmezí 124–149 mm. Průměrná hodnota největší šířky mozkovny u skeletů mužského pohlaví je 137.19 mm se směrodatnou odchylkou 4.81 mm, přičemž směrodatná odchylka představuje 3.51% aritmetického průměru. 25% naměřených hodnot je menších nebo rovných 134 mm, 50% naměřených hodnot je menších nebo rovných 137 mm a 75% naměřených hodnot je menších nebo rovných 140 mm. Mezikvartilové rozpětí má rozsah 6 mm. Hodnota koeficientu šikmosti, 0.0841, ukazuje na kladně zešikmená data s prodlouženým pravým koncem. Hodnota koeficientu šikmosti je však tak blízka nule, že zmíněný efekt zešikmení nebude téměř znatelný. Hodnota koeficientu špičatosti, -0.2957, ukazuje na plošší charakter dat. ♣

Příklad 3.10. 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 3.10

Oba znaky X a Y jsou intervalového typu. Ke stanovení míry závislosti mezi těmito znaky použijeme Pearsonův korelační koeficient r_{12} . Tento koeficient nabývá hodnoty mezi -1 a 1, tj. $r_{12} \in (-1; 1)$, přičemž kladné znaménko koeficientu určuje přímý směr *lineární* závislosti a záporné znaménko určuje nepřímý směr *lineární* závislosti. Stupnice těsnosti závislosti mezi dvěma znaky podle hodnoty Pearsonova korelačního koeficientu je uvedena v tabulce 4 společně se stupnicí pro Spearmanův koeficient pořadové korelace.

Hodnotu Pearsonova korelačního koeficientu spočítáme příkazem `cor()` se specifikací argumentu `method = 'pearson'`. První dva argumenty příkazu budou vektory naměřených hodnot znaků X (`skull.BM`) a Y (`skull.LM`).

```

125 skull.BM <- data[data$sex == 'm', 'skull.B']
126 skull.LM <- data[data$sex == 'm', 'skull.L']
127 (r12 <- cor(skull.BM, skull.LM, method = 'pearson'))

```

Interpretace výsledků: Pearsonův korelační koeficient nabývá hodnoty 0.1682. Mezi největší šířkou a délkou mozkovny u skeletů mužského pohlaví existuje nízký stupeň přímé lineární závislosti. ♣

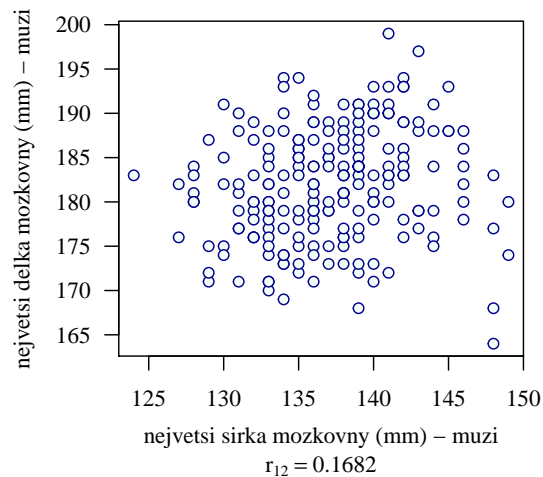
Příklad 3.11. Dvourozměrný tečkový diagram

Výslednou míru závislosti mezi znaky $X = \text{největší šířka mozkovny}$ a $Y = \text{největší délka mozkovny}$ pro skelety mužského pohlaví vizualizujeme pomocí dvourozměrného tečkového diagramu sestrojeného v rámci příkladu ???. Do diagramu doplníme akorát popisek s hodnotou Pearsonova korelačního koeficientu r_{12} .

Řešení příkladu 3.11

Dvourozměrný tečkový diagram vykreslíme příkazem `plot()`, přičemž první dva argumenty budou vektory naměřených hodnot znaku X (`skull.BM`) a znaku Y (`skull.LM`), viz str. ???. Argumentem `xlab = ''` zabráníme vypsání popisku osy x , který následně doplníme do grafu samostatně (příkaz `mtext()`) pod osu x (argument `side`) do výšky 2.3 (argument `line`). Analogicky přidáme do grafu řádek s hodnotou korelačního koeficientu. Text řádku generujeme příkazem `bquote()`, kde `r[12]` je syntaxe zápisu r_{12} , symbol `==` odpovídá syntaxi znaménka `=`. `.(r12)` zajistí vypsání hodnoty uložené v proměnné `r12`, tj. 0.1682.

```
129 r12 <- round(r12, digits = 4)
130 plot(skull.BM , skull.LM , pch = 21, col = 'darkblue', bg = 'mintcream',
131       xlab = '', ylab = 'nejvetsi delka mozkovny (mm) - muzi', las = 1)
132
133 mtext('nejvetsi sirka mozkovny (mm) - muzi', side = 1, line = 2.3)
134 mtext(bquote(r[12] == .(r12)), side = 1, line = 3.5)
```



♣

Dataset 4: Délkové rozměry klíčních kostí

Hodnocený soubor představují osteometrická data klíční kosti (*clavicula*) anglického souboru dokumentovaných skeletů (Parsons, 1916; soubor 03-paired-means-clavicle2.txt). Konkrétně jde o délku klíční kosti z pravé a levé strany těla v párovém uspořádání. Jednotlivé kosti bez druhostranné kosti nebyly do souboru zařazeny.

Popis proměnných:

- `id` – pořadové číslo jednice;
- `sex` – pohlaví (m – muž, f – žena);

- `length.R` – délka kosti z pravé strany (mm);
- `length.L` – délka kosti z levé strany (mm).

