3 Grafický a číselný popis rozložení dat

Data analyzujeme s cílem porozumět nějakému problému. Porozumění vzniká z kombinace znalostí o kontextu, jak data vznikla, a schopnosti využít statistické grafy a numerické výpočty. Například údaje z lékařské studie znamenají velmi málo, pokud nevíme, s jakým cílem se studie provedla a jak měření různých biomedicínských parametrů přispívají k tomuto cíli. Na druhé straně měření mnoha stovek jedinců má malou hodnotu i pro lékařského experta, pokud údaje pomocí statistických prostředků neupravíme, nezobrazíme a netransformujeme do několika dobře zvolených numerických charakteristik – popisných statistik.

Každá množina dat obsahuje informaci o skupině objektů. Účelem analýzy dat je přehledně zpřístupnit data graficky, tabulkově a výpočtem různých statistických charakteristik tak, aby byly dobře patrné jejich statistické vlastnosti a umožnilo se také srovnání různých podskupin dat a kategorií, které jsou předem dány nebo je výzkumník vytváří v průběhu analýzy dat. Porovnáváme mezi sebou data pro muže a ženy, pro sportovce a nesportovce, data naměřená v různých časových okamžicích apod. Informace jsou organizovány pomocí proměnných. Při analýze zohledňujeme skutečnost, že proměnné mohou být různého typu (diskrétní, spojité) a v různém měřítku.

V této kapitole se zaměříme na jednorozměrný popis a analýzu proměnných (každou proměnnou hodnotíme zvlášť), posouzení vzájemných vztahů mezi proměnnými pomocí popisných metod budeme probírat především v kapitolách o regresní a korelační analýze a o analýze kategoriálních proměnných.

Při rozhodování, co budeme pomocí dat počítat nebo jak je budeme zobrazovat, mohou hrát roli čtyři aspekty účelu analýzy.

Explorace. Popis dat není statická záležitost – především v procesu explorace, kdy hledáme v datech zajímavé konfigurace a vztahy. Počítače a software nám umožňují prozkoumávat data při hledání předběžných hypotéz. Z této filozofie vychází i moderní oblast statistické analýzy „data mining“, dolování smysluplných závěrů z dat. Podrobněji se o technikách explorace zmíníme ve zvláštním oddíle, zde uvedeme jen dvě základní strategie:

- začínáme se zkoumáním jednotlivých proměnných, teprve pak analyzujeme jejich vztahy;

- začínáme nejdříve zobrazovat data pomocí grafů, pak přidáme numerické charakteristiky specifických aspektů dat.

Na exploraci se váže **kontrola dat**, jíž jsme zdůraznili už v předchozí kapitole. Grafické metody jsou pro diagnostiku chyb v údajích zvlášť vhodné. Nápadité zobrazení nám o datech může hodně prozradit – např. zda neobsahují špatně zapsané nebo změřené údaje. Zobrazení pomáhá odhalit přítomnost odlehlých hodnot, jež mohou zcela zkreslit výsledky další analýzy.

Odhadování. Z konceptu inferenční statistiky a statistického usuzování se odvozuje druhý účel popisné statistiky. Produkty popisné statistiky, zvláště různé numerické charakteristiky, tvoří základní číselné kameny pro odhadování populačních charakteristik. Další metody – modelování dat, přezkušování modelů a provádění zobecnění apod. – nám umožní odvodit, jak jsou odhady charakteristik přesné, a testovat hypotézy o populačních hodnotách.

Komunikace. Nejzřejmějším důvodem pro popis dat je komunikace. Je zapotřebí zobrazit data tak, aby se jejich důležité vlastnosti efektivně zprostředkovaly příjemci informací. Tomuto účelu slouží jak grafy nebo tabulky, tak numerické charakteristiky. Příklady takové přístupu najdeme v běžných médiích ve zprávách o nezaměstnanosti, rozdělení obliby politických stran, hospodářské situaci, při analýze sportovních výsledků apod.

Princip, který řídí popisnou analýzu dat, je následující:

1. Nejdříve se pokusíme zobrazit data graficky, případně tabulkou.
2. Hledáme základní konfigurace a tendence v datech, případně odchylky od nich.
3. Přidáváme numerické charakteristiky různých aspektů dat.
4. Často se nám podaří vystihnout stručným způsobem základní konfiguraci dat pomocí pravděpodobnostního modelu.

3.1 Způsoby zobrazení dat

Východiskem každé statistické analýzy jsou zachycená primární data nějakého pozorování nebo experimentu. Důležitými prostředky v předběžné, explorační analýze i při konečné prezentaci dat jsou grafické metody a znázornění dat pomocí tabulek. Rozhodnutí, zda zobrazit údaje pomocí obrázku nebo tabulkou, je do jisté míry věcí citu. Grafické metody jsou celkově vhodné pro ukázání širších kvalitativních vlastností dat. Tabelační metody jsou určitě vhodnější, jestliže vybrané údaje chceme uvést v přesném tvaru nebo když se očekává, že tyto údaje budou zapotřebí k dalším výpočtům. O použití grafů a tabulek se také zmíníme při výkladu o prezentaci výsledků (kap. 15).

3.1.1 Metody zobrazení kvalitativních a ordinálních dat

Nominální a ordinální data se zobrazují mnoha způsoby v závislosti na počtu a typu kategorií uvažovaného znaku. Při malém počtu pozorování je možné některé kategorie znaku sloučit. Jako zobrazovací prostředky se používají tabulky s procenty, koláčové a sloupcové grafy. Uvádíme zobrazení (tab. 3.1 a obr. 3.1) pro ordinální proměnnou *prospěch z matematiky* pro data z tabulky 2.9 (s. 77).

3.1.2 Metody zobrazení kvantitativních dat

Stručně uvedeme jednorozměrné popisné grafické a tabelační metody pro soubor kvantitativních měření jedné proměnné. V tomto případě si statistický soubor dat můžeme představit jako n-tici reálných čísel, v níž se jednotlivé prvky mohou opakovat, přičemž pořadí, jak byly prvky získány, nepřikládáme žádný význam. Například {2; 8; 9; 10; 1; 0; 5} je statistický soubor o 7 prvcích ($n = 7$). Obecně takovou n-tici zachycujeme symbolem $x_1, x_2, \ldots, x_n$. Pro náš příklad je $x_1 = 2$, $x_2 = 8, \ldots, x_n = 5$.

Tabulka četností, relativních četností a kumulativních četností je základní numerické zobrazení, při kterém se v souboru přítomné hodnoty kvantitativní proměnné setřídí a pro každou hodnotu se zjistí její absolutní i relativní četnost, dále absolutní a relativní kumulativní četnost. Četnosti se mohou zobrazit graficky. Údajům o výkonech ve skoku do dálky z tabulky 2.9 odpovídá tabulka 3.2. Aby nebyla tabulka rozsáhlá, volí se vhodně délka intervalů k vytvoření tříd, do nichž se setřídí příslušné hodnoty. Tabulka má tolik řádek, kolik tříd se vytvořilo. Čím je interval delší, tím má tabulka méně řádků.

Tabulky slouží pro první přehled získaných měření. Tohoto cíle se snad ještě lépe dosáhne použitím grafických prostředků. Grafické zobrazení vytváří geometrický obraz dat. Přitom se využívají body, plochy, úsečky nebo různé další obrazce. Nejznámější způsob zobrazení hodnot jedné proměnné se nazývá **histogram**. V tomto případě osa X odpovídá hodnotám proměnné a osa Y absolutním nebo relativním četnostem. Pro dobré zobrazení je důležité zvolit optimálně počet tříd, které pokryjí celé rozmezí hodnot. Čím je dat méně, tím by měl být také menší počet tříd. Pro malé rozsahy výběrů se nevyplatí histogram sestavovat.

