3 Charakteristiky polohy a variability a štatistická grafika

Každé rozdelenie pravdepodobnosti býva charakterizované parametrami, ktoré sa odhadujú z realizácií (dát). Tieto parametre označujeme ako **číselné charakteristiky rozdelenia** (**štatistiky**). Patria medzi ne **charakteristiky polohy a variability** ako napr. stredná hodnota, medián a ostatné kvantily, rozptyl, rozpätie a pod.

Stredná (očakávaná) hodnota (prvý začiatočný moment) rozdelenia náhodnej veličiny X, ozn. $E[X]$, vypočítame ako (Casella a Berger, 2002)

$$E[X] = \begin{cases} \sum_{i=1}^{n} x_i \Pr(X = x_i), & \text{ak } X \text{ je diskrétna náhodná premenná s realizáciami } x_1, x_2, \ldots, x_n, \\ \int_{-\infty}^{\infty} x f_X(x) dx, & \text{ak } X \text{ je spojitá náhodná premenná s hustotou } f_X(x). \end{cases}$$

Rozptyl (druhý centrálny moment) $Var[X] = E[(X - E[X])^2]$, ktorého kladnú odmocninu $SD[X]$, vyjadrujúcu „rozptýlenie“ X okolo $E[X]$, nazývame **smerodajná odchýlka**.

Príklad 110 (charakteristiky normálneho rozdelenia) *Majme náhodný výber $X_1, X_2, ..., X_n$ z normálneho rozdelenia $N(\mu, \sigma^2)$. Potom $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ nazývame **výberová stredná hodnota** a $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ **výberový rozptyl***

Kedže X je náhodná premenná, $\overline{X}$ nie je aritmetický priemer, ale tzv. výberový priemer; aritmetický priemer $\overline{x}$ je realizácia výberového priemeru. Podobne S^2 je tzv. výberový rozptyl; $\widehat{\sigma}^2 = s^2$ je jeho realizácia. Tiež $SD[X] = S$ je výberová smerodajná odchýlka a nie smerodajná odchýlka; smerodajná odchýlka $\widehat{\sigma} = s$. Podiel $V_k = S^2/\overline{X}$ nazývame **výberový koeficient variácie**.

Príklad 111 (charakteristiky binomického rozdelenia) *Ak X pochádza z binomického rozdelenia, ozn. $Bin(N, p)$, potom $E[X] = Np$ je **stredná hodnota** a $Var[X] = Np(1-p)$ **rozptyl** náhodnej veličiny X.*

Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z nejakého rozdelenia a $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ **usporiadaný náhodný výber** (vo vzrastajúcom poradí) s rozsahom n, kde $X_{(i)}$ nazývame **poriadkové štatistiky**. Sú to náhodné premenné, pre ktoré platí $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. Potom (Casella a Berger, 2002)

$$X_{(1)} = \min_{1 \le i \le n} X_i,$$

$$X_{(2)} = \text{druhé najmenšie } X_i \ (X_i \text{ v poradí druhé})$$

$$\vdots$$

$$X_{(n-1)} = X_i \text{ v poradí } (n-1)\text{-vé},$$

$$X_{(n)} = \max_{1 \le i \le n} X_i.$$

Výberový medián je charakterizovaný číslom $Q_2 = \widetilde{X}$, ktoré rozdeľuje náhodný výber tak, že približne polovica X_i je menšia ako táto hodnota a polovica je väčšia ako táto hodnota. Definujeme ho ako[24]

$$\widetilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{ak } n \text{ je nepárne}, \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{ak } n \text{ je párne}. \end{cases}$$

[24]Ak n je nepárne, potom medián je prostredná hodnota, t.j. hodnota nachádzajúca sa v poradí ako prostredná alebo na mieste X, ktoré zodpovedá poradiu $(n+1)/2$. Ak n je nepárne, potom medián je uprostred zoradených hodnôt, t.j. hodnota na mieste X, ktoré zodpovedá priemeru X s poradiami $n/2$ a $n/2+1$.

Majme číslo $p \in (0,1)$. Potom **výberovým** $100p$**-tym percentilom náhodného výberu** $X_{(\{np\})}$ bude také X_i, kde približne np hodnôt X_i bude menších ako $X_{(\{np\})}$ a $n(1-p)$ väčších ako $X_{(\{np\})}$. Ak $p = 0.5$, ide o 50. percentil alebo medián. Označenie $\{x\}$ v indexe znamená *najbližšie celé číslo*, kde $i - 0.5 \leq x < i + 0.5$, kde $\{x\} = i$. Potom

$$\widetilde{X}_p = X_{(\{np\})} = \begin{cases} X_{(\{np\})}, & \text{ak } \frac{1}{2n} < p \leq 0.5, \\ X_{(n+1-\{n(1-p)\})}, & \text{ak } 0.5 < p < 1 - \frac{1}{2n}. \end{cases}$$

Príklad 112 (výberový $100p$**-ty percentil)** *Ak* $n = 12$*, potom výberovým* 65. *percentilom je* $X_{(9)}$*, pretože* $n(1-p) = 12 \times (1 - 0.65) = 4.2$ *a* $n + 1 - \{n(1-p)\} = 12 + 1 - 4 = 9$.

Často používanými percentilmi[25] sú **výberový dolný kvartil** (25. percentil) $Q_1 = \widetilde{X}_{0.25}$ a **výberový horný kvartil** (75. percentil) $Q_3 = \widetilde{X}_{0.75}$.

Na výpočet rozptylu nejakého percentilu musíme poznať pravdepodobnostnú funkciu poriadkovej štatistiky, jej hustotu a distribučnú funkciu. Výpočet môžeme výhodne zjednodušiť, keď vieme, aké je asymptotické rozdelenie poriadkovej štatistiky.

Definícia 27 (pravdepodobnostná funkcia poriadkovej štatistiky) *Majme náhodný výber* $X_1, X_2, \ldots, X_n$ *z nejakého diskrétneho rozdelenia s pravdepodobnostnou funkciou* $f_X(x_i) = p_i$*, kde* $x_1 < x_2 < \ldots < x_n$*. Nech* $P_0 = 0$*,* $P_1 = p_1$*,* $P_2 = p_1 + p_2$*, ...,* $P_i = p_1 + p_2 + \ldots + p_i$*. Nech* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ *sú poriadkové štatistiky. Potom (Bickel a Doksum, 2006)*

$$\Pr(X_{(j)} \leq x_i) = \Pr(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

a

$$\Pr(X_{(j)} = x_i) = \Pr(X_{(j)} \leq x_i) - \Pr(X_{(j)} \leq x_{i-1}) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right],$$

kde Y *je náhodná premenná počtu hodnôt z* $X_1, X_2, \ldots, X_n$*, ktoré sú menšie alebo rovné ako* x_i*. Nech* $\{X_j \leq x_i\}$ *je priaznivá udalosť a* $\{X_j > x_i\}$ *je nepriaznivá udalosť. Ak* $i = 1$*, potom* $\Pr(X_j = x_i) = \Pr(X_j \leq x_i)$*, pretože* $\Pr_0 = 0$*. Teda* Y *je počet priaznivých udalostí v* n *pokusoch, t.j.* Y*=card*$\{X_j \leq x_i\}$*. Pravdepodobnosť úspechu je potom* $\Pr(X_j \leq x_i)$ *pre každý pokus, pretože pokusy sú rovnako rozdelené. Priaznivá a nepriaznivá udalosť* j*-teho pokusu je nezávislá od výsledku iného pokusu, pretože* X_j *sú nezávislé od ostatných* X_i*. Teda* $Y \sim Bin(n, P_i)$.

Ozn. card$\{\cdot\}$ znamená veľkosť (kardinalitu) množiny.

Ak $X_1, X_2, \ldots, X_n$ je náhodný výber zo spojitého rozdelenia, zhody neprichádzajú do úvahy a pravdepodobnosť, že nejaké dve alebo viaceré X_j sú rovnaké, je nula. Teda $\Pr(X_{(1)} < X_{(2)} < ... < X_{(n)}) = 1$ a výberový priestor pre $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ je $\mathcal{X} = \{(x_1, x_2, \ldots, x_n) : x_1 < x_2 < \ldots < x_n\}$.

Definícia 28 (hustota a distribučná funkcia poriadkovej štatistiky) *Nech* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ *sú poriadkové štatistiky náhodného výberu* $X_1, X_2, \ldots, X_n$ *s distribučnou funkciou* $F_X(x)$ *a hustotou* $f_X(x)$*. Potom hustota poriadkovej štatistiky* $X_{(j)}$ *je rovná (Bickel a Doksum, 2006)*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{n-j}$$

[25] Všeobecne definujeme **kvantil** ako hodnotu skúmanej veličiny, ktorá delí náhodný výber na dve časti pod a nad kvantilom; číslo v dolnom indexe hovorí o časti náhodného výberu pod kvantilom. Podľa toho rozlišujeme aj jeho typ, napr. kvartil, decil, percentil a pod.

a distribučná funkcia

$$F_{X_{(j)}}(x) = \Pr(X_{(j)} \leq x) = \Pr(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}.$$

Definícia 29 (asymptotické rozdelenie poriadkovej štatistiky) *Nech $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ sú poriadkové štatistiky náhodného výberu $X_1, X_2, \ldots, X_n$. Majme pravdepodobnosť α, kde $F(t_\alpha) = \alpha$. Asymptoticky platí, že $\sqrt{n}(\frac{j}{n} - \alpha)$ konverguje k 0. Potom je poriadková štatistika $X_{(j)}$ normálne rozdelená so strednou hodnotou $E[X_{(j)}] = t_\alpha$ a rozptylom $\sigma^2_{X_{(j)}} = \frac{\alpha(1-\alpha)}{f^2(t_\alpha)n}$. Ak $X \sim N(\mu, \sigma^2)$, potom $\sigma^2_{X_{(j)}} = \sigma^2 \frac{\pi^2}{24 \ln n}$ (Casella a Berger, 2002).*

Príklad 113 (rozptyl poriadkovej štatistiky) *Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 29.*

Príklad 114 (rozptyl poriadkovej štatistiky, $X \sim N(\mu, \sigma^2)$) *Pomocou definície 29 odvoďte rozptyl poriadkovej štatistiky, ak $X \sim N(\mu, \sigma^2)$.*

Definícia 30 (stredná hodnota a rozptyl mediánu) *Stredná hodnota mediánu $X_{(\frac{n+1}{2})}$ je rovná $E[X_{(\frac{n+1}{2})}] = \widetilde{\mu}$ a rozptyl mediánu $\sigma^2_{X_{(\frac{n+1}{2})}} = \frac{1}{4f^2(\widetilde{\mu})n}$, kde n je nepárne. Ak $X \sim N(\mu, \sigma^2)$, potom $\sigma^2_{X_{(\frac{n+1}{2})}} = \sigma^2 \frac{\pi}{2n}$ (Casella a Berger, 2002).*

Príklad 115 (rozptyl mediánu) *Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 30.*

Príklad 116 (rozptyl mediánu, $X \sim N(\mu, \sigma^2)$) *Pomocou definície 30 odvoďte rozptyl poriadkovej štatistiky, ak $X \sim N(\mu, \sigma^2)$.*

Ak má náhodná premenná X normálne rozdelenie, výpočet rozptylu mediánu sa zjednoduší, t.j. stačí poznať σ a n. Znalosť rozptylu mediánu je potrebná na výpočet $100 \times (1 - \alpha)\%$ intervalu spoľahlivosti pre medián (pozri kapitolu 4 Testovanie hypotéz). Popis rozdelenia mediánu, ak n je párne, je nad rámec tejto knihy.

Rozpätie náhodného výberu je vzdialenosť medzi najmenšou a najväčšou poriadkovou štatistikou, t.j. $R = X_{(n)} - X_{(1)}$.

Definícia 31 (hodnoty distribučnej funkcie v kvantiloch) *Empirická distribučná funkcia $F_n(x)$ je definovaná nasledovne*

$$F_n(x) = \begin{cases} 0, & ak\ x < X_{(1)}, \\ \frac{i}{n}, & ak\ X_{(i)} \leq x < X_{(i+1)}, \\ 1, & ak\ x \geq X_{(n)}. \end{cases}$$

*Majme transformáciu $T_{(1)} = F_n(X_{(1)})$, $T_{(2)} = F_n(X_{(2)}), \ldots, T_{(n)} = F_n(X_{(n)})$. Potom $T_{(1)}$, $T_{(2)}$, $\ldots, T_{(n)}$ sú **poriadkové štatistiky**. Potom platí*

$$\lim_{n\to\infty} \Pr(\sup_{\forall x \in \mathcal{Y}} [F_n(x) - F(x)] n^{1/2} \leq \lambda) = \Phi(\lambda),$$

kde $F(X)$ je teoretická distribučná funkcia a $\Phi(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$. Potom $100 \times (1 - \alpha)\%$ pás spoľahlivosti pre $F_n(x)$ definujeme ako $F_n(x) \pm \lambda_\alpha 1/n^{1/2}$, kde $\Phi(\lambda_\alpha) = 1 - \alpha$ a $Var[F_n(x)] = 1/n$ (Kolmogorov, 1933; Smirnov, 1933; Wilks, 1948). Potom môžeme tvrdiť, že $F(X)$ patrí do $100 \times (1 - \alpha)\%$ pásu spoľahlivosti a zároveň je medzi nulou a jednotkou s pravdepodobnosťou $1 - \alpha$.