Příklad 3.12. Základní číselné charakteristiky pro intervalový znak

Načtete datový soubor `03-paired-means-clavicle2.txt` a vypíšete první čtyři řádky z načtené tabulky. Zjistěte, zda datový soubor obsahuje neznámé hodnoty a případně je z načteného souboru odstraňte. Zaměřte se pouze na znak X = délka levé klíční kosti pro skelety ženského pohlaví. Vytvořte tabulku základních číselných charakteristik pro znak X .

Řešení příkladu 3.12

Načtení datového souboru provedeme příkazem `read.delim()`, první čtyři řádky tabulky vypíšeme příkazem `head()` se specifikací argumentu `n = 4`.

```
135 data <- read.delim('03-paired-means-clavicle2.txt')
136 head(data, n = 4)
```

	id	sex	length.R	length.L	
137	1	66	m	126	130
138	2	69	m	158	159
139	3	71	m	153	151
140	4	72	m	145	147
141					

Pomocí funkce `is.na()` zjistíme, zda datový soubor obsahuje neznámé hodnoty.

```
142 sum(is.na(data))
```


```
[1] 0
```

Datová tabulka neobsahuje žádné neznámé hodnoty. Pomocí podmnožinového operátoru `[]` nyní vybereme z tabulky `data` pouze údaje o levé klíční kosti (`length.L`) u skeletů ženského pohlaví. Naměřené hodnoty si příkazem `sort()` vzestupně seřadíme.

```
144 length.LF <- data[data$sex == 'f', 'length.L']
145 length.LF <- sort(length.LF)
146 length(length.LF)
```

```
[1] 50
```

Datová tabulka obsahuje údaje o délkách levostranných klíčních kostí u 50 skeletů ženského pohlaví.

Znak X = délka levé klíční kosti pro skelety ženského pohlaví je příkladem intervalového typu dat. V tabulce základních číselných charakteristik budou obsaženy následující charakteristiky: aritmetický průměr, rozptyl, směrodatná odchylka, koeficient variace, minimální hodnota, dolní kvartil, medián, horní kvartil, maximální hodnota, mezikvartilové rozpětí, koeficient šikmosti a koeficient špičatosti. Nejprve provedeme ruční výpočet každé z těchto číselných charakteristik a následně uskutečníme kontrolu pomocí softwaru .

Začneme výpočtem aritmetického průměru m , tj.

$$m = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{50} (121 + 127 + \dots + 162 + 162) = \frac{6927}{50} = 138.54.$$

Rozptyl s^2 vypočítáme jako

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 \\ &= \frac{1}{50} ((121 - 138.54)^2 + (127 - 138.54)^2 + \dots + (162 - 138.54)^2 + (162 - 138.54)^2) \\ &= \frac{1}{50} ((-17.54)^2 + (-11.54)^2 + \dots + 23.46^2 + 23.46^2) \\ &= \frac{3582.42}{50} \doteq 70.5684. \end{aligned}$$

Směrodatnou odchylku s stanovíme jako odmocninu z rozptylu, tj.

$$s = \sqrt{s^2} = \sqrt{70.5684} \doteq 8.4005.$$

Koeficient variace v je dopočítáme jako podíl směrodatné odchylky a aritmetického průměru vynásobený 100%, tj.

$$v = \frac{s}{m} \times 100\% = \frac{8.4005}{138.54} \times 100\% = 0.060636 \times 100\% \doteq 6.0636\%.$$

Minimální naměřenou hodnotu nalezneme na první pozici v posloupnosti seřazených naměřených hodnot.

148 `length.LF [1]`

```
[1] 121
```

149

$$x_{\min} = 121.$$

V případě výpočtu dolního kvartilu postupujeme analogicky jako v příkladech 3.5 a 3.9. Koeficient $\alpha = 0.25$, počet naměřených hodnot $n = 50$. Součin $c = n \times \alpha = 50 \times 0.25 = 12.5$ není celé číslo, tedy c zaokrouhlíme na nejbližší vyšší celé číslo, tj. 13, a dolní kvartil bude rovný hodnotě umístěné na 13. pozici v seřazeném vektoru `skull.LF`.

150 `length.LF [13]`

```
[1] 134
```

151

$$x_{0.25} = x_{(13)} = 134.$$

Pro výpočet mediánu $x_{0.50}$ je $\alpha = 0.50$ a počet naměřených hodnot $n = 50$. Součin $c = n \times \alpha = 50 \times 0.50 = 25$ je celé číslo, proto medián stanovíme jako průměr hodnot umístěných na 25. a 26. pozici v seřazeném vektoru `skull.LF`.

152 `length.LF [25]`

```
[1] 137
```

153

154 `length.LF [26]`

```
[1] 138
```

155

$$x_{0.50} = \frac{x_{(25)} + x_{(26)}}{2} = \frac{137 + 138}{2} = 137.5.$$

V případě výpočtu horního kvartilu je $\alpha = 0.75$ a počet naměřených hodnot $n = 50$. Součin $n \times \alpha = 50 \times 0.75 = 37.5$, není celé číslo, tedy c zaokrouhlíme na nejbližší vyšší celé číslo, tj. 38, a horní kvartil bude rovný hodnotě umístěné na 38. pozici v seřazeném vektoru `skull.LF`.

156 `length.LF [38]`

```
[1] 142
```

157

$$x_{0.75} = x_{(38)} = 142.$$

Maximální naměřenou hodnotu nalezneme na poslední pozici v posloupnosti seřazených naměřených hodnot.