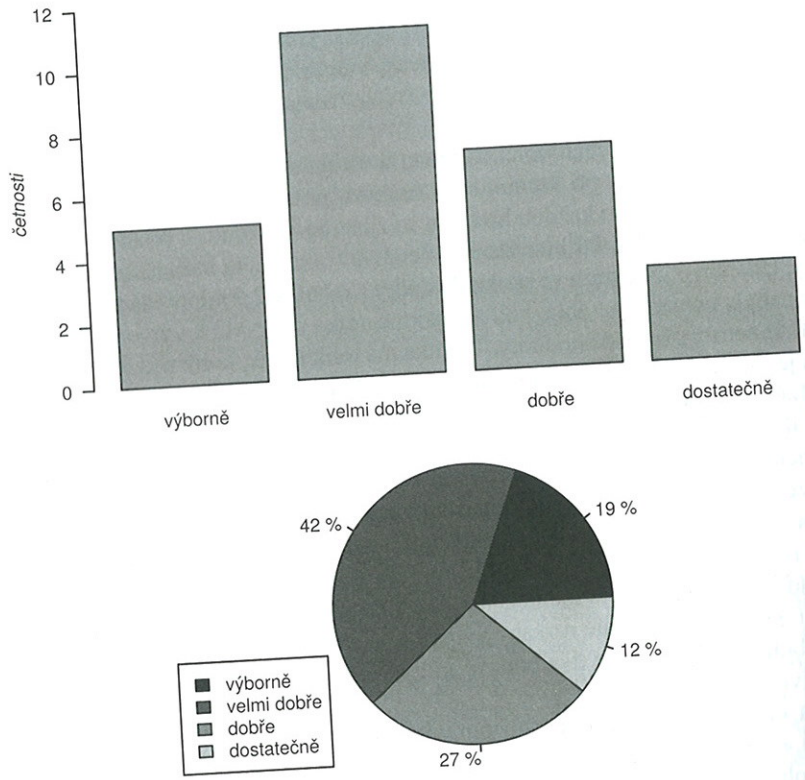
Obrázek 3.2 obsahuje dva bodové grafy výsledků ve skoku dalekém, zvlášť pro dívky a chlapce.

Na obrázku 3.3 je příklad histogramu pro data v tabulce 2.9. Další graf (obr. 3.4) ukazuje odpovídající kumulativní relativní četnosti. Na obrázku 3.4 jsou zobrazeny kumulativní četnosti z tabulky 2.9 (s. 77).

Tab. 3.1 Absolutní a relativní četnosti hodnot znaku *Prospěch z matematiky*

	výborně	velmi dobře	dobře	dostatečně	SUMA
n_i	5	11	7	3	26
$f_i = 100 \times n_i / n$	19,23	42,31	26,92	11,54	100

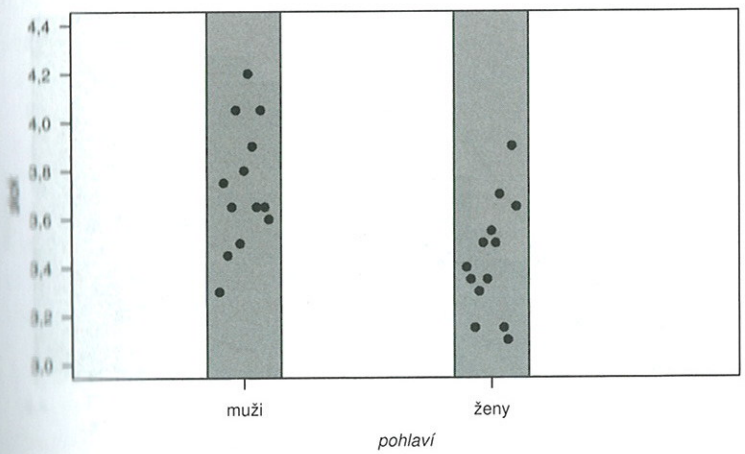
Obr. 3.1 Příklady zobrazení znaku *Prospěch z matematiky* (sloupcový a koláčový graf)



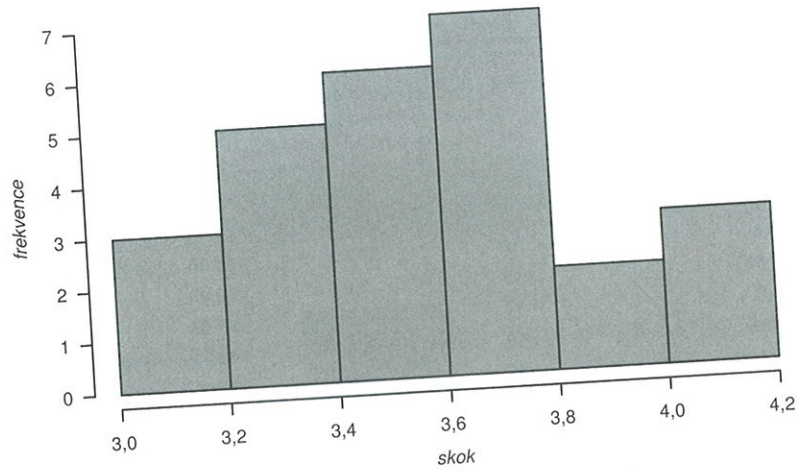
Tab. 3.2 Příklad základní úpravy primárních dat – tabulka četností a kumulativních četností pro data z tabulky 2.9

Skok daleký [m]	Počet	Kumulativní počet	Procenta	Kumulativní procenta	Graf procent
3,1	1	1	3,85	3,85	|
3,15	2	3	7,69	11,54	|||
3,3	2	5	7,69	19,23	|||
3,35	2	7	7,69	26,92	|||
3,4	1	8	3,85	30,77	|
3,45	1	9	3,85	34,62	|
3,5	3	12	11,54	46,15	||||
3,55	1	13	3,85	50,00	|
3,6	1	14	3,85	53,85	|
3,65	4	18	15,38	69,23	|||||
3,7	1	19	3,85	73,08	|
3,75	1	20	3,85	76,92	|
3,8	1	21	3,85	80,77	|
3,9	2	23	7,69	88,46	|||
4,05	2	25	7,69	96,15	|||
4,2	1	26	3,85	100,00	|

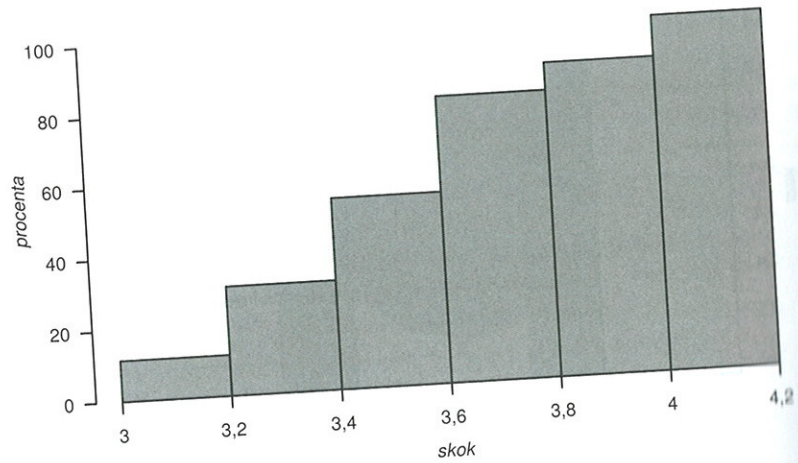
Obr. 3.2 Bodový graf



Obr. 3.3 Histogram četností pro data o skoku dalekém z tabulky 3.2



Obr. 3.4 Kumulativní četnosti pro data o skoku dalekém z tabulky 3.2

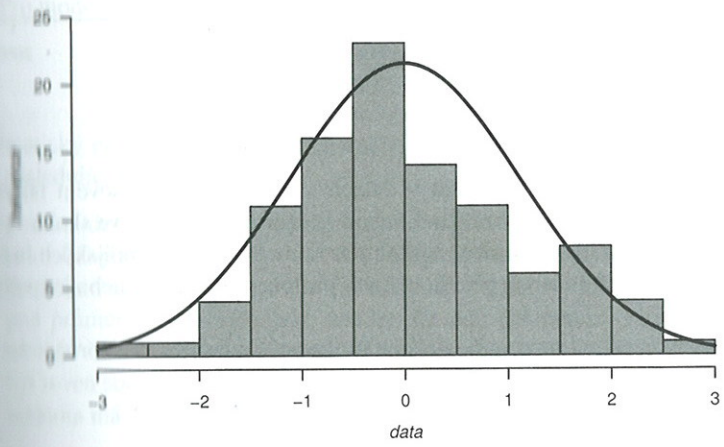


Při popisování a analýze toho, co graf zobrazuje, si všímáme nejdříve základní tvarové konfigurace a pak deviací od tohoto tvaru. Hodnotíme:

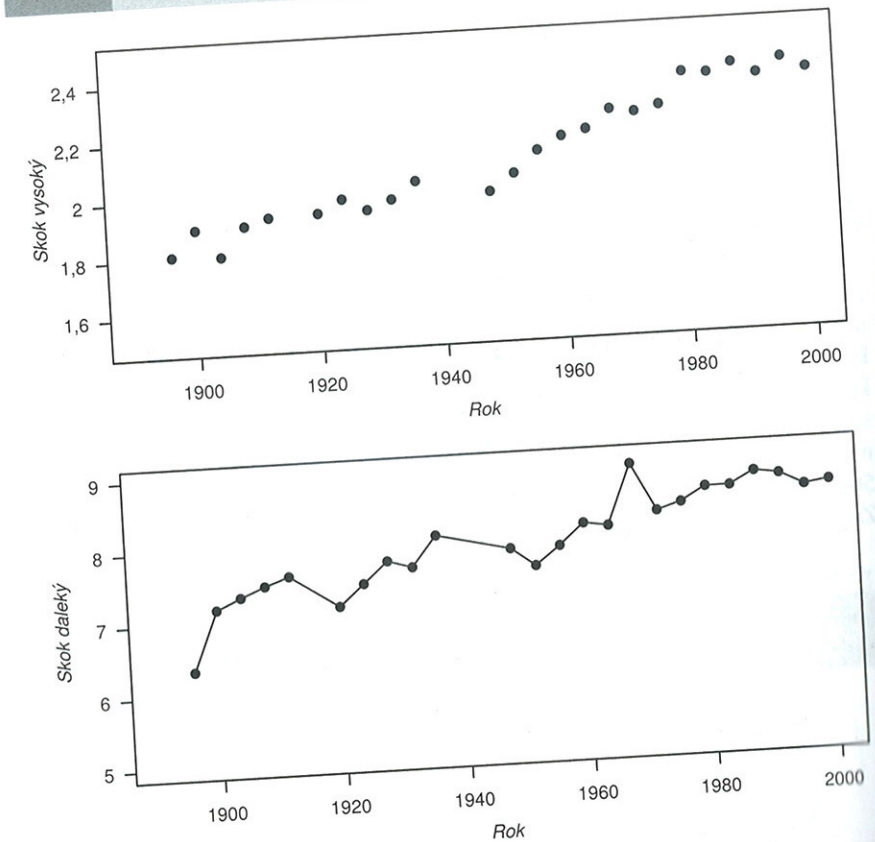
- *zhuštění* – kde se nalézá místo nebo místa nejvyšší četnosti hodnot;
- *shluky* – existuje jeden nebo více shluků dat v grafu;
- *mezery* – jsou v grafu intervaly nebo oblasti bez hodnot;
- *odlehlé hodnoty* – existují v grafu údaje podstatně rozdílné od zbytku dat;
- *tvar rozdělení* – lze popsat jednoduše tvar rozdělení dat?

Například u histogramu určíme základní tvar rozdělení a identifikujeme přítomnost odlehlých hodnot, které evidentně nepatří k základnímu tvaru histogramu. Hledáme přiléhavé vyjádření popisu základního tvaru. Histogram může mít symetrický tvar, nebo může být zešikmen na pravou, resp. levou stranu, jestliže jeho pravá, resp. levá strana je mnohem delší než levá, resp. pravá strana. Také může mít jeden, dva nebo více vrcholů. Histogram prokládáme někdy ideální křivkou, jež se nazývá *hustota*. Tvar histogramu porovnáváme často s hustotou, která se nazývá gaussovská křivka nebo normální křivka. Gaussovská křivka je symetrická křivka zvonovitého tvaru (viz kap. 4). Data s tímto rozdělením se nazývají normálně rozdělená data. Na obrázku 3.5 je histogramem znázorněno 50 údajů s průměrnou hodnotou nula a s proloženou ideální gaussovskou křivkou. Znázorňují se procentuální podíly v jednotlivých intervalech.

Obr. 3.5 Normálně rozdělená data s proloženou gaussovskou křivkou, procentuální zastoupení



Obr. 3.6 Příklad zobrazení trendu – mistrovské výkony ve skoku vysokém a dalekém na OH



Pokud chceme znázornit trend v datech v závislosti na časovém faktoru, použijeme graf trendu. Na obrázku 3.6 jsou znázorněny výkony ve skoku vysokém a dalekém, za něž sportovec získal zlatou medaili na olympijských hrách. Data doplňujeme proloženou přímkou, jinou proloženou křivkou nebo je spojíme úsečkou.

V této knize poznáme mnoho dalších možností grafického znázornění dat.

3.2 Míry centrální tendence

Statistické zpracování dat pomocí tabulek a grafů usnadňuje jejich vizuální analýzu a celkové posouzení datové konfigurace. Pro další zpracování však potřebujeme data vhodně kondenzovat. Proto se počítají různé číselné charakteristiky – **popisné statistiky**, které zachycují různé aspekty dat. Jedná se především o charakteristiky centrální tendence a rozptýlenosti, ale i o další charakteristiky jako šikmost nebo špičatost rozdělení dat.

Míry centrální tendence se snaží charakterizovat typickou hodnotu dat. (Říká se jim také střední hodnoty, resp. míry střední hodnoty nebo míry polohy – protože určují, kde na číselné ose je vzorek rozložen.) Nejznámější z nich jsou aritmetický průměr, medián a modus.

3.2.1 Aritmetický průměr

Aritmetický průměr je definován jako součet všech naměřených údajů vydělený jejich počtem. Označujeme ho pomocí symbolu $\bar{x}$ nebo M. Výpočet má tedy podobu:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Poznamenejme, že stejný výpočet vyjadřují zkrácené zápisy:

$$\bar{x} = \frac{\sum_i x_i}{n} \qquad \text{nebo} \qquad \bar{x} = \frac{\sum x_i}{n}$$

kde znak $\sum$ symbolizuje součet hodnot x_i pro všechny možné hodnoty indexu i.

Pro modelová data $\{2; 8; 9; 10; 1; 0; 5\}$ má průměr hodnotu

$$\bar{x} = \frac{2 + 8 + 9 + 10 + 1 + 0 + 5}{7} = 5.$$

Aritmetický průměr je optimální charakteristikou typické hodnoty množiny dat pro následující vlastnosti:

1. Součet odchylek měření od průměru se rovná nule – např. pro data z příkladu jsou odchylky $\{-3; 3; 4; 5; -4; -5; 0\}$ a jejich součet je číslo nula.
2. Fyzikálně si aritmetický průměr představujeme jako těžiště dat – součet dat pod průměrem je stejný jako součet dat nad průměrem, oba součty jsou v rovnováze. Součet vzdáleností od průměru hodnot nižších než průměr má být roven součtu vzdáleností od průměru hodnot vyšších než průměr. Každá hodnota má stejnou váhu.

3. Výraz $\sum (x_i - b)^2$ je nejmenší vzhledem k parametru b, jestliže b se rovná aritmetickému průměru. Výraz $\sum (x_i - b)^2$ jistým způsobem charakterizuje celkovou chybu, které se dopouštíme, když chceme nahradit všechny údaje jednou hodnotou b. Tvrzení vyjadřuje, že $\bar{x}$ odhaduje data s nejmenší chybou, přičemž za míru chyby považujeme kvadratickou odchylku.

Pokud máme několik průměrů spočítaných z různých podmnožin dat a známe příslušné počty měření n_i, lze vypočítat celkový průměr ze všech dat jako **vážený průměr:**

$$\bar{x} = \frac{\sum n_i \bar{x}_i}{\sum n_i}$$

Podobně se počítá průměr pro data zadaná četnostním způsobem pomocí frekvenční tabulky, v níž pro hodnoty x_i jsou ještě zadané jejich četnosti výskytu f_i:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

3.2.2 Medián a modus

Medián (označovaný Me nebo $\tilde{x}$) znamená hodnotu, jež dělí řadu *podle velikosti seřazených výsledků* na dvě stejně početné poloviny. Jestliže n je sudé číslo, pak Me je jakékoli číslo z intervalu $(x_{n/2}, x_{n/2+1})$. Jednoznačněji

$$Me = 0{,}5(x_{n/2} + x_{n/2+1}).$$

Jestliže n je liché číslo, pak

$$Me = x_{(n+1)/2}.$$

Pro modelová data seřazená podle velikosti $\{0; 1; 2; 5; 8; 9; 10\}$ zjistíme hodnotu mediánu 5. Medián je na rozdíl od aritmetického průměru málo citlivý k odlehlým hodnotám. Představme si třeba jakkoli velikou změnu nejmenší hodnoty směrem dolů: medián zůstane stejný, ale jistě se změní průměr. Medián Me má optimální vlastnost v tom smyslu, že minimalizuje součet absolutních odchylek měření od zvoleného čísla. Tato vlastnost je analogická vlastnosti aritmetického průměru $\bar{x}$, který minimalizuje součet kvadratických odchylek.

Modus nebo modální hodnota je hodnota, jež se v datech vyskytuje nejčastěji. Tato charakteristika nalézá uplatnění především u kategoriálních dat. Symbolicky se označuje $\hat{x}$ nebo Mo. V případě spojitých dat se odečítá pomocí sestrojeného histogramu, kdy se spočítá jako průměr z krajních hodnot intervalu, který obsahuje nejvíc dat. Pokud existuje v histogramu více vrcholů, udáváme je všechny. Říkáme pak, že rozdělení je dvou-, tří- nebo obecně vícevrcholové.

3.2.3 Použití měr centrální tendence

Rozhodnutí, kterou charakteristiku průměru nebo polohy použijeme při popisu dat, závisí na cílech a předpokladech analýzy. První omezení představuje úroveň měřítka měření. Aritmetický průměr se jen zřídka používá pro hodnocení dat, jejichž měřítko není intervalové nebo poměrové. Pro tyto dvě měřítka často uvádíme všechny střední hodnoty a všímáme si, proč se liší. Jestliže jsou data symetricky rozdělená, všechny tyto charakteristiky jsou přibližně stejné. Uvedeme základní pokyny pro užití středních hodnot.

Aritmetický průměr se má používat:

- jestliže data jsou získána minimálně v intervalovém měřítku (tzn. průměr neužíváme pro údaje kategoriální);
- jestliže je rozdělení symetrické;
- jestliže chceme použít statistické testy.

Medián se má použít:

- jestliže data jsou získána minimálně v ordinálním měřítku;
- jestliže chceme znát střed rozdělení dat;
- jestliže data mohou obsahovat odlehlé hodnoty;
- jestliže rozdělení dat je silně zešikmené.

Modus se má použít:

- jestliže rozdělení má více vrcholů;
- jestliže chceme získat o rozdělení jenom základní přehled;
- jestliže se slovem „průměrně" míní nejčastější hodnota.

3.3 Míry rozptýlenosti

Náhodně proměnlivé údaje nestačí charakterizovat jenom střední hodnotou. Omezenost středních hodnot spočívá v tom, že udávají pouze to, kolem jaké hodnoty se data „centrují", resp. které hodnoty jsou nejčastější. Data se stejnou střední hodnotou mohu mít různou rozptýlenost. Velikost proměnlivosti dat zachycujeme vhodně vybranou mírou rozptýlenosti dat. Existuje mnoho měr rozptýlenosti a záleží na okolnostech, kterou nebo které použijeme. Numerické charakteristiky tvaru rozdělení dat mají důležitý význam při kondenzaci dat do několika málo popisných údajů; pamatujme však, že nejlepší představu o datech nám poskytuje graf.

3.3.1 Variační rozpětí

Přestože se maximální a minimální hodnota uvádějí pravidelně při popisu dat, **variační rozpětí** R se počítá zřídka, ačkoli je jeho zjištění jednoduché:

$$R = x_{\max} - x_{\min}$$

Nevýhodou variačního rozpětí je velká citlivost vůči odlehlým hodnotám.

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má R hodnotu 10.

3.3.2 Rozptyl a směrodatná odchylka

Obě tyto charakteristiky popíšeme v jednom odstavci, protože spolu úzce souvisejí. Oběma je společná vlastnost, že na rozdíl od variačního rozpětí R využívají při výpočtu všechny údaje a obě se vztahují k aritmetickému průměru – měří rozptýlenost dat kolem aritmetického průměru dat. Dávají větší váhu extrémnějším hodnotám než průměrná absolutní odchylka.

Rozptyl je definován jako průměrná kvadratická odchylka měření od aritmetického průměru, přičemž při průměrování této odchylky dělíme číslem $(n - 1)$:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Při větších rozsazích není rozdíl mezi dělením číslem n nebo $n - 1$ významný. Dělení číslem n se použije, jestliže rozptyl počítáme pro všechny prvky populace. Při výpočtech někdy vycházíme ze vzorce

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n - 1},$$

který vede ke stejné hodnotě.

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má rozptyl hodnotu

$$s^2 = \frac{(2-5)^2 + (8-5)^2 + (9-5)^2 + (10-5)^2 + (1-5)^2 + (0-5)^2 + (5-5)^2}{6}$$

$$= 16{,}66.$$

Rozptyl se především používá v inferenční statistice při výpočtu různých testovacích statistik. Počítá se pomocí čtverců odchylek dat od průměru, proto má jiný rozměr než původní data.

Směrodatná odchylka s je odmocnina z rozptylu a vrací míru rozptýlenosti do měřítka původních dat:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má směrodatná odchylka hodnotu

$$s = \sqrt{\frac{(2-5)^2+(8-5)^2+(9-5)^2+(10-5)^2+(1-5)^2+(0-5)^2+(5-5)^2}{6}} = 4{,}08.$$

Při pokusu porozumět výpočtu směrodatné odchylky si všímáme jednotlivých operací:

1. Nejdříve vypočteme jednotlivé odchylky od průměru $(x_i - \bar{x})$, které pro daný údaj vyjadřují, jak se liší od typické hodnoty.
2. Čtverec odchylky (umocnění na druhou) převádí záporné odchylky na kladná čísla a zároveň větším odchylkám dává větší váhu. Například odchylce −2 dává váhu 4, ale odchylce 3 dává váhu 9.
3. Součet (suma) čtverců odchylek zachycuje všechny odchylky jedním číslem.
4. Dělením číslem $(n - 1)$ počítáme průměr s kvadratických odchylek.
5. Odmocnina převádí druhou mocninu do původního měřítka dat.

Základní vlastnosti směrodatné odchylky:

- směrodatná odchylka měří rozptýlenost kolem průměrů a má se používat jenom tehdy, když průměr je vhodný jako míra střední hodnoty (viz s. 95);
- $s = 0$ pouze tehdy, když se všechna data rovnají stejné hodnotě, jinak $s > 0$;
- stejně jako průměr $\bar{x}$ je i směrodatná odchylka s silně ovlivněna extrémními hodnotami – jedna nebo dvě odlehlé hodnoty zvětšují silně s;
- jestliže je rozdělení dat silně zešikmené, směrodatná odchylka neposkytuje dobrou informaci o rozptýlenosti dat – v takovém případě používáme kvantilové míry, které vysvětlíme v další části odstavce.

Jestliže chceme posoudit relativní velikost rozptýlenosti dat vzhledem k průměru, použijeme **koeficient variace** neboli **variační koeficient** VK. Počítáme ho, když chceme porovnat rozptýlenost dat skupin měření stejné proměnné s různým průměrem, nebo v těch případech, kdy se mění velikost směrodatné odchylky tak, že je přímo závislá na úrovni měřené proměnné ($s = k\bar{x}$, kde k je konstanta):

$$VK = \frac{s}{\bar{x}}$$

Pro modelová data {2; 8; 9; 10; 1; 0; 5} má koeficient variace hodnotu

$$VK = \frac{4{,}05}{5} = 0{,}85$$

Někdy se uvádí v procentech:

$$VK = \frac{s}{\bar{x}} \times 100\%$$

Pro naše modelová data má VK hodnotu 85 %.

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:

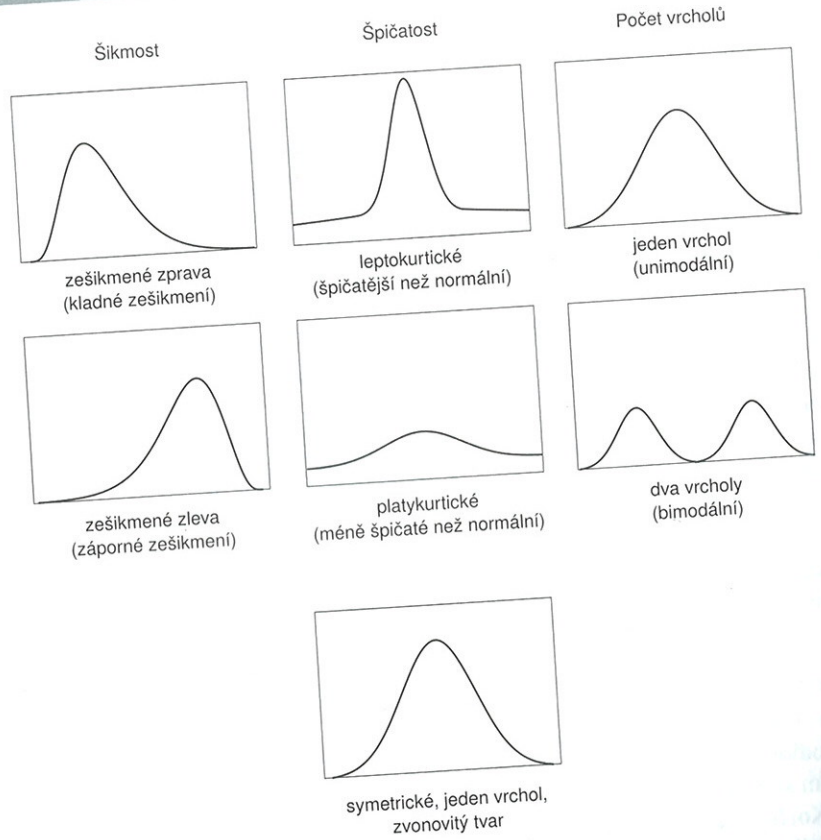
$$S_2 = \frac{m_4}{m_2^2} - 3$$

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).

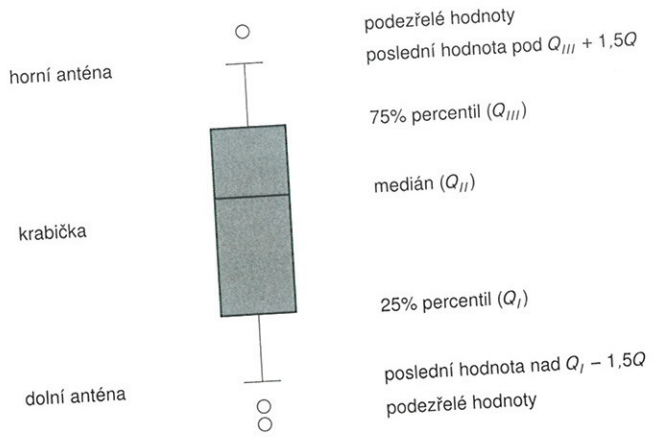
Tab. 3.3 Příklad popisu dat pomocí pěti hodnot

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

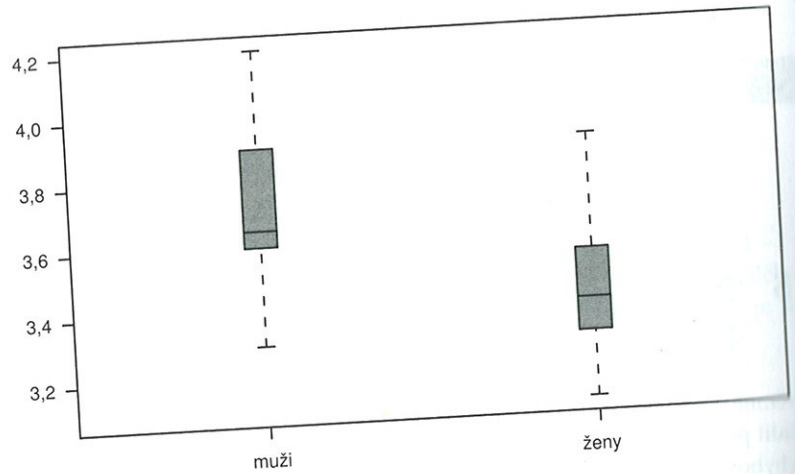
3.6 Zkoumání přítomnosti odlehlých hodnot a rezistentní odhady

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

Obr. 3.8 Konstrukce krabicového grafu s anténami (Q je interkvartilové rozpětí)



Obr. 3.9 Příklad krabicového grafu – výkony ve skoku dalekém



propočty a úvahy provedeme pro množinu měření bez odlehlých hodnot a s odlehlými hodnotami a posuzujeme rozdílnost získaných závěrů. Další strategie vychází z používání rezistentních odhadů, které nejsou citlivé k odlehlým hodnotám.

Upozorníme na příklady rezistentních odhadů střední hodnoty a rozptýlenosti. Vycházíme z poznatku, že aritmetický průměr je velmi citlivý vůči krajním hodnotám. Při podezření, že výběr obsahuje odlehlé hodnoty, se ho snažíme nahradit jiným výpočtem. Medián Me je rezistentní odhad střední hodnoty. Také se používá odhad pomocí percentilového průměru: $(2Me + Q_{III} + Q_I)/4$.

Useknutý průměr (trimmed mean) znamená, že při výpočtu aritmetického průměru zcela ignorujeme určité procento extrémních hodnot. Winsorizovaný průměr (winsorized mean) popisuje ve své knížce Havránek (1993). Vychází se z toho, že určité části krajních hodnot z množiny dat se přiřadí jedna zvolená méně extrémní hodnota. Z takto upravených hodnot se v dalším kroku počítá aritmetický průměr. Rezistentní odhady rozptýlenosti vycházejí z percentilových charakteristik. Například směrodatnou odchylku jako parametr normálního rozdělení lze odhadnout pomocí interkvartilového rozpětí nebo mediánové absolutní odchylky:

$$\hat{\sigma} = \frac{Q_{III} - Q_I}{1{,}35} = \frac{MAD}{0{,}67}$$

Existují statistické testy, kterými odlehlé hodnoty posuzujeme. Jejich užitečnost je sporná. Proto uvedeme pro hledání odlehlých hodnot v množině dat jedné proměnné jenom základní pravidla, která mají orientační charakter. V kapitole o regresní analýze uvedeme pravidlo o odlehlé hodnotě při prokládání regresní přímky.

Obecné pravidlo říká, že při deseti a více kvantitativních měřeních můžeme vyřadit jednu hodnotu, pokud neleží v intervalu $\bar{x} \pm 3s$, přičemž obě použité výběrové statistiky počítáme bez podezřelé hodnoty. Poznamenejme však, že problém odlehlých hodnot je nutné posuzovat v daném kontextu obezřetně. Při vylučování extrémních hodnot nesmíme postupovat mechanicky. Pro některé množiny dat jsou odlehlé hodnoty typickým jevem, což je nutné zohlednit v celé další statistické analýze.

Krabicový graf s anténami využívá pro identifikaci odlehlých hodnot kritériem, jež se opírá o velikost interkvartilového rozpětí Q. Pokud měření je od dolního nebo horního kvartilu vzdáleno více než $3/2Q$, označí se jako odlehlá hodnota. Jiné kritérium vychází ze znalosti mediánové absolutní odchylky MAD. Jestliže posuzované měření x^* je vzdáleno od mediánu o více než $5MAD$, měření klasifikujeme jako odlehlé.

PŘÍKLAD 3.1

Vyhledání odlehlých hodnot pomocí mediánu absolutních odchylek

Pravidlo využívající *MAD* demonstrujeme na datech, které jsme již seřadili podle velikosti:

2,8; 3,1; 3,7; 5,4; 6,2; 6,9; 7,2; 8,9; 12,7; 22,2; 29,8

Medián *Me* má hodnotu 6,9. Absolutní odchylka prvního čísla v řadě 2,8 od této hodnoty je 4,1; podobně vypočítáme všechny absolutní odchylky od *Me*:

4,1; 3,8; 3,2; 1,5; 0,7; 0,0; 0,3; 2,0; 5,8; 15,3; 22,9

Když je uspořádáme podle velikosti a vypočítáme *MAD*, získáme hodnotu *MAD* = 3,2. Za podezřelou hodnotu x^* považujeme číslo 29,8. Zjistíme poměr vzdálenosti x^* od *Me* a *MAD*, jenž má hodnotu 22,9/3,2 = 7,16, proto klasifikujeme údaj 29,8 jako odlehlou hodnotu. Ještě přezkoušíme číslo 22,2. Pro něj má hodnotící poměr hodnotu 15,3/3,2 = 4,78, takže zkoumanou hodnotu nezařadíme jako odlehlou. Další postup má odhalit příčinu, která vedla k údaji 29,8.

3.7 Transformace dat, standardizace

Existují různé možnosti, co s daty dělat, když jsou již v počítači. Často se používá globální úprava dat pomocí transformací. K nejpoužívanějším patří standardizace. Stručně charakterizujeme ty nejjednodušší z nich.

Funkční transformace

Přičítání nebo odečítání konstanty. Jedná se o jednoduchou akci, kdy ke všem datům přičteme kladnou nebo zápornou konstantu. Nejobvyklejší taktikou je odečtení aritmetického průměru od všech získaných skórů dané proměnné. Dostáváme tzv. centrovaná data nebo odchylky od průměru:

$$odchylka = (naměřený\ údaj) - (průměr)$$

Získáváme tak přehled, jak jsou jednotlivé údaje vzdálené od průměru. Také lze použít místo průměru medián nebo jinou míru centrální tendence. Průměr, medián a modus takto transformovaných dat se změní stejně jako původní údaje.

Násobení (dělení) konstantou. Tato operace se často nazývá škálování (toto slovo však má i jiné významy). Používá se např. při přechodu mezi použitými jednotkami měření (mezi kilogramy a gramy, metry a centimetry apod.). Také

pro tuto transformaci platí, že průměr, medián a modus transformovaných dat se změní stejně jako původní údaje.

Standardizace kombinuje odečítání a násobení. Standardizace se provádí podle předpisu

$$standardizovaný\ údaj = \frac{naměřený\ údaj - průměr\ dat}{směrodatná\ odchylka}.$$

Symbolicky tuto transformaci vyjádříme ve tvaru

$$x' = \frac{x - \bar{x}}{s_x}.$$

K transformaci lze také použít kvartilové charakteristiky

$$x' = \frac{x - \tilde{x}}{Q}.$$

Standardizace znamená, že průměr (nebo medián) standardizovaných dat je 0 a jejich směrodatná odchylka (nebo interkvartilové rozpětí) je 1. Rozdělení, která jsou takto standardizována, se mnohem snadněji srovnávají a někdy i kombinují. Standardizovaná data se často nazývají též standardizované skóry.

Data se symetrickým rozdělením standardizovaná průměrem a směrodatnou odchylkou jsou symetricky rozdělená kolem nuly a jejich hodnoty se pohybují přibližně v rozmezí od −3 do 3. Hodnoty mimo tyto meze se prověřují, zda nemají charakter odlehlých hodnot (outliers).

Jiné funkční transformace. Dosud uvedené úpravy dat jsou příkladem nejjednodušších funkčních transformací. Data se transformují různými dalšími funkcemi jako logaritmus, odmocnina, obrácená hodnota apod. Těmito transformacemi často linearizujeme jinak nelineární vztahy nebo upravujeme tvar rozdělení dat, aby se více podobalo rozdělení popsanému Gaussovou křivkou. O některých transformacích se zmíníme ještě v odstavci o explorační analýze, v kapitolách o regresní analýze a analýze rozptylu.

Standardizaci pomocí průměru a směrodatné odchylky lze vyjádřit jako transformaci pomocí lineární funkce $y = a + bx$, kde $a = -\bar{x}/s_x$ a $b = 1/s_x$. I z tohoto důvodu je důležité vědět, jaký má vliv jednoduchá lineární transformace $y = a + bx$ ($b > 0$) na hodnotu některých základních popisných statistik.

průměr $\bar{y} = a + b\bar{x}$ směrodatná odchylka $s_y = bs_x$

šikmost $S_{1y} = S_{1x}$ špičatost $S_{2y} = S_{2x}$

Lineární transformace nemění typ tvaru rozdělení dat.

Převod hodnot na pořadové hodnoty a percentily

V těchto dvou transformacích přiřazujeme naměřené hodnotě její pořadí nebo percentilovou hladinu. Nová hodnota udává relativní pozici původní hodnoty v celé množině dat vzhledem k relaci řazení podle velikosti.

Transformace do pořadí – označujeme ji někdy R – převádí daný údaj do intervalu 1 až n. Jsou-li všechny údaje různé, najdeme nejmenší údaj x_i a přiřadíme mu číslo $R_i = 1$, a tak postupujeme, dokud nepřiřadíme všem prvkům jejich pořadová čísla. Obecně můžeme říci, že přiřazujeme údaji x_j číslo R_j, což je počet x_i, jež jsou menší nebo rovny udaji x_j. Pokud jsou některé x_j stejné, pak jim přiřazujeme průměrné pořadí, které odpovídá této skupince shodných hodnot.

Percentilová transformace převádí údaje do intervalu 0–100. Každému údaji je přiřazena percentilová hladina, jež odpovídá relativnímu počtu údajů (vynásobenému číslem 100), které jsou menší než tento údaj nebo stejné. Percentilová hodnota 50 odpovídá mediánu a hodnota 100 maximu původních dat u všech proměnných.

Tab. 3.4 Příklad převodu dat na pořadové hodnoty a percentily

Pořadí	Skok daleký [m]	Standardizovaná hodnota	Percentil	Pořadí	Skok daleký [m]	Standardizovaná hodnota	Percentil
1	3,10	−1,68	3,85	14	3,60	0,06	53,85
2	3,15	−1,50	9,62	15	3,65	0,23	63,46
3	3,15	−1,50	9,62	16	3,65	0,23	63,46
4	3,30	−0,98	17,31	17	3,65	0,23	63,46
5	3,30	−0,98	17,31	18	3,65	0,23	63,46
6	3,35	−0,81	25,00	19	3,70	0,41	73,08
7	3,35	−0,81	25,00	20	3,75	0,58	76,92
8	3,40	−0,63	30,77	21	3,80	0,75	80,77
9	3,45	−0,46	34,62	22	3,90	1,10	86,54
10	3,50	−0,29	42,31	23	3,90	1,10	86,54
11	3,50	−0,29	42,31	24	4,05	1,62	94,23
12	3,50	−0,29	42,31	25	4,05	1,62	94,23
13	3,55	−0,11	50,00	26	4,20	2,14	100,00

PŘÍKLAD 3.2

Transformace dat na hodnoty pořadové, standardizované a percentilové

Pro vzorová data ve skoku dalekém uvedeme jejich pořadí bez úpravy na stejné hodnoty, standardizované hodnoty a percentilové hodnoty. Pro lepší přehled jsou záznamy již seřazeny podle výkonu ve skoku dalekém (tabulka 3.4). Je patrné, že percentilové hodnoty jsou mezi čísly 0 a 100, standardizované hodnoty jsou mezi čísly −3 a +3. Pořadí jsou čísla v intervalu 1 až n. Při úpravě pořadí pro stejné hodnoty bychom přiřadili např. oběma číslům 3,15 v souboru dat pořadí 2,5 a třem hodnotám 3,50 pořadí 11.

3.8 Explorační analýza dat

V tomto odstavci vysvětlíme ideu explorační analýzy a uvedeme příklady. Přitom zmíníme i některé statistické pojmy a techniky, které zatím nebyly dostatečně vymezeny. Vrátíme se k nim a podrobně je probereme v dalších kapitolách.

Explorační analýza dat (EDA) je skupina statistických technik a určitý přístup k analýze numerických dat, zdůrazňující grafické a tabelační znázorňování dat, metody rezistentní k odlehlým hodnotám, snahu odhalit v datech nápadné konfigurace a schopnost navrhovat deskriptivní modely dat. Předpokládá se, že každá analýza dat začíná pečlivým prozkoumáním struktury dat, jež nám poskytne přehled o chování proměnných, abychom mohli přistoupit ke komplexnější analýze. Doporučuje se však explorovat data i na konci analytického procesu, abychom neopomněli příležitost použít k analýze jiné perspektivy, které by mohly podpořit náš základní teoretický pohled nebo naznačit nové vztahy, jež by měly být zkoumány. Platí, že všechny hypotézy navržené v explorační fázi je nutné přezkoušet pomocí nových dat.

Základní metodologická práce o EDA pochází od Johna Tukeye (1977), jenž navrhl několik nových technik a inovoval některé starší metody popisné statistiky. Pomocí nich vytvořil systematickou strategii analýzy dat.

Tukey přirovnává explorační analýzu dat k práci detektiva, který vyhledává důkazy o vině nebo nevině podezřelého, přičemž není jeho úkolem definitivně rozhodnout. Detektiv sleduje události a jednotlivé aktéry, generuje hypotézy a sestavuje předběžné zprávy. Nezajímá ho zcela přesné rozlišení pravých a nepravých výpovědí. Klade si otázky a navrhuje odpovědi. V podobné situaci jako při provádění explorační analýzy dat je také státní zástupce, který zkoumá dostupná fakta, než se rozhodne, zda případ předloží k soudu. Statistické usuzování

nebo *inferenční statistika* – známější etapa statistického zkoumání, jež rigorózně využívá technik statistického testování hypotéz – se přirovnává v této metafoře k soudnímu tribunálu, ve kterém nakonec rozhoduje soudce o vině a nevině.

Statistická analýza není jenom testování hypotéz (kap. 5); to bývá závěrečný krok. Výzkum vyžaduje, abychom poznávali nashromážděná data a učili se z nich ve všech fázích procesu tvorby modelu a teorií. Na začátku zkoumáme, zda nalezené rozmezí hodnot u dané proměnné má smysl. Jsou kategorie diskrétní proměnné rovnoměrně zastoupeny? Jak se liší jednotlivé skupiny dat? Výzkumníka např. nezajímá při analýze dvojrozměrných dat jenom verifikace hypotézy v regresní analýze, že směrnice proložené přímky se rovná 0 (úkol technik statistického usuzování). Častěji mu jde o širší okruh otázek: Jak dál postupovat, když jsou získány hodnoty pouze pro dvě úrovně nezávisle proměnné, a ne v celém jejím možném rozsahu? Obsahuje materiál data, která se značně liší od ostatních? Podobá se tvar datové konfigurace bodů elipse nebo spíše banánu? Co je na datech zvláštní?

Výzkumník si uvědomuje skutečnost, že použití určité statistické metody závisí na mnoha předpokladech. Snaží se prozkoumat anomálie v datech, odlehlé hodnoty nebo nelineárnost. Jestliže výzkumník pracuje tímto způsobem, jedná se o explorační analýzu.

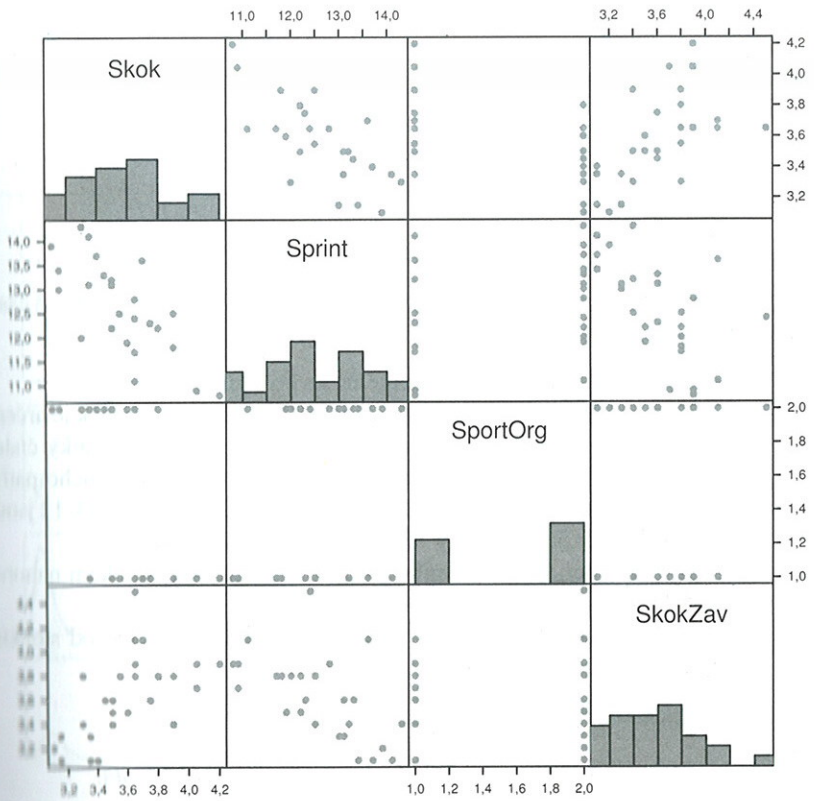
Tukey charakterizuje oba hlavní typy analýzy (inferenční a explorační) tak, že

a) oba způsoby jsou důležité;
b) explorační analýza (EDA) se provádí jako první;
c) každá studie by měla kombinovat oba způsoby.

Za čtyři základní prvky explorační analýzy se považují vizualizace dat; analýza reziduálních hodnot; transformace dat; robustní a rezistentní procedury.

- **Vizualizace dat** se řídí zásadou „Obraz je více než tisíc slov". Pomocí grafu je jednodušší detekovat konfigurace a struktury. Grafickými metodami hledáme odlehlé hodnoty, rozeznáváme shluky v datech, kontrolujeme rozdělení dat a jejich předpoklady, zkoumáme vztahy mezi proměnnými, srovnáváme průměry diferencí a zkoumáme data závislá na čase. Poznali jsme a ještě poznáme mnoho různých způsobů grafického znázornění. Velmi často se pro úsporné zobrazení rozložení dat a jejich porovnání ve skupinách používá krabičkový graf (viz kap. 3.5). Za užitečný se také považuje mnohonásobný dvojrozměrný bodový graf s histogramy. Na obr. 3.10 jsou zobrazeny údaje z modelové matice dat pomocí tzv. maticového grafu. Pro zobrazení jednorozměrných dat se používá také graf stonku a listu (stem-leaf plot), který navrhl Tukey. Od histogramu se liší tím, že máme stále přehled o jednotlivých hod-

Obr. 3.10 Maticový graf pro explorační analýzu modelových dat



notách. V histogramu jsou zobrazeny pouze body, v **grafu stonku a listů** (stem-leaf plot) se zobrazují významné číslice jednotlivých hodnot. Pro jeho sestrojení musíme zvolit, od které pozice budeme číslice hodnot zobrazovat, v případě potřeby také statistický krok. Obrázek 3.11 zobrazuje výkony ve sprintu chlapců a dívek z tabulky 2.9 pomocí jednoduchého komparujícího grafu stonku a listů. Pomocí něho lze rekonstruovat všechna data. Z grafu lze např. odvodit, že v souboru dat existuje jeden údaj 14,1 u dívek a jeden údaj 14,3 u chlapců. U dívek nalezneme dva údaje 12,5. Tímto grafem snadno zobrazíme data pro výběry do rozsahu $n = 100$.

Obr. 3.11 Příklad grafu stonku a listů pro časy ve sprintu z tabulky 2.9, s. 77

```
Jednotka škálování 0,1

Chlapci          Dívky
   998 | 10 | 0
   871 | 11 |
   432 | 12 | 02558
    31 | 13 | 0124679
     3 | 14 | 1
```

Uvedeme kroky zhotovení jednoduchého grafu stonku a listů (např. pravá část obrázku 3.11):

1. Určíme, kterou číslici čísel považujeme za poslední platné místo.
2. Přiřadíme každé pozorování (číslo) do jednoho ze stonků. Stonek je určen platnými číslicemi bez poslední číslice (v grafu 3.11 jsou stonky čísla 10;11;12;13 a 14). Na každém stonku je tolik listů, kolik do něho patří čísel. Každý list je určen poslední platnou číslicí listu (pro stonek 12 jsou to číslice 0; 2; 5; 5; 8 pro listy 12,0; 12,2; 12,5; 12,5; 12,8).
3. Sestavíme stonky do vertikálního sloupce s nejmenším stonkem nahoře a od shora po pravé straně nakreslíme vertikální čáru.
4. Každý list (pozorování) stonku doplníme do řádku napravo od stonku v pořadí danou velikostí číslice listu.

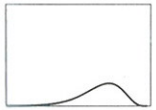
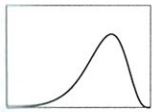
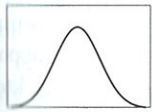
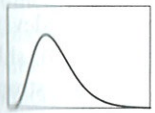
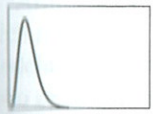
- **Analýza reziduálních hodnot** vychází ze základní představy

$$data = predikce\ modelem(funkce) + reziduální\ hodnota.$$

Predikce modelem poukazuje na očekávanou hodnotu sledované proměnné. Reziduální hodnota nebo chyba je hodnota, o níž se liší získaný údaj od naší původní představy. Vyhodnocením těchto chyb výzkumník posuzuje adekvátnost modelu, který sestrojil. Popsaný koncept využíváme často v regresní analýze nebo analýze rozptylu a popisujeme jej v kapitole 7.3.2. Modelování bez analýzy reziduálních hodnot není dostatečně přesvědčivé. Tento typ analýzy neslouží pouze k ověřování validity modelu, ale i k jeho vylepšování, protože pomáhá odhalit aspekty vztahů, které model nezohledňuje.

- V EDA používáme hlavně **nelineární transformace** dat a měníme tak jejich tvar rozdělení. Jiný cíl transformace dat je linearizace nelineárních vztahů (viz kap. 7.3.6) nebo dosažení konstantního rozptylu v analýze rozptylu (viz

Obr. 3.12 Transformace dat k normalitě pomocí funkce x^q

Problém	Transformace	Efekt
	$X = X^3$	snižuje (extrémní) zešikmení zleva
	$X = X^2$	snižuje zešikmení zleva
	$X = X^1$	žádný účinek
	$X = X^{1/2} = \sqrt{X}$	snižuje zešikmení zprava
	$X = \ln X$	snižuje zešikmení zprava
	$X = -X^{-1/2} = -1/\sqrt{X}$	snižuje (extrémní) zešikmení zprava

kap. 9.1.1). Obrázek 3.12 ukazuje tvary rozdělení dat a transformace, kterými je upravíme tak, aby bylo rozdělení transformovaných dat více podobné normálnímu.

- Parametrické testy statistických hypotéz jsou založeny na odhadech, jež jsou citlivé k odlehlým hodnotám nebo k zešikmení rozdělení dat. V EDA se používají rezistentní odhady jako medián, percentilový průměr, seříznutý nebo useknutý průměr, které jsme popsali v předchozím odstavci.

Je potřebné poznamenat, že pojmy **rezistentní metoda** a **robustní metoda** se trochu liší, přestože oba pojmy se zaměňují. Koncept robustnosti se týká odolnosti statistických metod proti porušení jejich předpokladů. Rezistence

znamená, že odhad je imunní vůči odlehlé hodnotě. EDA se více zajímá o rezistenci, kdežto v testování hypotéz usilujeme spíše o robustnost.

EDA je důležitá ve studiích, kde se zkoumá mnoho proměnných navzdory tomu, že nebyla na začátku uvažována nějaká propracovaná teorie, která by ospravedlnila jejich zahrnutí. Výzkumník pouze „cítil", že je nutné tyto proměnné sledovat. Tomu odpovídá i nutnost použít rezistentní metody odhadu, protože data se často získají ad hoc za nereplikovatelných podmínek, kdy vztah proměnných k teoretickým konstruktům má pouze vágní povahu. Vychází se přitom z filozofie, že prvním úkolem analýzy je objevovat, ne hodnotit. Úkoly hodnocení se ponechávají na jinou dobu. Explorační analýza dat se věnuje formulování modelu a navrhování hypotéz použitím empirických dat. Tato fáze analýzy je důležitá proto, že umožňuje odhalit nečekané a na první pohled těžko zjistitelné vlastnosti dat, čímž se získává hlubší vhled do zkoumané problematiky. Do vědy nepatří pouze tradiční mechanické postupy, ale i hra s daty s cílem vytvářet návrhy, které dají datům smysl.

V rámci EDA lze použít i jiné techniky, než navrhl Tukey. Jde především o vícerozměrné metody, jako mnohonásobná regresní analýza (kap. 10), shluková analýza (kap 13.6), metody vytváření regresních a klasifikačních stromů (kap. 13.3) nebo metody explorační faktorové analýzy (kap. 13.8.1).

Souhrn

Cílem popisné statistiky je organizace a popis dat, jež byla získána v rámci pozorování a experimentu. To zahrnuje identifikaci odlehlých hodnot, znázorňování dat a jejich porovnávání pomocí tabulek a grafů, numerickou redukci velkého množství dat pomocí vhodně navržených popisných charakteristik polohy, rozptýlenosti a tvaru rozdělení dat. Tyto techniky se používají také k explorační analýze dat, kdy odhalujeme neočekávané aspekty dat a prozkoumáváme jejich strukturu. Pro tento účel byly navrženy speciální techniky jako krabicový graf nebo graf stonku a listů a další metody pro grafickou exploraci a zobrazení dat.

Poznatky o rozdělení dat získané popsanými technikami interpretujeme ve výzkumné zprávě v kontextu úlohy a vzniku dat a hledáme pro ně vysvětlení. Jsou také důležité pro další typy analýz.

Numerický souhrn dat určité proměnné má obsahovat popis centrální hodnoty a rozptýlenosti. Centrální tendenci různým způsobem popisují aritmetický průměr, medián nebo modus. Jestliže použijeme medián, pak rozptýlenost dat popisujeme dolním a horním kvartilem, případně interkvartilovým rozpětím. Souhrn pomocí pěti hodnot sestává z mediánu, kvartilů, minimální a maximální hodnoty. Krabicový graf se sestrojují pomocí těchto pěti hodnot.

Aritmetický průměr i směrodatná odchylka jsou dobrým popisem symetrických rozdělení. Využívají se v nejznámějším teoretickým pravděpodobnostním modelu – v normálním rozdělení. Jsou však ovlivňovány odlehlými hodnotami nebo zešikmením rozdělení dat.

Medián a dolní a horní kvartil nejsou tak ovlivněny odlehlými hodnotami. První a třetí kvartil spolu s extrémními hodnotami charakterizují oba konce rozdělení dat. Tyto charakteristiky patří do třídy rezistentních statistik.

Nejdůležitější mírou rozptýlenosti je rozptyl s^2 a směrodatná odchylka s. Obě tyto míry nabývají nulové hodnoty, pokud jsou všechny údaje stejné.

Upřesnění popisu tvaru rozdělení dat získáme pomocí koeficientů špičatosti a šikmosti.

O grafických prostředcích explorační analýzy dat pojednávají podrobně Anscombe (1996), Havránek (1993) nebo Tukey (1977).