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

Na exploraci se váže **kontrola dat**, již 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.

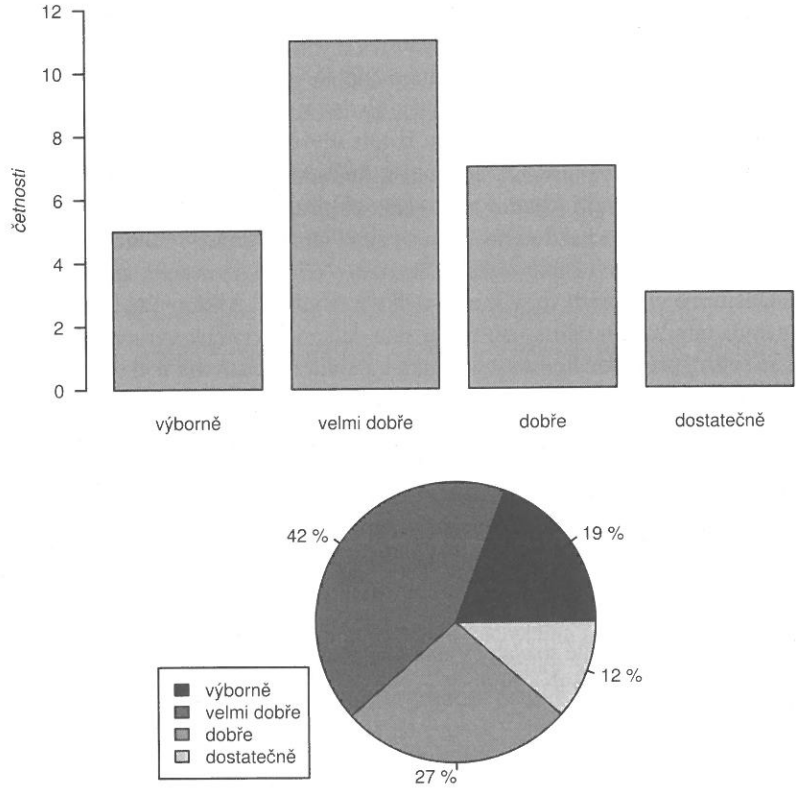
Obrázek 3.2 obsahuje dva bodové grafy výsledků ve skoku dalekém, zvlášť pro dívky a chlapce.

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

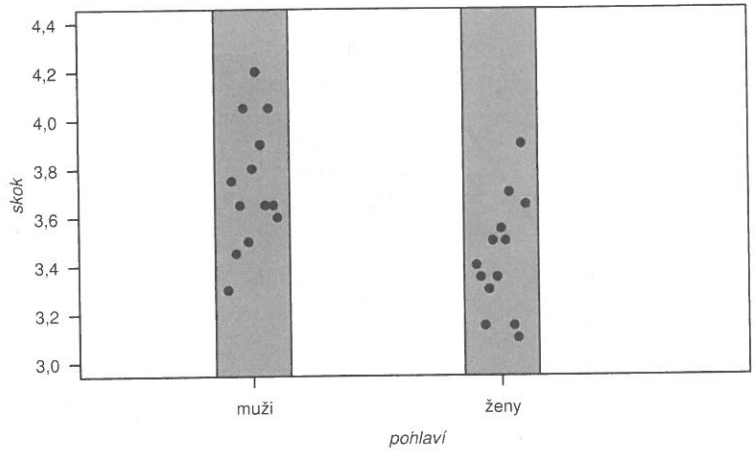
Obr. 3.1 Příklady zobrazení znaku *Prospěch z matematiky* (sloupcový a koláčový graf)



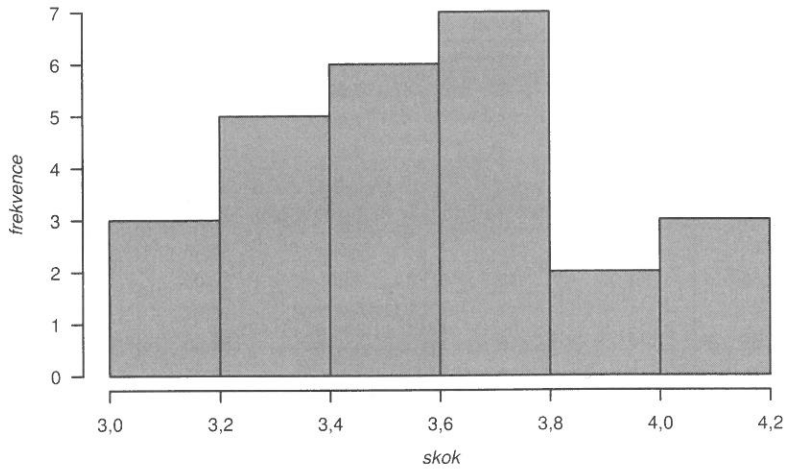
Tab. 3.2 Příklad základní úpravy primárních dat – tabulka četností a kumulativních četností pro data z tabulky 2.9