3.1 Charakteristiky polohy

Realizácie budeme označovať ako $x_1, x_2, \ldots, x_n$, **usporiadané realizácie** budú $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Potom môžeme definovať nasledovné odhady charakteristík polohy (výberové charakteristiky polohy) spolu s ich anglickými ekvivalentami:

- **výberové minimum** $X_{\min}$, ktorého realizácia $x_{\min} = x_{(1)}$;
- **výberové maximum** $X_{\max}$, ktorého realizácia $x_{\max} = x_{(n)}$;
- **výberový aritmetický priemer** $\overline{X}$, ktorého realizácia $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{n_j} x_j f_j, n_j \leqslant n$, kde f_j sú frekvencie (počty) prislúchajúcich x_j a $n = \sum_j f_j$;
- **výberový modus** X_{mod}, ktorého realizácia x_{mod} je najčastejšie sa vyskytujúca hodnota (pri diskrétnej premennej ide o hodnotu x, v ktorej má pravdepodobnmostná funkcia svoje maximum; pri spojitej premennej ide o hodnotu x, v ktorej má hustota svoje maximum);
- **výberový medián** $\widetilde{X}$(robustný odhad polohy), ktorého realizácia

$$\widetilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{ak } n \text{ je nepárne,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{ak } n \text{ je párne;} \end{cases}$$

  rozdelenie je *symetrické*, ak $\overline{x} = \widetilde{x} = x_{\mathrm{mod}}$, rozdelenie je *pozitívne zošikmené* (pravostranne), ak $\overline{x} > \widetilde{x} > x_{\mathrm{mod}}$ a rozdelenie je *negatívne zošikmené* (ľavostranne), ak $\overline{x} < \widetilde{x} < x_{\mathrm{mod}}$;
- **výberové kvartily** poznáme tri
  - **prvý (dolný) kvartil** Q_1, ktorého realizácia $\widetilde{x}_{0.25}$ predstavuje hodnotu, od ktorej je 1/4 dát menšia a 3/4 dát sú väčšie,

$$\Pr\left[x_{\min}, \widetilde{x}_{0.25}\right] = \Pr\left[X \leq \widetilde{x}_{0.25}\right] = \frac{1}{4}, \Pr\left[\widetilde{x}_{0.25}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.25}\right] = \frac{3}{4};$$

  - **druhý kvartil** (medián) Q_2, ktorého realizácia $\widetilde{x}_{0.5} = \widetilde{x}$ je hodnota, od ktorej je 1/2 dát menšia a 1/2 dát je väčšia,

$$\Pr\left[x_{\min}, \widetilde{x}_{0.5}\right] = \Pr\left[X \leq \widetilde{x}_{0.5}\right] = \frac{1}{2}, \Pr\left[\widetilde{x}_{0.5}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.5}\right] = \frac{1}{2};$$

  - **tretí (horný) kvartil** Q_3, ktorého realizácia $\widetilde{x}_{0.75}$ predstavuje hodnotu, od ktorej je 1/4 dát väčšia a 3/4 dát sú menšie,

$$\Pr\left[x_{\min}, \widetilde{x}_{0.75}\right] = \Pr\left[X \leq \widetilde{x}_{0.75}\right] = \frac{3}{4}, \Pr\left[\widetilde{x}_{0.75}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.75}\right] = \frac{1}{4};$$

- **výberové decily** $\widetilde{X}_k$, ktorých realizácie $\widetilde{x}_k$ delia súbor na desatiny, t.j. $k/10$ dát je pod decilom a $(10-k)/10$ nad decilom, kde $k \in \{0, 1, \ldots, 10\}$;
- **výberové percentily** $\widetilde{X}_p$ (čítame ako $100p$-percentil[26]), ktorých realizácie $\widetilde{x}_p$ definujeme ako

$$\widetilde{x}_p = \begin{cases} x_{(k+1)} & \text{pre } k \neq np, \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{pre } k = np, \end{cases}$$

  kde $k = \lfloor np \rfloor$, čo je celá časť čísla np (niekedy sa používa definícia cez $\{x\}$ v indexe, čo znamená najbližšie celé číslo);

[26]Výraz „$100p$-percentil“ čítame ako „stokrát p-ty percentil“. Ak $p = 0.75$, potom $100 \times 0.75 = 75$, čo čítame ako „75. percentil“.

- **výberový päťčíselný súhrn** $(X_{\min}, Q_1, Q_2, Q_3, X_{\max})$, ktorého realizáciu označujeme ako $(x_{\min}, \widetilde{x}_{0.25}, \widetilde{x}_{0.50}, \widetilde{x}_{0.75}, x_{\max})$.

Príklad 117 (výšky 10-ročných dievčat) *Majme výšky $n = 12$ náhodne vybraných 10-ročných dievčat v cm usporiadaných podľa veľkosti (**poradia** ozn. ako r_i pre $x_{(i)}$; pri rovnakých pozorovaniach hovoríme o **strednoporadiach**; strednoporadie sa vypočíta ako priemer poradí realizácií s rovnakou hodnotou).*

Riešenie (pozri tabuľku 16)

Tabuľka 16: Zoradné realizácie x_i a ich poradia r_i pre výšky 10-ročných dievčat

i	1	2	3	4	5	6	7	8	9	10	11	12
$x_{(i)}$	131	132	135	141	141	141	141	142	143	146	146	151
r_i	1	2	3	5.5	5.5	5.5	5.5	8	9	10.5	10.5	12

$\overline{x} \doteq 140.83$, $\widetilde{x} = \frac{1}{2}\left(x_{(6)} + x_{(7)}\right) = 141$, $Q_1 = \widetilde{x}_{0.25} = \frac{1}{2}\left(x_{(3)} + x_{(4)}\right) = 138$, kde $k = \lfloor 12 \times 0.25 \rfloor = 3$, $Q_3 = \widetilde{x}_{0.75} = \frac{1}{2}\left(x_{(9)} + x_{(10)}\right) = 144.5$, kde $k = \lfloor 12 \times 0.75 \rfloor = 9$.

Čo sa stane so spomínanými charakteristikami polohy, keď zmeníme mierku (škálu), napr. gramy na kilogramy alebo namiesto hmotnosti použijeme logaritmus hmotnosti?

Nech $a, b \in \mathbb{R}$ sú nejaké dané konštanty, a je posunutie a b škála. Potom $y_i = a + bx_i$ a pre priemer[27]

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\left(na + b\sum_{i=1}^{n} x_i\right) = a + b\overline{x} = \overline{a + bx}.$$

Pokiaľ $b \in \mathbb{R}^+$, usporiadanie hodnôt x_i sa pri transformácii na $y_i = a + bx_i$ nezmení, teda

$$a + bx_{(1)} \leq a + bx_{(2)} \leq ... \leq a + bx_{(n)}.$$

Pre medián v tomto prípade platí[28]

$$\widetilde{y} = a + b\widetilde{x} = \widetilde{a + bx}.$$

Ľahko sa dá nahliadnuť, že usporiadanie zachová každá rastúca funkcia $g(x)$, teda platí

$$g\left(x_{(1)}\right) \leq g\left(x_{(2)}\right) \leq ... \leq g\left(x_{(n)}\right)$$

a pre medián bude platiť $g(\widetilde{x}) \doteq \widetilde{g}(x)$. Pre nepárne n platí predchádzajúci vzťah presne, označenie „približnosti“ potrebujeme pre párne n, kde $x_{\left(\frac{n}{2}\right)} < x_{\left(\frac{n}{2}+1\right)}$. V tomto prípade je však 1/2 hodnôt $g(x_i)$ menšia ako $\widetilde{g}(x)$. Teda špeciálne môžeme medián logaritmu (napr. hmotnosti) spočítať ako logaritmus mediánu (napr. hmotnosti). Pokiaľ dôjde v pozorovaniach k posunutiu, dôjde k rovnakému posunutiu aj v charakteristike polohy. Ak zmeníme mierku, potom stačí urobiť rovnakú úpravu aj u charakteristiky polohy.

Robustnou charakteristikou strednej hodnoty (odolnejšou na odľahlé pozorovania) je (Tukey, 1962)

[27]Rovnosť znamená, že priemer posunutej a preškálovanej veličiny y je rovný posunutému a preškálovanému priemeru pôvodnej veličiny x.

[28]Rovnosť znamená, že medián posunutej a preškálovanej veličiny y je rovný posunutému a preškálovanému mediánu pôvodnej veličiny x.

- **výberový γ-urezaný aritmetický priemer** $\overline{X}_g$, ktorého realizáciou je $\overline{x}_g$ a vypočíta sa ako

$$\overline{x}_g = \frac{1}{n-2g}\left(x_{(g+1)} + x_{(g+2)} + \ldots + x_{(n-g)}\right),$$

kde $g = \{\gamma n\}, g = \lfloor \gamma n \rfloor, \gamma = 0.1, 0.2$. Viac ako $\gamma 100$ % pozorovaní[29] musí byť nahradených, aby sa tento priemer zmenil na malý alebo veľký v porovnaní s pôvodným [[30]*breakdown point* angl $\overline{x}_t$ je teda γ],

- **výberový γ-winsorizovaný priemer** $\overline{X}_w$, ktorého realizácia $\overline{x}_w$ je definovaná ako

$$\overline{x}_w = \frac{1}{n}\left((g+1)x_{(g+1)} + x_{(g+2)} + \ldots + (g+1)x_{(n-g)}\right).$$

Viac ako $\gamma 100$ % pozorovaní musí byť nahradených, aby sa tento priemer zmenil na malý alebo veľký v porovnaní s pôvodným [*breakdown point* $\overline{x}_w$ je teda γ]. angl

3.2 Charakteristiky variability

Definujeme nasledovné základné odhady charakteristík variability (výberové charakteristiky variability) spolu s ich anglickými ekvivalentami:

- **výberový rozptyl** S^2, ktorého realizáciou je

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2;$$

pri lineárnej transformácii sa rozptyl mení nasledovne[31]

$$s_y^2 = s_{a+bx}^2 = b^2 s_x^2,$$

t.j.

$$\begin{aligned} s_y^2 = s_{a+bx}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(a + bx_i - \overline{a+bx}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(a + bx_i - (a + b\overline{x})\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(b\left(x_i - \overline{x}\right)\right)^2 = b^2 s_x^2; \end{aligned}$$

výpočtová podoba rozptylu

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{n-1}\left(\sum_{j=1}^{n_j} x_j^2 f_j - n\overline{x}^2\right), n_j \leq n,$$

kde f_j sú frekvencie (počty) prislúchajúcich x_j a $n = \sum_j f_j$;

[29] Výraz „$\gamma 100$ % pozorovaní“ čítame ako „gama krát stopercent pozorovaní“.

[30] *Breakdown point* hovorí o počte pozorovaní, ktoré potrebujeme na to, aby sme výrazne zmenili hodnotu charakteristiky polohy. Pre γ-urezaný a γ-winsorizovaný aritmetický priemer ide o γn pozorovaní, pre medián ide o $n/2$ pozorovaní a pre aritmetický priemer stačí iba jedno pozorovanie (preto hovoríme, že aritmetický priemer je veľmi citlivý na odľahlé pozorovania).

[31] Rovnosť znamená, že rozptyl posunutej a preškálovanej veličiny y je rovný násobku druhej mocniny škály a rozptylu pôvodnej veličiny x.

- **výberová smerodajná odchýlka** S, ktorej realizáciou je

$$s_x = \sqrt{s_x^2};$$

pri lineárnej transformácii sa smerodajná odchýlka mení nasledovne[32]

$$s_y = s_{a+bx} = |b|\, s_x,$$

teda, ak pripočítame ku všetkým pozorovaniam rovnakú konštantu, miera variability sa nezmení; zmena mierky (u pomerovej mierky zmena jednotiek) má za následok rovnakú úpravu jednotlivých pozorovaní i miery variability v podobe smerodajnej odchýlky;

- **výberový koeficient variácie** V_k, ktorého realizácia v_k predstavuje normalizovanú podobu výberového rozptylu (inverzia *signal-to-noise ratio*; podiel variability na priemere) angl

$$v_k = \frac{s_x}{\overline{x}};$$

bezrozmerná veličina, zvyčajne vyjadrovaná v percentách, t.j. $100 \times (s_x/\overline{x})$ % a môže sa používať len pre realizácie, ktorých rozsah nadobúda kladné hodnoty; používa sa pri porovnávaní variability súborov s nerovnakými priemermi (napr. pri porovnaní variability výšky detí určitého veku s výškou dospelých určitého veku alebo pri porovnaní variability premenných meraných v rôznych jednotkách);

- **výberový rozptyl aritmetického priemeru** $S^2_{\overline{X}}$, ktorého realizáciou je

$$s^2_{\overline{x}} = \frac{s_x^2}{n};$$

- **výberová stredná chyba priemeru** (**štandardná chyba**) $S_{\overline{X}}$, ktorej realizáciou je

$$s_{\overline{x}} = \frac{s_x}{\sqrt{n}};$$

- **výberový koeficient šikmosti** B_1, ktorého realizáciou je

$$b_1 = \frac{n^{-1}\sum_{i=1}^n (x_i - \overline{x})^3}{\left[n^{-1}\sum_{i=1}^n (x_i - \overline{x})^2\right]^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^n (x_i - \overline{x})^3}{\left[\sum_{i=1}^n (x_i - \overline{x})^2\right]^{3/2}},$$

kde rozdelenie je *symetrické*, ak $b_1 = 0$, *pozitívne zošikmené* (hustota na ľavej strane stúpa strmšie ako na pravej), ak $b_1 > 0$ a *negatívne zošikmené* (hustota na pravej strane stúpa strmšie ako na ľavej), ak $b_1 < 0$;

- **výberový koeficient špicatosti** B_2, ktorého realizáciou je

$$b_2 = \frac{n^{-1}\sum_{i=1}^n (x_i - \overline{x})^4}{\left[n^{-1}\sum_{i=1}^n (x_i - \overline{x})^2\right]^2} - 3 = \frac{n\sum_{i=1}^n (x_i - \overline{x})^4}{\left[\sum_{i=1}^n (x_i - \overline{x})^2\right]^2} - 3,$$

kde rozdelenie je normálne (*mezokurtické*), ak $b_2 = 0$, špicaté (*leptokurtické*), ak $b_2 > 0$ a ploché (*platykurtické*), ak $b_2 < 0$;

[32]Rovnosť znamená, že smerodajná odchýlka posunutej a preškálovanej veličiny y je rovná násobku absolútnej hodnoty škály a smerodajnej odchýlky pôvodnej veličiny x.