158 `length.LF [50]`

```
[1] 162
```

159

$$x_{\max} = 162.$$

Mezikvartilové rozpětí IQR získáme odečtením hodnoty dolního kvartilu od hodnoty horního kvartilu, tj.


$$IQR = x_{0.75} - x_{0.25} = 142 - 134 = 8. \quad (8)$$

Koeficient šikmosti b_1 vypočítáme pomocí vzorce 6, tj.

$$\begin{aligned}
 b_1 &= \frac{1}{n} \frac{\sum_{i=1}^n (x_i - m)^3}{s^3} \\
 &= \frac{1}{50} \frac{(121 - 138.54)^3 + (127 - 138.54)^3 + \dots + (162 - 138.54)^3 + (162 - 138.54)^3}{8.485786^3} \\
 &= \frac{1}{50} \frac{(-17.54)^3 + (-11.54)^3 + \dots + 23.46^3 + 23.46^3}{8.485786^3} \\
 &= \frac{24\,867.09}{30\,552.47} = 0.8139141 \doteq 0.8139.
 \end{aligned}$$

Koeficient špičatosti b_2 stanovíme pomocí vzorce 7, tj.

$$\begin{aligned}
 b_2 &= \frac{1}{n} \frac{\sum_{i=1}^n (x_i - m)^4}{s^4} - 3 \\
 &= \frac{1}{50} \frac{(121 - 138.54)^4 + (127 - 138.54)^4 + \dots + (162 - 138.54)^4 + (162 - 138.54)^4}{8.424642^4} - 3 \\
 &= \frac{1}{50} \frac{(-17.54)^4 + (-11.54)^4 + \dots + 23.46^4 + 23.46^4}{8.485786^4} - 3 \\
 &= \frac{963\,700.5}{259\,261.7} - 3 = 0.7170956 \doteq 0.7171.
 \end{aligned}$$

Všechny výše zmíněné základní charakteristiky vypočítáme nyní pomocí funkcí zabudovaných v softwaru . Na závěr všechny hodnoty vložíme do jedné tabulky (`data.frame()`), kterou vypíšeme se zaokrouhlením na čtyři desetinná místa (`round()`).

```

160 m <- mean(length.LF)
161 s2 <- 1 / 50 * sum((length.LF - m)^2)
162 s <- sqrt(s2)
163 v <- s / m * 100
164
165 min <- min(length.LF)
166 x0.25 <- quantile(length.LF, probs = 0.25, type = 2)
167 x0.50 <- quantile(length.LF, probs = 0.50, type = 2)
168 x0.75 <- quantile(length.LF, probs = 0.75, type = 2)
169 IQR <- IQR(length.LF, type = 2)
170 max <- max(length.LF)
171
172 sikmost <- e1071::skewness(length.LF, type = 3)
173 spicatost <- e1071::kurtosis(length.LF, type = 3)
174
175 tab <- data.frame(m, var = s2, s, v, min, dolni.k = x0.25, median = x0.50,
176                  horni.k = x0.75, max, IQR, sikmost, spicatost,
177                  row.names = 'f-L')
178 round(tab, digits = 4)

```

	m	var	s	v	min	dolni.k	median	horni.k	max	IQR	sikmost	spicatost
f-L	138.54	70.5684	8.4005	6.0636	121	134	137.5	142	162	8	0.8139	0.7171

179
180

Interpretace výsledků: Délka levé klíční kosti u skeletů ženského pohlaví v datovém souboru se pohybuje v rozsahu od 121 mm do 162 mm. Průměrná hodnota délky levé klíční kosti u skeletů ženského pohlaví je 138.54 mm se směrodatnou odchylkou 8.40 mm, přičemž směrodatná odchylka představuje 6.06% aritmetického průměru. 25% naměřených hodnot je menších nebo rovných 134 mm, 50% naměřených hodnot je menších nebo rovných 137.5 mm a 75% naměřených hodnot je menších nebo rovných 142 mm. Mezikvartilové rozpětí pro délku levé klíční kosti je 8 mm. Hodnota koeficientu šikmosti, 0.8139, ukazuje na výrazněji kladně zešikmená data s prodlouženým pravým koncem. Hodnota koeficientu špičatosti, 0.7171, ukazuje na strmější charakter dat. ♣

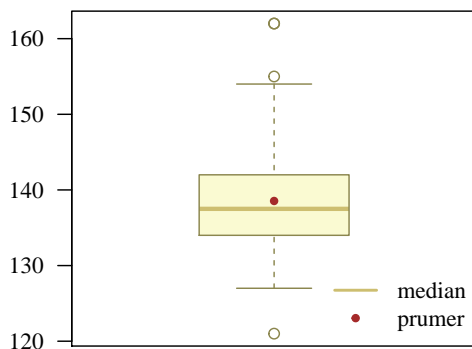
Příklad 3.13. Krabicový diagram

Pro znak $X = \text{délka levé klíční kosti}$ u žen sestrojte krabicový diagram. Do grafu doplňte hodnotu aritmetického průměru a vypište legendu.

Řešení příkladu 3.13

Krabicový diagram vykreslíme analogicky jako v příkladu 3.6 příkazem `boxplot()`. Příkazem `mtext()` doplníme do grafu popisek osy x na řádek 1.5. Hodnotu aritmetického průměru zaneseme do grafu pomocí funkce `points()`, kde prvním argumentem bude hodnota aritmetického průměru, kterou máme vloženou v proměnné `m`. Vykreslený bod bude kulatého tvaru s plným vnitřkem (`pch = 20`) v hnědé barvě (`col`). Nakonec do grafu doplníme legendu příkazem `legend()`, kde prvním argumentem specifikujeme pozici legendy vpravo dole (`'bottomright'`). První člen legendy bude plná čára (`lty = c(1, NA)`) o tloušťce 2 (`lwd = c(2, NA)`). Druhý člen legendy bude ve tvaru kulatého bodu s plným vnitřkem (`pch = c(NA, 20)`). Barvy a popisky obou členů legendy specifikujeme argumenty `col` a `legend`. Nakonec odstraníme černý rámeček okolo legendy nastavením argumentu `bty = 'n'`.

```
181 boxplot(length.LF, type = 2, xlab = '', las = 1,
182         col = 'lightgoldenrodyellow', border = 'khaki4', medcol = 'lightgoldenrod3')
183 mtext('delka leve klicni kosti (mm) - zeny', side = 1, line = 1.5)
184
185 points(m, pch = 20, col = 'brown')
186 legend('bottomright', lty = c(1, NA), pch = c(NA, 20), lwd = c(2, NA),
187       col = c('lightgoldenrod3', 'brown'),
188       legend = c('median', 'prumer'), bty = 'n')
```



delka leve klicni kosti (mm) – zeny



Příklad 3.14. Charakteristika závislosti pro znaky intervalového typu

Zaměříme se nyní na znaky $X = \text{délka levé klíční kosti}$ a $Y = \text{délka pravé klíční kosti}$ u skeletů ženského pohlaví najednou. Určete míru závislosti mezi znaky X a Y .

Řešení příkladu 3.14

Oba znaky X a Y jsou intervalového typu. Ke stanovení míry závislosti mezi těmito znaky použijeme Pearsonův korelační koeficient r_{12} , který spočítáme příkazem `cor()` se specifikací argumentu `method = 'pearson'`. Prvními dvěma argumenty příkazu jsou vektory naměřených hodnot znaků X (`length.LF`) a Y (`length.RF`).

```
189 length.LF <- data[data$sex == 'f', 'length.L']
190 length.RF <- data[data$sex == 'f', 'length.R']
191 (r12 <- cor(length.LF, length.RF, method = 'pearson'))
```

```
[1] 0.9296909
```

192

Interpretace výsledků: Pearsonův korelační koeficient nabývá hodnoty 0.9297. Mezi délkou pravé a levé klíční kosti u skeletů ženského pohlaví existuje velmi vysoký stupeň přímé lineární závislosti. S rostoucí délkou pravé klíční kosti roste délka levé klíční kosti a naopak. ♣

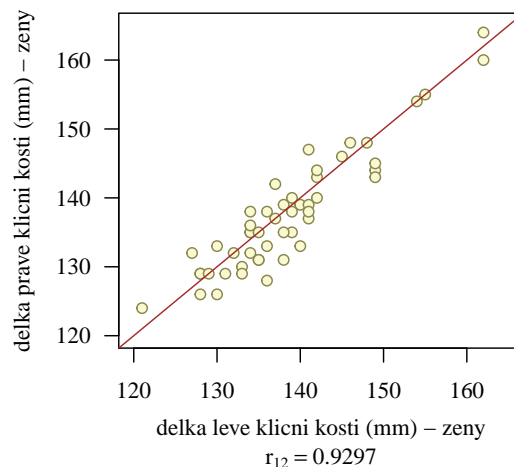
Příklad 3.15. Dvourozměrný tečkový diagram

Výslednou míru závislosti mezi znaky $X = \text{délka levé klíční kosti}$ a $Y = \text{délka pravé klíční kosti}$ u skeletů ženského pohlaví vizualizujte pomocí dvourozměrného tečkového diagramu. Do diagramu doplňte popisek s hodnotou Pearsonova korelačního koeficientu r_{12} a referenční přímku $x = y$.

Řešení příkladu 3.15

Analogicky jako v příkladu 3.11 vykreslíme dvourozměrný tečkový diagram příkazem `plot()`. Rozsah obou os, x i y stanovíme stejný (`xlim = c(120, 165)`, `ylim = c(120, 165)`). Řádek s hodnotou korelačního koeficientu r_{12} doplníme do grafu pomocí příkazu `mtext()`, přičemž obsah řádku vygenerujeme pomocí funkce `bquote()`. Z grafu vidíme, že naměřené hodnoty obou znaků X a Y se pohybují v podobných rozsazích, navíc mezi nimi pozorujeme přímý lineární trend, který máme podložený vysokou hodnotou Pearsonova korelačního koeficientu. Pro zvýraznění lineárního trendu mezi oběma znaky dokreslíme do grafu referenční lineární přímku $x = y$ (příkaz `abline()`) se sklonem 1 (argument `b = 1`) procházející bodem 0 (argument `a = 0`). Vykreslená přímka bude mít tloušťku 1 (`lwd = 1`) a hnědou barvu (`col = 'brown'`).

```
193 r12 <- round(r12, digits = 4)
194 plot(length.LF , length.RF , pch = 21, xlim = c(120, 165),
195       ylim = c(120, 165), col = 'khaki4', bg = 'lightgoldenrodyellow',
196       xlab = '', ylab = 'delka prave klicni kosti (mm) - zeny', las = 1)
197
198 mtext('delka leve klicni kosti (mm) - zeny', side = 1, line = 2.3)
199 mtext(bquote(r[12] == .(r12)), side = 1, line = 3.5)
200 abline(a = 0, b = 1, lwd = 1, col = 'brown')
```



3.4 Příklady k samostatnému procvičování

Příklad 3.16. Charakteristika polohy nominálního znaku

V rámci sekce ?? jsme jako mezivýstup příkladu ?? získali kontingenční tabulku simultánních absolutních četností znaků $X = \text{počet starších sourozenců}$ a $Y = \text{porodní hmotnost novorozence}$ (viz tabulka 5). Najděte vhodnou charakteristiku polohy pro znak *počet starších sourozenců*.

Tabulka 5: Simultánní absolutní četnosti znaků *počet starších sourozenců* a *porodní hmotnost novorozence*

	nízká	norma	vysoká
žádný	123	456	11
jeden	91	399	21
dva	26	138	11

Řešení příkladu 3.16

zadny	jeden	dva
590	511	175

201
202

Interpretace výsledků: Nejvíce novorozenců v datovém souboru bylo prvorozených s četností výskytu 590. ♣

Příklad 3.17. Charakteristika závislosti mezi dvěma nominálními znaky

Zaměřme se nyní na oba znaky $X = \text{počet starších sourozenců}$ a $Y = \text{porodní hmotnost novorozence}$ najednou. Určete míru závislosti mezi znaky X a Y . Výslednou míru závislosti porovnejte s mírou závislosti stanovenou v rámci příkladu 3.7 na základě hodnoty Spearmanova koeficientu pořadové korelace r_S . Který z obou koeficientů bychom upřednostnili pro charakterizaci vztahu mezi počtem starších sourozenců a porodní hmotností novorozence a proč?

Řešení příkladu 3.17

[1] 0.06940097

203

Interpretace výsledků: Mezi počtem starších sourozenců a porodní hmotností novorozence existuje zanedbatelný stupeň závislosti.

Odpověď na otázku: Znak $X = \text{počet starších sourozenců}$ je originálně proměnnou ordinálního typu, znak $Y = \text{porodní hmotnost novorozence}$ je originálně proměnnou spojitého typu. Kategorizací obou proměnných, tedy jejich převodem na proměnné nominálního typu, přicházíme o informace, které původní proměnné poskytují. Preferovanou charakteristikou závislosti je v tomto případě Spearmanův koeficient pořadové korelace ($r_S = 0.0476$; velmi nízký stupeň pořadové závislosti), který přistupuje k oběma proměnným jako k ordinálním, a pracuje tedy s širším množstvím informací než Cramérův koeficient. ♣

Příklad 3.18. Charakteristika polohy nominálního znaku

Načtete datový soubor 22-multinom-palmar-lines.txt. Z tabulky vyselektujte pouze údaje týkající se znaků $X = \text{barva vlasů}$ a $Y = \text{zakončení palmárních linií}$ u mužů. Změňte záhlaví tabulky tak, aby názvy jednotlivých variant znaku $X = \text{barva vlasů}$ byly: světlé, střední a tmavé; a názvy jednotlivých variant znaku $Y = \text{zakončení palmárních linií}$ byly: vysoké, střední a nízké. Stanovte vhodnou charakteristiku polohy pro znak X i pro znak Y .

Řešení příkladu 3.18

	vysoke	stredni	nizke
svetle	6	6	4
stredni	20	15	7
tmave	18	12	12

204
205
206
207

Charakteristika polohy pro barvu vlasů

svetle	stredni	tmave
16	42	42

208
209

Interpretace výsledků: Nejvíce mužů v datovém souboru mělo střední nebo tmavou barvu vlasů ($n_{\text{střední}} = n_{\text{tmavé}} = 42$).

Charakteristika polohy pro zakončení palmárních linií

```
vysoke  stredni  nizke
    44      33     23
```

210
211

Interpretace výsledků: Nejvíce mužů v datovém souboru mělo vysoké zakončení palmárních linií s četností výskytu 44. ♣

Příklad 3.19. Charakteristika závislosti mezi dvěma nominálními znaky

Zaměříme se nyní na oba znaky $X = \text{barva vlasů}$ a $Y = \text{zakončení palmárních linií}$ u mužů najednou. Určete míru závislosti mezi znaky X a Y . Míru závislosti mezi barvou vlasů a zakončením palmárních linií u mužů porovnejte s mírou závislosti mezi barvou vlasů a zakončením palmárních linií u žen (viz příklad 3.4). Zauvažujte, jak byste výsledek srovnání odborně zdůvodnili.


Řešení příkladu 3.19

```
[1] 0.1014841
```

212

Interpretace výsledků: Mezi barvou vlasů a zakončením palmárních linií u mužů existuje slabý stupeň závislosti. Stejný závěr jsme stanovili také pro vztah mezi barvou vlasů a zakončením palmárních linií u žen. ♣

Příklad 3.20. Základní číselné charakteristiky pro intervalový znak

Načtěte datový soubor 17-anova-newborns.txt, odstraňte z načtených dat NA hodnoty a zjistěte dimenzi datové tabulky. Zaměřte se pouze na znak $X = \text{porodní hmotnost novorozence}$. S pomocí softwaru  vytvořte tabulku základních číselných charakteristik pro znak X . Pro hodnoty kvantilů proveďte také ruční výpočet. Dále sestrojte krabicový diagram pro znak X a zanešte do něj hodnotu aritmetického průměru. Zamyslete se nad propojením diagramu s charakteristikami polohy a variability.

Řešení příkladu 3.20

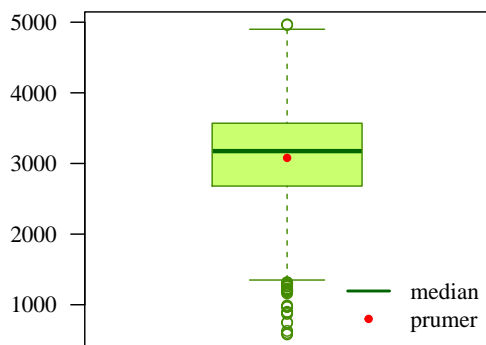
```
[1] 1382 4
```

213

Datový soubor obsahuje údaje o 1382 novorozencích, přičemž u každého novorozence máme záznamy o čtyřech znacích.

```
      m      var      s      v min dolni.k median horni.k  max IQR sikmost
hmt 3078.94 485440.5 696.7356 22.6291 580 2680 3175 3570 4970 890 -0.6094
      spicatost
hmt 0.4937
```

214
215
216
217



porodní hmotnost novorozence (g)

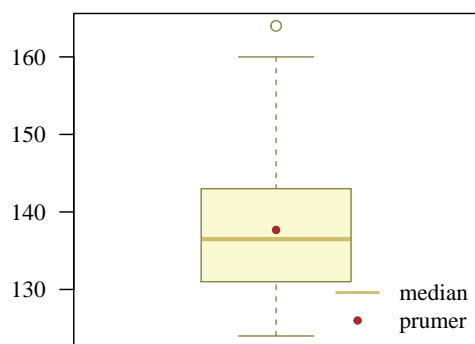
Interpretace výsledků: Porodní hmotnost novorozenců v datovém souboru nabývá hodnot v rozmezí 580–4970 g. Průměrná hodnota porodní hmotnosti je 3078.94 g se směrodatnou odchylkou 696.74 g, která představuje 22.63% aritmetického průměru. 25% naměřených hodnot je menších nebo rovných 2680 g, 50% naměřených hodnot je menších nebo rovných 3175 g a 75% naměřených hodnot je menších nebo rovných 3570 g. Mezikvartilové rozpětí pro porodní hmotnost novorozenců má rozsah 890 g. Hodnota koeficientu šikmosti, -0.6094, ukazuje na záporně zešíkmená data s prodlouženým levým koncem. Hodnota koeficientu špičatosti, 0.4937, ukazuje na strmý charakter dat. ♣

Příklad 3.21. Základní číselné charakteristiky pro intervalový znak

Načtěte datový soubor 03-paired-means-clavicle2.txt, zjistěte, zda datový soubor obsahuje neznámé hodnoty a případně je z načteného souboru odstraňte. Zaměřte se pouze na znak $X = \text{délka pravé klíční kosti}$ pro skelety ženského pohlaví. S pomocí softwaru vytvořte tabulku základních číselných charakteristik pro znak X . Pro hodnoty kvantilů proveďte také ruční výpočet. Dále sestrojte krabicový diagram pro znak X a zanešte do něj hodnotu aritmetického průměru.

Řešení příkladu 3.21


	m	var	s	v	min	dolni.k	median	horni.k	max	IQR	sikmost	spicatost	
f-R	137.68	73.5376	8.5754	6.2285	124	131	136.5	143	164	12	0.971	0.7501	218
													219



delka prave klíční kosti (mm) – ženy

Interpretace výsledků: Délka pravé klíční kosti u skeletů ženského pohlaví nabývá hodnot v rozmezí 124 mm až 164 mm. Průměrná délka pravé klíční kosti u skeletů ženského pohlaví v datovém souboru je 137.68 mm se směrodatnou odchylkou 8.58 mm, přičemž směrodatná odchylka představuje 6.23% aritmerického průměru. 25% naměřených hodnot je menších nebo rovných 131 mm, 50% naměřených hodnot je menších nebo rovných 136.5 mm a 75% naměřených hodnot je menších nebo rovných 143 mm. Mezikvartilové rozpětí má rozsah 12 mm. Hodnota koeficientu šikmosti, 0.9710, ukazuje na výrazněji kladně zešikmená data s prodlouženým pravým koncem. Hodnota koeficientu špičatosti, 0.7501, ukazuje na strmý charakter dat. ♣

Příklad 3.22. Základní číselné charakteristiky pro intervalový znak

Načtete datový soubor 03-paired-means-clavicle2.txt. Zaměřte se na znak $X = \text{délka levé klíční kosti}$ pro skelety mužského pohlaví. Pomocí softwaru  vytvořte tabulku základních číselných charakteristik pro znak X . Pro hodnoty kvantilů proveďte také ruční výpočet.


Řešení příkladu 3.22

	m	var	s	v	min	dolní.k	median	horní.k	max	IQR	šikmost	spicatost
muzi-L	153.6	96.96	9.8468	6.4107	130	147	154.5	158	176	11	0.2093	-0.2896

220
221

Interpretace výsledků: Naměřené délky levých klíčních kostí u skeletů mužského pohlaví nabývají hodnot v rozsahu 130–176 mm. Průměrná délka levé klíční kosti u skeletů mužského pohlaví v datovém souboru je 153.60 mm se směrodatnou odchylkou 9.85 mm, přičemž směrodatná odchylka představuje 6.41% aritmerického průměru. 25% naměřených hodnot je menších nebo rovných 147 mm, 50% naměřených hodnot je menších nebo rovných 154.5 mm a 75% naměřených hodnot je menších nebo rovných 158 mm. Mezikvartilové rozpětí naměřených hodnot má rozsah 11 mm. Hodnota koeficientu šikmosti, 0.2093, ukazuje na kladně zešikmená data s prodlouženým pravým koncem. Hodnota koeficientu špičatosti, -0.2896, ukazuje na plošší charakter dat. ♣

Příklad 3.23. Základní číselné charakteristiky pro intervalový znak

Načtete datový soubor 03-paired-means-clavicle2.txt. Zaměřte se na znak $Y = \text{délka pravé klíční kosti}$ pro skelety mužského pohlaví. S pomocí softwaru  vytvořte tabulku základních číselných charakteristik pro znak Y . Pro hodnoty kvantilů proveďte také ruční výpočet.

Řešení příkladu 3.23

	m	var	s	v	min	dolní.k	median	horní.k	max	IQR	šikmost
muzi-R	151.74	118.5124	10.8863	7.1743	126	143	153	160	175	17	-0.057
	spicatost										
muzi-R	-0.646										

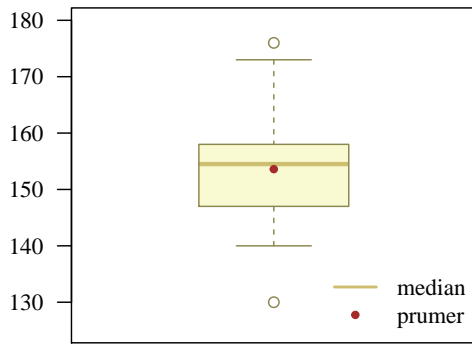
222
223
224
225

Interpretace výsledků: Délka pravé klíční kosti u skeletů mužského pohlaví nabývá hodnot v rozsahu od 126 mm do 175 mm. Průměrná délka pravé klíční kosti u skeletů mužského pohlaví v datovém souboru je 151.74 mm se směrodatnou odchylkou 10.89 mm, přičemž směrodatná odchylka představuje 7.17% aritmerického průměru. 25% naměřených hodnot je menších nebo rovných 143 mm, 50% naměřených hodnot je menších nebo rovných 153 mm a 75% naměřených hodnot je menších nebo rovných 160 mm. Mezikvartilové rozpětí má rozsah 17 mm. Hodnota koeficientu šikmosti, -0.057, ukazuje na záporně zešikmená data s tendencí k prodlouženému levému konci. Hodnota koeficientu je však tak malá, že zešikmení dat nebude okem skoro viditelné. Hodnota koeficientu špičatosti, -0.6460, ukazuje na plochý charakter dat. ♣

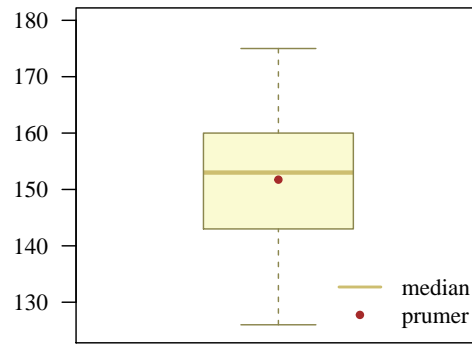
Příklad 3.24. Krabicový diagram

Vykreslete krabicový diagram (a) pro znak $X = \text{délka levé klíční kosti}$; (b) pro znak $Y = \text{délka pravé klíční kosti}$ pro skelety mužského pohlaví.

Řešení příkladu 3.24



delka leve klíční kosti (mm) – muži



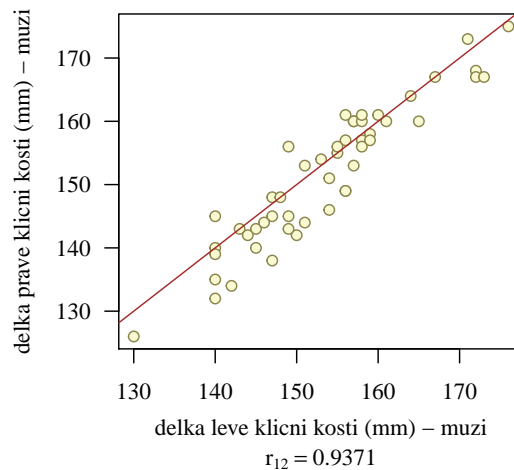
delka prave klíční kosti (mm) – muži



Příklad 3.25. Charakteristika závislosti pro znaky intervalového typu

Zaměříme se nyní na znaky $X = \text{délka levé klíční kosti}$ a $Y = \text{délka pravé klíční kosti}$ u skeletů mužského pohlaví najednou. Určete míru závislosti mezi znaky X a Y . Míru závislosti mezi znaky vizualizujte pomocí dvourozměrného tečkového diagramu. Do diagramu doplňte popisek s hodnotou korelačního koeficientu a referenční přímkou $x = y$. Míru závislosti mezi délkou pravé a levé klíční kosti u skeletů mužského pohlaví porovnejte s mírou závislosti stanovenou u skeletů ženského pohlaví (viz příklad 3.14). Zauvažujte, jak byste výsledek srovnání odborně zdůvodnili.

Řešení příkladu 3.25




Interpretace výsledků: Pearsonův korelační koeficient nabývá hodnoty 0.9371. Mezi délkou pravé a levé klíční kosti u skeletů mužského pohlaví existuje velmi vysoký stupeň přímé lineární závislosti. S rostoucí délkou pravé klíční kosti roste délka levé klíční kosti a naopak. Stejný závěr jsme stanovili také pro vztah mezi délkou pravé a levé klíční kosti u skeletů ženského pohlaví ($r_{12} = 0.9297$).



Příklad 3.26. 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 $X = \text{největší délka mozkovny}$ pro skelety ženského pohlaví. S pomocí softwaru  vytvořte tabulku základních číselných charakteristik pro znak X . Pro hodnoty kvantilů proveďte také ruční výpočet. Vraťte se k histogramu a krabicovému diagramu znaku *největší délka mozkovny* pro skelety ženského pohlaví sestrojených v rámci příkladu ???. Prozkoumejte, jak se vypočítané charakteristiky polohy, variability a nesymetrie projeví v tvaru histogramu a krabicového diagramu. Které číselné charakteristiky byste hledali v histogramu a které naopak v krabicovém diagramu?

Řešení příkladu 3.26


	m	var	s	v	min	dolni.k	median	horni.k	max	IQR	sikmost	spicatost
f-D	174.5321	38.3224	6.1905	3.5469	157	170	175	178	188	8	-0.0383	-0.2611

226
227

Interpretace výsledků: Naměřené hodnoty největší délky mozkovny pro skelety ženského pohlaví se pohybují v rozmezí 157–188 mm. Průměrná hodnota největší délky mozkovny u skeletů ženského pohlaví je 174.53 mm se směrodatnou odchylkou 6.19 mm, přičemž směrodatná odchylka představuje 3.55% aritmetického průměru. 25% naměřených hodnot je menších nebo rovných 170 mm, 50% naměřených hodnot je menších nebo rovných 175 mm a 75% naměřených hodnot je menších nebo rovných 178 mm. Mezikvartilové rozpětí má rozsah 8 mm. Hodnota koeficientu šikmosti, -0.0383, ukazuje na téměř neznatelně záporně zešikmená data. Hodnota koeficientu špičatosti, -0.2611, ukazuje na plošší charakter dat.

Odpověď na otázku: Pomocí histogramu můžeme vizualizovat hodnotu aritmetického průměru, rozptylu, resp. směrodatné odchylky, koeficientu šikmosti a špičatosti. Pomocí krabicového diagramu vizualizujeme minimální a maximální naměřenou hodnotu, dolní kvartil, medián, horní kvartil a mezikvartilové rozpětí, šikmost, špičatost a v neposlední řadě také aritmetický průměr, je-li v krabicovém diagramu zaznamenán. ♣

Příklad 3.27. Základní číselné charakteristiky pro intervalový znak

Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat neznámé hodnoty. Zaměřte se pouze na znak $Y = \text{největší šířka mozkovny}$ pro skelety ženského pohlaví. Pomocí softwaru  vytvořte tabulku základních číselných charakteristik pro znak Y . Hodnoty kvantilů stanovte také ručním výpočtem. Vraťte se k histogramu a krabicovému diagramu znaku *největší šířka mozkovny* pro skelety ženského pohlaví sestrojených v rámci příkladu ???. Prozkoumejte, jak se vypočítané charakteristiky polohy, variability a nesymetrie projeví v tvaru histogramu a krabicového diagramu.

Řešení příkladu 3.27

	m	var	s	v	min	dolni.k	median	horni.k	max	IQR	sikmost	spicatost
f-S	134.1468	21.85	4.6744	3.4845	118	131	134	137	146	6	0.0297	0.4235

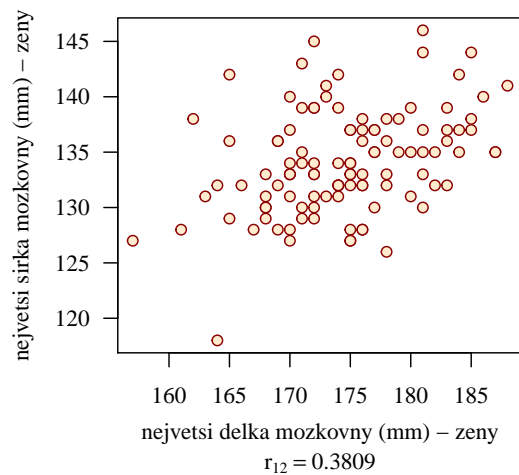
228
229

Interpretace výsledků: Naměřené hodnoty největší šířky mozkovny pro skelety ženského pohlaví se pohybují v rozmezí 118–146 mm. Průměrná hodnota největší šířky mozkovny u skeletů ženského pohlaví je 134.15 mm se směrodatnou odchylkou 4.67 mm, přičemž směrodatná odchylka představuje 3.48% aritmetického průměru. 25% naměřených hodnot je menších nebo rovných 131 mm, 50% naměřených hodnot je menších nebo rovných 134 mm a 75% naměřených hodnot je menších nebo rovných 137 mm. Mezikvartilové rozpětí má rozsah 6 mm. Hodnota koeficientu šikmosti, 0.0297, ukazuje na téměř neznatelně kladně zešikmená data. Hodnota koeficientu špičatosti, 0.4235, ukazuje na strmější charakter dat. ♣

Příklad 3.28. Charakteristika závislosti pro znaky intervalového typu

Zaměřme se nyní na znaky $X = \text{největší délka mozkovny}$ a $Y = \text{největší šířka mozkovny}$ pro skelety ženského pohlaví najednou. Určete míru závislosti mezi znaky X a Y . Míru závislosti mezi znaky vizualizujte pomocí dvou-rozměrného tečkového diagramu (viz příklad ??). Do diagramu doplňte popisek s hodnotou korelačního koeficientu. Míru závislosti mezi největší délkou a šířkou mozkovny u skeletů ženského pohlaví porovnejte s mírou závislosti stanovenou u skeletů mužského pohlaví (viz příklad 3.11). Zauvažujte, jak byste výsledek srovnání odborně zdůvodnili.

Řešení příkladu 3.28



Interpretace výsledků: Mezi největší šířkou a délkou mozkovny pro skelety ženského pohlaví existuje mírný stupeň přímé lineární závislosti ($r_{12} = 0.3809$). Naproti tomu mezi největší šířkou a délkou mozkovny pro skelety mužského pohlaví existuje pouze nízký stupeň přímé lineární závislosti ($r_{12} = 0.1682$). ♣