Kapitola 1

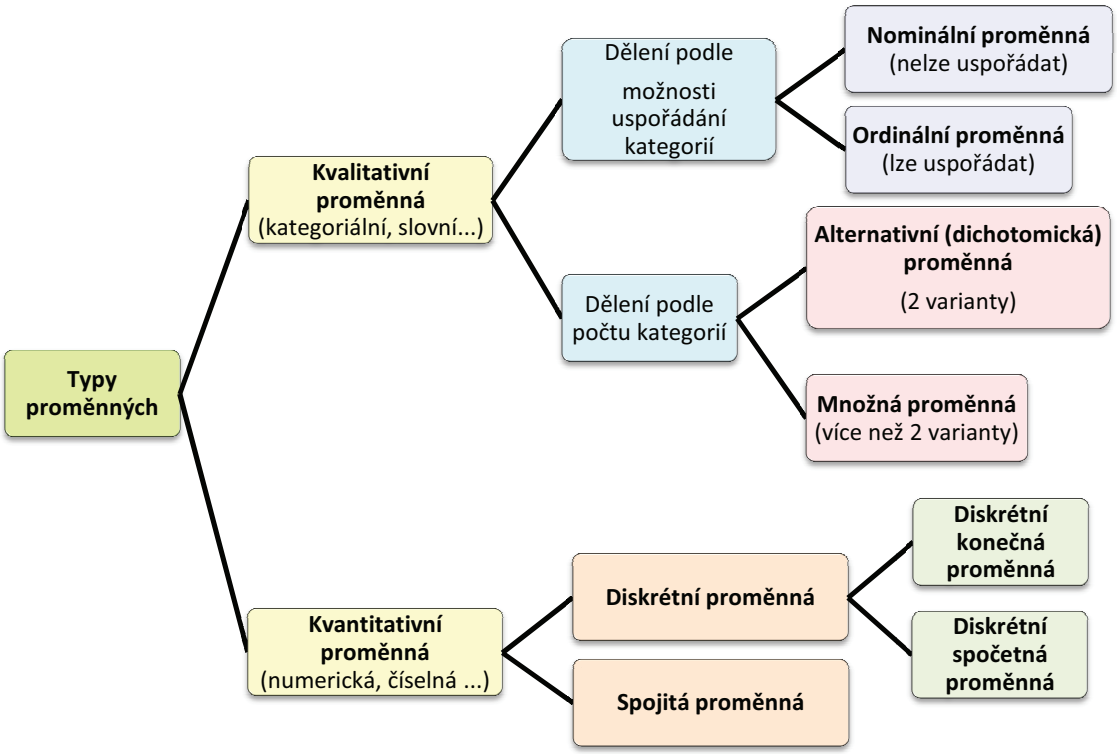
Explorační analýza proměnných

Cíle

Po prostudování této kapitoly budete znát

- základní pojmy explorační (popisné) statistiky
- typy datových proměnných
- statistické charakteristiky a grafickou demonstraci kvalitativních proměnných
- statistické charakteristiky a grafickou demonstraci kvantitativních proměnných

Údaje, které u výběrového souboru sledujeme, nazýváme **proměnné** (znaky, veličiny) a jejich jednotlivé hodnoty **varianty** proměnné. **Explorační (popisná) statistika** bývá prvním krokem k odhalení informací skrytých ve velkém množství proměnných a jejich variant. To znamená uspořádání proměnných do názornější formy a jejich popis několika málo hodnotami, které by obsahovaly co největší množství informací obsažených v původním souboru. Vzhledem k tomu, že způsob zpracování proměnných závisí především na jejich typu, seznámíme se nyní se základním dělením proměnných do různých kategorií. Toto dělení je prezentováno na následujícím obrázku.



Obr. 1.1: Demonstrace základních proměnných

- **Proměnná kvalitativní** (kategoriální, slovní,...) je proměnná, kterou nemůžeme měřit, můžeme ji pouze zařadit do tříd. Varianty kvalitativní proměnné nazýváme kategoriemi, jsou vyjádřeny slovně a podle vztahu mezi jednotlivými kategoriemi se dělí na dvě základní podskupiny.
  - **Proměnná nominální** nabývá rovnocenných variant; nelze je smysluplně porovnávat ani seřadit (např. pohlaví, národnost, značka hodinek...)
  - **Proměnná ordinální** tvoří přechod mezi kvalitativními a kvantitativními proměnnými; jednotlivým variantám lze přiřadit pořadí a vzájemně je porovnávat nebo seřadit (např. známka ve škole, velikost oděvů (S, M, L))

Jiným způsobem dělení kvalitativních proměnných je dělení podle počtu variant, jichž proměnné mohou nabývat.

- **Proměnná alternativní** nabývá pouze dvou různých variant (např. pohlaví, zapnuto/vypnuto, živý/mrtvý...)
- **Proměnná množná** nabývá více než dvou různých variant (např. vzdělání, jméno, barva očí...)

- **Proměnné kvantitativní** jsou proměnné měřitelné. Jsou vyjádřeny číselně a dělí se na
  - **Proměnné diskrétní** nabývající konečného nebo spočetného množství variant.
    - **Proměnné diskrétní konečné** – nabývají konečného počtu variant (např. známka z matematiky)
    - **Proměnné diskrétní spočetné** – nabývají spočetného množství variant (např. věk v letech, výška v centimetrech, váha v kilogramech...)
  - **Proměnné spojité** nabývající libovolných hodnot z R nebo z nějaké podmnožiny R (např. výška, váha, vzdálenost měst...)

Průvodce studiem

Tak, základní definice máme za sebou, proto můžeme přejít k věcem praktičtějším. Představte si situaci, že máte k dispozici statistický soubor o poměrně velkém rozsahu a stojíte před otázkou co s ním, jak jej co nejvýstižněji popsat a znázornit. Číselné hodnoty, kterými takovýto rozsáhlý soubor hodnot proměnné „nahradíme", postihují základní vlastnosti tohoto souboru a my jim budeme říkat ***statistické charakteristiky (statistiky)****. V následujících kapitolách se dozvíte, jak určit statistické charakteristiky pro různé typy proměnných a jak rozsáhlejší statistické soubory znázornit. Jdeme na to!*

1.1 Statistické charakteristiky kvalitativních proměnných

V tuto chvíli již víme, že kvalitativní proměnná má dva základní typy – nominální a ordinální.

1.1.1 Nominální proměnná

Nominální proměnná nabývá v rámci souboru různých, avšak rovnocenných kategorií. Počet těchto kategorií nebývá příliš vysoký, a proto první statistickou charakteristikou, kterou k popisu proměnné použijeme je četnost.

- **Četnost n_i** (absolutní četnost, angl. „frequency“) je definována jako počet výskytu dané varianty kvalitativní proměnné.

  V případě, že kvalitativní proměnná ve statistickém souboru o rozsahu n hodnot nabývá k různých variant, jejichž četnosti označíme $n_1, n_2, \ldots, n_k$, musí zřejmě platit

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n.$$

Chceme-li vyjádřit, jakou část souboru tvoří proměnné s některou variantou, použijeme pro popis proměnné relativní četnost.

- **Relativní četnost p_i** (angl. „relative frequency“) je definována jako

$$p_i = \frac{n_i}{n}, \qquad \text{popř. } p_i = \frac{n_i}{n} \cdot 100\,[\%].$$

  (Druhý vzorec použijeme v případě, chceme-li relativní četnost vyjádřit v procentech.) Pro relativní četnosti musí platit

$$p_1 + p_2 + \ldots + p_k = \sum_{i=1}^{k} p_i = 1, \quad \text{popř. } 100\,\%.$$

Při zpracování kvalitativní proměnné je vhodné četnosti i relativní četnosti uspořádat do tzv. **tabulky rozdělení četnosti** (angl. „frequency table“) – Tab. 1.1.

Tab. 1.1: Tabulka rozdělení četností pro nominální proměnnou

TABULKA ROZDĚLENÍ ČETNOSTI		
Hodnoty x_i	**Absolutní četnosti**	**Relativní četnosti**
	n_i	**p_i**
x_1	n_1	p_1
x_2	n_2	p_2
x_k	n_k	p_k
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$

Poslední charakteristikou, kterou si pro popis nominální proměnné uvedeme, je modus.

- **Modus** definujeme jako název varianty proměnné vykazující nejvyšší četnost.

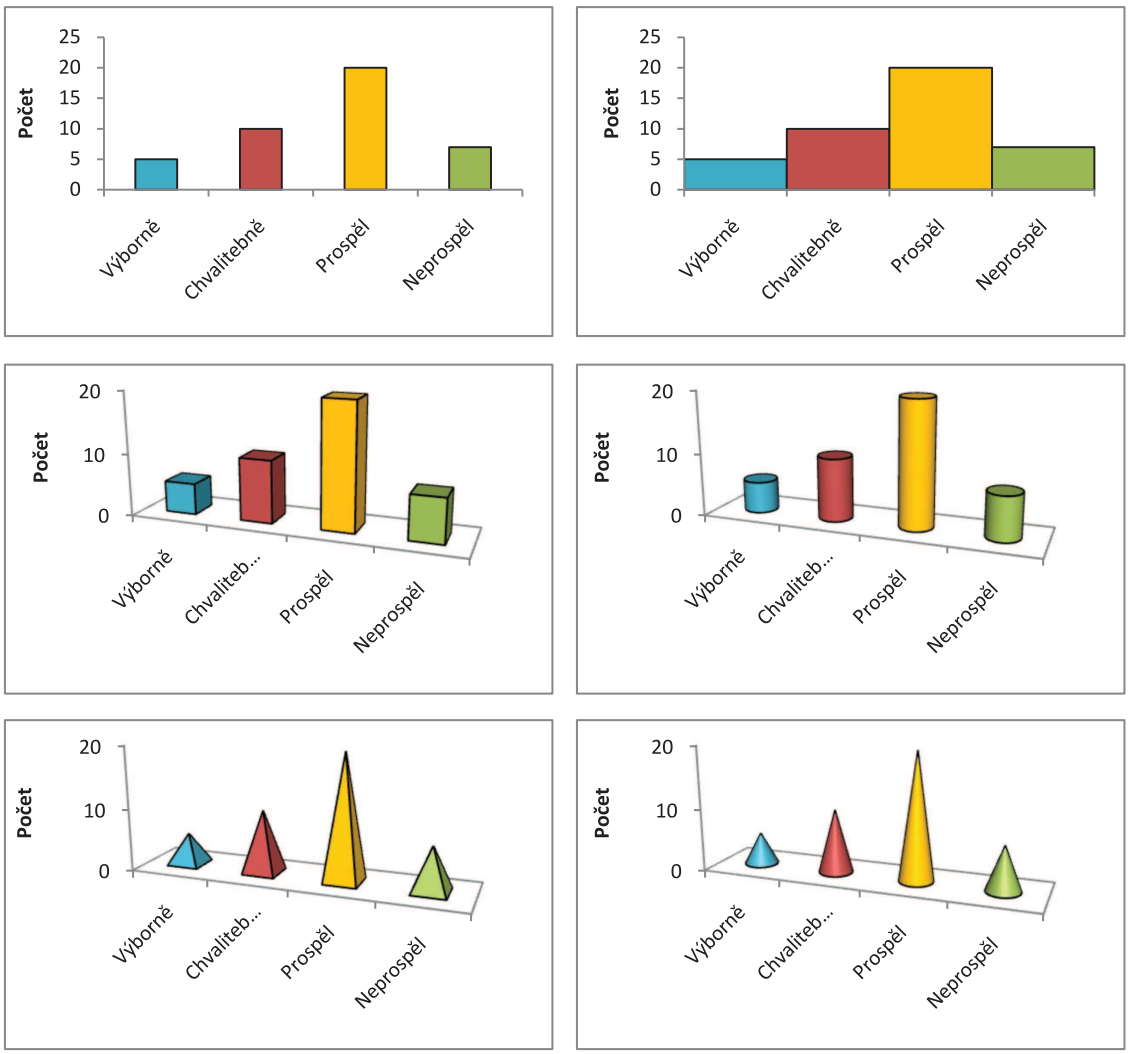
  Modus tedy můžeme chápat jako typického reprezentanta souboru. V případě, že se ve statistickém souboru vyskytuje více variant s maximální četnosti, modus neurčujeme.

1.1.2 Grafické znázornění kvalitativní proměnné

Pro větší názornost analýzy proměnných se ve statistice často užívají **grafy**. Pro nominální proměnnou jsou to tyto dva typy:

- **Histogram** (také sloupcový graf, angl. „bar chart“)
- **Výsečový graf** (také koláčový graf, angl. „pie chart“)

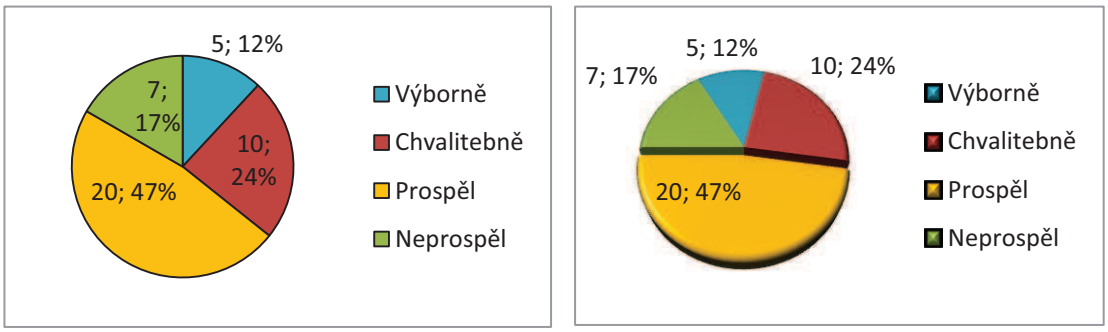
Histogram je klasickým grafem, v němž na jednu osu vynášíme varianty proměnné a na druhou osu jejich četnosti. Jednotlivé hodnoty četností jsou pak zobrazeny jako výšky sloupců (obdélníků, popř. hranolů, kuželů...)



Obr. 1.2: Ukázky histogramů

Výsečový graf prezentuje relativní četnosti jednotlivých variant proměnné, při-

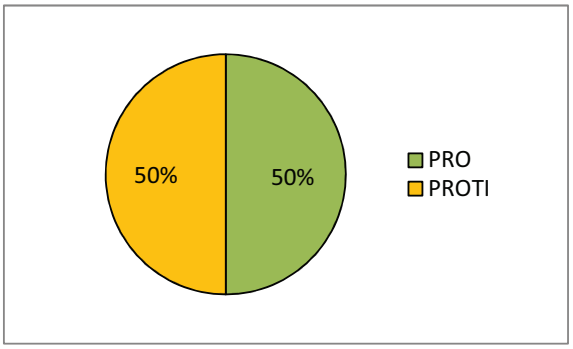
čemž jednotlivé relativní četnosti jsou úměrně reprezentovány plochami příslušných kruhových výsečí. (Změnou kruhu na elipsu dojde k trojrozměrnému efektu.)



Obr. 1.3: Ukázky výsečových grafů

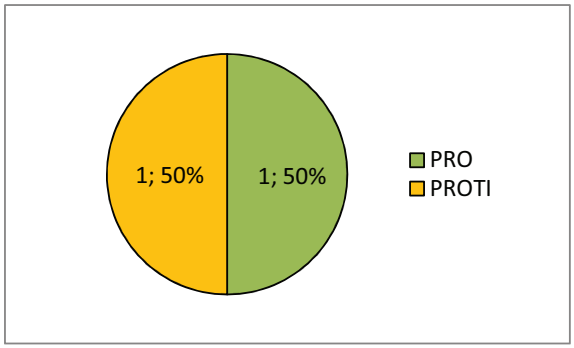
POZOR!!! V případě výsečového grafu si dejte zvláštní pozor na popis grafu. Jednotlivé výseče nestačí označit relativními četnostmi bez uvedení četnosti absolutních, popř. bez uvedení celkového počtu pozorování, to by mohlo vést k matení (ať už záměrnému nebo nechtěnému) toho, komu je graf určen. Zamyslete se nad následující ukázkou.

Příklad k zamyšlení: Minulý týden jsme zpracovali anketu týkající se názoru na zavedení školného na vysokých školách. Výsledky prezentuje následující graf.



Obr. 1.4: Chybná prezentace výsečového grafu

Co vy na to? Zajímavé výsledky, že? A věřte, nevěřte – pravdivé. A nyní graf doplníme tak, jak jsme doporučili.



Obr. 1.5: Správná prezentace výsečového grafu

Co si myslíte nyní? Z druhého grafu je patrné, že byli dotazování pouze dva lidé, jeden byl pro a druhý proti. Jaká je vypovídací schopnost takové ankety? Jaký je nyní Váš názor na prezentované výsledky? A závěr? Vytvářejte pouze takové grafy, jejichž interpretace je zcela jasná a je-li Vám výsečový graf bez uvedení absolutních četností předkládán, ptejte se vždy, zda je důvod v neznalosti autora nebo zda je to jeho záměr.

Průvodce studiem

Teď přišel čas na ověření, zda jste porozuměli předcházejícímu výkladu. Následující příklad se pokuste vyřešit samostatně, ukázkové řešení použijte ke kontrole svého postupu.

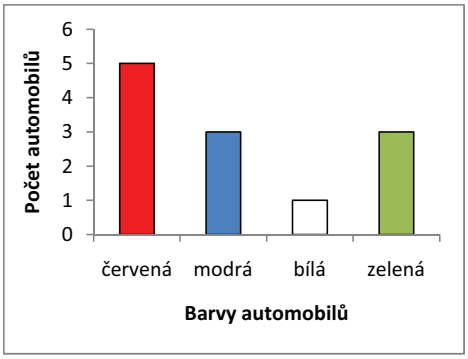
Příklad 1.1. Níže uvedená data představují částečný výsledek pozorování zaznamenaný při průzkumu zatížení jedné z ostravských křižovatek, a sice barvu projíždějících automobilů. Data vyhodnoťte a graficky znázorněte.

červená, modrá, zelená, modrá, červená, zelená, červená, červená, modrá, zelená, bílá, červená

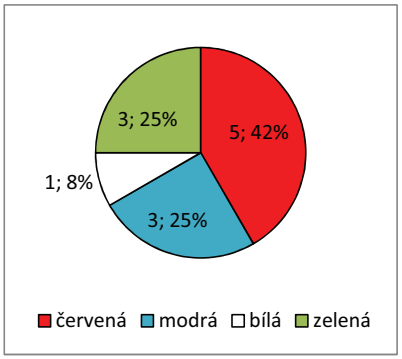
Řešení. Je zřejmé, že se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že barvy automobilů nemá smysl seřazovat, víme, že se jedná o proměnnou nominální. Pro její popis proto zvolíme tabulku četností, určíme modus a barvu projíždějících automobilů znázorníme prostřednictvím histogramu a výsečového grafu.
Modus = červená (tj. v zaznamenaném vzorku se vyskytlo nejvíce červených automobilů)

Tab. 1.2: Tabulka rozdělení četností pro pozorované barvy automobilů

TABULKA ROZDĚLENÍ ČETNOSTI		
Barvy projíždějících automobilů	**Absolutní četnost**	**Relativní četnost**
	$\mathbf{n_i}$	$\mathbf{p_i}$
červená	5	$5/12 = 0{,}42$
modrá	3	$3/12 = 0{,}25$
bílá	1	$1/12 = 0{,}08$
zelená	3	$3/12 = 0{,}25$
Celkem	12	1,00



Obr. 1.6: Pozorované barvy automobilů - histogram



Obr. 1.7: Pozorované barvy automobilů - výsečový graf

Celkem bylo pozorováno 12 automobilů. ▲

1.1.3 Ordinální proměnná

Ordinální proměnná, stejně jako proměnná nominální, nabývá v rámci souboru různých slovních variant, avšak tyto varianty mají přirozené uspořádání, tj. můžeme určit, která je „menší“ a která „větší“.

Pro popis ordinální proměnné se používají stejné statistické charakteristiky a grafy jako pro popis proměnné nominální (četnost, relativní četnost, modus + histogram, výsečový graf), rozšířené o další dvě charakteristiky (kumulativní četnost, kumulativní relativní četnost), které berou v úvahu uspořádání ordinální proměnné.

- **Kumulativní četnost** $\boldsymbol{m_i}$ (angl. „cumulative frequency“) definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné i-té variantě.

Uvažte např. proměnnou „známka ze statistiky“, která nabývá variant: „výborně“, „velmi dobře“, „prospěl“, „neprospěl“, pak např. kumulativní četnost pro variantu „prospěl“ bude rovna počtu studentů, kteří ze statistiky získali známku „prospěl“ nebo lepší.

Jsou-li jednotlivé varianty uspořádány podle své „velikosti“(„$x_1 < x_2 < \ldots < < x_k$“), platí

$$m_i = \sum_{j=1}^{i} n_j$$

Je tedy zřejmé, že kumulativní četnost k-té („nejvyšší“) varianty je rovna rozsahu proměnné – $m_k = n$.

Druhou speciální charakteristikou určenou pouze pro ordinální proměnnou je kumulativní relativní četnost.

- **Kumulativní relativní četnost F_i** (angl. „cumulative relative frequency“) vyjadřuje jakou část souboru tvoří hodnoty nabývající i-té a nižší varianty.

$$F_i = \sum_{j=1}^{i} p_j,$$

což není nic jiného než relativní vyjádření kumulativní četnosti:

$$F_i = \frac{m_i}{n}.$$

Obdobně jako pro nominální proměnné, můžeme i pro proměnné ordinální prezentovat statistické charakteristiky pomocí tabulky rozdělení četnosti. Ta obsahuje ve srovnání s tabulkou rozdělení četností pro nominální proměnnou navíc hodnoty kumulativních a kumulativních relativních četností.

Tab. 1.3: Tabulka rozdělení četností pro ordinální proměnnou

TABULKA ROZDĚLENÍ ČETNOSTÍ				
Hodnoty x_i	**Absolutní četnost**	**Relativní četnost**	**Kumulativní četnost**	**Kumulativní relativní četnost**
	n_i	**p_i**	**m_i**	**F_i**
x_1	n_1	p_1	$m_1 = n_1$	$F_1 = p_1$
x_2	n_2	p_2	$m_2 = n_1 + n_2 = m_1 + n_2$	$F_2 = p_1 + p_2 = F_1 + p_2$
x_k	n_k	p_k	$m_k = m_{k-1} + n_k = n$	$F_k = F_{k-1} + p_k = 1$
Celkem	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	-----	-----

1.1.4 Grafické znázornění ordinální proměnné

Co se týče grafické prezentace ordinální proměnné, zmínili jsme histogram a výsečový graf. Ani jeden z těchto grafů však nezaznamenává uspořádání jednotlivých variant. K tomu nám slouží polygon kumulativních (resp. kumulativních relativních) četností, kterému se říká Lorenzova křivka, popř. Paretův graf.

Lorenzova křivka (polygon kumulativních četností, Galtonova ogiva, S křivka) je spojnicovým grafem, který získáme tak, že na vodorovnou osu vynášíme jednotlivé varianty proměnné v pořadí od „nejmenší“ do „největší“ a na svislou osu příslušné hodnoty kumulativních četností. Znázorněné body spojíme úsečkami.

Všimněte si, že směrnice (sklon) polygonu kumulativních četností je tím nižší, čím nižší je rozdíl mezi četnostmi jednotlivých variant.

Obr. 1.8: Lorenzova křivka

1.1.5 Paretova analýza

V různých odvětvích lidské činnosti (ekonomie, sociologie, řízení jakosti, ...) se setkáváme s Paretovým principem, který lze formulovat tak, že 80% následků pramení z 20% příčin (20% lidí vlastní 80% celkového bohatství, 80% závad je způsobeno 20% všech příčin, ...). V praxi pak bývá snahou nalézt toto malé spektrum příčin (životně důležitá menšina), které tak významně ovlivňuje výsledek. Tento postup, který si vysvětlíme na níže uvedeném příkladu, se nazývá Paretova analýza.

Příklad 1.2. V závodě je na jednom ze zařízení pozorována častá poruchovost a z toho plynoucí ztráty a prostoje. Management podniku se chystá zavést inovace, které by napomohly snížit tuto poruchovost. Na pracovišti byla v období 27. 10. 2009 – 6. 11. 2009 sledována a zaznamenávána příčina závad na daném zařízení. Byly zaznamenány tyto typy závad:

A – netěsnost
B – porucha ložiska
C – přehřátí
D – selhání přepěťové ochrany
E – deformace
F – chyba obsluhy
G – jiná závada

Analyzujte závady zaznamenané v tabulce.

Řešení.

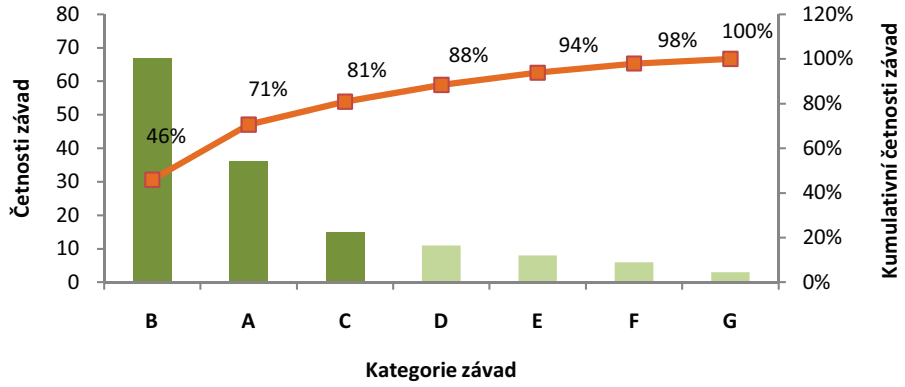
Datum	Závada
27.10.2009	B
27.10.2009	C
27.10.2009	A
27.10.2009	B
27.10.2009	A
27.10.2009	A
28.10.2009	B
28.10.2009	B
29.10.2009	D

Z ukázky datového souboru je zřejmé, že máme k dispozici chronologický záznam závad. Naším úkolem je tyto závady analyzovat a navrhnout ty z nich, jejichž odstraněním se dosáhne požadovaného snížení poruchovosti zařízení.

Závady budeme analyzovat jako ordinální proměnnou seřaditelnou podle četností výskytu. K Paretově analýze pak využijeme tabulku četnosti závad a tzv. **Paretův graf**, který je sloučením histogramu proměnné seřazené podle četnosti výskytu (od největší četnosti výskytu po nejmenší) a příslušného polygonu kumulativních četnosti – Lorenzovy křivky.

Tab. 1.4: Tabulka rozdělení četností závad

Závada	Četnost	Kumulativní četnost	Relativní četnost	Kumulativní rel. četnost
B	67	67	46%	46%
A	36	103	25%	71%
C	15	118	10%	81%
D	11	129	8%	88%
E	8	137	5%	94%
F	6	143	4%	98%
G	3	146	2%	100%
Celkem	146		100%	



Obr. 1.9: Paretův graf závad

Na základě Tab. 1.4 a grafu (Obr. 1.9) lze okamžitě identifikovat, že rozhodující podíl na poruchovosti zařízení mají závady typu B (46% všech závad). Skupina závad B, A, C pak zapříčiňuje 81% všech poruch.

Obdobným způsobem bychom mohli popsat vliv různých závad na ztráty apod. ▲

Průvodce studiem

A znovu si můžete ověřit, zda dokážete správně aplikovat nabyté vědomosti.

Příklad 1.3. Následující data představují velikosti triček prodaných při výprodeji firmy TRIKO.

S, M, L, S, M, L, XL, XL, M, XL, XL, L, M, S, M, L, L, XL, XL, XL, L, M

a) Data vyhodnoťte a graficky znázorněte.
b) Určete kolik procent lidí si koupilo tričko velikosti nejvýše L.

Řešení.

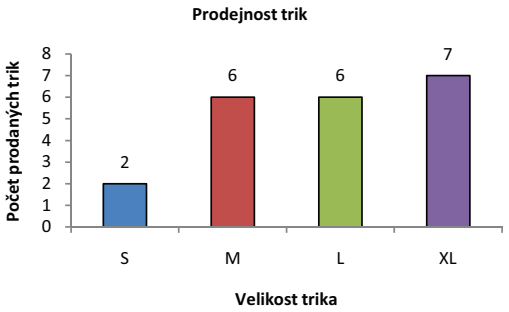
ad a) Zřejmě se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že velikosti triček lze seřadit, jde o proměnnou ordinální. Pro její popis proto použijeme tabulku četností pro ordinální proměnnou, v níž varianty velikosti triček budou seřazeny od nejmenší po největší (S, M, L, XL) a modus.

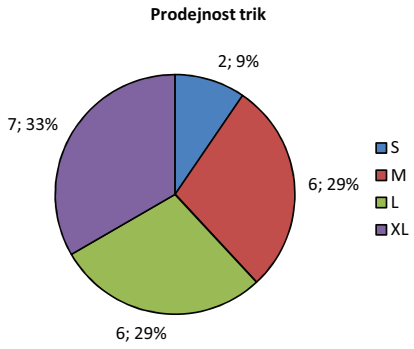
Tab. 1.5: Tabulka rozdělení četností prodejnosti triček podle velikosti

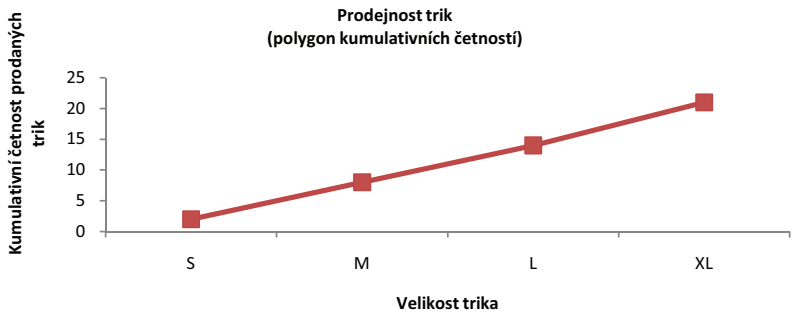
TABULKA ROZDĚLENÍ ČETNOSTÍ				
Velikosti triček	**Absolutní četnost**	**Relativní četnost**	**Kumulativní četnost**	**Kumulativní relativní četnost**
	n_i	p_i	m_i	F_i
S	3	$3/22 = 0{,}14$	3	$3/22 = 0{,}14$
M	6	$6/22 = 0{,}27$	$3+6=9$	$9/22 = 0{,}41$
L	6	$6/22 = 0{,}27$	$9+6=15$	$15/22 = 0{,}68$
XL	7	$7/22 = 0{,}32$	$15+7=22$	$22/22 = 1{,}00$
Celkem	22	1,00	-----	-----

Modus = XL (nejvíce lidí si koupilo tričko velikosti XL)

Grafický výstup bude tvořit histogram, výsečový graf a Lorenzova křivka. Jelikož nechceme používat Paretův princip, Paretův graf vytvářet nebudeme.







ad b) Na tuto otázku nám dá odpověď relativní kumulativní četnost pro variantu L, která určuje jaká část prodaných triček byla velikosti L a nižších. Tj. 68% zákazníků si koupilo tričko velikosti L a menší.

▲

1.2 Statistické charakteristiky numerických proměnných

Pro popis numerické proměnné můžeme použít většinu statistických charakteristik užívaných pro popis proměnné ordinální (četnost, relativní četnost, kumulativní četnost, kumulativní relativní četnost), což doplníme dalšími dvěma skupinami charakteristik - mírami polohy a mírami variability.

- **Míry polohy** určující typické rozložení hodnot proměnné (jejich rozmístění na číselné ose).
- **Míry variability** určující variabilitu (rozptyl) hodnot kolem své typické polohy.

1.2.1 Míry polohy a variability

Snad nejpoužívanějšími mírami polohy jsou průměry proměnných. Průměry představují průměrnou nebo typickou hodnotu výběrového souboru. Zřejmě nejznámějším průměrem pro kvantitativní proměnnou je

- **Aritmetický průměr $\bar{x}$** (angl. „mean“)

  Jeho hodnotu získáme pomocí známého vztahu

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

  kde: x ... jednotlive hodnoty proměnné,
  n ... rozsah výběrového souboru (počet hodnot proměnné).

  Jsou-li hodnoty analyzované proměnné uspořádány do tabulky četností, používáme pro výpočet aritmetického průměru vztah

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i},$$

kde četnosti n_i představují váhu, která je přisuzována jednotlivým hodnotám proměnné x_i. Takto vypočítaný aritmetický průměr se nazývá **vážený aritmetický průměr**.

Známé jsou i **vlastnosti aritmetického průměru**.

1. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0,$

   neboli: součet všech odchylek hodnot proměnné od jejich aritmetického průměru je roven nule, což znamená, že aritmetický průměr kompenzuje vliv náhodných chyb na proměnnou.

2. $\forall a \in \mathbb{R} : \frac{\sum_{i=1}^{n} (a+x_i)}{n} = a + \bar{x},$

   neboli: přičteme-li ke všem hodnotám proměnné stejné číslo, zvětší se o toto číslo rovněž aritmetický průměr.

3. $\forall b \in \mathbb{R} : \frac{\sum_{i=1}^{n} (bx_i)}{n} = b\bar{x},$

   neboli: vynásobíme-li všechny hodnoty proměnné stejným číslem, zvětší se stejným způsobem rovněž aritmetický průměr.

Příklad 1.4. Učitel matematiky na gymnáziu přiřazuje jednotlivým výsledkům studentů váhy následujícím způsobem.

	Váha
Zkoušení a dílčí testy	1
Opakovací testy	2
Kompozice	3

U studenta Masaříka má učitel za 1. pololetí záznam:

Zkoušení:	2
Dílčí testy:	3, 2, 1, 3
Opakovací testy:	2, 3, 1
Kompozice:	3, 2

Určete výslednou průměrnou známku studenta.

Řešení. Jde o klasický případ užití váženého průměru, kdy význam jednotlivých známek je oceněn jejich váhami.

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}$$

$$\bar{x} = \frac{2 \cdot 1 + 3 \cdot 1 + 2 \cdot 1 + 1 \cdot 1 + 3 \cdot 1 + 2 \cdot 2 + 3 \cdot 2 + 1 \cdot 2 + 3 \cdot 3 + 2 \cdot 3}{1 + 1 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3} = \frac{38}{17} \doteq 2,2$$

Vzhledem k tomu, že vážený průměr známek studenta Masaříka je 2,2, měl by tento student na pololetní vysvědčení dostat z matematiky 2.

▲

Přestože to tak na první pohled vypadá, aritmetický průměr nemusí být vždy pro výpočet průměru výběrového souboru nejvhodnější.

- **Harmonický průměr**

  Pro výpočet průměru v případech, kdy proměnná má charakter části z celku (úlohy o společné práci, ...), používáme průměr harmonický, který je definován vztahem

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

  Máme-li údaje setříděné do tabulky četností, používáme dle níže uvedeného vztahu **vážený harmonický průměr**.

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

Příklad 1.5. Totožná součástka se vyrábí na dvou automatech. Starší z nich vyrobí 1 kus každých 6 minut, nový každé 3 minuty. Jak dlouho trvá v průměru výroba jedné součástky?

Řešení. Jde o typickou úlohu o společné práci. Pro určení průměrné doby trvání výroby součástky proto použijeme harmonický průměr.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{2}{\frac{1}{6} + \frac{1}{3}} = 4\ [min]$$

Výroba jedné součástky trvá průměrně 4 minuty.

▲

- **Geometrický průměr**
  Pracujeme-li s kladnou proměnnou představující relativní změny (růstové indexy, cenové indexy...), používáme tzv. **geometrický průměr**, který je definován jako *n-tá* odmocnina ze součinu hodnot proměnné.

$$\bar{x}_G = \sqrt[n]{x_1 \cdot \ x_2 \cdot \ \ldots \ \cdot x_n}$$

  Stejně jako v předchozích případech lze zapsat rovněž vzorec pro **vážený geometrický průměr**.

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ \ldots \ \cdot x_n^{n_k}},$$

  kde

$$n = \sum_{i=1}^{k} n_i.$$

Příklad 1.6. Předloni byla výše ročního platu zaměstnance ve firmě 200 000 Kč, loni 220 000 Kč a letos 250 000 Kč. Jaký je průměrný koeficient růstu jeho platu?

Řešení. **Koeficient růstu** k_t je podíl dvou hodnot kladné proměnné.

$$k_t = \frac{x_t}{x_{t-1}},$$

kde x_t ... hodnota proměnné x v aktuálním období t,
x_{t-1} ... hodnota proměnné x v předchozím období $t-1$.

Často se koeficient růstu uvádí v procentech, pak hovoříme o **relativním přírůstku** $\boldsymbol{\sigma_t}$.

$$\sigma_t = (k_t - 1) \cdot 100 = \frac{x_t - x_{t-1}}{x_{t-1}} \cdot 100 \ [\%]$$

	Plat [Kč]	Koeficient růstu	Relativní přírůstek [%]
předloni	200 000		
loni	220 000	$\frac{220\,000}{200\,000} = 1{,}100$	10,0%
letos	250 000	$\frac{250\,000}{220\,000} = 1{,}136$	13,6%

Koeficient růstu představuje relativní změnu, pro výpočet průměru proto použijeme geometrický průměr.

$$\bar{k}_t = \sqrt{1,100 \cdot 1,136} = 1,118$$

Plat zaměstnance během posledních třech let rostl průměrně o 11,8% ročně.

▲

Vzhledem k tomu, že průměr se stanovuje ze všech hodnot proměnné, nese maximum informací o výběrovém souboru. Na druhé straně je však velmi citlivý na tzv. **odlehlá pozorování**, což jsou hodnoty, které se mimořádně liší od ostatních a dokážou proto vychýlit průměr natolik, že přestává daný výběr reprezentovat. K identifikaci odlehlých pozorování se vrátíme později.

Mezi míry polohy, které jsou na odlehlých pozorováních méně závislé, patří

- **Modus**

  Pozor! v případě modu budeme rozlišovat mezi diskrétní a spojitou kvantitativní proměnnou. **Pro diskrétní proměnnou** definujeme modus jako hodnotu nejčetnější varianty proměnné (podobně jako u kvalitativní proměnné).

  Naproti tomu **u spojité proměnné** považujeme za modus $\hat{x}$ hodnotu kolem níž je největší koncentrace hodnot proměnné. Mnohdy mluvíme o typické hodnotě proměnné. Pro určení této hodnoty využijeme tzv. **shorth** (čti „šórt“ a skloňuj podle hrad), což je nejkratší interval, v němž leží alespoň 50% hodnot proměnné (v případě výběru o rozsahu $n = 2k (k \in \mathbb{N})$ (sudý počet hodnot), leží v shorthu k hodnot–což je 50% $(n/2)$ hodnot proměnné, v případě výběru o rozsahu $n = 2k + 1 (k \in \mathbb{N})$ (lichý počet hodnot), leží v shorthu $k + 1$ hodnot - což je o 1 více než je 50% hodnot proměnné). **Modus** pak definujeme jako střed shorthu.

  Z předcházejících definic vyplývá, že délka shorthu (horní mez – dolní mez) je jednoznačně dána, to však nemusí platit pro jeho umístění a tudíž ani pro modus. Pokud lze modus určit jednoznačně, mluvíme o **unimodální proměnné**, má-li proměnná dva mody, nazýváme ji **bimodální**. Existence dvou a více modu ve výběru obvykle signalizuje nesourodost (heterogenitu) hodnot proměnné. Tuto nesourodost bývá možné odstranit rozdělením souboru na podsoubory - roztříděním podle některého jiného znaku (např. bimodální znak výška člověka lze roztřídit podle pohlaví na dva unimodální znaky - výška žen a výška mužů).

Průvodce studiem

Zdála se Vám pasáž o modu kvantitativní proměnné příliš složitá? Pokusíme se ji nyní osvětlit na jednoduchém příkladu, který Vám snad případné nejasnosti ozřejmí.

Příklad 1.7. Následující data představují věk hudebníků vystupujících na přehlídce dechových orchestrů. Proměnnou věk považujte za spojitou. Určete průměr, shorth a modus věku hudebníků.

22 82 27 43 19 47 41 34 34 42 35

Řešení. **a) Určení průměru:**

V tomto případě jednoznačně použijeme aritmetický průměr (proměnná věk nepředstavuje ani část celku ani relativní změnu).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{22+82+27+43+19+47+41+34+34+42+35}{11} = 38,7 \ let$$

Průměrný věk hudebníka vystupujícího na přehlídce dechových orchestrů je 38,7 let.

Prohlédněte si ještě jednou zadaná data a promyslete si nakolik je průměrný věk reprezentativní statistikou daného výběru (pozor na odlehlá pozorování).

b) Určení shorthu:

Náš výběrový soubor má 11 hodnot, z čehož vyplývá, že v shorthu bude ležet 6 z nich (rozsah souboru je 11 (lichý počet hodnot), 50% z toho je 5,5 *(5,5 hodnoty se špatně určuje, že?)* a nejbližší vyšší přirozené číslo je 6 – neboli: $\lceil \frac{n}{2} \rceil = \lceil \frac{11}{2} \rceil = \lceil 5,5 \rceil = 6$).

A další postup?

- Hodnoty proměnné seřadíme.
- Určíme délky všech 6-ti členných intervalů, v nichž $x_1 < x_{i+1} < \ldots < x_{i+5}$ pro $i = 1, 2, \ldots, n-5$.
- Nejkratší z těchto intervalů prohlásíme za shorth (délka intervalu $= x_{i+5} - x_i$)

Originální data	**Seřazená data**	**Délky 6-ti členných intervalů**
22	19	16 (= 35–19)
82	22	19 (= 41–22)
27	27	15 (= 42–27)
43	**34**	**9 (= 43–34)**
19	**34**	13 (= 47–34)
47	**35**	47 (= 82–35)
41	**41**	
34	**42**	
34	**43**	
42	47	
35	82	

Z tabulky je zřejmé, že nejkratší interval má délku 9, čemuž odpovídá jediný interval: $\langle 34; 43 \rangle$.

Shorth = $\langle 34; 43 \rangle$, což můžeme interpretovat např. tak, že polovina hudebníků je ve věku 34 až 43 let (jde přitom o nejkratší interval ze všech možných).

c)Určení modu:

Modus je definován jako střed shortu.

$$\widehat{x} = \frac{34 + 43}{2} = 38,5 \; let$$

Modus = 38,5 let, tj. typický věk hudebníka vystupujícího na této přehlídce dechových orchestrů je 38,5 let.

▲

Pro podrobnější vyjádření rozložení hodnot proměnné v rámci souboru slouží statistiky nazývané **výběrové kvantily.**

- **Výběrové kvantily** (angl. quantile, resp. percentile)

  Výběrové kvantily jsou statistiky, které charakterizují polohu jednotlivých hodnot v rámci proměnné. Podobně jako modus, jsou i výběrové kvantily rezistentní (odolné) vůči odlehlým pozorováním. Obecně je výběrový kvantil (dále jen kvantil) chápán jako hodnota, která rozděluje výběrový soubor na dvě části – první z nich obsahuje hodnoty, které jsou menší než daný kvantil, druhá část obsahuje hodnoty, které jsou větší nebo rovny danému kvantilu. Pro určení kvantilu je proto nutné výběr uspořádat od nejmenší hodnoty k největší.

Kvantil proměnné x, který odděluje *100p*% menších hodnot od zbytku souboru, tj. od *100(1-p)*% hodnot, nazýváme ***100p* %-ním kvantilem** a značíme jej $\boldsymbol{x_p}$.
V praxi se nejčastěji setkáváme s následujícími kvantily:

- **Kvartily**

  Dolní kvartil $\boldsymbol{x_{0,25}}$ = 25%-ní kvantil (rozděluje datový soubor tak, že 25% hodnot je menších než tento kvartil a zbytek, tj. 75% větších (nebo rovných))

  Medián $\boldsymbol{x_{0,5}}$ = 50%-ní kvantil (rozděluje datový soubor tak, že polovina (50%) hodnot je menších než medián a polovina (50%) hodnot větších (nebo rovných))

  Horní kvartil $\boldsymbol{x_{0,75}}$ = 75%-ní kvantil (rozděluje datový soubor tak, že 75% hodnot je menších než tento kvartil a zbytek, tj. 25% větších (nebo rovných))

  Kvartily dělí výběrový soubor na 4 přibližně stejně četné části.

- **Decily**–$x_{0,1}; x_{0,2}; ...; x_{0,9}$

  Decily dělí výběrový soubor na 10 přibližně stejně četných části.

- **Percentily**–$x_{0,01}; x_{0,02}; \ldots; x_{0,99}$

  Percentily dělí výběrový soubor na 100 přibližně stejně četných části.

A nyní se dostáváme k tomu, **jak se kvantily určují.**

1. Výběrový soubor uspořádáme podle velikosti.
2. Jednotlivým hodnotám proměnné přiřadíme pořadí, a to tak, že nejmenší hodnota bude mít pořadí 1 a nejvyšší hodnota pořadí n (rozsah souboru).
3. 100p%- ní kvantil je roven hodnotě proměnné s pořadím z_p, kde

   $$z_p = np + 0.5$$

   Není-li z_p celé číslo, pak daný kvantil určíme jako průměr prvků s pořadím $\lfloor z_p \rfloor$ a $\lceil z_p \rceil$.

POZOR! Zejména v souvislosti s hodnocením normovaných testů (SCIO testy, biometrické normy,...) se často setkáváme s vyjádřením „Patříte do p. percentilu“, přičemž p je celé číslo mezi 1 a 100. Je tím myšleno, že nejméně (p-1)% a zároveň méně než p% účastníků testu dosáhlo nižšího hodnocení než vy.

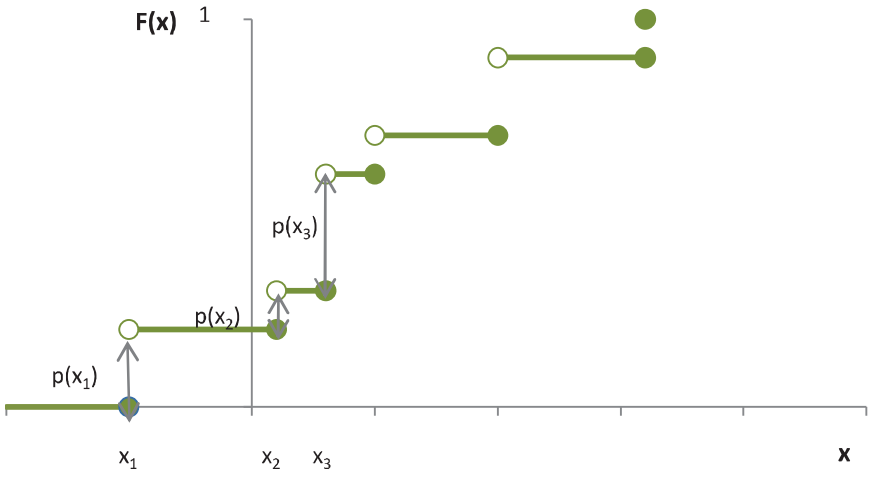
(Např. „Patříte do 80. percentilu“ znamená, že nejméně 79% (a nejvýše 80%) účastníku testu dosáhlo nižšího výsledku než vy.)

Za zmínku zajisté stojí i **vztah mezi kvantily a relativní kumulativní četnosti**. Zřejmě lze říci, že hodnota p udává relativní kumulativní četnost kvantilu x_p, tj. relativní četnost těch hodnot proměnné, které jsou menší než kvantil x_p. Kvantil a relativní kumulativní četnost jsou tedy inverzní pojmy. Grafické nebo tabulkové znázornění setříděné proměnné a příslušných kumulativních četností se označuje jako **distribuční funkce kumulativní četnosti**, popř. **empirická distribuční funkce**. Ujasněme si nyní, jak empirickou distribuční funkci pro kvantitativní proměnnou určit.

- **Empirická distribuční funkce F(x) pro kvantitativní proměnnou**

  Označme si $p(x_i)$ relativní četnost hodnoty x_i seřazeného výběrového souboru $x_1 < x_2 < \ldots < x_n$. Pro empirickou distribuční funkci F(x) pak platí:

$$F(x) \quad = \quad \begin{cases} 0 & \text{pro } x \leqq x_i \\ \sum_{i=1}^{j} F(x) & \text{pro } x_j < x \leqq x_{j+1}, 1 \leqq j \leqq n-1 \\ 1, & \text{pro } x_n < x \end{cases}$$



Obr. 1.10: Empirická distribuční funkce

Empirická distribuční funkce je monotónně rostoucí, zleva spojitou funkcí, která „skáče“ podle relativních četností příslušných jednotlivým hodnotám proměnné. Zjevně tedy platí, že

$$p(x_i) = \lim_{x \to x_i} F(x) - F(x_i)$$

Prostřednictvím kvantilů jsou definovány i další dvě statistiky kvantitativní proměnné – interkvartilové rozpětí a MAD.

- **Interkvartilové rozpětí IQR**

  Tato statistika je mírou variability souboru a je definována jako vzdálenost mezi horním a dolním kvartilem:

  $$IQR = x_{0.75} - x_{0.25}$$

- **MAD**

  Název MAD je zkratkou anglické definice – **m**edian **a**bsolute **d**eviation from the median, čili česky: medián absolutních odchylek od mediánu

  Jak jej tedy určíme?

  1. Výběrový soubor uspořádáme podle velikosti
  2. Určíme medián souboru
  3. Pro každou hodnotu souboru určíme absolutní hodnotu její odchylky od mediánu
  4. Absolutní odchylky od mediánu uspořádáme podle velikosti
  5. Určíme medián absolutních odchylek od mediánu, tj. MAD

Průvodce studiem

Zdá se Vám, že za sebou máte moc teorie? Abyste se ujistili, že nic není tak černé jak vypadá, zkuste pokračovat v předcházejícím řešeném příkladu.

Příklad 1.8. Pro data z řešeného příkladu 1.7 určete

a) všechny kvartily,

b) interkvartilové rozpětí,

c) MAD,

d) zakreslete empirickou distribuční funkci.

Tab. 1.6: Přiřazení pořadí hodnotám proměnné

Originální data	**Seřazená data**	**Pořadí**
22	19	1
82	22	2
27	27	3
43	34	4
19	34	5
47	35	6
41	41	7
34	42	8
34	43	9
42	47	10
35	82	11

Řešení. ad a)Naším úkolem je určit dolní kvartil $x_{0,25}$, medián $x_{0,5}$ a horní kvartil $x_{0,75}$. Budeme dodržovat postup doporučený pro určování kvantilů, to znamená – data seřadit a přiřadit jim pořadí. Výsledek prvních dvou bodů postupu ukazuje Tab.1.6.

A můžeme přejít k bodu 3, tj. stanovit pořadí hodnot proměnné pro jednotlivé kvartily a tím i jejich hodnoty.

Dolní kvartil $x_{0,25}$: $p = 0,25; n = 11 \Rightarrow z_p = 11 \cdot 0,25 + 0,5 = 3,25$,
Dolní kvartil je tedy průměrem prvků s pořadím 3 a 4. $x_{0,25} = \dfrac{27+34}{2} = 30,5$ let,
tj. 25% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 30,5 let (75% z nich má 30,5 let a více).

Medián $x_{0,5}$: $p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow x_{0,5} = 35$ *let*,
tj. polovina hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 35 let (50% z nich má 35 let a více).

Horní kvartil $x_{0,75}$: $p = 0,75; n = 11 \Rightarrow z_p = 11 \cdot 0,75 + 0,5 = 8,75$
Horní kvartil je tedy průměrem prvků s pořadím 8 a 9.$x_{0,75} = \dfrac{42+43}{2} = 42,5$ let,
tj. 75% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 42,5 let (25% z nich má 42,5 let a více).

ad b) **Interkvartilové rozpětí IQR:** $\text{IQR} = x_{0,75} - x_{0,25} = 43 - 27 = 16$.

Jak již bylo zmíněno, praktická interpretace IQR neexistuje.

Tab. 1.7: Postup při výpočtu statistiky MAD

Originální data x_i	**Seřazená data y_i**	Absolutní hodnoty odchylek seřazených dat od jejich mediánu $\lvert y_i - x_{0,5}\rvert$	Seřazené absolutní hodnoty odchylek seřazených dat od jejich mediánu $\boldsymbol{M_i}$
22	19	16=$\lvert 19-35\rvert$	0
82	22	13=$\lvert 22-35\rvert$	1
27	27	8=$\lvert 27-35\rvert$	1
43	34	1=$\lvert 34-35\rvert$	6
19	34	1=$\lvert 22-35\rvert$	7
47	35	0=$\lvert 35-35\rvert$	8
41	41	6=$\lvert 41-35\rvert$	8
34	42	7=$\lvert 42-35\rvert$	12
34	43	8=$\lvert 43-35\rvert$	13
42	47	12=$\lvert 47-35\rvert$	16
35	82	47=$\lvert 22-35\rvert$	47

ad c) **MAD** Chceme-li určit tuto statistiku, budeme postupovat přesně podle toho,

co skrývá zkratka v názvu – medián absolutních odchylek od mediánu. Provedení uvedeného postupu ukazuje Tab 1.7.

$x_{0,5}$=35

$MAD = M_{0,5}$,
$p = 0,5; n = 11 \Rightarrow z_p = 11 \cdot 0,5 + 0,5 = 6 \Rightarrow M_{0,5} = 8$,

(MAD je medián absolutních odchylek od mediánu, tj. 6. hodnota seřazeného souboru absolutních odchylek od mediánu).MAD = 8.

ad d) Zbývá poslední úkol – sestrojit **empirickou distribuční funkci**. Připomeňme si proto její definici a postupujme podle ní.

$$F(x)=\begin{cases} 0 & \text{pro } x \leqq x_i \\ \sum_{i=1}^{j} F(x) & \text{pro } x_j < x \leqq x_{j+1}, 1 \leqq j \leqq n-1 \\ 1 & \text{pro } x_n < x \end{cases}$$

Do tabulky si zapíšeme seřazené hodnoty proměnné, jejich četnosti, relativní četnosti a z nich odvodíme empirickou distribuční funkci.

Tab. 1.8: Postup výpočtu empirické distribuční funkce

Originální data x_i
22
82
27
43
19
47
41
34
34
42
35

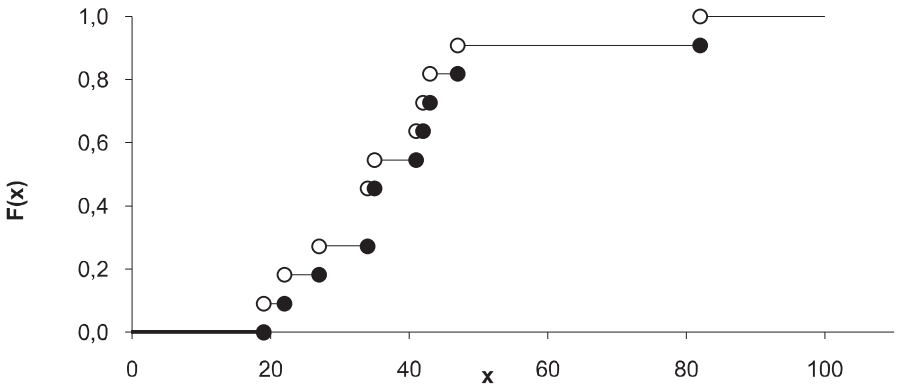
Seřazené hodnoty x_i	Absolutní četnosti seřazených hodnot n_i	Relativní četnosti seřazených hodnot p_i	Empirická dist. funkce $F(x_i)$
19	1	1/11	0
22	1	1/11	1/11
27	1	1/11	2/11
34	2	2/11	3/11
35	1	1/11	5/11
41	1	1/11	6/11
42	1	1/11	7/11
43	1	1/11	8/11
47	1	1/11	9/11
82	1	1/11	10/11

Z definice emp. dist. funkce *F(x)* tedy plyne, že pro všechna x menší než 19 je *F(x)* rovna nule, pro x větší než 19 a menší nebo rovna 22 je *F(x)* rovna 1/11, pro x větší než 22 a menší nebo rovna 27 je *F(x)* rovna 1/11 + 1/11, atd. Pro $x > 82$ je *F(x)*=1. Shrneme do Tab. 1.9.

Tab. 1.9: Empirická distribuční funkce

x	$(-\infty; 19\rangle$	$(19; 22\rangle$	$22; 27\rangle$	$(27; 34\rangle$	$(34; 35\rangle$
F(x)	0	1/11	2/11	3/11	5/11

x	$(35; 41\rangle$	$(41; 42\rangle$	$(42; 43\rangle$	$(43; 47\rangle$	$(47; 82\rangle$	$(82; \infty\rangle$
F(x)	6/11	7/11	8/11	9/11	10/11	11/11



Obr. 1.11: Empirická distribuční funcke-graf

▲

Průvodce studiem

Zvládli jste to? Gratuluji. Pokud jste s příkladem měli nějaké problémy, doporučuji vám, abyste pasáž o kvantilech a empirické distribuční funkci znovu důkladně prostudovali – není to naposled, co se s těmito pojmy setkáváte.

Až dosud jsme se zabývali převážně statistickými charakteristikami umožňujícími popis polohy proměnné, tj. mírami polohy. Průměry, modus, stejně jako medián vyjadřují pomyslný „střed“ proměnné, neříkají však nic o rozložení jednotlivých hodnot proměnné kolem tohoto „středu“, tj. o variabilitě proměnné. Je zřejmé, že čím větší je rozptýlenost hodnot proměnné kolem jejího pomyslného „středu“, tím menší je schopnost tohoto „středu“ reprezentovat proměnnou.

Následující statistické charakteristiky nám umožňují popis variability (rozptýlenosti) výběrového souboru, neboli popis rozptylu jednotlivých hodnot kolem středu proměnné – nazýváme je tedy mírami variability. Z dosud zmíněných statistických charakteristik zařazujeme mezi míry variability shorth a interkvartilové rozpětí.

- **Výběrový rozptyl** s^2 (čti „s kvadrát“, angl. sample variance) je nejrozšířenější mírou variability výběrového souboru. Určujeme jej podle vztahu

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

Vidíme, že výběrový rozptyl je dán podílem součtu kvadrátu odchylek jednotlivých hodnot od průměru a rozsahu souboru sníženého o jedničku.

Mezi základní **vlastnosti výběrového rozptylu** patří:

1. Výběrový rozptyl konstantního souboru je roven nule,což znamená, že jsou-li všechny hodnoty proměnné stejné, má soubor nulovou rozptýlenost.
2. $$\forall a \in \mathbb{R} : \left(\left(s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \right) \wedge (y_i = a + x_i) \right) \Rightarrow$$
$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}((a + x_i) - (a + \bar{x}))^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = s^2$$
což znamená, že přičteme-li ke všem hodnotám proměnné libovolnou konstantu, výběrový rozptyl proměnné se nezmění.

3.
$$\forall b \in \mathbb{R} : ((s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \wedge y_i = bx_i)) \Rightarrow$$
$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}((bx_i) - (b\bar{x}))^2}{n-1} = \frac{\sum_{i=1}^{n} b^2(x_i - \bar{x})^2}{n-1} = b^2 s^2$$
což znamená, že vynásobíme-li všechny hodnoty proměnné libovolnou konstantou (b), výběrový rozptyl proměnné se zvětší kvadrátem této konstanty (b^2 krát)

Nevýhodou použití výběrového rozptylu jakožto míry variability je to, že jednotka této charakteristiky je druhou mocninou jednotky proměnné. Např. je-li proměnnou denní tržba uvedena v Kč, bude výběrový rozptyl této proměnné vyjádřen v $K\check{c}^2$. Následující míra variability tuto vlastnost nemá.

- **Výběrová směrodatná odchylka s** (angl. sample standard deviation) je definována jako kladná odmocnina výběrového rozptylu
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Nevýhodou výběrového rozptylu i výběrové směrodatné odchylky je skutečnost, že neumožňují porovnávat varibilitu proměnných vyjádřených v různých jednotkách. Která proměnná má větší variabilitu – výška nebo hmotnost dospělého člověka? Na tuto otázku nám dá odpověď tzv. variační koeficient.

- **Variační koeficient** V_x (angl. coefficient of variation)

  vyjadřuje relativní míru variability proměnné x. Podle níže uvedeného vztahu jej lze stanovit pouze pro proměnné, které nabývají výhradně kladných hodnot. Variační koeficient je bezrozměrný. Uvádíme-li jej v [%], hodnotu získanou z definičního vzorce vynásobíme 100%.
$$V_x = \frac{V}{\bar{x}}, \text{ popř. } V_x = \frac{V}{\bar{x}} \cdot 100[\%]$$

Příklad 1.9. Firma vyrábějící tabulové sklo vyvinula méně nákladnou technologii pro zlepšení odolnosti skla vůči žáru. Pro testování bylo vybráno 5 tabulí skla a rozřezáno na polovinu. Jedna polovina pak byla ošetřena novou technologií, zatímco druhá byla ponechána jako kontrolní. Obě poloviny pak byly vystaveny zvyšujícímu se působení tepla, dokud nepraskly. Výsledky jsou uvedeny v Tab. 1.10. Porovnejte obě technologie pomocí základních charakteristik explorační statistiky (průměru a rozptylu, popř. směrodatné odchylky).

Tab. 1.10: Tavná teplota skla při použití staré a nové technologie

Mezní teplota (sklo prasklo) [^{o}C]	
Stará technologie x_i	**Nová technologie y_i**
475	485
436	390
495	520
483	460
426	488

Řešení. Nejprve se pokusíme porovnat obě technologie pouze za pomocí průměru. Vzhledem k tomu, že proměnná „mezní teplota“ nevyjadřuje ani část celku ani relativní změny, volíme průměr aritmetický.

Průměr pro starou technologii vychází

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{475 + 436 + \ldots + 426}{5} \doteq 463\,[^oC]$$

Průměr pro novou technologii:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{485 + 390 + \ldots + 488}{5} \doteq 469\,[^oC]$$

Na základě vypočtených průměrů bychom mohli říci, že novou technologii doporučujeme, poněvadž mezní teplota je při nové technologii o 6^{o}C vyšší.

A jaký závěr vyvodíme, doplníme-li k základním informacím míry variability?

Stará technologie:

Výběrový rozptyl:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1} = \frac{(475-463)^2 + (436-463)^2 + \ldots + (426-463)^2}{5-1} \doteq 916\,[^oC^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1}} = \sqrt{\frac{(475-463)^2 + \ldots + (426-463)^2}{5-1}} \doteq 31\,[^oC].$$

Nová technologie:

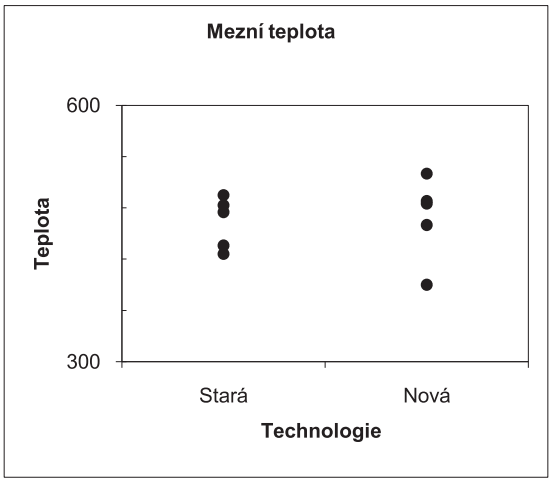
Výběrový rozptyl:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - y)^2}{n-1} = \frac{(485-469)^2 + (390-469)^2 + \ldots + (488-469)^2}{5-1} \doteq 2384\,[^oC^2]$$

Výběrová směrodatná odchylka:

$$s_y = \sqrt{s_y^2} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y)^2}{n-1}} == \sqrt{\frac{(485-469)^2 + \ldots + (488-469)^2}{5-1}} \doteq 49\,[^oC].$$

Výběrový rozptyl (výběrová směrodatná odchylka) vyšel pro novou technologii mnohem vyšší než pro technologii starou. Co to znamená? Podívejte se na grafické znázornění naměřených dat na Obr. 1.12.



Obr. 1.12: Srovnání technologií teplot pro starou a novou technologii

Mezní teploty pro novou technologii jsou mnohem rozptýlenější, tzn. že tato technologie není ještě dobře zvládnutá a její použití nám nezaručí zkvalitnění výroby. V tomto případě může dojít k silnému zvýšení, ale také k silnému snížení mezní teploty – proto by se měla nová technologie ještě vrátit do vývoje.

Zdůrazněme, že tyto závěry jsou stanoveny pouze na základě explorační analýzy. Pro rozhodnutí takovýchto případů nám statistika nabízí exaktnější metody (testování hypotéz), s nimiž se seznámíte později.

▲

Vzpomínáte si ještě na zmínku o odlehlých pozorováních? Dozvěděli jste se, že za odlehlá pozorování považujeme ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. vypovídací hodnotu průměru. Nyní se dozvíte, jak odlehlé hodnoty identifikovat.

- **Identifikace odlehlých pozorování**(angl. outliers)

  Ve statistické praxi se obvykle můžete setkat s několika způsoby identifikace odlehlých pozorování. My ukážeme tři z nich.

  1. **Vnitřní hradby:** Za odlehlé pozorování lze považovat takovou hodnotu x_i, která je od dolního, resp. horního kvartilu vzdálená více než 1,5 násobek interkvartilového rozpětí. Tedy:

     $$[(x_i < x_{0,25} - 1,5 \cdot IQR) \vee (x_i > x_{0,75} + 1,5 \cdot IQR)] \Rightarrow$$
     $$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

  2. **z-souřadnice (*z-skóre*):** Za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota z-souřadnice je větší než 3, tj. hodnota, která je od průměru vzdálenější než 3s. Tedy:

     $$z - skóre_i = \frac{x_i - \bar{x}}{s}$$

     $$|z - skóre_i| > 3 \Rightarrow \left|\frac{x_i - \bar{x}}{s}\right| > 3 \Rightarrow |x_i - \bar{x}| > 3s \Rightarrow$$
     $$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

  3. **$x_{0,5}$-souřadnice ($x_{0,5}$ – *skóre*):** Za odlehlé pozorování lze považovat takovou hodnotu x_i, jejíž absolutní hodnota mediánové souřadnice je větší než 3, tj. hodnota, která je od mediánu vzdálenější než $3 \cdot 1,483 \cdot$ MAD. Tedy:

     $$x_{0,5} - skóre_i = \frac{x_i - x_0,5}{1,483MAD}$$

     $$|x_{0,5} - skóre_i| > 3 \Rightarrow \left|\frac{x_i - x_{0,5}}{1,483MAD}\right| > 3 \Rightarrow |x_i - x_{0,5}| > 3 \cdot 1,483MAD \Rightarrow$$
     $$\Rightarrow x_i \text{ je odlehlým pozorováním}$$

  V konkrétním případě můžete pro identifikaci odlehlých pozorování zvolit libovolné z těchto tří pravidel. Za zmínku stojí, že z-souřadnice je „méně přísná“ k odlehlým pozorováním než mediánová souřadnice. Je to proto, že z-souřadnice se určuje na základě průměru a výběrové směrodatné odchylky, jež jsou silně ovlivněny hodnotami odlehlých pozorování. Naproti tomu mediánová souřadnice se určuje na základě mediánu a MADu, které jsou vůči odlehlým pozorováním odolné.

Někteří statistici rozdělují odlehlá pozorování do dvou skupin – na **odlehlá pozorování** a **extrémní pozorování**. Pro toto rozlišení využívají pojmů vnitřní a vnější hradby. Definice hradeb vychází z pravidla pro identifikaci odlehlých pozorování pomocí IQR.

Vnitřní hradby:	dolní mez:	$h_D = x_{0,25} - 1,5IQR$
	horní mez:	$h_H = x_{0,75} + 1,5IQR$
Vnější hradby:	dolní mez:	$H_D = x_{0,25} - 3IQR$
	horní mez:	$H_H = x_{0,75} + 3IQR$

Pozorování ležící mimo vnější hradby pak nazýváme extrémní, pozorování ležící vně vnitřních hradeb, avšak uvnitř hradeb vnějších nazýváme odlehlá.

Pokud o některé hodnotě proměnné rozhodneme, že je odlehlým pozorováním, je nutné rozlišit o jaký typ odlehlosti se jedná. V případě, že odlehlost pozorování je způsobena:

- hrubými chybami, překlepy, prokazatelným selháním lidí či techniky ...
- důsledky poruch, chybného měření, technologických chyb ...

tzn., známe-li příčinu odlehlosti a předpokládáme-li, že již nenastane, jsme oprávněni tato pozorování vyloučit z dalšího zpracování. V ostatních případech je nutno zvážit, zda se vyloučením odlehlých pozorování nepřipravíme o důležité informace o jevech vyskytujících se s nízkou četností.

Dalšími charakteristikami popisujícími kvantitativní proměnnou jsou **výběrová šikmost** a **výběrová špičatost**. Vzorce podle nichž se určují tyto charakteristiky jsou poměrně složité a proto se podle nich „ručně“ většinou nepočítá, jsou součástí většiny statistických programů.

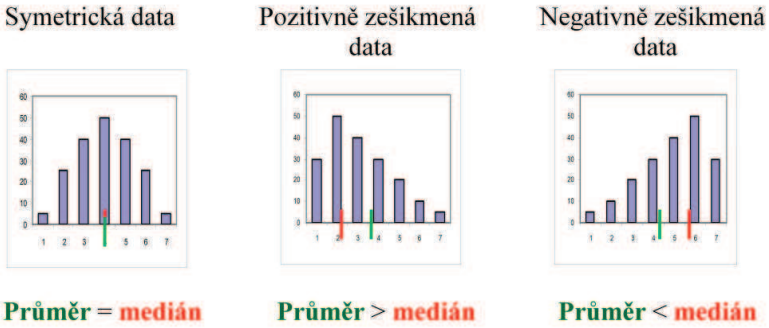
- **Výběrová šikmost *a*** (angl. skewness)

  vyjadřuje asymetrii rozložení hodnot proměnné kolem jejího průměru. Výběrová šikmost je definována vztahem:

$$a = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

  A jak výběrovou šikmost interpretujeme?

  $a = 0$ ... hodnoty proměnné jsou kolem jejího průměru rozloženy symetricky
  $a > 0$ ... u proměnné převažují hodnoty menší než průměr
  $a < 0$ ... u proměnné převažují hodnoty větší než průměr



Souvislost mezi šikmostí a charakteristikami polohy

Symetrické rozdělení:	$\bar{x} = x_{0,5}$
Pozitivně zešikmené rozdělení:	$\bar{x} > x_{0,5}$
Negativně zešikmené rozdělení:	$\bar{x} < x_{0,5}$

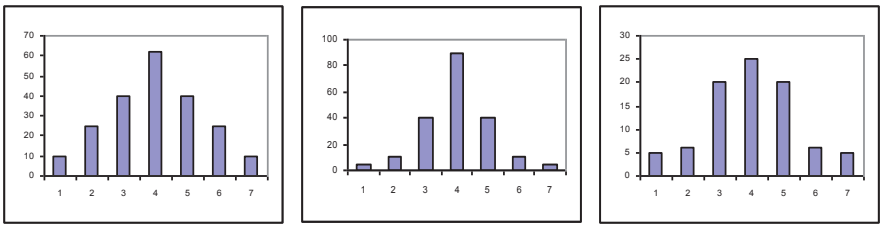
- **Výběrová špičatost *b*** (angl. kurtosis)

vyjadřuje koncentraci hodnot proměnné kolem jejího průměru. Výběrová špičatost je definována vztahem

$$b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}.$$

A jak výběrovou výběrovou špičatost?

$b = 0$ ... špičatost odpovídá normálnímu rozdělení (bude definováno později)
$b > 0$ ... špičaté rozdělení proměnné
$b < 0$ ... ploché rozdělení proměnné



b=0 b>0 b<0

1.3 Přesnost statistických charakteristik kvantitativních proměnných

V této chvíli jste se seznámili s řadou statistických charakteristik. Vzniká otázka, s jakou přesností máme tyto číselné charakteristiky uvádět. Je zřejmé, že počet platných cifer by měl korespondovat s přesností měření. Víme-li, například, že nejistota měření určité proměnné je jeden kilogram, nemá smysl průměr této proměnné uvádět s přesností na gramy.

Platí jednoduché pravidlo.

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme **nahoru** na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily...) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

Příklady chybně zapsaných hodnot číselných charakteristik vidíte v Tab. 1.11.

Tab. 1.11: Příklady chybného zápisu číselných charakteristik

	Délka [m]	Váha [kg]	Teplota [^{0}C]
Průměr	2,26	127,6	14 567
Medián	2,675	117,8	13 700
Směrodatná odchylka	0,78	23,7	1 200 (před zaokrouhlením 1235)
Proč je zápis chybný?	*Různý počet des. míst.*	*3 platné cifry u směrodatné odchylky.*	*Nejnižší zapsaný řád průměru (jednotky) neodpovídá nejnižšímu zapsanému řádu směrodatné odchylky (stovky).*

Jak by měl zápis vypadat správně ukazuje Tab.1.12.

Tab. 1.12: Příklady správného zápisu číselných charakteristik

	Délka [m]	Váha [kg]	Teplota [^{0}C]
Průměr	2,26	128	14 600
Medián	2,68	118	13 700
Směrodatná odchylka	0,78	24	1 200

Průvodce studiem

Tak, a máte to takřka vše za sebou – všechny číselné charakteristiky, které budete využívat pro popis kvantitativní proměnné jsou definovány. Zbývá nám jediné – ukázat si jak můžeme kvantitativní proměnnou znázornit graficky. Tak vzhůru do toho, neboť o nic složitého nejde.

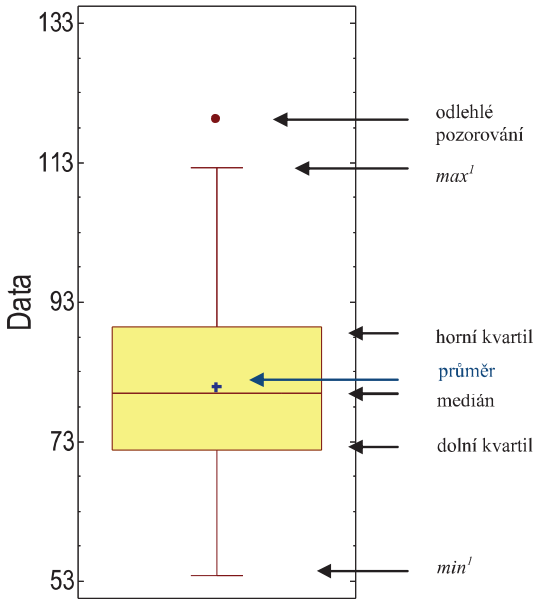
1.3.1 Grafické znázornění kvalitativní proměnné

- Krabicový graf(angl. Box plot)

Krabicový graf se ve statistice využívá od roku 1977, kdy jej poprvé prezentoval americký statistik J. W. Tukey. Nazval jej „box with whiskers plot“ – krabicový graf s vousama. Grafická podoba tohoto grafu se v různých aplikacích mírně liší. Jednu z jeho verzí vidíte na uvedeném obrázku.

Odlehlá pozorování jsou znázorněna jako izolované body, konec horního (popř. konec dolního) vousu představují maximum (popř. minimum) proměnné po vyloučení odlehlých pozorování,„víko“ krabice udává horní kvartil, „dno“ dolní kvartil, vodorovná úsečka uvnitř krabice označuje medián.

Z polohy mediánu vzhledem ke „krabici“ lze dobře usuzovat na symetrii vnitřních 50% dat a my tak získáváme dobrý přehled o středu a rozptýlenosti proměnné.



Obr. 1.13: Krabicový graf

Pozn.: Z popisu krabicového grafu je zřejmé, že jeho konstrukci začínáme zakreslením odlehlých pozorování a až poté vyznačujeme ostatní číselné charakteristiky proměnné (min_1, max_1, kvartily a shorth).

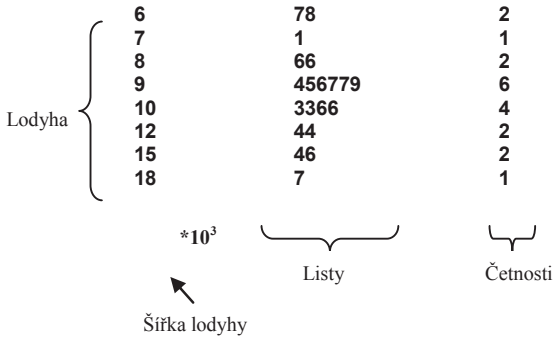
- **Číslicový histogram** (Lodyha s listy, angl. Stem and leaf plot)

Jak jsme si ukázali, výhodou krabicového grafu je jeho jednoduchost, někdy nám však chybí informace o konkrétních hodnotách proměnné. Chtěli bychom proto nějak přehledně zapsat číselné hodnoty výběru a k tomu nám slouží právě číslicový histogram. Navíc nám tento graf dává dobrou představu o šikmosti proměnné.

Představme si proměnnou představující průměrné měsíční platy zaměstnanců ve státní správě.

Průměrný měsíční plat [Kč]
10 654, 9 765, 8 675, 12 435, 9 675, 10 343, 18 786, 15 420, 8 675, 7 132, 6 732, 6 878, 15 657, 9 754, 9 543, 9 435, 10 647, 12 453, 9 987, 10 342.

A vy nyní stojíte před problémem jak tato data znázornit. Pokud se nad touto otázkou trochu zamyslíme, zjistíme, že pro naší informaci nejsou tak důležité koruny ani desetikoruny rozdílu. V tomto případě se nám jedná přinejmenším o stokoruny. Co kdybychom tedy informaci o „nedůležitých“ řádech zanedbali a znázornili setříděná data pouze na základě vyšších řádů? My jsme se rozhodli, že důležitý řád jsou pro nás stokoruny. Hodnoty stojící o řád výš (v našem případě tisíce) zapíšeme setříděné pod sebe, tak, že tvoří jakýsi stonek (**lodyhu**), přičemž pod graf uvedeme tzv. **šířku lodyhy**, která udává koeficient, jímž se hodnoty uvedené v grafu násobí.



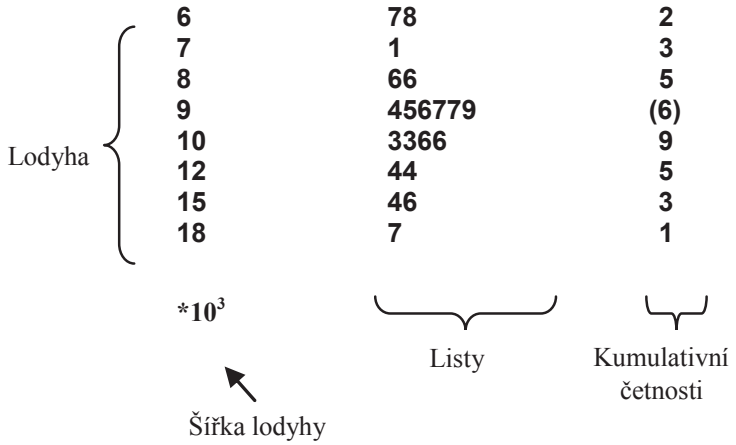
Obr. 1.14: Číslicový histrogram

Druhý sloupec grafu, **listy**, budou tvořit číslice, reprezentující zvolený „důležitý“ řád, zapisované do příslušných řádků (opět seřazené podle velikosti). A konečně – třetí sloupec udává absolutní četnosti příslušné daným řádkům.

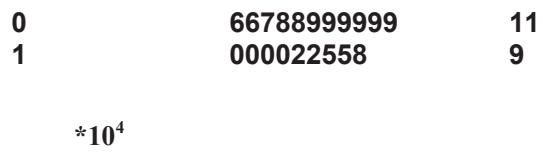
Jste ze slovního popisu poněkud zmateni? Prohlédněte si důkladně obrázek reprezentující číslicový histogram na Obr. 1.14. Např. první řádek reprezentuje dvě hodnoty – (6.7 a 6.8)*103 Kč, tj. 6700 Kč a 6800 Kč (koruny a desetikoruny jsme zanedbali), šestý řádek reprezentuje také dvě hodnoty – (12.4 a 12.4)*103 Kč, tj. dvě osoby s průměrným měsíčním příjmem 12400 Kč, atd. Už je to jasnější, dokázali byste tento graf sestrojit sami?

Existují různé modifikace číslicového histogramu. Např. zobrazované četnosti mohou být kumulativní, přičemž v řádku, v němž se nachází medián, se uvádí absolutní četnost (v závorce) a směrem k tomuto řádků se četnosti kumulují jednak od nejnižších hodnot, jednak od nejvyšších hodnot.

Konečně můžete namítnout, že způsobu konstrukce číslicového histogramu je pro jeden případ vždy několik. Nikde není dáno, který řád proměnné je pro zaznamenání důležitý a který už je zanedbatelný. (Srovnávali jsme platy dobře, když jsme je zaznamenali s přesnosti na stokoruny? Nestačilo znázornit číslicový histogram vzhledem k tisícikorunám?) Toto rozhodnutí leží vždy na tom, kdo data zpracovává. Můžeme uvést jen jedno pravidlo – dlouhé lodyhy s krátkými listy a krátké lodyhy s dlouhými listy svědčí o nevhodné volbě měřítka.



Obr. 1.15: Číslicový histrogram



Obr. 1.16: Nevhodná volba číslicového histogramu

Shrnutí:

Σ

Kvalitativní – Kategoriální proměnná

a) Nominální proměnná – nemá smysl uspořádání

Základní statistiky pro popis nominální proměnné:

- četnost
- relativní četnost
- modus

Grafické zobrazení nominální proměnné:

- histogram
- výsečový graf

b) Ordinální proměnnná – má smysl uspořádání

Základní statistiky pro popis ordinální proměnné:

- četnost
- relativní četnost
- kumulativní četnost
- relativní kumulativní četnost
- modus

Grafické zobrazení ordinální proměnné:

- histogram
- výsečový graf
- Lorenzova křivka
- Paretův graf

Paretův princip – 80% následků pramení z 20% příčin

Paretova analýza – postup vedoucí k nalezení „životně důležité menšiny“ (spektra příčin ovlivňujících rozhodujícím způsobem následky)

Kvantitativní – Numerická proměnná

Míry polohy

- Průměr $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Mopdus (střed shortu)
- Kvantily (dolní kvartil, medián, horní kvartil, ...)

Míry variability

- Variační rozpětí $x_{max} - x_{min}$
- Interkvartilové rozpětí $IQR = x_{0,75} - x_{0,25}$
- Výběrová směrodatná odchylka $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- Variační koeficient $V_x = \frac{V}{\bar{x}}$, *popř.* $V_x = \frac{V}{\bar{x}} \cdot 100[\%]$

Míry šikmosti a špičatosti

- Výběrová šikmost $\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$
- Výběrová špičatost $\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$

Směrodatnou odchylku jakožto míru nejistoty měření zaokrouhlujeme **nahoru** na jednu, maximálně dvě platné cifry a míry polohy (průměr, kvantily ...) zaokrouhlujeme tak, aby nejnižší zapsaný řád odpovídal nejnižšímu zapsanému řádu směrodatné odchylky.

Identifikace odlehlých pozorování

- Vnitřní hradby: dolní mez: $h_D = x_{0,25} - 1,5IQR$; horní mez: $h_H = x_{0,75} + 1,5IQR$
- Z – souřadnice $z - skóre_i = \frac{x_i - \bar{x}}{s}$
- Mediánová souřadnice $x_{0,5} - skóre_i = \frac{x_i - x_{0,5}}{1,483MAD}$

Grafické zobrazení numerické proměnné:

- Empirická distribuční funkce
- Krabicový graf (angl. Box plot)
- Číslicový histogram (lodyha s listy, angl. Stem and leaf)

Kontrolní otázky

?

1. Test ze Statistiky píše velké množství studentů. Představte si, že každý z nich odpoví správně přesně na polovinu otázek. V tomto případě bude směrodatná odchylka počtu správných odpovědí
   a) rovna průměru,
   b) rovna mediánu,
   c) rovna nule,
   d) směrodatnou odchylku nelze určit bez dalších informací.
   e) dvojnásobku módu.

2. Největší kumulativní absolutní četnost v množině čísel se rovná
   a) součtu všech absolutních četností,
   b) 1,
   c) dvojnásobku průměru,
   d) dvojnásobku mediánu,
   e) dvojnásobku módu.

3. Několik studentů píše test ze Statistiky s 10-ti otázkami. Nejhorší výsledek jsou 3 správné odpovědi, nejlepší výsledek je 10 správných odpovědí. Jakou hodnotu má medián?
   a) 7 $(= 10 - 3)$
   b) 6,5$(= \frac{3+10}{2})$
   c) Medián nelze určit, pokud neznáme konkrétní výsledky jednotlivých žáků.

4. Představte si, že jste absolvovali normovaný test (např. SCIO test) a že Vám sdělili, že patříte do 91. percentilu. To znamená, že
   a) 90 žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy.
   b) 90 žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy.
   c) 90% žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy.
   d) 90% žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy.

5. Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že
   a) medián mzdy je vyšší než průměrná mzda,
   b) medián mzdy je nižší než průměrná mzda,
   c) medián mzdy je stejný jako průměrná mzda,
   d) o vztahu mezi mediánem mzdy a průměrnou mzdou nelze rozhodnout.

6. Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že
   a) mzdy mají kladnou šikmost,

b) mzdy mají zápornou šikmost,

c) mzdy mají kladnou špičatost,mzdy mají zápornou špičatost,

d) vztah mezi průměrem a 60% kvantilem nevypovídá nic o šikmosti ani o špičatosti dat.

7. Lékař Petře sdělil, že patří do 3. percentilu ohledně BMI (Body mass index – poměr váhy (kg) ke kvadrátu výšky (m)). Petra má pravděpodobně

a) podváhu,

b) normální váhu,

c) nadváhu,

d) bez dalších informací nelze usuzovat na Petřinu váhu.

8. Představte si, že jste absolvovali normovaný test (např. SCIO test). Měl(a) jste lepší výsledek než 85 studentů ze 100. To znamená, že

a) patříte do 99. decilu,

b) patříte do 95. decilu,

c) patříte do 10. decilu,

d) patříte do 9. decilu,

e) patříte do 2. kvartilu.

9. Pro srovnání variability váhy a výšky je možné použít

a) průměr,

b) rozptyl,

c) směrodatnou odchylku,

d) variační koeficient,

e) šikmost.

10. Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, průměrný plat ve firmě se zvýší

a) o 100,- Kč,

b) o 1000,- Kč,

c) průměrný plat se nezmění.

11. Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, průměrný plat ve firmě se zvýší

a) dvojnásobně,

b) čtyřnásobně,

c) průměrný plat se nezmění.

12. Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, průměrný plat ve firmě se zvýší
    a) o 20%,
    b) o 400%,
    c) o 40%,
    d) o 44%,
    e) Průměrný plat se nezmění.

13. Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, rozptyl platů ve firmě se zvýší
    a) o 100,- Kč,
    b) o 1000,- Kč,
    c) rozptyl platů se nezmění.

14. Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, rozptyl platů ve firmě se zvýší
    a) dvojnásobně,
    b) čtyřnásobně,
    c) rozptyl platů se nezmění.

15. Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, rozptyl platů ve firmě se zvýší
    a) o 20%,
    b) o 400%,
    c) o 40%,
    d) o 44%,
    e) Rozptyl platů se nezmění.

16. Největší kumulativní relativní četnost se rovná
    a) dvojnásobku průměru,
    b) dvojnásobku mediánu,
    c) dvojnásobku módu,
    d) součtu všech jednotlivých hodnot absolutních četností,
    e) 1.

17. Určete, zda jsou následující tvrzení pravdivá.
    a) Geometrický průměr je definován pro proměnné, které nabývají pouze kladných hodnot.Jedna čtvrtina hodnot je větší než 25% kvantil, zatímco tři čtvrtiny hodnot jsou menší.

b) Mají-li dvě proměnné stejný průměr a stejný rozptyl, mají stejný variační koeficient.

c) Mzdy v ČR mají kladnou šikmost. (V ČR mají zhruba 2/3 lidí podprůměrný plat.)

d) Nejčetnější hodnota v souboru se nazývá medián.

e) Rozptyl má vždy kladnou hodnotu.

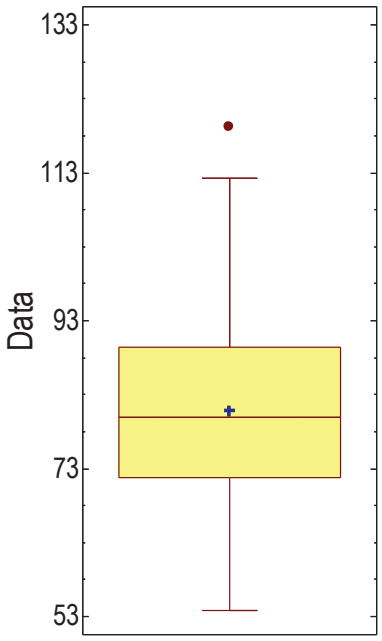
18. V grafu na Obr. 17, modrý křížek označuje

a) medián

b) průměr

c) modus

d) Interkvartilové rozpětí (IQR)



Obr. 1.17: Proměnná x

19. Určete zda jsou následující tvrzení pravdivá. Proměnná znázorněna na Obr. 17

a) neobsahuje odlehlá pozorování,

b) má kladnou šikmost,

c) je kladná,

d) má více než polovinu hodnot větších než 83.

20. Na atletických závodech mládeže žáci soutěžili ve 4 kategoriích. Určete, který výrok je nepravdivý.
   a) Na obrázku je znázorněn histogram a nejméně soutěžících bylo ve skoku do dálky.
   b) Celkem ve čtyřech kategoriích soutěžilo 80 žáků.
   c) Modus = hod koulí.
   d) Modus = 30.

Obr. 1.18: Zastoupení žáků na atletických závodech

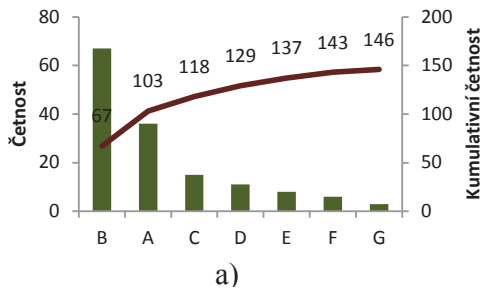
21. Následující graf Stem&leaf reprezentuje množství peněz, které studenti jedné třídy vybrali na humanitární účely.

0	11555889	8
1	112344555	(9)
2	005	6
3	025	3

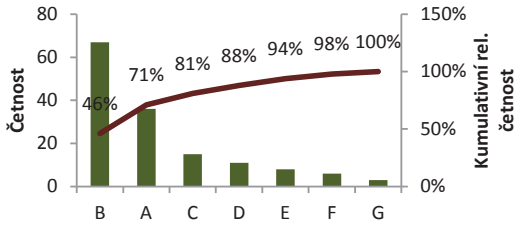
Multiply by 10^2

Které z následujících výroků jsou určitě nepravdivé?
   a) 10 studentů věnovalo méně než 120 Kč.
   b) Medián vybrané částky činí 120 Kč.
   c) Na humanitární účely přispělo v této třídě 23 studentů.
   d) Přispívající studenti věnovali na humanitární účely částky od 1,- Kč do 35,- Kč.

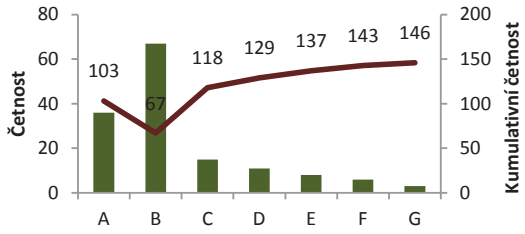
22. Určete, na kterém obrázku je zobrazen Paretův graf.



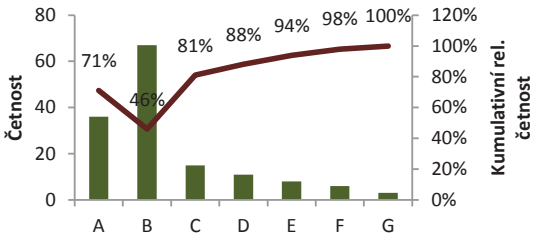
a)



b)



c)



d)

Úlohy k řešení

1. Zemědělské družstvo dostalo 1 000 kuřat s průměrnou váhou 1,37 kg. Cena byla 50,- Kč za kilogram. Během dne se prodalo 300 kuřat za 24 000,- Kč. Jaká byla průměrná váha neprodaných kuřat?

2. V jisté společnosti je průměrný plat 13 500,- Kč. 30% pracovníků s nejnižším platem má průměrně 9 000,- Kč. Na začátku roku došlo ke zvýšení platů pracovníků této skupiny jednotně o 500,- Kč. O kolik % vzrostl průměrný plat v celé společnosti následkem uvedeného zvýšení platu?

3. