MUNI
ECON

Popisná statistika

Povinná literatura: Mann (2016), Kapitola 3

Z čeho studovat třetí lekci?

Povinná literatura:	Mann (2016), kap. 03
Příprava na cvičení:	Leaflet 04 Koncepty a procedury, cv. 04, kap. 03
Příprava na zkoušku:	Mann (2016), kap. 03 Leaflet 04 Sbírka úloh, kap. 03 Koncepty a procedury, cv. 04, kap. 03

Motivační vstup

Nástupní plat absolventů v prvním zaměstnání dle fakult MU (2022)

Fakulta	Průměr	Fakulta	Průměr
FI	63 990 Kč	FSS	29 339 Kč
LF	44 438 Kč	PřF	26 165 Kč
ESF	32 529 Kč	PdF	25 649 Kč
FSpS	31 250 Kč	FF	25 023 Kč
PrF	31 602 Kč	MU	32 586 Kč

Na základě výše uvedené tabulky, která vychází z dat Masarykovy univerzity pro rok 2022, jsou uvedeny průměrné nástupní platy absolventů jednotlivých fakult. Jak můžeme vidět, mezi těmito fakultami měli absolventi Fakulty informatiky (FI) nejvyšší nástupní plat ve výši 63 990 Kč, zatímco absolventi Pedagogické fakulty (PdF) měli nejnižší nástupní plat ve výši 25 649 Kč. Je zřejmé, že mezi nástupními platy absolventů různých fakult existují značné rozdíly.

Průměr

Průměr pro neseskupená data je získán dělením součtu všech hodnot počtem všech hodnot v datové sadě.

Průměr pro populaci:

$$\mu = \frac{\sum x}{N}$$

Průměr pro vzorek:

$$\bar{x} = \frac{\sum x}{n}$$

kde $\sum x$ je součet všech hodnot; N je velikost populace; n je velikost vzorku; μ je průměr populace; a $\bar{x}$ je průměr vzorku.

Příklad 1

Tabulka 3.1 udává celkový zisk (v milionech dolarů) 10 amerických společností za rok 2014 (www.fortune.com). Určete průměrný zisk za rok 2014 u těchto 10 společností.

Table 3.1 **2014 Profits of 10 U.S. Companies**

Company	Profits (million of dollars)
Apple	37,037
AT&T	18,249
Bank of America	11,431
Exxon Mobil	32,580
General Motors	5346
General Electric	13,057
Hewlett-Packard	5113
Home Depot	5385
IBM	16,483
Wal-Mart	16,022

Příklad 1: Řešení

$$\begin{aligned}\sum x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \\ &= 37\,037 + 18\,249 + 11\,431 + 32\,580 + 5\,346 + 13\,057 \\ &\quad + 5\,113 + 5\,385 + 16\,483 + 16\,022 = 160\,703\end{aligned}$$

$$\overline{x} = \frac{\sum x}{n} = \frac{160\,706}{10} = 16,070.3 = \$16\,070.3 \text{ milionu}$$

Těchto 10 společností vydělalo v roce 2014 v průměru **16 070.3 milionu dolarů**.

Příklad 2: Vliv odlehlé hodnoty na průměr

Máme seznam cen osmi náhodně vybraných domů ze všech domů nabízených k prodeji:

Cena domu			
$245 670	$176 200	$450 394	$310 160
$360 280	$272 440	$393 610	$3 874 480

Všimněte si, že cena posledního domu je $3 874 480, což je **odlehlá hodnota**. Ukažte, jak zahrnutí této odlehlé hodnoty ovlivňuje hodnotu průměru.

Příklad 2: Řešení

Pokud nezahrneme cenu nejdražšího domu (odlehlé hodnoty), průměr cen ostatních sedmi domů je:

Průměr bez odlehlé hodnoty = $\frac{2\,208\,754}{7} = \$315\,536.29$

Nyní, abychom viděli vliv odlehlé hodnoty na hodnotu průměru, zahrneme cenu nejdražšího domu a najdeme průměrnou cenu osmi domů. Tento průměr je

Průměr s odlehlou hodnotou = $\frac{2\,208\,754+3\,874\,480}{8} = \$760\,404.25$

=> Takže když zahrneme cenu nejdražšího domu, průměr se více než zdvojnásobí, protože vzroste z 315 536,29 dolarů na 760 404.25 dolarů.

Případová studie

Nástupní plat absolventů v prvním zaměstnání dle fakult MU (2022)

Fakulta	Průměr	Průměr bez odlehlých pozorování
FI	63 990 Kč	56 656 Kč
LF	44 438 Kč	40 993 Kč
ESF	32 529 Kč	32 052 Kč
FSpS	31 250 Kč	30 500 Kč
PrF	31 602 Kč	30 078 Kč
MU	32 586 Kč	29 809 Kč
FSS	29 339 Kč	27 728 Kč
PřF	26 165 Kč	25 060 Kč
PdF	25 649 Kč	25 808 Kč
FF	25 023 Kč	24 730 Kč

Medián

Medián je hodnota, která dělí datový soubor, jenž byl seřazen vzestupně, na dvě stejné poloviny.

=> Pokud má datový soubor lichý počet hodnot, medián je dán hodnotou prostředního členu v seřazeném datovém souboru.

=> Pokud má datový soubor sudý počet hodnot, medián je dán průměrem dvou prostředních hodnot v seřazeném datovém souboru.

Výpočet mediánu se skládá ze dvou kroků:

1. Seřadit daný soubor dat vzestupně.
2. Najít hodnotu, která rozdělí seřazený soubor dat na dvě stejné části. Tato hodnota poté určuje medián.

Příklad 3

Tabulka 3.2 uvádí odměny generálních ředitelek 11 amerických společností za rok 2014 (USA TODAY, 1. května 2015).

Table 3.2 **Compensations of 11 Female CEOs**

Company & CEO	2014 Compensation (millions of dollars)
General Dynamics, Phebe Novakovic	19.3
GM, Mary Barra	16.2
Hewlett-Packard, Meg Whitman	19.6
IBM, Virginia Rometty	19.3
Lockheed Martin, Marillyn Hewson	33.7
Mondelez, Irene Rosenfeld	21.0
PepsiCo, Indra Nooyi	22.5
Sempra, Debra Reed	16.9
TJX, Carol Meyrowitz	28.7
Yahoo, Marissa Mayer	42.1
Xerox, Ursula Burns	22.2

Najděte medián těchto údajů.

Příklad 3: Řešení

Pro výpočet mediánu provádíme následující dva kroky.

Krok 1: Seřadíme daná data vzestupně takto:

16.2 16.9 19.3 19.3 19.6 21.0 22.2 22.5 28.7 33.7 42.1

Krok 2: Existuje 11 datových hodnot. Šestá hodnota dělí těchto 11 hodnot na dvě stejné části. Tedy šestá hodnota udává medián, jak je ukázáno níže.

16.2 16.9 19.3 19.3 19.6 **21.0** 22.2 22.5 28.7 33.7 42.1

↑

Medián

Medián kompenzací za rok 2014 pro těchto 11 generálních ředitelek je 21.0 milionu dolarů.

Medián vs průměr

Medián udává střed histogramu, přičemž polovina datových hodnot je vlevo od mediánu a polovina vpravo od mediánu.

=> Výhodou použití mediánu jako míry centrální tendence je, že není ovlivněn odlehlými hodnotami.

=> V důsledku toho je medián upřednostňován před průměrem jako míra středu pro datové soubory, které obsahují odlehlé hodnoty.

Případová studie

Nástupní plat absolventů v prvním zaměstnání dle fakult MU (2022)

Fakulta	Průměr	Průměr bez odlehlých pozorování	Medián
FI	63 990 Kč	56 656 Kč	47 500 Kč
LF	44 438 Kč	40 993 Kč	36 250 Kč
ESF	32 529 Kč	32 052 Kč	31 000 Kč
FSpS	31 250 Kč	30 500 Kč	31 250 Kč
PrF	31 602 Kč	30 078 Kč	28 750 Kč
MU	32 586 Kč	29 809 Kč	28 500 Kč
FSS	29 339 Kč	27 728 Kč	28 000 Kč
PřF	26 165 Kč	25 060 Kč	25 000 Kč
PdF	25 649 Kč	25 808 Kč	26 500 Kč
FF	25 023 Kč	24 730 Kč	25 500 Kč

Modus

Modus je hodnota, která se v datovém souboru vyskytuje s největší frekvencí.

Zásadním nedostatkem modu je, že datová sada může mít žádný, jeden, dva nebo více než jeden modus. Naproti tomu bude mít vždy pouze jeden průměr a pouze jeden medián.

- Unimodální - Datová sada s pouze jedním modem.
- Bimodální - Datová sada se dvěma mody.
- Multimodální - Datová sada s více než dvěma mody.

Poznámka: Jednou z výhod modu je, že lze vypočítat pro oba druhy dat – kvantitativní i kvalitativní – zatímco průměr a medián lze vypočítat pouze pro kvantitativní data.

Příklad 4

Zadání: Najděte modus. Následující data udávají rychlost (v km za hodinu) osmi aut, která byla zastavena na vesnici pro překročení rychlosti.

77 82 74 81 79 84 74 78

Řešení: V této datové sadě se hodnota 74 vyskytuje dvakrát a každá z ostatních hodnot pouze jednou. Protože se 74 vyskytuje nejčastěji, je to modus.

Modus = 74 km za hodinu

Příklad 5: (Dataset bez modu)

Zadání: Najděte modus. Příjmy pěti náhodně vybraných rodin v loňském roce byly 76 150 dolarů, 95 750 dolarů, 124 985 dolarů, 87 490 dolarů a 53 740 dolarů.

Řešení: Datová sada nemá modus, protože každá hodnota v této datové sadě se vyskytuje pouze jednou.

Příklad 6: (Modus u kvalitativních dat)

Zadání: Najděte modus. Hodnost pěti studentů, kteří jsou členy studentského senátu na vysoké škole, jsou postupně: senior, sophomore, senior, junior a senior.

Řešení: Protože kategorie senior se vyskytuje častěji než ostatní kategorie, je to modus pro tuto datovou sadu. Pro tuto datovou sadu nemůžeme vypočítat průměr a medián.

Oříznutý průměr

Po odstranění k % hodnot z každého konce seřazené datové sady je průměr zbývajících hodnot nazýván **k% oříznutý průměr**.

Výpočet:

1. Pro výpočet oříznutého průměru datové sady nejprve seřadíme daná data vzestupně.
2. Poté odstraníme k % hodnot z každého konce seřazené datové sady, kde k je jakékoliv kladné číslo, jako je 5%, 10% a tak dále.
3. Průměr zbývajících hodnot se nazývá k% oříznutý průměr.

Příklad 7

Zadání: Vypočítejte 10% oříznutý průměr. Následující údaje uvádějí utracené peníze deseti vybranými studenty za knihy v roce 2015.

890 1354 1861 1644 87 5403 1429 1993 938 2176

Řešení:

Pro výpočet ořezaného průměru nejprve seřadíme daná data:

87 890 938 1354 1429 1644 1861 1993 2176 5403

Pro výpočet 10% oříznutého průměru odstraníme 10% hodnot z každého konce seřazených dat: 10% z 10 hodnot = 10*(0.10) = 1.

Tedy odstraníme jednu hodnotu z každého konce seřazených dat. Po odstranění dvou hodnot, jedné z každého konce, nám zůstane následujících osm hodnot:

890 938 1354 1429 1644 1861 1993 2176

Příklad 7: Řešení

$$\sum x = 890+938+1354+1429+1644+1861+1993+2176 = 12\,285$$

$$10\%\ \text{Ořezaný průměr} = \frac{12\,285}{8} = 1535.625 \doteq 1535.63$$

Tedy, pokud odstraníme 10 % hodnot z každého konce seřazených dat pro tento příklad, můžeme říci, že studenti za knihy v roce 2015 utratili v průměru $1535.63.

Vzhledem k tomu, že v této datové sadě lze považovat $87 a $5403 za odlehlé hodnoty, je rozumné tyto dvě hodnoty odstranit a vypočítat ořezaný průměr pro zbývající hodnoty namísto výpočtu průměru ze všech 10 hodnot.

Vážený průměr

Pokud má každá hodnota v datové sadě odlišnou četnost výskytu a každé takové hodnotě je přiřazena určitá váha odpovídající jejímu zastoupení v souboru dat, pak pro určení středu této datové sady použijeme výpočet **váženého průměru**.

Výpočet:

1. Pro výpočet váženého průměru datové sady označíme proměnnou x a váhy w.
2. Sečteme všechny váhy a tuto sumu označíme $\sum w$. Poté vynásobíme každou hodnotu x příslušnou hodnotou w.
3. Součet výsledných součinů dává $\sum xw$. Dělení $\sum xw$ sumou $\sum w$ dává vážený průměr.

=> Vážený průměr $= \frac{\sum(xw)}{\sum w}$

Příklad 8

Zadání: Aneta si letos v červnu čtyřikrát koupila benzín do svého auta. Koupila 10 galonů za cenu 2.60 dolaru za galon, 13 galonů za cenu 2.80 dolaru za galon, 8 galonů za cenu 2.70 dolaru za galon a 15 galonů za cenu 2.75 dolaru za galon. Jaká je průměrná cena, kterou Aneta letos v červnu zaplatila za benzín?

Řešení:

Tabulka: Ceny a množství zakoupeného benzínu

Cena x	Galony benzínu w	xw
2.60	10	26.00
2.80	13	36.40
2.70	8	21.60
2.75	15	41.25
	$\Sigma w = 46$	$\Sigma xw = 125.25$

Příklad 8: Řešení

Proměnná představuje cenu benzínu za galon, označíme ji x. Váhy jsou množství zakoupených galonů, tyto váhy označíme w.

Vynásobíme každou hodnotu x příslušnou hodnotou w a sečtením výsledných hodnot získáme $\sum xw$. Nakonec dělíme $\sum xw$ sumou $\sum w$, abychom našli vážený průměr.

$$\textbf{Vážený průměr} = \frac{\sum xw}{\sum w} = \frac{125.25}{46} = \$2.72$$

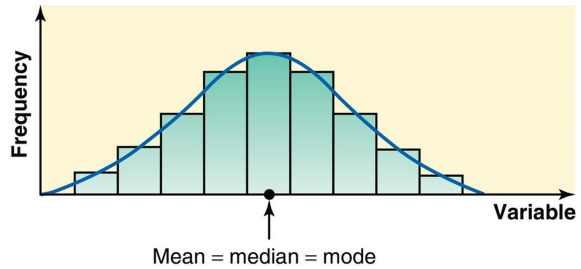
Aneta v červnu v průměru zaplatila 2.72 dolarů za galon benzínu.

Vztahy mezi průměrem, mediánem a modem

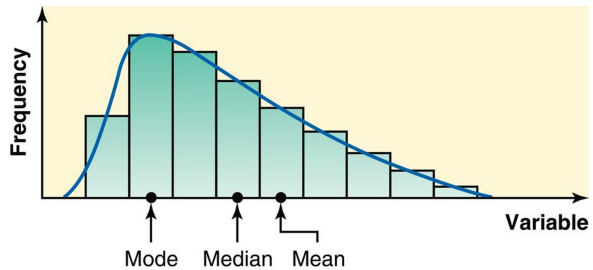
1. Pro **symetrický histogram s jedním vrcholem** (viz obrázek 3.2) jsou hodnoty průměru, mediánu a modu shodné a nacházejí se ve středu rozdělení.
2. U **histogramu zešikmeného doprava** (viz obrázek 3.3) je hodnota průměru největší, hodnota modu nejmenší a hodnota mediánu leží mezi těmito dvěma. (Všimněte si, že modus se vždy vyskytuje v bodě vrcholu.) Hodnota průměru je v tomto případě největší, protože je citlivá na odlehlé hodnoty, které se vyskytují v pravém ocasu. Tyto odlehlé hodnoty táhnou průměr doprava.
3. Pokud je **histogram zešikmen doleva** (viz obrázek 3.4), je hodnota průměru nejmenší a hodnota modu největší, přičemž hodnota mediánu leží mezi těmito dvěma. V tomto případě odlehlé hodnoty v levém ocasu táhnou průměr doleva.

Vztahy mezi průměrem, mediánem a modem

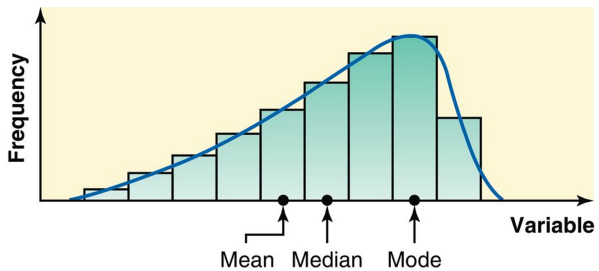
Obrázek: 3.2 Symetrický histogram s jedním vrcholem:



Obrázek: 3.3 Doprava zešikmený histogram:



Obrázek: 3.4 Doleva zešikmený histogram:



Rozpětí

Nalezení rozpětí pro neseskupená data

Rozpětí = Největší hodnota − Nejmenší hodnota

Nevýhody:

- Rozpětí, stejně jako průměr, má nevýhodu, že je ovlivněn odlehlými hodnotami. V důsledku toho není rozpětí dobrým ukazatelem míry rozptylu pro datový soubor, který obsahuje odlehlé hodnoty.
- Jeho výpočet je založen pouze na dvou hodnotách: největší a nejmenší. Při výpočtu rozpětí se ignorují všechny ostatní hodnoty v datovém souboru. Tím pádem není rozpětí velmi uspokojivým ukazatelem rozptylu.

Příklad 9

Zadání: Tabulka udává celkové plochy ve čtverečních mílích čtyř států USA. Najděte rozpětí pro tento soubor dat.

Stát	Arkansas	Louisiana	Oklahoma	Texas
Celková plocha	53 182	49 651	69 903	267 277

Řešení:

$$\begin{aligned} \text{Rozpětí} &= \text{Největší hodnota} - \text{Nejmenší hodnota} \\ &= 267\,277 - 49\,651 \\ &= 217\,626 \text{ čtverečních mil} \end{aligned}$$

Celkové plochy těchto čtyř států se pohybují v rozpětí **217 626 čtverečních mil**.

Rozptyl a směrodatná odchylka

Směrodatná odchylka - nejpoužívanější ukazatel rozptýlenosti dat.

Hodnota směrodatné odchylky ukazuje, jak úzce jsou hodnoty datového souboru shlukovány kolem průměru.

Obecně platí, že nižší hodnota směrodatné odchylky u datového souboru značí, že hodnoty tohoto souboru jsou rozloženy v relativně menším rozpětí kolem průměru. Naopak, vyšší hodnota směrodatné odchylky u datového souboru značí, že hodnoty tohoto souboru jsou rozloženy v relativně větším rozpětí kolem průměru.

Rozptyl a směrodatná odchylka

Směrodatná odchylka je získána jako kladná část druhé odmocniny z **rozptylu**.

Rozptyl vypočítaný pro data celé populace je označen σ^2 (čte se jako sigma na druhou), a rozptyl vypočítaný pro vzorek dat je označen s^2.

Směrodatná odchylka vypočítaná pro data celé populace je označena σ, a směrodatná odchylka vypočítaná pro vzorek dat je označena s.

Rozptyl a směrodatná odchylka

Základní vzorce pro výpočet rozptylu a směrodatné odchylky

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \quad \text{a} \quad s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \quad \text{a} \quad s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

kde σ^2 je rozptyl populace, s^2 je výběrový rozptyl, σ je směrodatná odchylka populace a s je výběrová směrodatná odchylka.

Rozptyl a směrodatná odchylka

Upravené vzorce pro výpočet rozptylu a směrodatné odchylky

$$\sigma^2 = \frac{\sum x^2 - \left(\frac{(\sum x)^2}{N}\right)}{N} \quad \text{a} \quad s^2 = \frac{\sum x^2 - \left(\frac{(\sum x)^2}{n}\right)}{n-1}$$

$$\sigma = \sqrt{\frac{\sum x^2 - \left(\frac{(\sum x)^2}{N}\right)}{N}} \quad \text{a} \quad s = \sqrt{\frac{\sum x^2 - \left(\frac{(\sum x)^2}{n}\right)}{n-1}}$$

kde σ^2 je rozptyl populace, s^2 je výběrový rozptyl, σ je směrodatná odchylka populace a s je výběrová směrodatná odchylka.

Rozptyl a směrodatná odchylka

Dvě fakta:

1. Hodnoty rozptylu a směrodatné odchylky nikdy nejsou záporné. To znamená, že čitatel ve vzorci pro výpočet rozptylu by nikdy neměl dát zápornou hodnotu. Obvykle jsou hodnoty rozptylu a směrodatné odchylky kladné, ale pokud má datový soubor nulovou variabilitu, pak jsou jak rozptyl, tak směrodatná odchylka rovněž nula.
2. Měřící jednotky rozptylu jsou vždy kvadrátem (čtverečním útvarem) měřicích jednotek původních dat. Důvodem je to, že pro výpočet rozptylu jsou původní hodnoty umocňovány na druhou. Měřicí jednotky směrodatné odchylky jsou stejné jako měřicí jednotky původních dat, protože směrodatná odchylka je získána druhou odmocninou rozptylu.

Rozptyl a směrodatná odchylka

Upozornění na numerické chyby:

- Všimněte si, že Σx^2 není totéž co $(\Sigma x)^2$. Hodnota Σx^2 je získána umocněním hodnot x a následným jejich sečtením. Hodnota $(\Sigma x)^2$ je získána umocněním hodnoty Σx.

Příklad 10

Zadání: Nechť x označuje odměny (v milionech dolarů) pro ženy ve funkci generální ředitelky amerických společností v roce 2014. Výpočet Σx a Σx^2 je uveden v tabulce 3.6.

Table 3.6 Calculation of Σx and Σx^2

x	x^2
19.3	372.49
16.2	262.44
19.6	384.16
19.3	372.49
33.7	1135.69
21.0	441.00
22.5	506.25
16.9	285.61
28.7	823.69
42.1	1772.41
22.2	492.84
$\Sigma x = 261.5$	$\Sigma x^2 = 6849.07$

Určete výběrový rozptyl a výběrovou směrodatnou odchylku pro x.

Příklad 10: Řešení

Krok 1. **Určete Σx:** Součet hodnot v 1. sloupci tabulky je 261.5.

Krok 2. **Určete Σx^2:** Součet hodnot v 2. sloupci tabulky je 6 849.07.

Krok 3. **Určete výběrový rozptyl:**

$$s^2 = \frac{\sum x^2 - \left(\frac{(\sum x)^2}{n}\right)}{n-1} = \frac{6849.07 - \left(\frac{(261.5)^2}{11}\right)}{11-1} = \frac{632.5018}{10} = 63.2502$$

Krok 4. **Získání výběrové směrodatné odchylky:**

$$s = \sqrt{\frac{\sum x^2 - \left(\frac{(\sum x)^2}{n}\right)}{n-1}} = \sqrt{63.2502} = 7.952999 \approx 7.95 \text{ milionů}$$

=> Směrodatná odchylka odměn z roku 2014 těchto 11 žen ve funkci generálních ředitelek amerických společností je 7.95 milionů dolarů.

Příklad 11

Zadání: Máme zaznamenány výdělky pro všech šest zaměstnanců malé společnosti za rok 2015 (v tisících dolarech před zdaněním). Vypočtěte rozptyl a směrodatnou odchylku těchto dat.

Řešení: Nechť x označuje výdělky před zdaněním zaměstnance této společnosti za rok 2015. Hodnoty Σx a Σx^2 jsou vypočítány v tabulce:

	Z1	Z2	Z3	Z4	Z5	Z6	Σ
x	88.50	108.40	65.50	52.50	79.80	54.60	449.30
x^2	7832.25	11750.56	4290.25	2756.25	6368.04	2981.16	35978.51

$$\sigma^2 = \frac{\sum x^2 - \left(\frac{(\sum x)^2}{N}\right)}{N} = \frac{35\,978.51 - \left(\frac{(449.30)^2}{6}\right)}{6} = 388.90$$

$$\sigma = \sqrt{388.90} = 19.721 \text{ tisíc}$$

Populační směrodatná odchylka výdělků za rok 2015 všech šesti zaměstnanců této společnosti je 19.721 tisíc dolarů.

Koeficient variace

- Jednou z nevýhod směrodatné odchylky jako míry rozptylu je, že je to míra absolutní variability, a ne relativní variability.
- Někdy můžeme potřebovat porovnat variabilitu dvou různých datových souborů, které mají různé měrné jednotky. V takových případech je preferována míra relativní variability. Jednou takovou mírou je **koeficient variace**.

Koeficient variace

Koeficient variace vyjadřuje směrodatnou odchylku jako procento průměru a vypočítává se následovně:

Pro data z populace:

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Pro data ze vzorku:

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Poznámka: Koeficient variace nemá žádné jednotky měření, protože je vždy vyjádřen jako procento.

Příklad 12

Zadání: Roční platy všech zaměstnanců pracujících pro velkou společnost mají průměr 72 350 USD a směrodatnou odchylku 12 820 USD. Počet let vzdělání stejných zaměstnanců má průměr 15 let a směrodatnou odchylku 2 roky. Je relativní rozptýlenost dat v platech vyšší nebo nižší než rozptýlenost dat u let vzdělání těchto zaměstnanců? Odpovězte na otázku výpočtem koeficientu variace pro každou proměnnou.

Řešení:

CV pro platy = $\frac{\sigma}{\mu} \cdot 100\% = \frac{12\,820}{72\,350} \cdot 100\% = 17.72\%$

CV pro roky vzdělání = $\frac{\sigma}{\mu} \cdot 100\% = \frac{2}{15} \cdot 100\% = 13.33\%$

Vzhledem k tomu, že koeficient variace pro plat (17.72%) má vyšší hodnotu než koeficient variace pro roky vzdělání (13.33%), mají platy vyšší relativní variaci než roky vzdělání.

Parametry populace a statistiky vzorku

Číselná míry, jako je průměr, medián, modus, rozpětí, rozptyl nebo směrodatná odchylka, vypočítaná pro soubor dat populace, se nazývá **parametr populace**, nebo jednoduše **parametr**.

Shrnutá míra vypočítaná pro vzorek dat se nazývá **statistika vzorku**, nebo jednoduše **statistika**.

Chebyshevova věta

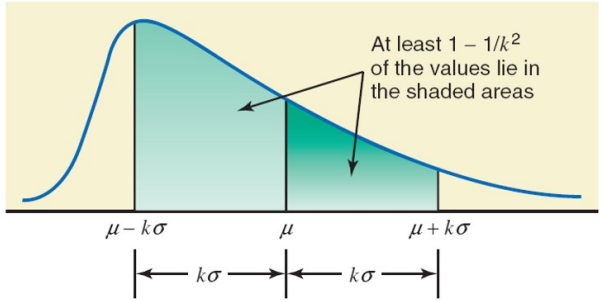
Pro libovolné číslo k větší než 1, leží alespoň $(1 - \frac{1}{k^2})$ hodnot datové sady do vzdálenosti k směrodatných odchylek od průměru.

Tabulka: Plochy pod distribuční křivkou dle Chebyshevovy věty

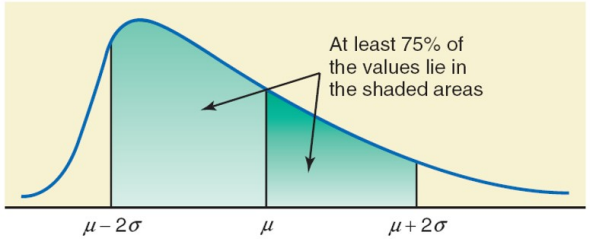
k	**Interval**	$1 - \frac{1}{k^2}$	**Minimální plocha v rámci k směrodatných odchylek**
1.5	$\mu \pm 1.5\sigma$	$1 - \frac{1}{1.5^2} = 1 - .44 = .56$	56%
2.0	$\mu \pm 2\sigma$	$1 - \frac{1}{2^2} = 1 - .25 = .75$	75%
2.5	$\mu \pm 2.5\sigma$	$1 - \frac{1}{2.5^2} = 1 - .16 = .84$	84%
3.0	$\mu \pm 3\sigma$	$1 - \frac{1}{3.0^2} = 1 - .11 = .89$	89%

Chebyshevova věta

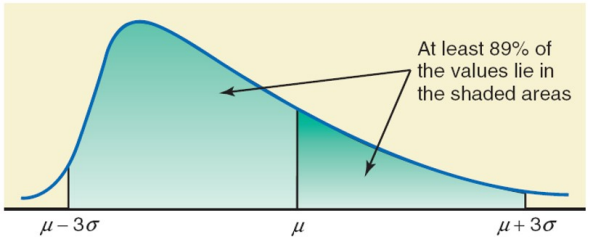
Obrázek: 3.5 Chebyshevova věta



Obrázek: 3.6 Chebyshevova věta: 2x směrodatná odchylka



Obrázek: 3.7 Chebyshevova věta: 3x směrodatná odchylka



Příklad 13

Zadání: Průměrný systolický krevní tlak u 4000 testovaných žen byl stanoven na 187 mm Hg se směrodatnou odchylkou 22. Použitím Chebyshevovy věty zjistěte, alespoň kolik procent žen v této skupině má systolický krevní tlak mezi 143 a 231 mm Hg.

Řešení:

Nechť $\mu = 187$ a $\sigma = 22$ jsou průměr a směrodatná odchylka systolického krevního tlaku těchto žen.

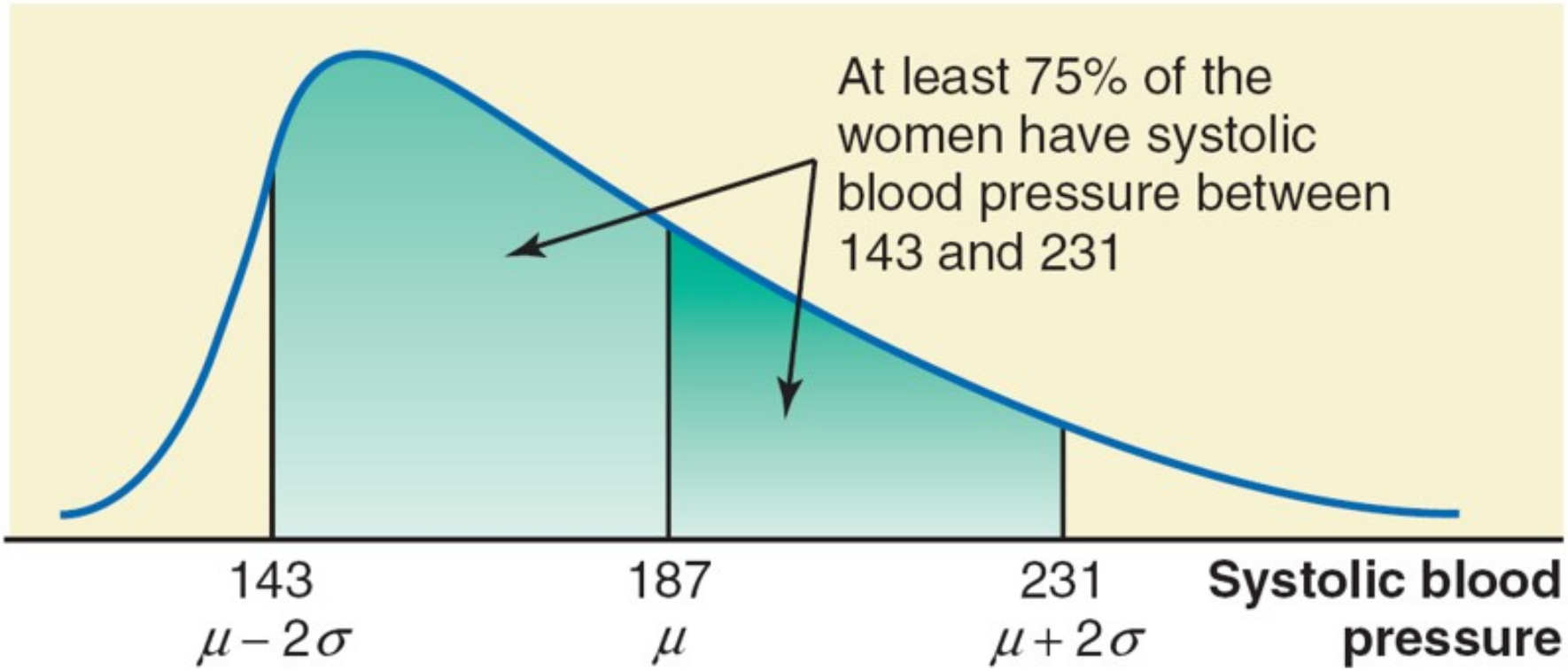
$$|143 - \mu| = |143 - 187| = -44 \quad \leftrightarrow \quad |231 - \mu| = |231 - 187| = 44$$

Hodnota k je získána dělením vzdálenosti mezi průměrem a každým bodem a směrodatnou odchylkou. Tedy $k = \frac{44}{22} = 2$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = .75 \text{ nebo } 75\%$$

Takže podle Chebyshevovy věty má alespoň 75% žen systolický krevní tlak mezi 143 a 231 mm Hg.

Příklad 13: Grafické řešení

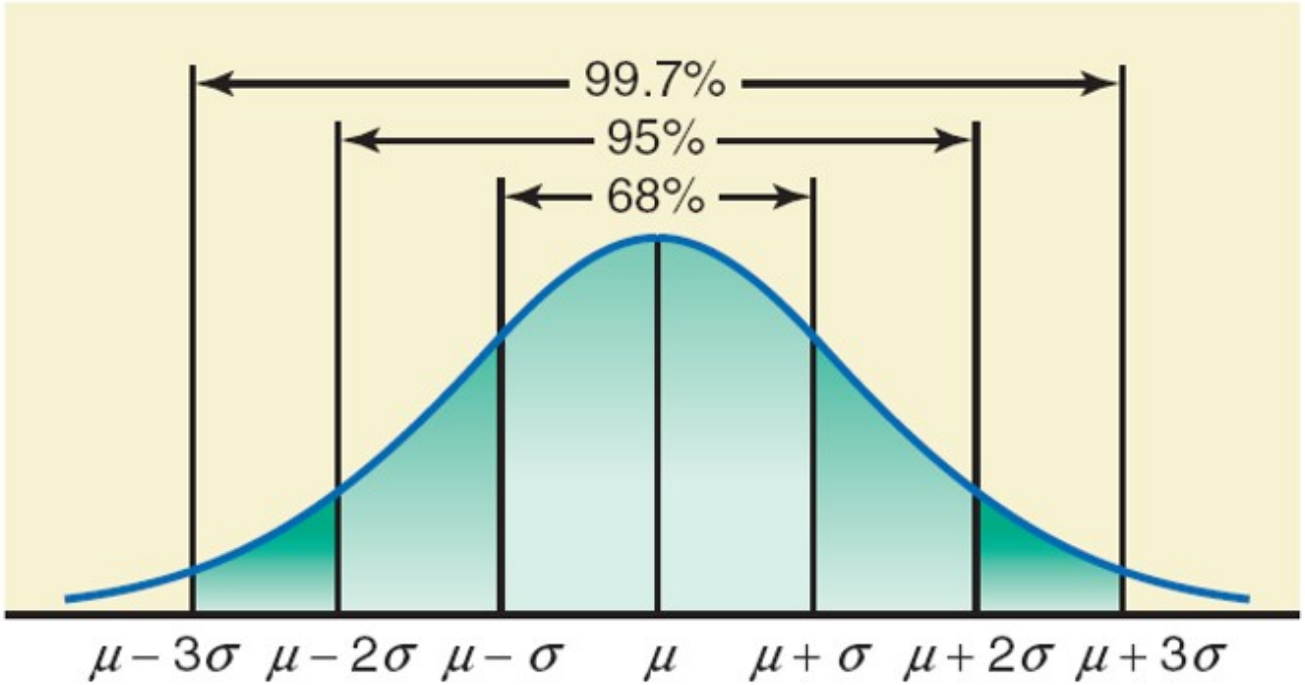


Empirické pravidlo

Pro distribuci s tvarem zvonu platí přibližně:

- 68% pozorování leží do jedné směrodatné odchylky od průměru.
- 95% pozorování leží do dvou směrodatných odchylek od průměru.
- 99.7% pozorování leží do tří směrodatných odchylek od průměru.

Empirické pravidlo graficky



Interval	Aproximativní plocha
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

Příklad 14

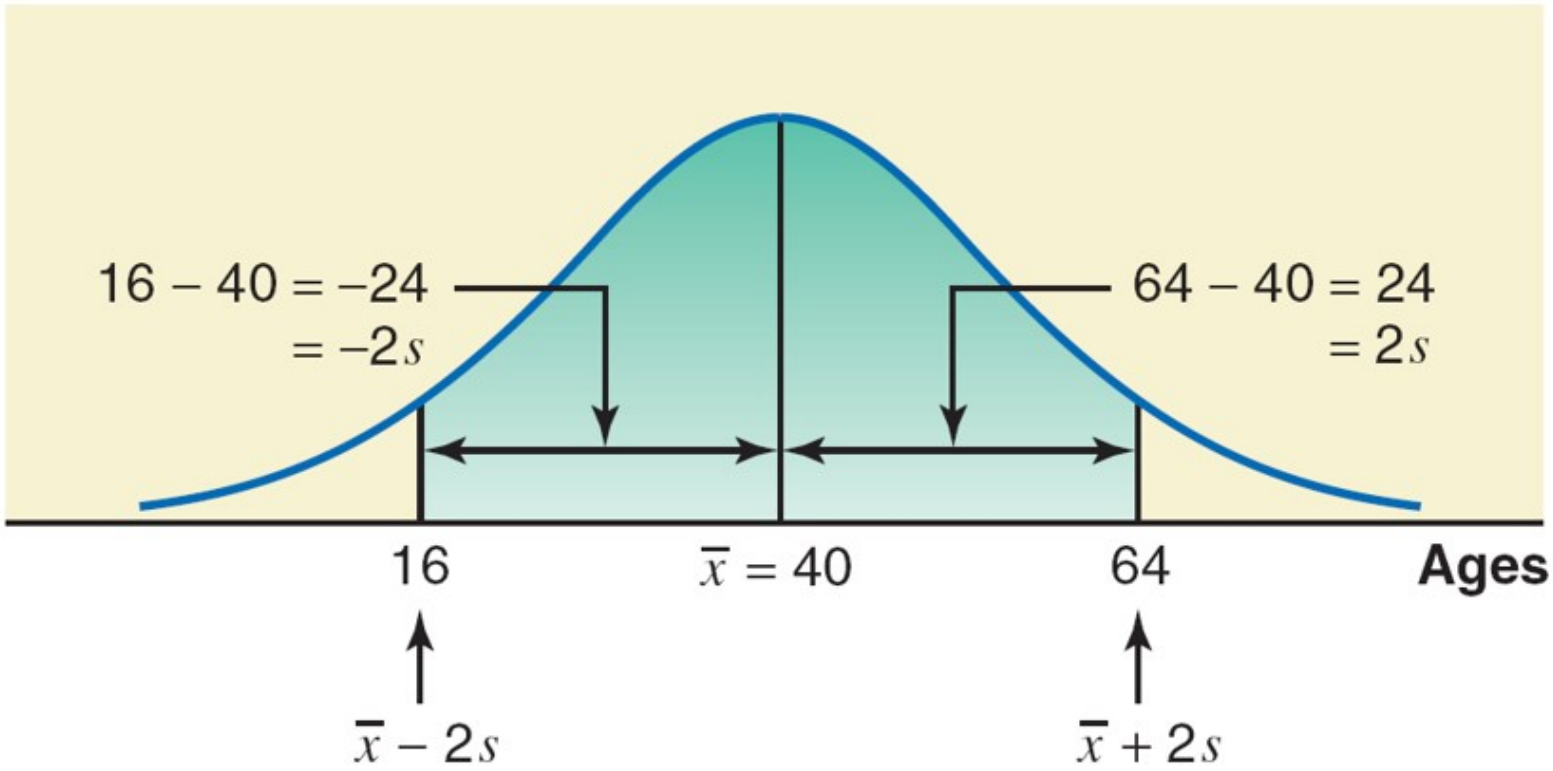
Zadání: Distribuce věku vzorku 5000 osob má tvar zvonu s průměrem 40 let a směrodatnou odchylkou 12 let. Určete přibližné procento osob, které jsou ve věku 16 až 64 let.

Řešení:

Z poskytnutých informací známe tvar rozdělení a jeho parametry $\bar{x} = 40$ a $s = 12$ let. Dále jsme schopni dopočítat, že každý z obou bodů, 16 a 64, je od průměru vzdálen 24 jednotek. Spočteme konstantu k jako k = 24/12 = 2.

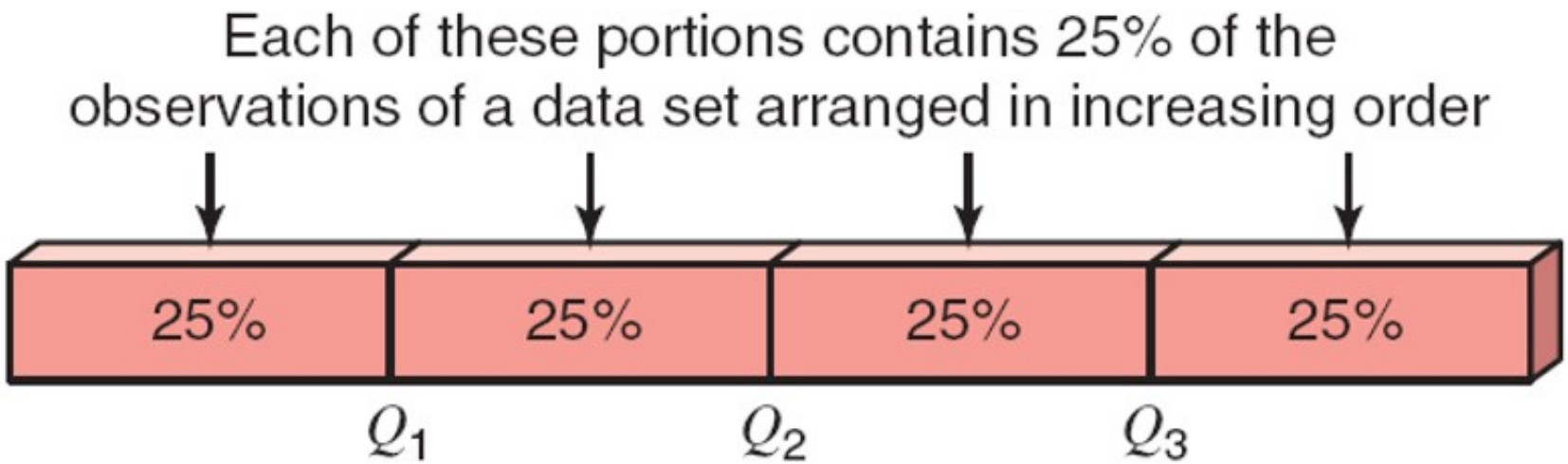
Protože plocha (pod křivkou ve tvaru zvonu) je ve vzdálenosti dvou směrodatných odchylek od průměru přibližně 95 %, tak přibližně 95 % lidí ve vzorku je ve věku 16 až 64 let.

Příklad 14: Grafické řešení



Kvartily a mezikvartilové rozpětí

Kvartily jsou tři shrnující ukazatele, které dělí seřazený soubor dat na čtyři stejné části. Druhý kvartil je to samé co medián datového souboru. První kvartil je hodnota středního členu mezi pozorováními, která jsou menší než medián, a třetí kvartil je hodnota středního členu mezi pozorováními, která jsou větší než medián.



Kvartily a mezikvartilové rozpětí

Rozdíl mezi třetím a prvním kvartilem udává **mezikvartilové rozpětí**.

Výpočet mezikvartilového rozpětí

IQR = Mezikvartilové rozpětí = Q3 − Q1

Příklad 15

Byl vybrán vzorek 12 studentů dojíždějících z koleje na vysokou školu. Následující data udávají typickou jednosměrnou dobu dojíždění (v minutách) pro těchto 12 studentů:

$$29, 14, 39, 17, 7, 47, 63, 37, 42, 18, 24, 55$$

(a) Najděte hodnoty tří kvartilů.

(b) Kde spadá doba dojíždění 47 minut ve vztahu ke třem kvartilům?

(c) Najděte mezikvartilové rozpětí.

Příklad 15: Řešení a)

Krok 1. Nejdříve seřadíme daná data vzestupně takto:

$$7, 14, 17, 18, 24, 29, 37, 39, 42, 47, 55, 63$$

Krok 2. Najdeme druhý kvartil, který je zároveň mediánem. V celkovém počtu 12 datových hodnot je medián mezi šestým a sedmým členem. Tedy medián a tím pádem druhý kvartil je dán průměrem šestého a sedmého hodnoty v seřazené datové sadě, to je průměr 29 a 37. Druhý kvartil (medián) je tedy:

$$Q_2 = \frac{29 + 37}{2} = 33$$

Příklad 15: Řešení a)

Krok 3. Najdeme medián hodnot dat, které jsou menší než Q_2, a to nám dá hodnotu prvního kvartilu. Hodnoty, které jsou menší než Q_2, jsou: $7, 14, 17, 18, 24, 29$. Hodnota, která dělí těchto šest datových hodnot na dvě stejné části, je dána průměrem dvou prostředních hodnot, 17 a 18. Tedy první kvartil je:

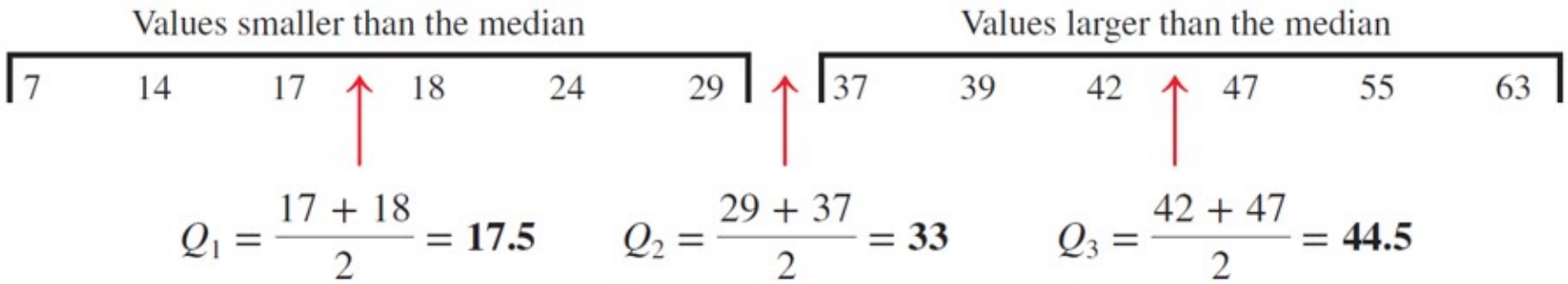
$$Q_1 = \frac{17 + 18}{2} = 17.5$$

Krok 4. Najdeme medián hodnot dat, které jsou větší než Q_2, a to nám dá hodnotu třetího kvartilu. Hodnoty, které jsou větší než Q_2, jsou: $37, 39, 42, 47, 55, 63$. Hodnota, která dělí těchto šest datových hodnot na dvě stejné části, je dána průměrem dvou prostředních hodnot, 42 a 47. Tedy třetí kvartil je:

$$Q_3 = \frac{42 + 47}{2} = 44.5$$

Příklad 15: Řešení a)

Řešení a) Nyní můžeme shrnout výpočet všech tří kvartilů:



Hodnota $Q_1 = 17.5$ minut indikuje, že 25% z 12 studentů v tomto vzorku dojíždí za méně než 17.5 minut a 75% z nich dojíždí za více než 17.5 minut. Podobně, $Q_2 = 33$ naznačuje, že polovina z 12 studentů dojíždí za méně než 33 minut a druhá polovina za více než 33 minut. Hodnota $Q_3 = 44.5$ minut ukazuje, že 75% těchto 12 studentů v tomto vzorku dojíždí za méně než 44.5 minut a 25% z nich dojíždí za více než 44.5 minut.

Příklad 15: Řešení b) + c)

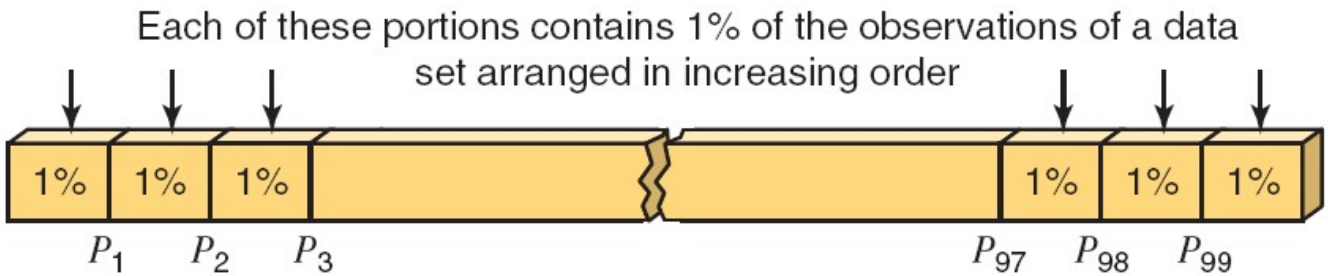
Řešení b) Pohledem na pozici měření 47 minut můžeme konstatovat, že tato hodnota spadá do nejvyšších 25% dob dojíždění.

Řešení c) Mezikvartilové rozpětí je dáno rozdílem mezi hodnotami třetího a prvního kvartilu. Tedy

$$\text{IQR} = \text{Mezikvartilové rozpětí} = Q_3 - Q_1 = 44.5 - 17.5 = 27 \text{ minut}$$

Percentily a percentilové pořadí

Percentil dělí seřazené hodnoty v datové sadě na sto stejně velkých částí. Existují percentily od 1 do 99, které poskytují detailní pohled na rozdělení dat.



Výpočet percentilů

Aproximativní hodnota k-tého percentilu, označovaného P_k, je $P_k =$ Hodnota $\left(\frac{k \cdot n}{100}\right)$ -tého členu v seřazeném datovém souboru, kde k označuje číslo percentilu a n reprezentuje velikost vzorku.

Příklad 16

Zadání: Vyjděme ze zadání Příkladu 15. Mějme opět 12 studentů a jejich čas dojíždění na vysokou školu:

$$29, 14, 39, 17, 7, 47, 63, 37, 42, 18, 24, 55.$$

Najděte hodnotu 70. percentilu. Stručně vysvětlete, co 70. percentil znamená.

Příklad 16: Řešení

Krok 1. Nejprve seřadíme zadaná data vzestupně takto:

$$7, 14, 17, 18, 24, 29, 37, 39, 42, 47, 55, 63$$

Krok 2. Najdeme $\left(\frac{k \cdot n}{100}\right)$ -tý prvek.

$$\frac{k \cdot n}{100} = \frac{(70) \cdot (12)}{100} = 8.4 \rightarrow \text{9-tý prvek}$$

Zde n=12 a k=70, protože chceme najít 70. percentil u 12 pozorování. Náš hledaný 70. percentil, P_{70}, je dán hodnotou 9. prvku v seřazené datové sadě. Poznamenejme, že jsme 8.4 zaokrouhlili nahoru na 9, což je standardní postup při výpočtu percentilu.

Příklad 16: Řešení

Krok 3. Najdeme hodnotu 9. prvku v seřazených datech. To nám dá hodnotu 70. percentilu, P_{70}.

$$7,\ 14,\ 17,\ 18,\ 24,\ 29,\ 37,\ 39,\ \mathbf{42},\ 47,\ 55,\ 63$$

$$P_{70} = \text{Hodnota 9. prvku} = \mathbf{42\ minuty}$$

=> Můžeme konstatovat, že přibližně 70 % z těchto 12 studentů dojíždí za méně nebo rovno 42 minutám.

Percentily a percentilové pořadí

Percentilové pořadí je hodnota vyjadřující pozici daného pozorování ve vztahu k ostatním hodnotám v datové sadě, kdy data jsou uspořádána od nejmenší po největší. Konkrétně percentilové pořadí určitého pozorování ukazuje, jaké procento hodnot v datové sadě je menší než dané pozorování.

Nalezení hodnoty percentilového pořadí

$$\text{Percentilové pořadí } x_i = \frac{\text{Počet hodnot menších než } x_i}{\text{Celkový počet hodnot v datové sadě}} \cdot 100\%$$

Příklad 17

Zadání: Vyjděme ze zadání Příkladu 15. Mějme opět 12 studentů a jejich čas dojíždění na vysokou školu:

$$29, 14, 39, 17, 7, 47, 63, 37, 42, 18, 24, 55.$$

Najděte percentilové pořadí 42 minut. Stručně interpretujte co tato hodnota věcně znamená.

Příklad 17: Řešení

Krok 1. Nejprve seřadíme zadaná data vzestupně:

$$7, 14, 17, 18, 24, 29, 37, 39, 42, 47, 55, 63$$

Krok 2. Zjistíme, kolik datových hodnot je menších než 42. V seřazených datech výše je osm datových hodnot, které jsou menší než 42.

Krok 3. Nalezneme percentilové pořadí 42 minut. Vzhledem k tomu, že 8 z 12 hodnot v dané datové sadě je menších než 42:

$$\text{Percentilové pořadí } 42 = \frac{8}{12} \cdot 100\% = 66.67\%$$

=> Zaokrouhlením této odpovědi na nejbližší celé číslo můžeme konstatovat, že asi 67% studentů v tomto vzorku dojíždí méně než 42 minut.

Krabicový diagram (Boxplot)

Krabicový diagram (Boxplot) je grafické znázornění rozdělení datové sady, které ukazuje její střed, rozptyl a šikmost. Je užitečný pro vizuální identifikaci rozpětí dat, mediánu, kvartilů a odlehlých pozorování.

- **Krabice (box):** Vnitřní krabice zobrazuje rozpětí mezikvartilového rozpětí (IQR), který je rozdílem mezi třetím kvartilem (Q3) a prvním kvartilem (Q1). Střed krabice označuje medián (druhý kvartil, Q2) datové sady.
- **Vousy (whiskers):** Čáry, které se táhnou z krabice směrem ven k extrémním hodnotám dat. Obvykle končí na hodnotě, která je v rámci 1.5x IQR od prvního nebo třetího kvartilu. Hodnoty ležící mimo toto rozpětí jsou často považovány za outliery.
- **Odlehlé pozorování:** Jednotlivé body, které leží mimo vousy, mohou být označeny jako potenciální odlehlé pozorování.

Příklad 18

Následující data zahrnují příjmy (v tisících dolarů) pro vzorek 12 domácností.

$$75, 69, 84, 112, 74, 104, 81, 90, 94, 144, 79, 98$$

Sestavte krabicový diagram pro tato data.

Příklad 18: Řešení

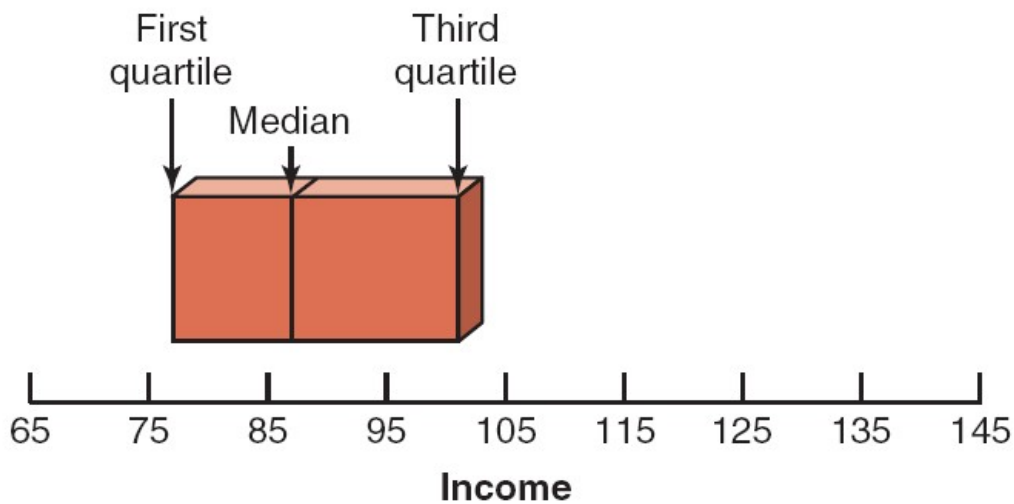
Krok 1. Nejprve seřadíme data vzestupně a vypočítáme medián, prvního kvartil, třetí kvartil a mezikvartilové rozpětí. Seřazená data jsou: 69, 74, 75, 79, 81, 84, 90, 94, 98, 104, 112, 144.
Medián = $\frac{84+90}{2} = 87$
Q_1 = $\frac{75+79}{2} = 77$
Q_3 = $\frac{98+104}{2} = 101$
IQR = $Q_3 - Q_1$ = $101 - 77 = 24$

Krok 2. Určíme body, které jsou 1.5·IQR pod Q_1 a 1.5·IQR nad Q_3.
1.5 krát IQR = $1.5 \cdot 24 = 36$
Dolní mez vousů = $Q_1 - 36 = 77 - 36 = 41$
Horní mez vousů = $Q_3 + 36 = 101 + 36 = 137$

Příklad 18: Řešení

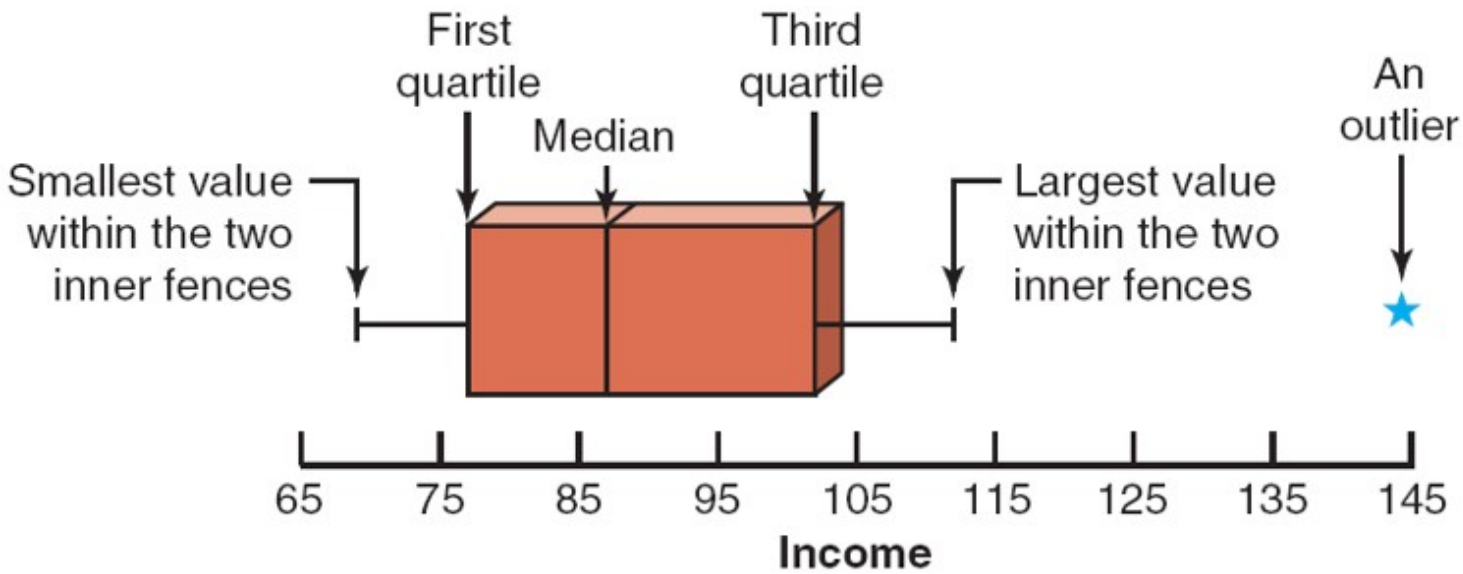
Krok 3. Určete nejmenší a největší hodnoty v daném datovém souboru mezi dvěma vnitřními vousy.
Nejmenší hodnota mezi dolní a horní mezí vousů = 69
Největší hodnota mezi dolní a horní mezí vousů = 112

Krok 4. Nakreslete vodorovnou čáru a označte na ní úrovně příjmů tak, aby byly pokryty všechny hodnoty v dané datové sadě. Výsledek tohoto kroku je znázorněn na obrázku:



Příklad 18: Řešení

Krok 5. Nakreslete dvě čáry, které spojí body nejmenší a největší hodnoty mezi dolní a horní mezí vousů s krabicí. V tomto příkladu jsou tyto hodnoty 69 a 112. Tím je krabicový diagram dokončen, výsledek je znázorněn na obrázku



Shrnutí přednášky:

Míry středu pro neseskupená data
- Průměr, medián, modus
- Alternativní výpočty průměru
- Vztahy mezi průměrem, mediánem a modem

Ukazatele variability pro neseskupená data
- Rozpětí
- Rozptyl a směrodatná odchylka
- Parametry populace a statistiky vzorku

Využití směrodatné odchylky
- Chebyshevova věta
- Empirické pravidlo

Charakteristiky polohy
- Kvartily a mezikvartilové rozpětí
- Percentily a percentilové pořadí

Krabicový diagram

Co si nastudovat na následující týden?

Příprava na cvičení: Leaflet 04
Koncepty a procedury, cv. 04, kap. 03

Povinná literatura: Mann (2016), Kapitola 4

Děkuji za pozornost!