- **výberová suma štvorcov** $\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, ktorej realizácia je

$$SS=\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2,$$

kde sa tento čitateľ rozptylu používa napr. v lineárnom regresnom modeli, v modeli ANOVA a pod.;

- **výberová suma absolútnych odchýlok** $\sum_{i=1}^{n}\left|X_i-\widetilde{X}_{0.5}\right|$, ktorej realizácia je

$$SAD=\sum_{i=1}^{n}\left|x_i-\widetilde{x}_{0.5}\right|;$$

- **výberový priemer absolútnych odchýlok** $\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\widetilde{X}_{0.5}\right|$, ktorého realizácia je

$$MAD=SAD/n;$$

- **výberové rozpätie** $X_{\max}-X_{\min}$, ktorého realizáciou je

$$D=x_{\max}-x_{\min};$$

- **výberové medzikvartilové rozpätie** Q_3-Q_1, ktorého realizáciou je

$$D_Q=\widetilde{x}_{0.75}-\widetilde{x}_{0.25};$$

kde rozdelenie je (medzi kvartilmi) *symetrické*, ak $\widetilde{x}_{0.75}-\widetilde{x}_{0.50}=\widetilde{x}_{0.50}-\widetilde{x}_{0.25}$, *pozitívne zošikmené*, ak $\widetilde{x}_{0.75}-\widetilde{x}_{0.50}>\widetilde{x}_{0.50}-\widetilde{x}_{0.25}$ a *negatívne zošikmené*, ak $\widetilde{x}_{0.75}-\widetilde{x}_{0.50}<\widetilde{x}_{0.50}-\widetilde{x}_{0.25}$;

- **výberové decilové rozpätie** $\widetilde{X}_{0.9}-\widetilde{X}_{0.1}$, ktorého realizáciou je

$$D_D=\widetilde{x}_{0.9}-\widetilde{x}_{0.1};$$

- **výberové percentilové rozpätie** $\widetilde{X}_{0.99}-\widetilde{X}_{0.01}$, ktorého realizáciou je

$$D_P=\widetilde{x}_{0.99}-\widetilde{x}_{0.01}.$$

Robustnými charakteristikami variability sú (Tukey, 1962)

- **výberový γ-urezaný rozptyl** S_g^2, ktorého realizácia s_g^2 sa vypočíta ako

$$s_g^2=\frac{1}{n-2g-1}\sum_{i=g+1}^{n-g}x_{(i)};$$

viac ako $\gamma 100$ % pozorovaní musí byť nahradených, aby sa tento rozptyl zmenil na veľký v porovnaní s pôvodným s^2 [*breakdown point* s_g^2 je γ]; platí $s_g^2<s^2$ pretože urezanie odstráni angl odľahlé hodnoty;

- **výberový γ-winsorizovaný rozptyl** S_w^2, ktorého realizáciu označujeme ako s_w^2; viac ako $\gamma 100$ % pozorovaní musí byť nahradených, aby sa tento rozptyl zmenil na veľký v porovnaní s pôvodným s^2 [*breakdown point* s_w^2 je γ]; platí $s_w^2<s^2$ pretože winsorizovanie priťahuje angl extrémne hodnoty bližšie k priemeru;

- **výberový kvartilový koeficient variácie** $V_{k,Q}$, ktorého realizáciu $v_{k,Q}$ vypočítame ako

$$v_{k,Q} = \frac{Q_3 - Q_1}{Q_2}.$$

Ďalšie robustné charakteristiky variability (výberové rozpätie) charakterizované pomocou upravených hraníc sú

- **výberové robustné minimum a maximum** („vnútorné hradby“) $X^*_{\min}$ a $X^*_{\max}$, ktorých realizácie sú definované ako

$$x^*_{\min} = B_D = \widetilde{x}_{0.25} - 1.5\,(\widetilde{x}_{0.75} - \widetilde{x}_{0.25}) = Q_1 - 1.5D_Q,$$
$$x^*_{\max} = B_H = \widetilde{x}_{0.75} + 1.5\,(\widetilde{x}_{0.75} - \widetilde{x}_{0.25}) = Q_1 - 1.5D_Q,$$

  kde prvky vybočujúce z hradieb sa považujú za *podozrivé, potencionálne odľahlé pozorovania*;

- **výberové robustné minimum a maximum** („vonkajšie hradby“) definované ako $Q_1 - 3(Q_3 - Q_1)$, $Q_3 + 3(Q_3 - Q_1)$, ktorých realizácie sú $B^*_D = \widetilde{x}_{0.25} - 3D_Q$, $B^*_H = \widetilde{x}_{0.75} + 3D_Q$
  - pokiaľ sú nejaké $x_i < B^*_D \vee x_i > B^*_H$, hovoríme, že ide o *vzdialené body*[33],
  - ak $x_i \in \langle B^*_D, B_D) \vee (B_H, B^*_H\rangle$, ide o *body vonkajšie*,
  - ak $x_i \in \langle B_D, B_H\rangle$, ide o *body vnútorné* alebo *body priľahlé mediánu*;
  - pre normálne rozdelenie platí $B_H - B_D = Q_3 + 1.5D_Q - Q_1 + 1.5D_Q = 4D_Q \doteq 4.2$; pravdepodobnosť, že $x_i \notin \langle B_D, B_H\rangle$ je potom 0.04;

- **výberové robustné miery šikmosti** B_{1Q} a B_{1O} a ich rozptyly za podmienky asymptotickej normality $B_{1\cdot}$, kde $\cdot = Q$ alebo O, ktorých realizácie sú definované nasledovne
  - kvartilový koeficient šikmosti

$$b_{1Q} = \frac{(\widetilde{x}_{0.75} - \widetilde{x}_{0.50}) - (\widetilde{x}_{0.50} - \widetilde{x}_{0.25})}{\widetilde{x}_{0.75} - \widetilde{x}_{0.25}}, Var_{as}(b_{1Q}) = 1.84,$$

  - oktilový koeficient šikmosti

$$b_{1O} = \frac{(\widetilde{x}_{0.875} - \widetilde{x}_{0.50}) - (\widetilde{x}_{0.50} - \widetilde{x}_{0.125})}{\widetilde{x}_{0.875} - \widetilde{x}_{0.125}}, Var_{as}(b_{1O}) = 1.15.$$

3.3 Detekcia odľahlých pozorovaní

Homogénny náhodný výber je taký výber, v ktorom všetky $x_i, i = 1, 2, \ldots, n$, sú realizácie rovnakého rozdelenia pravdepodobnosti s konštantným rozptylom σ^2. K **nehomogenitám výberu** dochádza všade tam, kde sa vyskytujú výrazné nerovnomernosti v realizáciách, náhle sa menia podmienky experimentu a pod. Nehomogenita môže byť spôsobená aj nevhodne zvoleným výberom subjektov.

Špeciálnym prípadom ovplyvňujúcim homogenitu výberu sú **odľahlé pozorovania** (outliers). angl
Takéto pozorovania skresľujú odhady polohy (špeciálne aritmetického priemeru) a variability (hlavne rozptylu), takže môžu znehodnotiť ďalšiu štatistickú analýzu. Pri ich overovaní sa používa mnoho idealizovaných predpokladov. Musíme poznať ich predpokladaný počet, ich rozdelenie a tiež rozdelenie ostatných prvkov náhodného výberu. Navyše je nutné zostrojiť štatistický alebo pravdepodobnostný model, podľa ktorého sa odľahlé pozorovania chovajú. Testovanie odľahlých pozorovaní bez doplnkových informácií je teda málo spoľahlivé.

Jednoduchou technikou, kedy sa predpokladá, že dáta majú normálne rozdelenie, je **modifikácia vnútorných hradieb** B_D a B_H na

[33] Ozn. $\vee$ znamená „alebo“ a ozn. $\wedge$ znamená „a súčasne“.

$$B_D^{mod} = \widetilde{x}_{0.25} - kD_Q, B_H^{mod} = \widetilde{x}_{0.75} + kD_Q,$$

kde sa parameter k volí tak, aby pravdepodobnosť $\Pr(n,k)$ bola dostatočne vysoká, napr. 0.95. $\Pr(n,k)$ je pravdepodobnosť, že žiaden prvok z náhodného výberu z normálneho rozdelenia s rozsahom n nebude mimo intervalu $I = \langle B_D^{mod}, B_H^{mod} \rangle$. Ak $\Pr(n,k) = 0.95$ a $n \in \langle 8, 100 \rangle$ použijeme aproximáciu $k \approx 2.25 - 3.6/n$. Teda prvky mimo I sa považujú za odľahlé. Postup spomenutý vyššie je dostatočne robustný.

Definujme mieru rozptylu ako **kvantilovú odchýlku** $D_{Q^*} = 2D_Q$. Ak urobíme štandardizáciu, dostaneme $D_{Q_{st}^*} = 1$ a **štandardizovaný medián** bude $\widetilde{x}_{st} = 0$ a **štandardizovaná kvantilová funkcia** indikujúca tvar bude (Meloun a Militký, 2004)

$$Q_{st}(p) = \frac{\widetilde{x}_p - \widetilde{x}_{0.5}}{D_{Q^*}}.$$

Hodnoty kvantilov, pre ktoré platí $|Q_{st}(p)| \geq 1$, sú považované za vybočujúce (pre normálne rozdelenie) a hovoríme potom o **identifikátoroch dlhých „chvostov“** (koncov). Hodnoty $Q_{st}(p)$ môžeme použiť na

– identifikáciu miery šikmosti $SQ = Q_{st}(0.25) + Q_{st}(0.75)$, kedy je rozdelenie pravdepodobnosti symetrické medzi kvartilmi, ak SQ je rovné nule,

– identifikáciu dĺžky koncov, kedy

- $Q_{st}(0.95) < 0.5$ hovorí o krátkych koncoch,
- $Q_{st}(0.95) > 1$ hovorí o dlhých koncoch a
- pre stredne dlhé konce bude platiť $Q_{st}(0.95) \in \langle 0.5, 1.0 \rangle$.

Okrem vonkajších a vnútorných hradieb, môžeme odľahlé pozorovania jednoducho detegovať aj nasledovne (Rousseeuw a van Zomeren, 1990)

- $\left|X - \overline{X}\right| > 2s$,
- $\widehat{\sigma} = MAD/0.6745$, $\left|X - \widetilde{X}_{0.5}\right| > k\frac{MAD}{0.6745}$ (ako k sa najčastejšie používa 2 alebo 2.24).

Príklad 118 (odľahlé pozorovania; dĺžka kľúčnej kosti) *Za predpokladu normality náhodnej premennej X najväčšia dĺžka kosti klúčnej z pravej strany (`cla.L`; dáta: `more-samples-variances-clavicle.txt`), t.j. $X \sim N(\mu, \sigma^2)$, identifikujte odľahlé pozorovania pomocou (a) vnútorných hradieb B_D a B_H, (b) modifikovaných vnútorných hradieb B_D^{mod} a B_H^{mod}, (c) identifikujte dlhé „chvosty“ rozdelenia tejto premennej na základe štandardizovanej kvantilovej funkcie.*

3.4 Z-skóre

V antropológii nás veľmi často zaujímajú normované realizácie, nazývané tiež normované veličiny, a to z-**skóre**, ktoré definujeme ako

$$z_i = \frac{x_i - \overline{x}}{s_x}.$$

Dostaneme ich ako špeciálny prípad lineárnej transformácie $y = a + bx$, kde voľbou $b = 1/s_x$ a $a = -\overline{x}/s_x$. Potom aritmetický priemer z-skóre

$$\overline{z} = -\frac{\overline{x}}{s_x} + \frac{1}{s_x}\overline{x} = 0$$

a rozptyl z-skóre

$$s_z^2 = \left(\frac{1}{s_x}\right)^2 s_x^2 = 1,$$

Pomocou z-skóre môžeme vyjadriť aj koeficient šikmosti a špicatosti, kde

$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3,$$

a za podmienky asymptotickej normality

$$E\left[b_1\right] = 0, Var\left[b_1\right] = \frac{n-2}{(n+1)(n+3)};$$

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^4 = \frac{1}{n}\sum_{i=1}^{n} z_i^4$$

a za podmienky asymptotickej normality

$$E\left[b_2\right] = 3 - \frac{6}{n+1}, Var\left[b_2\right] = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Pokiaľ dáta pochádzajú z normálneho rozdelenia, budú mať oba koeficienty hodnoty približne nulové (pri b_2 po odčítaní konštanty 3).

Musíme si uvedomiť, že prítomnosť odľahlých pozorovaní v realizáciách ovplyvňuje výpočet priemeru a rozptylu, ktoré sú potrebné na výpočet z-skóre, a následne tak ovplyvňuje tiež hodnotu vlastného z-skóre. Ak je rozdelenie dát zošikmenené alebo je normalita porušená inak, nebude z-skóre odrážať situáciu vierohodne a jeho ďalšie použitie je problematické. Ak predpokladáme, že je rozdelenie znaku v populácii normálne, obmedzuje sa použie z-skóre napr. na zistenie, či nejaké pozorovanie (pacient) patrí svojimi charakteristikami do zdravej populácie.

Príklad 119 (z-skóre; šírka lebky) *Majme náhodnú premennú X šírka lebky (`skull.B`; mm; dáta: `one-sample-mean-skull-mf.txt`) u mužov. Za predpokladu asymptotickej normality X, t.j. $X \sim N(\mu, \sigma^2)$, vypočítajte z-skóre tejto premennej pomocou funkcie `mean()` a `sd()`. Výsledok skontrolujte pomocou funkcie `scale()`.*

3.5 Príklady na charakteristiky polohy a variability

Príklad 120 (argument minima) *Vygenerujte pseudonáhodné čísla $X \sim N(\mu, \sigma^2)$, kde $\mu = 0, \sigma^2 = 1$ a $n = 1000$. Vygenerované čísla ozn. $x_i, i = 1, 2, \ldots, 1000$. Nájdite numericky také c, ktoré minimalizuje (a) sumu štvorcov odchýlok $\sum_{i=1}^{1000}(x_i - c)^2$, t.j. $c_1 = \arg\min_{\forall c}\sum_{i=1}^{1000}(x_i - c)^2$ a (b) sumu absolútnych odchýlok $\sum_{i=1}^{100}|x_i - c|$, t.j. $c_2 = \arg\min_{\forall c}\sum_{i=1}^{1000}|x_i - c|$. Za c dosadzujte postupne (1) všetky $x_{(j)}$ ($x_{(j)}$ sú usporiadané x_i podľa veľkosti od najmenšieho po najväčšie) a vybrané charakteristiky polohy ako (2) aritmetický priemer, (3) nejaké kvantily $\widetilde{x}_p$, kde $p \in \langle 0, 1\rangle$ a pod. Nakreslite obrázok závislosti (a) sumy štvorcov odchýlok na $x_{(j)}$, t.j. body $[x_j, y_j]$, kde $y_j = \sum_{i=1}^{1000}(x_i - x_{(j)})^2$ a (b) sumy absolútnych odchýlok na $x_{(j)}$, t.j. body $[x_{(j)}, y_j]$, kde $y_j = \sum_{i=1}^{1000}|x_i - x_{(j)}|$. Podobné obrázky nakreslite aj pre $\widetilde{x}_p$ namiesto $x_{(j)}$.*

Príklad 121 (výšky 10-ročných dievčat, pokrač.) *Vypočítajte základné charakteristiky polohy a variability.*

Riešenie v R (pozri tabuľku 17 a 18)
Dáta:

```
x <- c(131, 132, 135, 141, 141, 141, 141, 142, 143, 146, 146, 151)
```

Minimum, maximum, medián, aritmetický priemer, prvý kvartil, druhý kvartil, tretí kvartil, kvartily (pomocou jednej funkcie), päťčíselný súhrn, rozptyl a smerodajná odchýlka:

```
min(x) # 131
max(x) # 151
median(x) # 141
mean(x) # 140.8333
priemer <- sum(x)/length(x) # 140.8333
q1 <- quantile(x,0.25,type=2) # 138
q2 <- quantile(x,0.50,type=2) # 141
q3 <- quantile(x,0.75,type=2) # 144.5
quantile(x,c(0.25,0.5,0.75),type=2) # 138.0 141.0 144.5
quantile(x,c(0,0.25,0.5,0.75,1),type=2) # 131.0 138.0 141.0 144.5 151.0
var(x) # 33.78788
sd(x) # 5.812734
```

Funkcie na výpočet rozptylu, smerodajnej odchýlky, štandardnej chyby, šikmosti a špicatosti:

```
"rozptyl" <- function(x) sum((x-mean(x))^2)/(length(x)-1)
"smerodch" <- function(x) sqrt(var(x))
"SE" <- function(x) sqrt(var(x) / length(x))
SE(x) # 1.677992
"sikmost" <- function(x) {(1/length(x))*sum(((x-mean(x))/(sqrt(var(x))))^3)}
sikmost(x) # -0.2121993
"spicatost" <- function(x) {(1/length(x))*sum(((x-mean(x))/
                            (sqrt(var(x))))^4) - 3}
spicatost(x) # -0.9029347
```

Suma štvorcov, priemerná absolútna odchýlka (priemer absolútnych odchýlok), suma absolútnych odchýlok, rozpätie, medzikvartilové rozpätie, zisťovanie symetrie, robustný výpočet minima a maxima („vnútorné hradby“), robustné rozpätie:

```
var(x)*(length(x)-1) # 371.6667
mad(x) # 5.1891
mad(x)*length(x) # 62.2692
c(min(x),max(x)) # 131 151
range(x) # 131 151
Dq <- quantile(x,0.75,type=2)-quantile(x,0.25,type=2) # 6.5
c(q3 - q2, q2 - q1) # 3.5 3.0
Bd <- q1-1.5*Dq; Bh <- q3+1.5*Dq
c(Bd,Bh) # 128.25 154.25
```

Funkcia na výpočet kvartilovej šikmosti:

```
"kvart.sikmost" <- function(x) {
                    q1 <- quantile(x,0.25,type=2)
                    q2 <- quantile(x,0.5,type=2)
                    q3 <- quantile(x,0.75,type=2)
                    ((q3 - q2) - (q2 - q1))/(q3 - q1)
                    }
kvart.sikmost(x) # 0.07692308
```

Funkcia na výpočet oktilovej šikmosti:

```
"oktil.sikmost" <- function(x)  {
                   q125 <- quantile(x,0.125,type=2)
                   q2 <- quantile(x,0.5,type=2)
                   q875 <- quantile(x,0.875,type=2)
                   ((q875 - q2) - (q2 - q125))/(q875  - q125)
                   }
oktil.sikmost(x) # -0.2857143
```

Funkcia na výpočet výberového γ-urezaného aritmetického priemeru a rozptylu (vytvorenie dátového vektora na ich výpočet):

```
"urezanie" <- function(x, gama = 0.1){
x <- na.omit(x) # odstranenie NA, ak sa v datach nachadzaju
n <- length(x)
g.min <- floor(gama*n)  # najvacsie cele cislo mensie ako gama*n
g.max <- floor((1-gama)*n) # najvacsie cele cislo mensie ako (1-gama)*n
x.sort <- sort(x) # zoradenie podla velkosti
x.min <- x.sort[g.min] # vybratie dolnej hranice
x.max <- x.sort[g.max] # vybratie hornej hranice
xg <- x[x > x.min & x < x.max]
return(xg)
}
```

Funkcia na výpočet výberového aritmetického priemeru a rozptylu winsorizovaného pomocou „vnútorných hradieb“ (vytvorenie dátového vektora na ich výpočet):

```
"winsorizacia" <- function(x){
x <- na.omit(x) # odstranenie NA
q1 <- quantile(x,0.25)
q3 <- quantile(x,0.75)
Dq <- q3 - q1
min.x <- q1 - 1.5*Dq
max.x <- q3 + 1.5*Dq
xw <- x
for (i in 1: length(x)) if (x[i] >= max.x) xw[i] <- max.x
for (i in 1: length(x)) if (x[i] <= min.x) xw[i] <- min.x
return(xw)
}
```

Porovnanie troch typov aritmetických priemerov a rozptylov:

```
xg <- urezanie(x)
xw <- winsorizacia(x)
TAB <- rbind(c(length(x), mean(x), sd(x)),
             c(length(xg), mean(xg), sd(xg)),
             c(length(xw), mean(xw), sd(xw)))
dimnames(TAB)[[1]] <- c("surove␣data","urezane␣data","winsorizovane␣data")
dimnames(TAB)[[2]] <- c("rozsah","priemer","sd")
TAB <- round(TAB,2)
```

Tabuľka 17: Rozsah, aritmetický priemer a smerodajná odchýlka pre surové, urezané a winsorizované dáta (výšky 10-ročných dievčat)

	rozsah	aritmetický priemer	smerodajná odchýlka
surové dáta	$n = 12$	$\overline{x} = 140.83$	$s^2 = 5.81$
urezané dáta	$n_g = 8$	$\overline{x}_g = 139.50$	$s_g^2 = 3.85$
winsorizované dáta	$n_w = 12$	$\overline{x}_w = 141.03$	$s_w^2 = 5.21$

Funkcia na výpočet niektorých základných charakteristík:

```
"zakl.char" <- function(x, type = 7){
# odstranenie chybajucich pozorovani
```

Tabuľka 18: Vybrané charakteristiky polohy a variability pre surové dáta (výšky 10-ročných dievčat)

charakteristika	n	$\overline{x}$	$\widetilde{x}_{\min}$	$\widetilde{x}_{0.25}$	$\widetilde{x}_{0.50}$	$\widetilde{x}_{0.75}$	$\widetilde{x}_{\max}$	b_1	b_2	s_x	$s_{\overline{x}}$
hodnota	12	140.8	131.0	138.0	141.0	144.5	151.0	−0.21	−0.90	5.81	1.68

```
x <- x[!is.na(x)]
n <- length(x) # rozsah
# kvantily p = 0,0.25,0.5,0.75 a 1
kvantily <- quantile(x,c(0,0.25,0.5,0.75,1),type=type)
priemer <- mean(x) # priemer
SD <- sd(x) # smerodajna odchylka
StEr <- SE(x) # standardna chyba
sikm <- sikmost(x) # sikmost
spic <- spicatost(x) # spicatost
# vsetky vysledky spolu
vysledky <- c(n,priemer,kvantily,sikm,spic,SD,StEr)
# priradenie nazvov vysledkom
names(vysledky) <- c("n","priem",names(kvantily),"sik","spic","sd","se")
# zaokruhlenie na dve desatinne miesta
vysledky <- round(vysledky,2)
return(vysledky)
}
zakl.char(x, type = 2)
```

Príklad 122 (základné charakteristiky polohy a variability) *Vypočítajte základné charateristiky polohy a variability pre premennú najväčšia dĺžka lebky (`skull.L`) a najväčšia šírka lebky (`skull.B`) u mužov; dáta: `one-sample-mean-skull-mf.txt`. Výsledok uložte pomocou funkcie `write.table()`.*

Riešenie v R (pozri tabuľku 19)

```
DATA <- read.table("one-sample-mean-skull-mf.txt",header=TRUE)
names(DATA) ## "id"      "pop"     "sex"     "skull.L" "skull.B"
attach(DATA)
ZCH1 <- zakl.char(skull.L[sex=="m"])
ZCH2 <- zakl.char(skull.B[sex=="m"])
ZCH <- rbind(ZCH1,ZCH2)
dimnames(ZCH)[[1]] <- c("skull.L","skull.B")
ZCH
write.table(ZCH,"skull-tab-01.txt")
```

Tabuľka 19: Vybrané charakteristiky polohy a variability pre najväčšiu dĺžku lebky

	n	$\overline{x}$	$\widetilde{x}_{\min}$	$\widetilde{x}_{0.25}$	$\widetilde{x}_{0.50}$	$\widetilde{x}_{0.75}$	$\widetilde{x}_{\max}$	b_1	b_2	s_x	$s_{\overline{x}}$
`skull.L`	217	182.04	164.00	177.00	182.00	187.00	199.00	-0.06	-0.48	6.37	0.43
`skull.B`	216	137.19	124.00	134.00	137.00	140.00	149.00	0.08	-0.30	4.82	0.33

Príklad 123 (šikmosť a špicatosť) *Naprogramujte funkcie na výpočet rozptylu šikmosti a špicatosti.*

Príklad 124 (základné charakteristiky polohy a variability) *Vypočítajte základné charakteristiky polohy a variability pre nasledujúce premenné:*
(a) šírko-dĺžkový index lebky vypočítaný ako podiel premenných šírka lebky (`skull.B`; v mm) a dĺžka lebky (`skull.L`; v mm; dáta: `one-sample-mean-skull-mf.txt`) u mužov;
(b) stranový rozdiel vertikálneho priemeru diafýzy kľúčnej kosti (`simd.R` a `simd.L`; v mm) na pravej aj ľavej strane tela (dáta: `paired-means-clavicle2.txt`);
(c) najväčšia výška mozgovne (`skull.pH`; mm) a morfologická výška tváre (`face.H`; mm; dáta: `one-sample-correlation-skull-mf.txt`).

3.6 Štatistická grafika

Pokiaľ chceme zobraziť základné a relevantné grafy (spolu s výpočtom základných charakteristík polohy a variability), hovoríme o **exploratórnej analýze dát (EDA)**; pozri Murrell (2011) a Kabacoff (2011). Grafická interpretácia výberového súboru je možná pomocou stĺpcového diagramu, spojnicového grafu (polygónu početnosti, frekvenčnej krivky a polygónu kumulatívnych početností), bodového grafu, kruhového diagramu, histogramu, empirickej distribučnej funkcie, krabicového diagramu a kvantilového diagramu.

3.6.1 Stĺpcový diagram

Stĺpcový diagram – vyjadruje číselné hodnoty pomocou obdĺžnikových stĺpcov, obyčajne v zvislej, no niekedy aj vo vodorovnej polohe; môže byť neškálovaný, t.j. v *absolútnej škále*, alebo škálovaný, t.j. v *relatívnej škále* (lepšie porovnanie v prípade sledovania viacerých súborov); špeciálnymi variantmi sú **veková pyramída** (strom života, znázorňuje vekové zloženie obyvateľstva) a **histogram**. Stĺpcový diagram nakreslíme pomocou funkcie `barplot(x)`.

Kombinácii stĺpcových diagramov usporiadaných nad sebou pre škálované kategoriálne dáta, t.j. dáta v podobe pravdepodobností, ktorých suma je rovná jednej, sa hovorí aj **spinogram**. Spinogram nakreslíme pomocou funkcií

```
199 library(vcd)
200 spine(x)
```

Príklad 125 (stĺpcový diagram; oči vs. vlasy) *Kontingenčná tabuľka predstavuje pravdepodobnosti výskytu rôznych farieb vlasov a očí v populácii. Ide o model multinomického rozdelenia, ozn. $Mult(\mathbf{p}, N)$, pretože všetky pravdepodobnosti dávajú v sume jednotku. Vypočítajte početnosti všetkých buniek tabuľky, ak máme v populácii* 1000 *jedincov. Prepočítajte pravdepodobnosti na model súčinového multinomického rozdelenia (pravdepodobnosti výskytu pre dve farby očí a tri farby vlasov pozri v tabuľke 20). Použite základné funkcie (pozri Spector, 2008; Venables a kol., 2013; Verzani, 2005) a skontrolujte pomocou funkcií `margin.table(oci)` a `prop.table(oci)`. Nakreslite stĺpcové diagramy pre oba typy súčinových multinomických rozdelení – pre početnosti, ako aj pre pravdepodobnosti. [Marginálne súčty možno pridať pomocou funkcie `addmargins(oci)`; keby sme mali surové dáta (hodnoty* 0 *a* 1 *pre každý subjekt), kontingenčnú tabuľku vytvoríme pomocou funkcie `ftable(data)`.]*

Riešenie v R (pozri tabuľky 21 až 24 a obrázok 20)

```
201 modre.oci <- c(0.12,0.22,0.06) # vektor pravdepodobnosti
202 hnede.oci <- c(0.15,0.34,0.11) # vektor pravdepodobnosti
203 oci <- rbind(modre.oci,hnede.oci) # dva vektory spojene do matice
204 sum(oci) # celkova (totalna) suma = 1
```

Tabuľka 20: Kontingenčná tabuľka 2×3 pravdepodobností výskytu pre dve farby očí a tri farby vlasov

oči/vlasy	blond	hnedé	ryšavé
modré	0.12	0.22	0.06
hnedé	0.15	0.34	0.11

```
# nazvy riadkov a stlpcov
dimnames(oci)[[1]] <- c("modre","hnede")
dimnames(oci)[[2]] <- c("blond","hnede","rysave")
round(addmargins(oci),2) # marginalne pravdepodobnosti

oci.pocty <- oci*1000 # fiktivne pocetnosti
addmargins(oci.pocty) # marginalne pocetnosti

sumy <- apply(oci.pocty,2,sum) # sumy po stlpcoch
oci.prav <- oci.pocty
oci.prav[1,] <- oci.pocty[1,]/sumy
oci.prav[2,] <- oci.pocty[2,]/sumy
apply(oci.prav,2,sum) # suma po stlpcoch = 1
round(addmargins(oci.prav,1),2) # marginalne pocetnosti po riadkoch

sumy1 <- apply(oci.pocty,1,sum) # sumy po riadkoch
oci.prav1 <- oci.pocty
oci.prav1[,1] <- oci.pocty[,1]/sumy1
oci.prav1[,2] <- oci.pocty[,2]/sumy1
oci.prav1[,3] <- oci.pocty[,3]/sumy1
apply(oci.prav1,1,sum) # suma po riadkoch = 1
round(addmargins(oci.prav1,2),4) # marginalne pocetnosti po stlpcoch

par(mfcol=c(2,2),mar=c(5.5,2,1,1))
barplot(oci.pocty,space=0)
barplot(oci.prav,space=0)
barplot(t(oci.pocty),space=0)
barplot(t(oci.prav1),space=0)
# mozne pridanie legendy (nezobrazena)
# legend("topright",c("modre oci","hnede oci"),fill=c("black","grey"))
```

Tabuľka 21: Kontingenčná tabuľka 2×3 pravdepodobností výskytu pre dve farby očí a tri farby vlasov spolu s marginálnymi pravdepodobnosťami (multinomické rozdelenie)

oči/vlasy	blond	hnedé	ryšavé	suma
modré	0.12	0.22	0.06	0.40
hnedé	0.15	0.34	0.11	0.60
suma	0.27	0.56	0.17	1.00

Tabuľka 22: Kontingenčná tabuľka 2×3 početností výskytu pre dve farby očí a tri farby vlasov spolu s marginálnymi početnosťami (multinomické rozdelenie)

oči/vlasy	blond	hnedé	ryšavé	suma
modré	120	220	60	400
hnedé	150	340	110	600
suma	270	560	170	1000

Tabuľka 23: Kontingenčná tabuľka 2×3 pravdepodobností výskytu pre dve farby očí a tri farby vlasov spolu s marginálnymi stĺpcovými početnosťami (súčinové multinomické rozdelenie; po stĺpcoch)

oči/vlasy	blond	hnedé	ryšavé
modré	0.4444	0.3929	0.3529
hnedé	0.5556	0.6071	0.6471
suma	1.0000	1.0000	1.0000

Tabuľka 24: Kontingenčná tabuľka 2×3 pravdepodobností výskytu pre dve farby očí a tri farby vlasov spolu s marginálnymi riadkovými početnosťami (súčinové multinomické rozdelenie; po riadkoch)

oči/vlasy	blond	hnedé	ryšavé	suma
modré	0.3000	0.5500	0.1500	1.0000
hnedé	0.2500	0.5667	0.1833	1.0000

3.6.2 Spojnicový graf, polygón početnosti a frekvenčná krivka

Spojnicový graf – znázorňuje priebeh časového radu a jeho špeciálnymi prípadmi sú **polygón početnosti**, **frekvenčná krivka**, **polygón kumulatívnych početností**.

Polygón početnosti – spojnicový diagram, v ktorom nad stredmi triednych intervalov I_i vztýčime kolmice, s výškou úmernou príslušným triednym početnostiam a koncové body kolmíc pospájame. Ich súradnice sú $[x_i^*, n_i]$. Ide o zobrazenie priebehu početností vnútri každého intervalu, ale plocha uzavretá spojnicou polygónu nie je presne (len približne) úmerná počtu pozorovaní v intervale.

Frekvenčná krivka – vznikne, keď koncové body polygónu početnosti pospájame hladkou krivkou; vystihuje celkom presne priebeh rozdelenia početnosti a plocha v každom mieste ohraničená krivkou je priamo úmerná počtu pozorovaní.

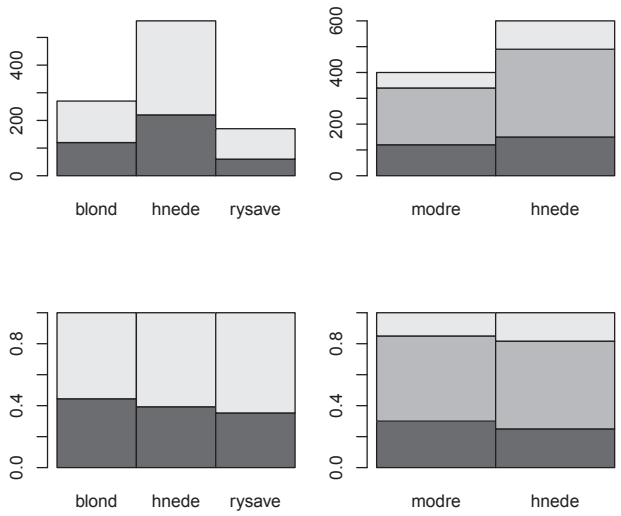
3.6.3 Bodový graf

Bodový (**rozptylový**) **graf** – zobrazuje namerané hodnoty v pravouhlej súradnicovej sústave (dvojdimenzionálnej, 2D; trojdimenzionálnej, 3D), pričom jednotlivé kategórie sa odlišujú pomocou rôznych značiek, farieb a pod.; často sa používa na zobrazenie závislosti dvoch znakov. Dvojdimenzionálny a trojdimenzionálny rozptylový graf nakreslíme pomocou funkcií

```
231 plot(x,y) # 2D
232 library(scatterplot3d)
233 scatterplot3d(x,y,z) # 3D
```

Argumenty funkcie `plot(x)`:

- `type=` argument kontrolujúci typ grafu
  - `type="p"` – body (prednastavená hodnota; default),
  - `type="l"` – čiary,
  - `type="b"` – body pospájané čiarami,
  - `type="s"` – schodovitá funkcia,
  - `type="n"` – prázdny obrázok;
- argumenty popisu osí a grafu
  - `xlab="retazec"` – popis x-ovej osi,
  - `ylab="retazec"` – popis y-ovej osi,



Obr. 20: Stĺpcové diagramy – početnosti (prvý riadok) pre vlasy (vľavo), pre oči (vpravo); pravdepodobnosti (druhý riadok) pre vlasy (vľavo), pre oči (vpravo)

  - `main="retazec"` – hlavný nadpis,
  - `sub="retazec"` – podnadpis pod x-ovou osou;

- farba `col="anglicky.nazov"` alebo kódovanie v RGB-škále (funkcia `rgb(cislo1, cislo2, cislo3)` vytvorí RGB-vektor z hodnôt intenzity); transformácia `col2rgb(anglicky.nazov)` vytvorí RGB-vektor z anglického názvu farby, kde RGB-vektor je textový vektor so 7 alebo 9 elementmi, kde za "#"nasleduje *red*, *blue* a *green* farebný kanál a voliteľne aj koeficient transparencie α v hexadecimálnej sústave (po preškálovaní na hodnoty $0, \ldots, 255$; default je `"black"`;
- veľkosť `cex=k`, $k \in \mathbb{R}$, prednastavená hodnota je 1
- typ bodov `pch=k`, číslo $k = 1, 2, \ldots, 20$, prednastavená hodnota je 1 (prázdny krúžok);
- typ čiar `lty=k`, číslo $k = 1, 2, \ldots, 20$, prednastavená hodnota je 1 (plná čiara).

Ďalšie funkcie súvisiace s bodovým grafom:

- `points(x,y)` – pridávanie bodov do obrázka;
- `lines(x,y)` – pridávanie čiar do obrázka;
- `text(x, y, labels)` – pridanie textu v bodoch špecifikovaných súradnicami `x` a `y`, kde popis `labels[i]` sa zobrazuje v bodoch (`x[i],y[i]`), prednastavené hodnoty sú `1:length(x)`;

  `plot(x,y, type="n"); text(x, y, names)`;
- `title(main,sub,xlab,ylab)` – dodatočné pridanie nadpisov a popisov osí;
- `legend(x,y,legend)` – dodatočné pridanie legendy, špecificky umiestnenej v súradniciach `x` a `y`, kde sú prednastavené nasledovné polohy `"bottomright"`, `"bottom"`, `"bottomleft"`, `"left"`, `"topleft"`, `"top"`, `"topright"`, `"right"` a `"center"`' argumenty funkcie
  - `fill="retazec"` – farba výplne (ne)orámovanej legendy,

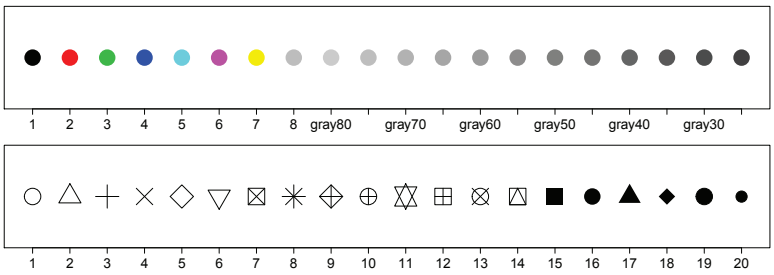
  - `col="retazec"` – farba nakreslených bodov alebo čiar,
  - `lty=k` – typ čiar, $k = 1, 2, \ldots, 20$,
  - `lwd=k` – šírka čiar, $k \in \mathbb{N}$, prednastavená hodnota je 1 (plná čiara),
  - `pch=k` – typ bodov, $k = 1, 2, \ldots, 20$, prednastavená hodnota je 1 (prázdny krúžok);

- `locator(n,type)` – určenie polohy konkrétneho bodu v grafe (napr. odľahlé pozorovanie) pomocou jedného bodového kliknutia myšou v jeho blízkosti, pričom funkcia bod nielen označí, ale aj vypočíta jeho súradnice `text(locator(1),"retazec")`; použitie v legende `legend(locator(1),...)`;
- `identify(x,y,labels)` identifikácia bodov, ak poznáme ich súradnice.

Príklad 126 (typy bodov a základné farby) *Nakreslite obrázok (a) základných dvadsiatich typov bodov a (b) ôsmich typov farieb a dvanástich odtieňov sivej.*

Riešenie v R (pozri obrázok 21)

```
windows(14,2.5)
par(mar=c(3,0.1,0.1,0.1))
plot(1:20,rep(0:1,10),type="n",sub="",xlab="",ylab="",bty="n",axes=FALSE)
points(1:20,rep(0.5,20),pch=1:20,cex=4)
axis(1,at=1:20,labels=1:20,cex.axis=1.5)
box()
windows(14,2.5)
par(mar=c(3,0.1,0.1,0.1))
plot(1:20,rep(0:1,10),type="n",sub="",xlab="",ylab="",bty="n",axes=FALSE)
points(1:8,rep(0.5,8),pch=16,col=1:8,cex=4)
siva <- paste("gray",rev(seq(25,80,by=5)),sep="")
points(9:20,rep(0.5,12),pch=16,col=siva,cex=4)
axis(1,at=1:20,labels=c(1:8,siva),cex.axis=1.5)
box()
```



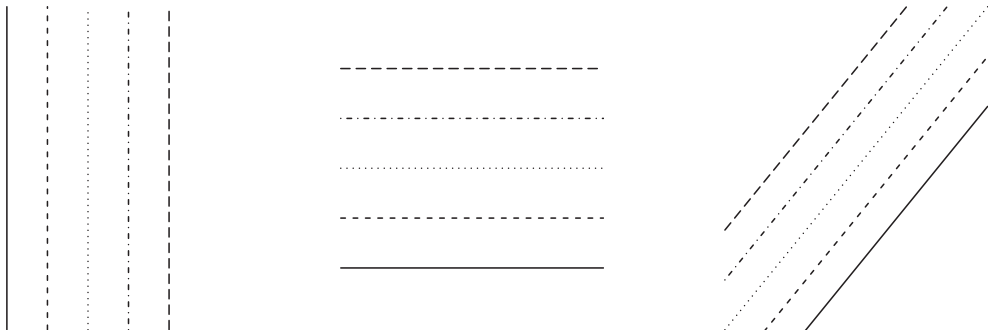
Obr. 21: Základné typy bodov (dolný riadok) a farieb (horný riadok)

Príklad 127 (typy bodov a základné farby) *Nakreslite obrázok zpiatich základných typov čiar, ktoré (a) smerujú zvislo, (b) smerujú vodorovne, (c) zvierajú s osou x uhol* 45°.

Riešenie v R (pozri obrázok 22)

```
windows(12,4)
par(mfcol=c(1,3))
plot(1,1,type="n",xlab="",ylab="",bty="n",axes=FALSE,xlim=c(-3,3),ylim=c(-3,3))
```

```
abline(v=-2:2,lty=1:5)
plot(1,1,type="n",xlab="",ylab="",bty="n",axes=FALSE,xlim=c(-3,3),ylim=c(-3,3))
abline(h=-2:2,lty=1:5)
plot(1,1,type="n",xlab="",ylab="",bty="n",axes=FALSE,xlim=c(-3,3),ylim=c(-3,3))
abline(a=-2,b=tan(pi/4),lty=1)
abline(a=-1,b=tan(pi/4),lty=2)
abline(a=0,b=tan(pi/4),lty=3)
abline(a=1,b=tan(pi/4),lty=4)
abline(a=2,b=tan(pi/4),lty=5)
```



Obr. 22: Základné typy čiar – zvislo, vodorovne a v uhle 45° (zľava doprava)

Príklad 128 (základy grafiky; dáta `iris`) *Nakreslite rozptylový graf dĺžky a šírky kališných lístkov pre všetky tri taxóny kosatcov pomocou (a) rôznych typov bodov a (b) rôznych farieb. Dokreslite do obrázku (c) regresné priamky pre každý taxón použitím rôzneho typu čiary. Pozri* ***`help(iris)`*** *ohľadom popisu premenných a ďalších detailov o dátach (Fisher, 1936/1971).*

Riešenie v R (pozri obrázok 23)

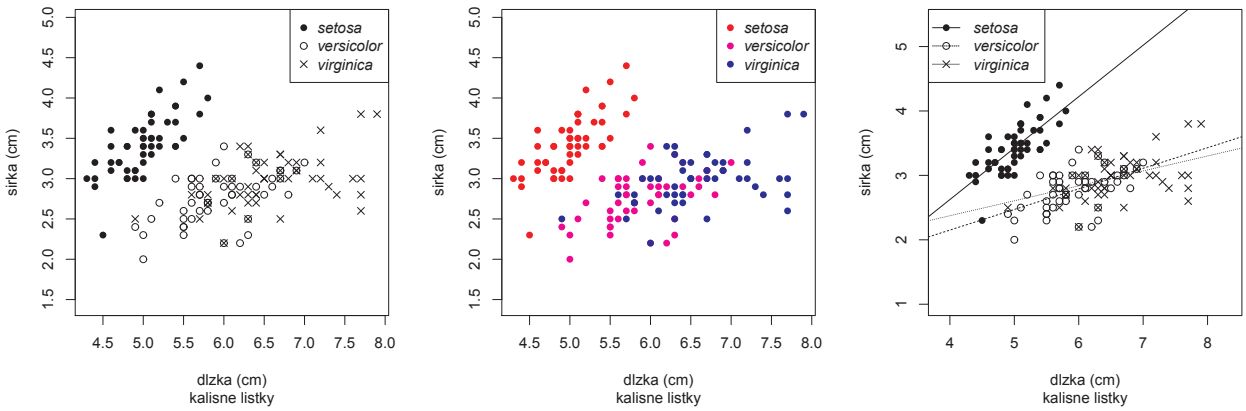
```
# zmena datoveho ramca na maticu
irisDATA <- as.matrix(iris[,1:4])
dimnames(irisDATA)[[2]]
# [1] "Sepal.Length" "Sepal.Width"
# [3] "Petal.Length" "Petal.Width"
irisLABELS <- iris[,5]
levels(irisLABELS)
# "setosa"     "versicolor" "virginica"
# rozptylovy graf [zle popisky osi]
plot(irisDATA[,"Sepal.Length"],irisDATA[,"Sepal.Width"])
# rozsahy oboch premennych
x.rozs <- range(irisDATA[,"Sepal.Length"])
y.rozs <- range(irisDATA[,"Sepal.Width"])
# typy bodov podla skupin
par(mfcol=c(1,3))
plot(irisDATA[,"Sepal.Length"],irisDATA[,"Sepal.Width"],type="n",
     xlab="",ylab="",asp=1)
points(irisDATA[irisLABELS=="setosa","Sepal.Length"],
       irisDATA[irisLABELS=="setosa","Sepal.Width"],pch=16)
points(irisDATA[irisLABELS=="versicolor","Sepal.Length"],
```

```
       irisDATA[irisLABELS=="versicolor","Sepal.Width"],pch=1)
points(irisDATA[irisLABELS=="virginica","Sepal.Length"],
       irisDATA[irisLABELS=="virginica","Sepal.Width"],pch=4)
title(xlab="dlzka␣(cm)",ylab="sirka␣(cm)",
       sub="kalisne␣listky")
legend("topright",c("setosa","versicolor","virginica"),text.font=3,pch=c(16,1,4))
# typy farieb podla skupin
plot(irisDATA[,"Sepal.Length"],irisDATA[,"Sepal.Width"],type="n",
       xlab="",ylab="",asp=1)
points(irisDATA[irisLABELS=="setosa","Sepal.Length"],
       irisDATA[irisLABELS=="setosa","Sepal.Width"],pch=16,col="red")
points(irisDATA[irisLABELS=="versicolor","Sepal.Length"],
       irisDATA[irisLABELS=="versicolor","Sepal.Width"],pch=16,
       col="magenta")
points(irisDATA[irisLABELS=="virginica","Sepal.Length"],
       irisDATA[irisLABELS=="virginica","Sepal.Width"],pch=16,col="blue")
title(xlab="dlzka␣(cm)",ylab="sirka␣(cm)",
       sub="kalisne␣listky")
legend("topright",c("setosa","versicolor","virginica"),text.font=3,
       pch=c(16,16,16),col=c("red","magenta","blue"))
# typy ciar  [presahuju mimo oblakov dat, co nie je statisticky spravne]
plot(irisDATA[,"Sepal.Length"],irisDATA[,"Sepal.Width"],type="n",
       xlab="",ylab="",asp=1,ylim=c(2.0-1,4.4+1))
points(irisDATA[irisLABELS=="setosa","Sepal.Length"],
       irisDATA[irisLABELS=="setosa","Sepal.Width"],pch=16)
points(irisDATA[irisLABELS=="versicolor","Sepal.Length"],
       irisDATA[irisLABELS=="versicolor","Sepal.Width"],pch=1)
points(irisDATA[irisLABELS=="virginica","Sepal.Length"],
       irisDATA[irisLABELS=="virginica","Sepal.Width"],pch=4)
# linearne regresne modely pre vsetky tri taxony zvlast
LM1 <- lm(irisDATA[irisLABELS=="setosa","Sepal.Width"] ~
          irisDATA[irisLABELS=="setosa","Sepal.Length"])
LM2 <- lm(irisDATA[irisLABELS=="versicolor","Sepal.Width"] ~
          irisDATA[irisLABELS=="versicolor","Sepal.Length"])
LM3 <- lm(irisDATA[irisLABELS=="virginica","Sepal.Width"] ~
          irisDATA[irisLABELS=="virginica","Sepal.Length"])
# ciary a ich typy pre linearny regresny model
abline(LM1,lty=1,lwd=2)
abline(LM2,lty=2,lwd=2)
abline(LM3,lty=3,lwd=2)
title(xlab="dlzka␣(cm)",ylab="sirka␣(cm)", sub="kalisne␣listky")
legend("topleft",c("setosa","versicolor","virginica"),text.font=3,pch=c(16,1,4),
       lty=c(1,2,3))
```

3.6.4 Kruhový diagram

Kruhový (**výsekový, koláčový**) **diagram** – zachytáva štruktúru dát takým spôsobom, že celá plocha kruhu predstavuje celý súbor a kruhové výseky jej jednotlivé časti, pričom polomery zvierajúce uhol 3.6° vymedzujú plochu odpovedajúcu 1 % celého obsahu. Kruhový diagram nakreslíme pomocou funkcie `pie(x)`. Argumenty funkcie `pie(x)`:

- `x` vektor relatívnych hodnôt (pravdepodobností), ktoré v súčte dávajú 1, teda i-ta položka bude `x[i]/sum(x)` kruhu (ale aj početností); graf začína horizontálnou čiarou doprava a pokračuje proti smeru hodinových ručičiek;
- `names` vektor mien prislúchajúcich jednotlivým položkám grafu;
- `col` vektor farieb, ktorými sú jednotlivé položky vyfarbené;
- `labels="retazec"` – pomenovania kruhových výsekov

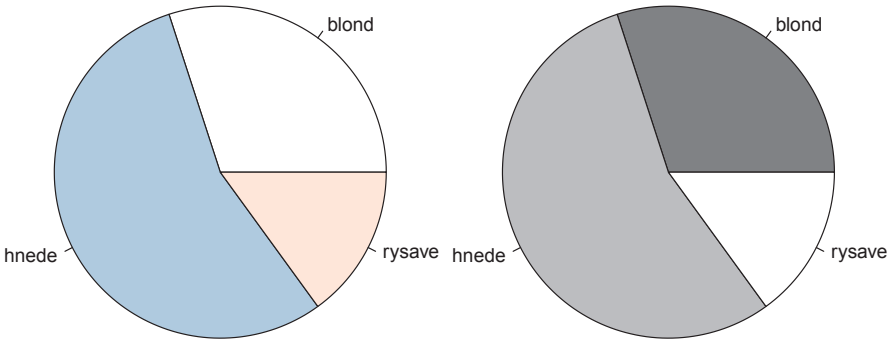


Obr. 23: Rozptylové grafy

Príklad 129 (kruhový diagram) *Vytvorte kruhový diagram znázorňujúci pravdepodobnosti výskytu rôznych farieb vlasov so súčasným výskytom modrých očí (dáta z príkladu oči vs. vlasy, tabuľka* `oci.prav1`*). Vytvorte rovnaký diagram v odtieňoch sivej farby (*`gray()`*). Použite tabuľku* `oci.prav1` *z príkladu 125.*

Riešenie v R (pozri obrázok 24)

```
windows(8,4)
par(mfcol=c(1,2),mar=c(0,0,0,0))
pie(oci.prav1[1,])
pie(oci.prav1[1,],col=gray(seq(0.4,1.0,length=3)))
```



Obr. 24: Kruhový diagram (dáta oči vs. vlasy)

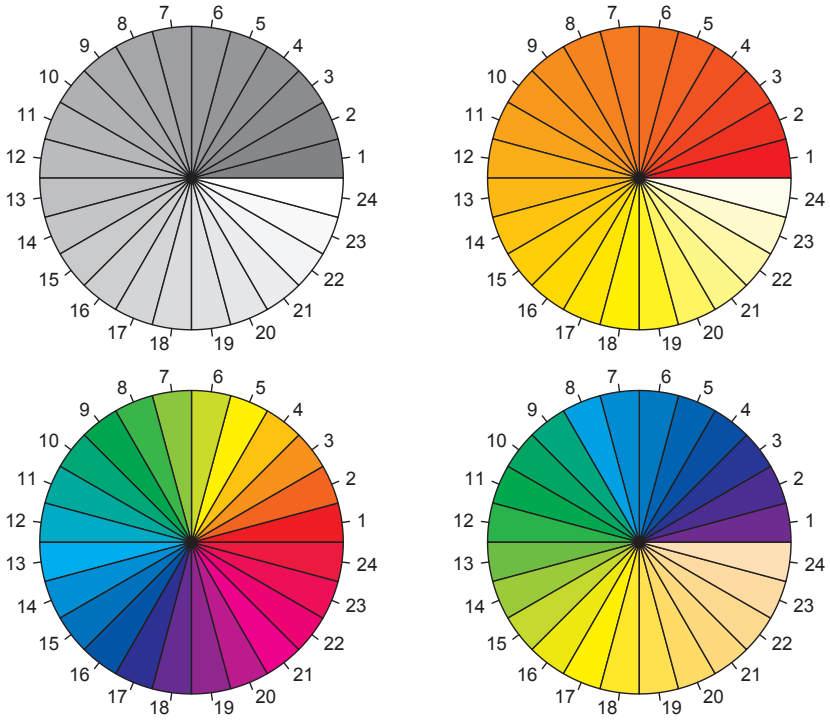
Príklad 130 (kruhový diagram) *Nakreslite tiež kruhový diagram 24*

- ***odtieňov sivej*** *(*`gray(sekvencia)`*, kde* `sekvencia` *sú čísla z intervalu* $\langle 0, 1\rangle$*),*
- ***odtieňov farieb dúhy*** *(*`rainbow(k)`*),*

- ***teplých farieb*** *(`heat.colors(k)`),*
- ***topografických farieb*** *(`topo.colors(k)` alebo `terrain.colors(k)`).*

Riešenie v (pozri obrázok 25)

```
windows(8,8)
par(mfcol=c(2,2),mar=c(0,0,0,0))
pie(rep(1,24),col=gray(seq(0.4,1.0,length=24)),radius=0.9)
pie(rep(1,24),col=rainbow(24),radius=0.9)
pie(rep(1,24),col=heat.colors(24),radius=0.9)
pie(rep(1,24),col=topo.colors(24),radius=0.9)
```



Obr. 25: Kruhový diagram (farebné škály)

3.6.5 Histogram

Histogram – predstavuje stĺpcový diagram s k stĺpcami, ktorých základňa sa rovná šírke intervalu $I_i = \langle x_i, x_{i+1}\rangle$ a výška i-teho stĺpca jeho početnosti ($i = 1, 2, \ldots, k$). Zobrazuje početnosti pozorovaní v jednotlivých intervaloch v *absolútnej škále* (na osi y sú zobrazené početnosti) a v *relatívnej škále* (obsah histogramu je rovný jednej). Histogram možno opísať pomocou *frekvenčnej tabuľky*, ktorá obsahuje početnosti a relatívne početnosti. Množstvo intervalov volí príslušný štatistický softvér alebo aj sám užívateľ. Potrebných je aspoň 12 triednych intervalov (ich počet nesmie klesnúť pod 6). Šírka jedného je minimálne $h_{\min} = 0.08(x_{\max} - x_{\min})$. Musí obsahovať minimálne 5 meraní. *Počet tried* histogramu je rovný $k = \log_2 n + 1 \doteq 2 + 3.3 \log_{10} n$ (**Sturgesova formula**; Becker a kol. (1988)), intervaly sú definované ako $\langle x_0, x_1\rangle, \langle x_1, x_2\rangle, \ldots, \langle x_{n-1}, x_n\rangle$. Šírka intervalov je teda $h = D/(\log_2 n + 1)$ *pre realizácie z normálneho rozdelenia.* Teraz už vlastne nepracujeme s realizáciami x_i, ale so stredmi intervalov $x_i^* = (x_i + x_{i+1})/2$. Počty hodnôt n_i, ktoré sa v intervale I_i nachádzajú, sa nazývajú *triedne početnosti.* Pokiaľ realizácie nemajú normálne rozdelenie, treba

použiť robustné algoritmy. Taktiež odľahlé pozorovania môžu dramaticky nafúknuť rozpätie, čo môže spôsobiť nárast šírky intervalov. Preto sa využívajú dva algoritmy ako kompromis medzi výchylkou (bias) a rozptylom realizácií pochádzajúcich z normálneho rozdelenia. Potom šírky triednych intervalov budú $h_1 = 3.49\widehat{\sigma}n^{-1/3}$, $\widehat{\sigma} = s$ (**Scottova formula**; Becker a kol. (1988)), $h_2 = 2D_Q n^{-1/3}$ (robustnejšia, **Freedman-Diaconisova formula**, ktorá je nezavislá od odľahlých pozorovaní a vyberá menšie intervaly ako Scottova formula; Venables a Ripley (2002)). Pre symetrické rozdelenia platí $h_3 = \lfloor 2\sqrt{n} \rfloor$ alebo $h_4 = \lfloor 2.46 \times (n-1)^{0.4} \rfloor$. Pokiaľ sa neočakáva príliš zošikmené rozdelenie, je šírka triednych intervalov h konštantná. V prípade komplikovanejších tvarov výberových rozdelení treba zväčšiť počet triednych intervalov alebo použiť špeciálne postupy na hľadanie nekonštantne dlhých triednych intervalov (Meloun a Militký, 2004). Histogram nakreslíme pomocou funkcie `hist(x)`. Argumenty funkcie `hist(x)`: angl

- `prob=FALSE` (prednastavená hodnota) – **v absolútnej škále**, kde na osi y sú početnosti;
- `prob=TRUE` – **v relatívnej škále** – suma obsahov stĺpcov (obdĺžnikov) je rovná jednej;
- `breaks="Sturges"` (**Sturgesova formula**, prednastavená hodnota), ďalšie možnosti sú `"Scott"` (**Scottova formula**) a `"FD"` alebo `"Freedman-Diaconis"` (**Freedman-Diaconisova formula**);
- `nclass=k`, $k \in \mathbb{N}$ – počet triednych intervalov;
- `plot=FALSE` – ak chceme vypísať číselné detaily, ale nechceme obrázok.

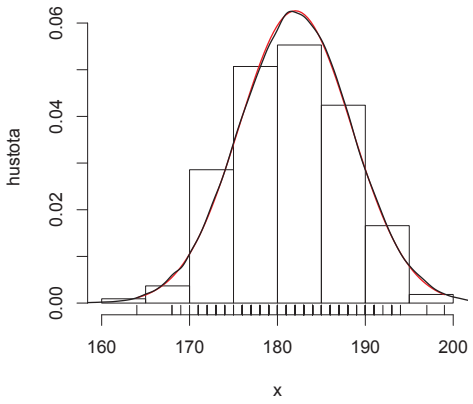
Príklad 131 (histogram a hustota normálneho rozdelenia) *Nakreslite histogram najväčšej dĺžky lebky (v mm)* ***skull.L*** *u mužov (dáta* ***one-sample-mean-skull-mf.txt****) a superponujte ho s (a) krivkou hustoty normálneho rozdelenia (červená farba) a (b) krivkou hustoty vypočítanej z dát (čierna farba). Pod bázu histogramu nakrelite tzv. „koberec“, ktorý charakterizuje polohu realizácií.*

Riešenie v R (pozri obrázok 26)

```
DATA <- read.table("one-sample-mean-skull-mf.txt",header=TRUE)
attach(DATA)
x <- na.omit(skull.L[sex=="m"]) # odstranenie chybajuceho pozorovania NA
priemer <- mean(x)
SD <- sd(x)
# 100 kvantilov od minima po maximum x a hodnoty hustoty v nich
kvant <- seq(min(x),max(x),length=100)
# doplnenie hustoty normalneho rozdelenia
hust <- dnorm(kvant,mean = priemer, sd=SD)
hist(x,ylim=c(0,range(hust)[2]),prob=TRUE,main="",xlab="x",ylab="hustota")
lines(kvant,hust,col="red",lwd=2)
# alternativne cez MC simulaciu
x.seq <- rnorm(100000,mean=mean(x),sd=sd(x))
lines(density(x.seq),lwd=2)
# koberec
rug(x)
```

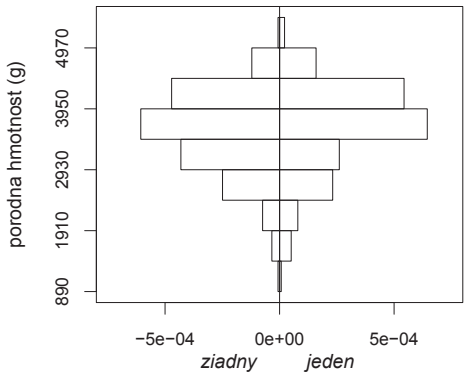
Príklad 132 (dva histogramy) *Nakreslite dva histogramy tak, aby sa svojimi bázami dotýkali. Aplikujte na dáta* ***two-samples-means-birth.txt****.*