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	|

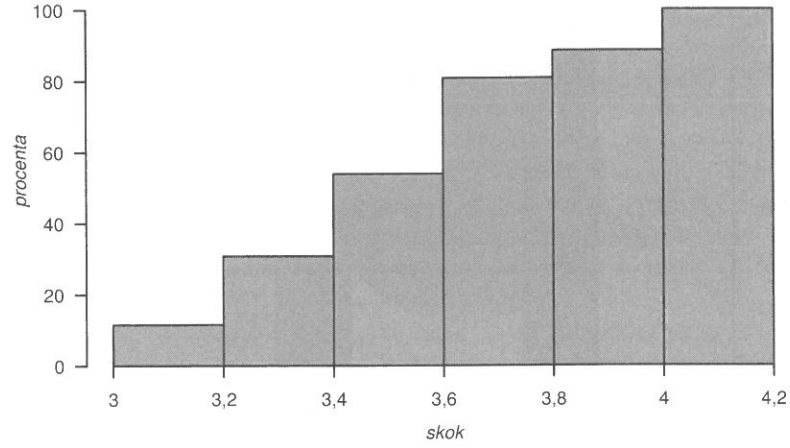
Obr. 3.2 Bodový graf



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

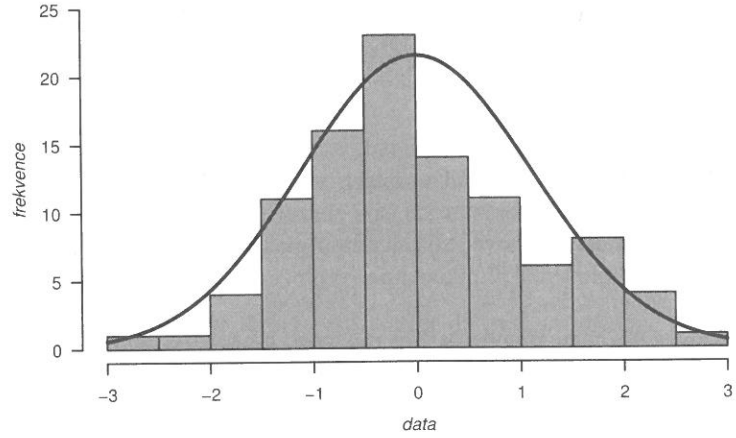


Při popisování a analýze toho, co graf zobrazuje, si všímáme nejdříve základní tvarové konfigurace a pak deviací od tohoto tvaru. Hodnotíme:

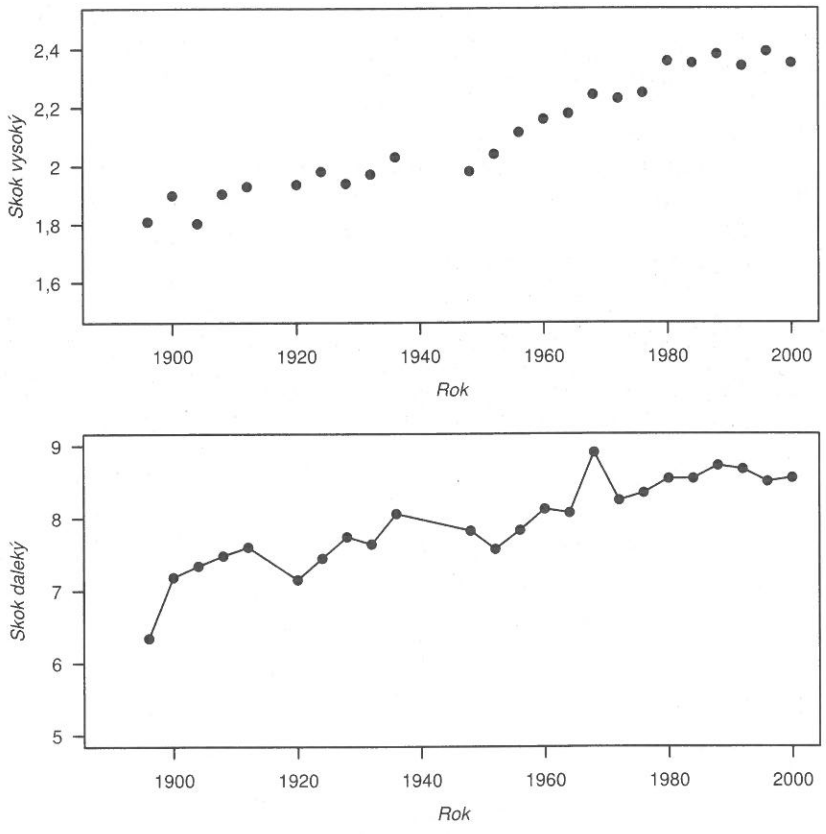
- *zhuštění* – kde se nalézá místo nebo místa nejvyšší četnosti hodnot;
- *shluky* – existuje jeden nebo více shluků dat v grafu;
- *mezery* – jsou v grafu intervaly nebo oblasti bez hodnot;
- *odlehlé hodnoty* – existují v grafu údaje podstatně rozdílné od zbytku dat;
- *tvar rozdělení* – lze popsat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