3.3.3 Míry rozptýlenosti založené na empirických kvantilech

Empirický kvantil je hodnota, pod níž leží definovaná část údajů. U empirického kvantilu udáváme jeho hladinu q a označujeme ho symbolem x_q. Parametr q je z intervalu hodnot $0 < q < 1$. Hladina q určuje relativní podíl údajů, které se nacházejí pod empirickým kvantilem x_q.

Pro data můžeme vypočítat mnoho různých empirických kvantilů. Některé z nich se však používají pravidelně. Slouží k popisu jednotlivých částí rozdělení dat a vypočítávají se z nich také míry rozptýlenosti.

Výpočet empirického kvantilu s hladinou q se děje tímto způsobem: Nechť $j = [qn]$, kde operace [.] znamená zaokrouhlování na nejbližší menší celé číslo. Jestliže $qn = [qn]$, pak $x_q = (x_j + x_{j+1})/2$, jinak $x_q = x_{j+1}$, kde x_j ($j = 1, 2, \ldots, n$) jsou data výběru seřazená podle velikosti.

Hladiny q někdy uvádíme v procentech. V tomto případě nalezené hodnoty označujeme jako **percentily** nebo přesněji empirické percentily na dané úrovni. Je tedy 25% percentil rovný kvantilu o hladině 0,25.

Percentily s hladinou 25 %, 50 % a 75 % nazýváme kvartily a označujeme je takto:

- Q_I je *první* neboli *dolní kvartil* ($q = 25\,\%$);
- Q_{II} je *druhý* neboli *medián* ($q = 50\,\%$) – ten již známe z výkladu o mírách centrální tendence;
- Q_{III} je *třetí* neboli *horní kvartil* ($q = 75\,\%$).

Pro data o výkonech v skoku dalekém z tabulky 2.9 (s. 77) mají kvartily hodnotu $Q_I = 3{,}35$, $Q_{II} = Me = 3{,}55$, $Q_{III} = 3{,}75$.

Při popisu krajních hodnot rozdělení udáváme percentily s hladinami buď 2,5 % a 97,5 %, anebo 5 % a 95 %. Tyto extrémní percentily se často používají při určování referenčních hodnot laboratorních údajů v biomedicíně.

Interkvartilové rozpětí $Q = Q_{III} - Q_I$ je charakteristikou rozptýlenosti, jež se používá spolu s kvartily k popisu tvaru dat, když se z nějakého důvodu nechceme opřít o průměrové charakteristiky, jako je aritmetický průměr nebo směrodatná odchylka. Z definice vyplývá, že v intervalu (Q_I, Q_{III}) se nachází 50 % údajů. Interkvartilové rozpětí má intuitivnější obsah než směrodatná odchylka a není na rozdíl od směrodatné odchylky tak citlivé vůči odlehlým hodnotám.

Pro data o výkonech v skoku dalekém z tabulky 2.9 má kvartilové rozpětí hodnotu $Q = Q_{III} - Q_I = 3{,}74 - 3{,}34 = 0{,}40$.

Mediánová absolutní odchylka je mírou rozptýlenosti vycházející z dvojnásobného použití výpočtu mediánu. Jedná se o míru rozptýlenosti, která – podobně jako interkvartilové rozpětí – není citlivá k odlehlým hodnotám. Spočítá se jako

medián z absolutních hodnot odchylek jednotlivých měření od mediánu. Označuje se někdy zkráceně *MAD – median absolute deviation*. Zkráceně vyjádříme výpočet této míry vzorcem:

$$MAD = Me\{|x_i - Me|\}$$

U údajů {0; 1; 2; 5; 8; 9; 10} jsme zjistili, že medián je 5. Absolutní diference mají hodnoty {5; 4; 3; 0; 3; 4; 5}. Seřadíme je podle velikosti a zjistíme z uspořádané sekvence {0; 3; 3; 4; 4; 5; 5} medián. *MAD* má tedy hodnotu 4.

3.4 Míry špičatosti a šikmosti

Tyto charakteristiky se používají méně často, ale obvykle společně. Slouží k jemnějšímu popisu specifických stránek dat. Hodnotíme pomocí nich také to, jak se rozdělení dat podobá normální (Gaussově) křivce. K výpočtu těchto charakteristik se přistupuje různě. Nejčastěji se využívají tzv. centrální momenty třetího a čtvrtého stupně. Centrální moment k-tého stupně m_k je obecně definován vzorcem

$$m_k = \frac{\sum (x_i - \bar{x})^k}{n}.$$

Šikmost S_1 měří zešikmenost, resp. nesymetrii dat a vypočítá se pomocí druhého a třetího momentu podle vzorce

$$S_1 = \frac{m_3}{m_2^{3/2}}.$$

$S_1 = 0$ platí přibližně pro rozdělení přibližně symetrické, $S_1 > 0$ pro rozdělení s prodlouženým pravým koncem, naopak $S_1 < 0$ pro rozdělení s prodlouženým levým koncem (obr. 3.7).