Riešenie v R (pozri obrázok 27)



Obr. 26: Histogram so superponovanou krivkou hustoty normálneho rozdelenia (červená farba) a hustoty vypočítanej z dát (čierna farba); pod histogramom je tzv. „koberec“

```
DATA <- read.table("two-samples-means-birth.txt",header=TRUE)
names(DATA)
# [1] "o.sib.N" "birth.W"
attach(DATA) # zmena typu objektu na faktor
o.sib.N.faktor <- as.factor(o.sib.N) # hladiny faktora
# oznacenie hladin a ich zmena
levels(o.sib.N.faktor)
o.sib.N.faktor1 <- factor(o.sib.N.faktor,labels=c("ziadny","jeden"))
rozsah <- range(birth.W)
library(Hmisc) # nacitanie kniznice
xx <- histbackback(split(birth.W,o.sib.N.faktor1),
                   probability=TRUE,xlab="",ylab="",axes=FALSE)
title(ylab="porodna␣hmotnost␣(g)")
axis(1,cex.axis=0.9)
axis(2,at=seq(0,8,length=5),labels=seq(rozsah[1],rozsah[2],length=5),
     cex.axis=0.9,las=1)
mtext("ziadny",side=1,line=2,at=-mean(xx$left),font=3)
mtext("jeden",side=1,line=2,at=mean(xx$right),font=3)
```

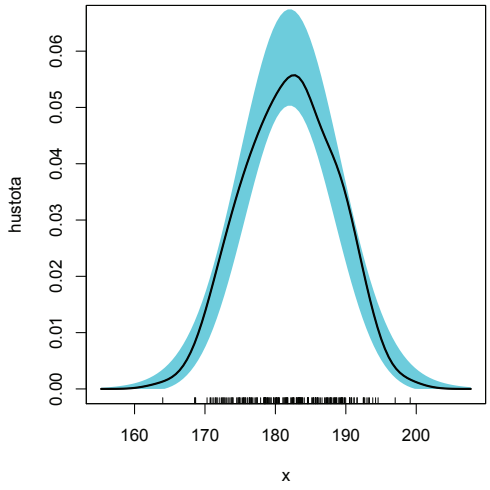


Obr. 27: Dva histogramy s priloženými bázami (základňami)

Príklad 133 (MC simulácia hustoty normálneho rozdelenia) *Superponujte hustotu najväčšej dĺžky lebky (v mm)* `skull.L` *(dáta* `one-sample-mean-skull-mf.txt`*) s hustotou normálneho rozdelenia vypočítanou pomocou MC simulácie s 95% pásom spoľahlivosti normálneho rozdelenia so strednou hodnotou rovnou aritmetickému priemeru a rozptylom rovným výberovému rozptylu (Bowman a Azzalini, 1997).*

Riešenie v R (pozri obrázok 28)

```
DATA <- read.table("one-sample-mean-skull-mf.txt",header=TRUE)
attach(DATA)
x <- na.omit(skull.L[sex=="m"])
library(sm)
yy <- sm.density(x, model="Normal")
windows(5,5)
par(mar=c(4.5,4.5,1,1))
sm.density(x,model="Normal",ylim=c(0,0.0655),ylab="hustota",cex.lab=1.2)
lines(yy$eval.points,yy$estimate,lwd=2)
```



Obr. 28: Hustota so superponovanou hustotou normálneho rozdelenia v podobe 95% pásom spoľahlivosti

3.6.6 Empirická distribučná funkcia

Histogram kumulatívnych početností (**súčtový histogram**) – namiesto početností budeme nad jednotlivými intervalmi I_i zakreslovať obdĺžniky s výškou rovnajúcou sa príslušným kumulatívnym početnostiam $N_i = \sum_{j=1}^{i} n_j$. Kumulatívne relatívne početnosti definujeme ako N_i/n, čomu zodpovedá **empirická distribučná funkcia**, definovaná pre zvolené číslo x ako relatívna početnosť v intervale $(-\infty, x\rangle$, teda ako N_i-tina hodnôt x_i menších alebo rovných ako x, t.j. (Wasserman, 2006)

$$\widehat{F}_n(x) = \frac{\#x_i < x}{n} = \sum_{i=1}^{n} I(x_i < x)/n,$$

kde $I(\cdot)$ je indikačná funkcia. Obrázok empirickej distribučnej funkcie nakreslíme pomocou funkcie `plot(ecdf(x),verticals = TRUE,do.points=FALSE)`. Argumenty funkcie `plot(ecdf(x))`:

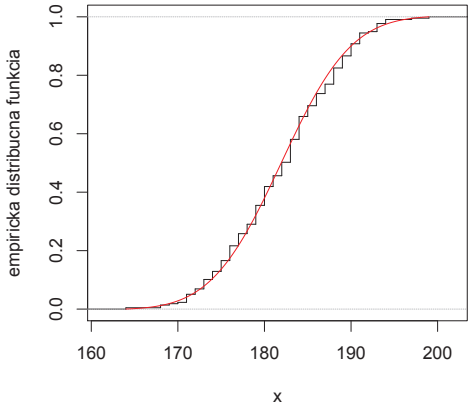
- `verticals = FALSE` je prednastavená hodnota; ak `verticals = TRUE`, je nakreslená schodovitá funkcia;

- `do.points = TRUE` je prednastavená hodnota; vkreslí do obrázka aj body, v ktorých je funkcia počítaná;
- `col.01line="gray70"` (prednastavená hodnota) – farba horizontálnych priamok v bodoch 0 a 1.

Príklad 134 (distribučná funkcia a hustota normálneho rozdelenia) *Nakreslite empirickú distribučnú funkciu najväčšej dĺžky lebky (v mm)* ***skull.L*** *u mužov (dáta* ***one-sample-mean-skull-mf.txt****) a superponujte ju s krivkou distribučnej funkcie normálneho rozdelenia červenou farbou.*

Riešenie v R (pozri obrázok 29)

```
DATA <- read.table("one-sample-mean-skull-mf.txt",header=TRUE)
attach(DATA)
x <- na.omit(skull.L[sex=="m"])
priemer <- mean(x)
SD <- sd(x)
ROZP <- range(x)
# zoradenych 200 hodnot (ekvidistantne) medzi min a max x
x1 <- seq(ROZP[1],ROZP[2],length=200)
# teoreticka CDF (normalne rozdelenie)
y <- dnorm(x1,mean=priemer,sd=SD)
y <- cumsum(y)/sum(y)
# zobrazenie krivky distribucnej funkcie
# teoretickeho normalneho rozdelenia
plot(ecdf(x),verticals=TRUE,do.points=FALSE,xlab="x",
     ylab="empiricka␣distribucna␣funkcia",main="")
lines(x1,y,col="red",lwd=2)
```



Obr. 29: Emirická distribučná funkcia superponovaná krivkou distribučnej funkcie normálneho rozdelenia (červená farba)

3.6.7 Krabicový diagram

Krabicový diagram – predstavuje grafické znázornenie päťčíselného súhrnu, t.j. zobrazuje v poradí zdola nahor hodnoty $x_{\min}, \widetilde{x}_{0.25}, \widetilde{x}_{0.50}, \widetilde{x}_{0.75}, x_{\max}$. Umožňuje tiež doplnenie hodnoty aritmetického

priemeru a tak zvýrazniť prípadné odchýlky od normality, identifikovať symetriu rozdelenia medzi kvartilmi, symetriu rozdelenia v koncoch rozdelenia, či odhaliť odľahlé pozorovania. Ak $\widetilde{x}_{0.50} = \overline{x}$ ide o **symetrické rozdelenie**, ak $\widetilde{x}_{0.50} < \overline{x}$, ide o **pravostranne zošikmené rozdelenie**, ak $\widetilde{x}_{0.50} > \overline{x}$, ide o **ľavostranne zošikmené rozdelenie**. Často sa používa na grafické porovnanie dvoch a viacerých skupín. *Šírka krabičiek* je proporčná k odmocnine z rozsahu výberového súboru $\sqrt{n}$. Ak hovoríme o **krabicových diagramoch so zárezom**, ide o zárez charakterizujúci 95% *empirický interval spoľahlivosti* (pozri kap. Testovanie hypotéz) mediánu $\widetilde{\mu}$, ozn. (d, h), kde

$$d = \widetilde{x} - 1.57\frac{D_Q}{\sqrt{n}}, h = \widetilde{x} + 1.57\frac{D_Q}{\sqrt{n}}.$$

Odhadom rozptylu mediánu je $\widehat{\sigma}^2_{\widetilde{x}} = D_Q/1.349$, kde vo všeobecnosti platí (pre akékoľvek rozdelenie pravdepodobnosti), že

$$\sigma^2_{\widetilde{x}} = \frac{1}{4nf^2\left(\widetilde{x}\right)},$$

kde f je hustota rozdelenia pravdepodobnosti (Casella a Berger, 2002). Pre normálne rozdelenie bude platiť $\sigma^2_{\widetilde{x}} = \sigma^2_x \frac{\pi}{2n}$, kde $\widetilde{X} \sim N\left(\widetilde{\mu}, \sigma^2_{\widetilde{x}}\right)$.

Obrázok krabicového diagramu nakreslíme pomocou funkcie `boxplot(x)`.

Argumenty funkcie `boxplot(x)`:

- `varwidth` je argument relatívnej šírky jednotlivých krabičiek; prednastavená hodnota je `FALSE` s rovnakou šírkou všetkých krabičiek; ak ju zmeníme na `TRUE`, šírka krabičiek bude odpovedať druhej odmocnine z počtu pozorovaní;
- `notch=TRUE` znamená zobrazenie zárezov krabičiek, ktoré odpovedajú 95% intervalom spoľahlivosti pre medián (prednastavená hodnota je `FALSE`);
- `col` určuje farbu vnútri krabičiek;
- `border` určuje farbu hraníc krabičiek;
- `names` je vektor pomenovaní pre jednotlivé zobrazované skupiny, ak ho vynecháme, použijú sa názvy z atribútu `names` z dátového rámca;
- `pch` typ bodu na zobrazenie odľahlých pozorovaní; prednastavená hodnota je 1;
- `horizontal=FALSE` je prednastavená hodnota;
- `plot=FALSE` – ak chceme iba vypísať číselné charakteristiky (prednastavená hodnota je `TRUE`).

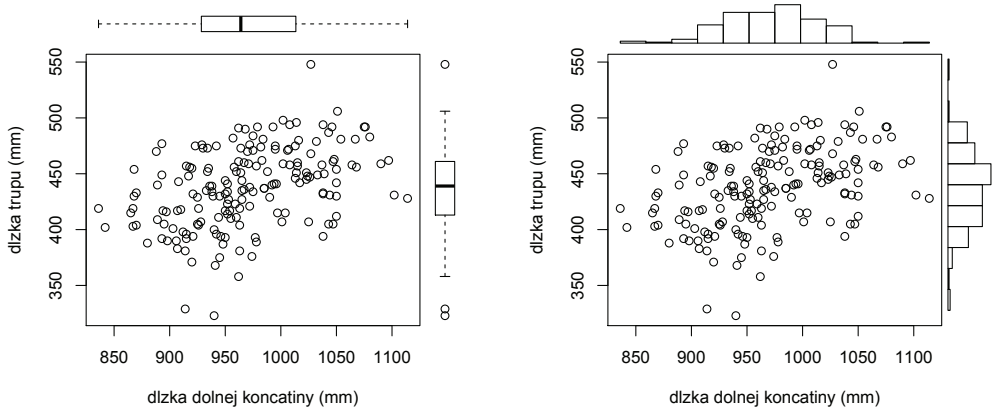
Najčastejšie používanou kombináciou argumentov je
`boxplot(x,varwidth=TRUE,notch=TRUE,outpch=16)`.
Hodnoty aritmetického priemeru sa do krabicových diagramov dokresľujú pomocou príkazu
`points(priemer, pch=16, col="red")`.

Príklad 135 (bodový graf a marginálne grafy) *Naprogramujte bodový graf dvoch premenných (1) s krabicovými diagramami a (2) histogramami pre marginálne dáta. Aplikujte na dáta* `two-samples-correlations-trunk.txt`.

Riešenie v R (pozri obrázok 30)

```
DATA <- read.table("two-samples-correlations-trunk.txt",header=TRUE)
names(DATA) # [1] "sex" "lowex.L" "tru.L"
attach(DATA)
par(fig=c(0,0.9,0,0.9))
plot(lowex.L, tru.L, xlab="dlzka␣dolnej␣koncatiny␣(mm)",
     ylab="dlzka␣trupu␣(mm)")
par(fig=c(0,0.9,0.55,1),new=TRUE)
boxplot(lowex.L,horizontal=TRUE,axes=FALSE)
par(fig=c(0.65,1,0,0.9),new=TRUE)
boxplot(tru.L,axes=FALSE)

H <- hist(tru.L,plot=FALSE)
k1 <- 0
k2 <- 5
par(fig=c(0,0.9,0,0.9))
plot(lowex.L,tru.L,xlab="dlzka␣dolnej␣koncatiny␣(mm)",
     ylab="dlzka␣trupu␣(mm)")
par(fig=c(0,0.9,0.55,1), new=TRUE)
plot(NULL,type="n",ylim=c(0,max(H$counts)+k1),xlim=c(range(H$breaks)),
     xlab="",ylab="",main="",bty="n",axes=FALSE)
rect(H$breaks[1:(length(H$breaks)-1)]+k1,0+k1,
     H$breaks[2:length(H$breaks)]+k1,H$counts+k1)
par(fig=c(0.65,1,0,0.9),new=TRUE)
plot(NULL,type="n",xlim=c(0,max(H$counts)+k2),ylim=c(range(H$breaks)),
     xlab="",ylab="",main="",bty="n",axes=FALSE)
rect(0+k2,H$breaks[1:(length(H$breaks)-1)]+k2,H$counts+k2,
     H$breaks[2:length(H$breaks)]+k2)
```



Obr. 30: Bodový graf s marginálnymi krabicovými diagramami (vľavo) a s histogramami (vpravo)

3.6.8 Kvantilový diagram

Kvantilový diagram (*qq*-diagram) – zobrazuje body so súradnicami $\left[\Phi^{-1}\left(i/\left(n+1\right)\right), x_{(i)}\right]$, kde $\Phi^{-1}\left(p\right)$ je kvantilová funkcia normovaného normálneho rozdelenia definovaná nasledovne

$$\Pr\left(Z \leq \Phi^{-1}\left(p\right)\right) = p.$$

Šikmosť $b_1 > 0$ sa prejaví v podobe **konvexného** usporiadania hodnôt, $b_1 < 0$ sa prejaví v podobe **konkávneho** usporiadania hodnôt. Taktiež zviditeľňuje **dĺžku „chvostov“** (ľavého a pravého konca krivky) rozdelenia, pričom **esovitým usporiadaním** bodov sa prejavujú krátke „chvosty“ a

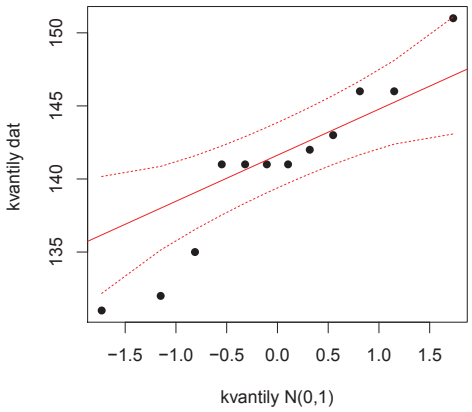
inverzne esovitým usporiadaním bodov dlhé „chvosty“. Zreteľná je tiež prípadná bimodalita rozdelenia. Štatisticky možno testovať normalitu rozdelenia pomocou simulácií z $N(0,1)$ a vytvorením 95% **Atkinsonovej obálky spoľahlivosti**. Pre normálne rozdelenie bude platiť $\sigma^2_{\widetilde{x}_p} = \sigma^2_x \frac{\pi^2}{24 \ln n}$, kde $\widetilde{X}_p \sim N\left(\widetilde{\mu}_p, \sigma^2_{\widetilde{x}_p}\right)$.

Kvantilový diagram vytvoríme pomocou funkcie `qqnorm(x)` a `qqline(x)`. Druhá funkcia dokreslí do grafu priamku prechádzajúcu bodmi charakterizujúcimi prvý a tretí kvartil teoretických a empirických kvantilov.

Príklad 136 (Atkinsonova obálka *qq*-diagramu) *Nakreslite* 95% *Atkinsonovu obálku spoľahlivosti (Atkinson, 1981; Flack a Flores, 1989). Aplikujte na dáta výšky* 10*-ročných dievčat.*

Riešenie v R (pozri obrázok 31)

```
library(car)
qqPlot(x,distribution="norm",xlab="kvantily␣N(0,1)",ylab="kvantily␣dat␣",
       main="",envelope=.95,col.lines="red",lwd=2,pch=16,cex=1,grid=FALSE)
```



Obr. 31: *qq*-diagram normovanej spojitej premennej výška 10-ročných dievčat (mm) so superponovanou obálkou normálneho rozdelenia

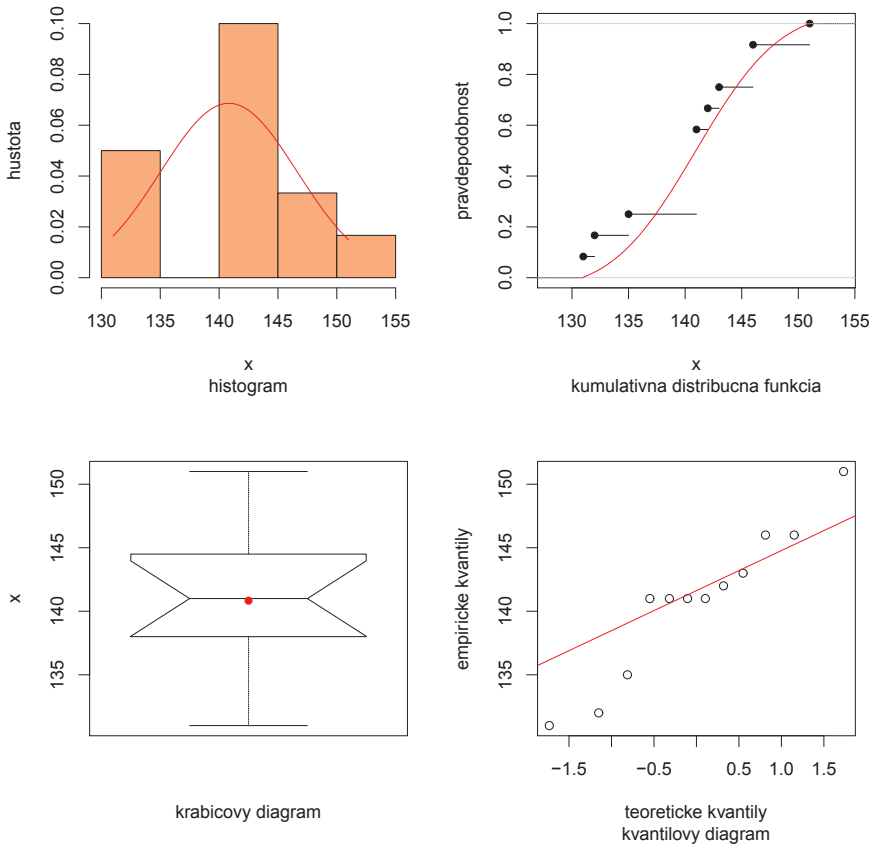
3.7 Príklady zo štatistickej grafiky

Príklad 137 (všetky základné grafy pre jeden výber ako funkcia) *Naprogramujte v* R *do jedného obrázka* 2×2 *štvoricu nasledovných grafov (1) histogram v relatívnej škále so superponovanou krivkou hustoty normálneho rozdelenia, (2) krabicový diagram so zakresleným priemerom, (3) empirickú (kumulatívnu) distribučnú funkciu superponovanú s teoretickou distribučnou funkciou normálneho rozdelenia a (4) qq-diagram so superponovanou qq-priamkou. Aplikujte na dáta výška* 10*-ročných dievčat.*

Riešenie v R (pozri obrázok 32)

```
"grafy.jeden.vyber" <- function(x){
par(mfcol=c(2,2),mar=c(5,5,3,1)) # nastavenie typu okna a jeho okrajov
priemer <- mean(x)
```

```
SD <- sd(x)
kvant <- seq(min(x),max(x),length=100)
hust <- dnorm(kvant,priemer,SD)
ROZP <- range(x)
y1 <- seq(ROZP[1],ROZP[2],length=200)
y <- dnorm(y1,mean=mean(y1),sd=sd(y1))
y <- cumsum(y)/sum(y)
# histogram a teoreticka hustota
hist(x,prob=TRUE,main="",xlab="x",ylab="hustota",col="lightsalmon")
title(sub="histogram")
lines(kvant,hust,col="red",lwd=2)
# krabicovy diagram
boxplot(x,varwidth=TRUE,notch=TRUE,outpch=16,xlab="krabicovy diagram",ylab="x")
points(priemer,pch=16,col="red")
# empiricka a teoreticka CDF
plot(ecdf(x),main="",xlab="x",ylab="pravdepodobnost")
lines(y1,y,col="red",lwd=2)
title(sub="kumulativna distribucna funkcia")
# kvantilovy diagram
qqnorm(x,main="",xlab="",ylab="")
qqline(x,col="red",lwd=2)
title(sub="kvantilovy diagram",xlab="teoreticke kvantily",
      ylab="empiricke kvantily")
}
# graf
grafy.jeden.vyber(x)
```



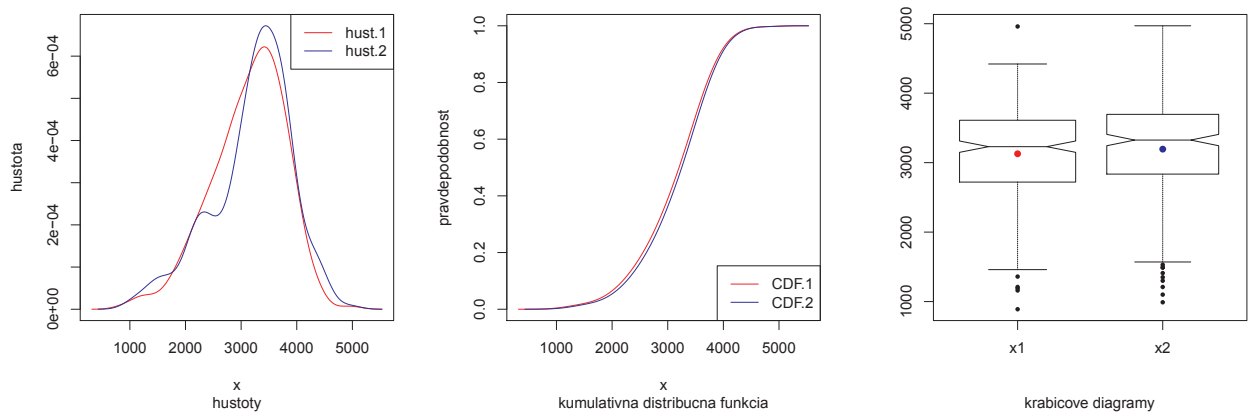
Obr. 32: Základná štvorica grafov pre spojitú premennú výška 10-ročných dievčat (mm)

Príklad 138 (všetky základné grafy pre dva výbery ako funkcia) *V R naprogramujte do jedného obrázka 1×3 trojicu nasledovných grafov (1) superponované krivky hustôt, (2) superponované krivky empirických (kumulatívnych) distribučných funkcií, (3) krabicové diagramy so zakresleným priemerom. Aplikujte na dáta* `two-samples-means-birth.txt`.

Riešenie v R (pozri obrázok 33)

```
"grafy.dva.vybery" <- function(x1,x2){
hust.1 <- density(x1)$y
x1.sekv <- density(x1)$x
CDF.1 <- cumsum(hust.1)/sum(hust.1)
hust.2 <- density(x2)$y
x2.sekv <- density(x2)$x
CDF.2 <- cumsum(hust.2)/sum(hust.2)
obe.sekv <- c(x1.sekv,x2.sekv)
obe.hust <- c(hust.1,hust.2)
obe.CDF <- c(CDF.1,CDF.2)
## dve hustoty
par(mfcol=c(1,3))
plot(obe.sekv,obe.hust,type="n",xlab="x",ylab="hustota",sub="hustoty")
lines(x1.sekv,hust.1,col="red",lwd=2)
lines(x2.sekv,hust.2,col="blue",lwd=2)
legend("topright",c("hust.1","hust.2"),lty=c(1,1),col=c("red","blue"),lwd=2)
## dve CDF
plot(obe.sekv,obe.CDF,type="n",xlab="x",ylab="pravdepodobnost",
    sub="kumulativna␣distribucna␣funkcia")
lines(x1.sekv,CDF.1,col="red",lwd=2)
lines(x2.sekv,CDF.1,col="blue",lwd=2)
legend("bottomright",c("CDF.1","CDF.2"),lty=c(1,1),col=c("red","blue"),lwd=2)
x <- c(x1,x2)
kod <- c(rep("x1",length(x1)),rep("x2",length(x2)))
## dva krabicove diagramy
boxplot(x ~ kod,varwidth=TRUE,notch=TRUE,outpch=16,sub="krabicove␣diagramy")
priem <- tapply(x,kod,mean)
points(priem,pch=16,col=c("red","blue"))
}
# grafy
DATA <- read.table("two-samples-means-birth.txt",header=TRUE)
names(DATA) # [1] "o.sib.N" "birth.W"
attach(DATA)
o.sib.N.faktor <- as.factor(o.sib.N) # zamena typu objektu na faktor
levels(o.sib.N.faktor) # kontrola hladin faktora
grafy.dva.vybery(birth.W[o.sib.N.faktor==0],birth.W[o.sib.N.faktor==1])
```

Príklad 139 (stĺpcový diagram) *(a) Vypočítajte podmienené pravdepodobnosti (pre každý riadok kontingenčnej tabuľky); t.j. za predpokladu súčinového multinomického rozdelenia. Premennú v riadkoch budeme označovat X (prediktor) a premennú v stĺpcoch ako Y (závisle premenná).*
(1) Y: farba dúhovky, X: radiálne útvary v štruktúre dúhovky (dáta: `multinom-iris-color.txt`*),*
(2) Y: zakončenie troch hlavných dlaňových línií, X: farba vlasov (dáta: `multinom-palmar-lines.txt`*),*
(3) Y: priľahlosť ušného laloka, X: pohlavie (dáta: `multinom-earlobe.txt`*),*
(4) Y: krvná skupina, X: mesto (dáta: `multinom-blood-groups.txt`*).*
(b) Nakreslite stĺpcové diagramy podmienených pravdepodobností pre (1) až (4).



Obr. 33: Základná trojica grafov pre spojitú premennú pre dáta `two-samples-means-birth.txt`

Príklad 140 (štatistická grafika) *Nakreslite (1) histogram v relatívnej škále so superponovanou krivkou hustoty normálneho rozdelenia, (2) krabicový diagram so zakresleným priemerom, (3) empirickú (kumulatívnu) distribučnú funkciu superponovanú s teoretickou distribučnou funkciou normálneho rozdelenia a (4) qq-diagram so superponovanou qq-priamkou pre nasledujúce premenné:*
(a) stranový rozdiel vertikálneho priemeru diafýzy kľúčnej kosti (`simd.R` a `simd.L`; v mm) na pravej a ľavej strane tela (dáta: `paired-means-clavicle2.txt`),
(b) najväčšia výška mozgovne (`skull.pH`; v mm; dáta: `one-sample-correlation-skull-mf.txt`) a
(c) morfologická výška tváre (`face.H`; v mm; dáta: `one-sample-correlation-skull-mf.txt`).

Príklad 141 (bodový graf a krabicové diagramy) *Nakreslite bodový graf spolu s krabicovými diagramami pre marginálne dáta pre premenné vertikálny priemer diafýzy kľúčnej kosti na pravej a ľavej strane tela (`simd.R` a `simd.L`; v mm; dáta: `paired-means-clavicle2.txt`).*

Príklad 142 (bodový graf a krabicové diagramy) *Nakreslite bodový graf spolu s krabicovými diagramami pre marginálne dáta pre*
(a) premenné najväčšia výška mozgovne (`skull.pH`; v mm) a morfologická výška tváre (`face.H`; v mm) a
(b) ich z-skóre (dáta: `one-sample-correlation-skull-mf.txt`).

Príklad 143 (krabicové diagramy) *Nakreslite krabicové diagramy pre obe pohlavia pre nasledovné premenné:*
(a) dĺžky lebky (`skull.H`; v mm; `sex`; dáta: `two-samples-means-skull.txt`),
(b) dĺžka dolnej končatiny (`lowex.L`; v mm; `sex`; dáta `two-samples-correlations-trunk.txt`),
(c) dĺžka trupu (`tru.L`; v mm; `sex`; dáta: `two-samples-correlations-trunk.txt`).

Príklad 144 (krabicové diagramy) *Nakreslite krabicové diagramy pre nasledovné premenné:*
(a) výška hornej časti tváre pre všetky populácie (`upface.H`*; v mm;* `pop`*; dáta:* `anova-means-skull.txt`*);*
(b) najväčsia dĺžka kľúčnej kosti pravej strany (`cla.L`*; v mm;* `population`*; dáta:* `more-samples-variances-clavicle.txt`*).*