Koeficient špičatosti S_2 měří odchylku špičatosti zkoumaného rozdělení od normálního rozdělení:

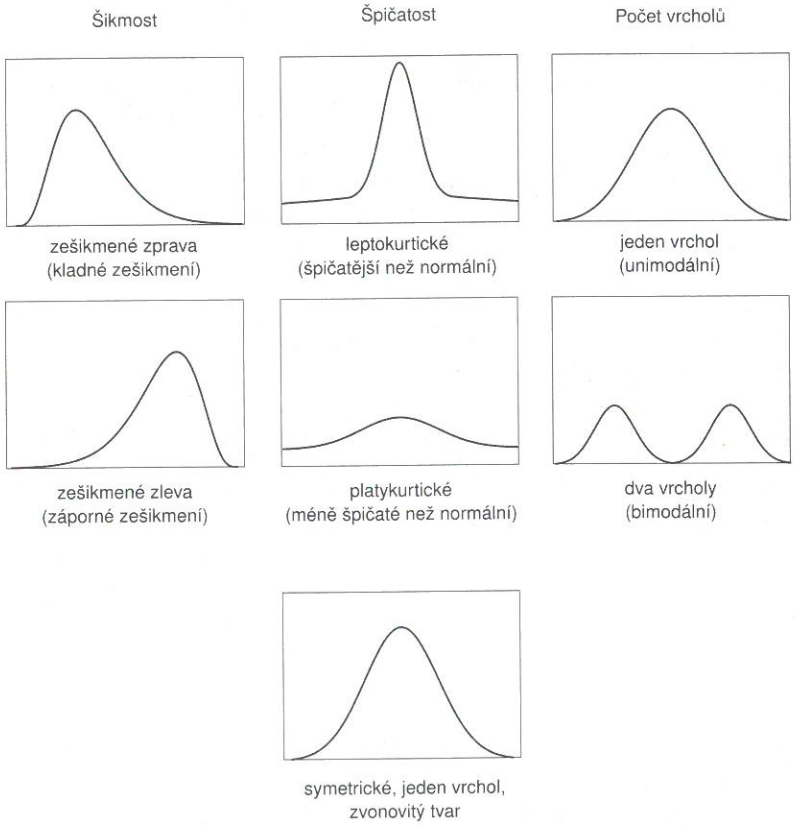
$$S_2 = \frac{m_4}{m_2^2} - 3$$

Takto vypočtená špičatost má pro normální rozdělení hodnotu 0. Symetrická rozdělení mohou mít stejný rozptyl, ale odlišnou špičatost. Plošší křivky ($S_2 > 0$) nazýváme platykurtické, špičatější křivky ($S_2 < 0$) leptokurtické.

Zešikmenost se také měří pomocí dalších koeficientů. U symetrických dat medián dělí na polovinu interkvartilové rozpětí. Tento poznatek je možné využít k definování koeficientu šikmosti KS pomocí kvartilů

$$KS = \frac{Q_{III} + Q_I - 2\tilde{x}}{Q},$$

Obr. 3.7 Tvary rozdělení



kde Q je interkvartilové rozpětí. Obecně platí, že KS nabývá hodnot od -1 do $+1$. Kladný, resp. záporný KS indikuje zleva, resp. zprava zešikmené rozdělení.

Statistik K. Pearson zavedl vlastní míru šikmosti SK, která zohledňuje skutečnost, že u zešikmených rozdělení se liší aritmetický průměr a medián:

$$SK = \frac{3\,(\bar{x} - Me)}{s}$$

3.5 Popis dat pomocí pěti hodnot a krabicový graf s anténami

Vhodným způsob k popisu jak centrální tendence dat, tak jejich rozptýlenosti je uvedení mediánu jako míry střední hodnoty, kvartilů a nejmenší a největší hodnoty (minima a maxima hodnot) pro popis rozptýlenosti. Pro hodnoty výkonů ve skoku dalekém z tabulky 2.9 uvádí tento souhrn tabulka 3.3. Těchto pět hodnot se využívá k sestrojení tzv. krabicového grafu s anténami (někdy se říká s tykadly nebo vousy). Krabicový graf je velmi oblíbeným prostředkem pro zobrazení dat. Je implementován ve všech solidnějších statistických programových systémech. Používá se pro znázornění jedné množiny dat, ale ještě častěji pro porovnávání několika skupin dat. Do jisté míry se podobá sloupkovému grafu s vyznačenými směrodatnými odchylkami pro porovnání rozptýlenosti měření. Také krabicový graf s anténami dovoluje posoudit a porovnat jak centrální tendence dat, tak jejich rozptýlenost. Navíc pomocí tohoto grafu posuzujeme i zešikmení a přítomnost odlehlých hodnot (outliers). Konstruuje se podle schématu na obrázku 3.8. Krabička obsahuje 50 % dat. Je rozdělena mediánem na dvě části. Její dolní hrana je určena dolním (prvním) kvartilem a horní hrana třetím kvartilem. Pokud je medián blízko jedné z horizontálních hran krabičky, rozdělení dat je zešikmené v opačném směru.

Zobrazíme data vzorové matice dat krabicovým grafem. Použijeme údaje pro skok daleký a porovnáme výkony chlapců a dívek. Graf neindikuje přítomnost odlehlých hodnot (obr. 3.9).

Tab. 3.3 Příklad popisu dat pomocí pěti hodnot

Minimum	Q_I	Medián	Q_{III}	Maximum
3,1	3,35	3,55	3,75	4,2

3.6 Zkoumání přítomnosti odlehlých hodnot a rezistentní odhady

Extrémně vysoké nebo nízké hodnoty přítomné v řadě měření mohou někdy vzbudit podezření, že jejich vznik není určen sledovanou náhodnou proměnnou, ale chybou zápisu nebo chybným měřením. Za určitých okolností se tato měření vyřazují ze zpracování. Jiná doporučená strategie spočívá v tom, že všechny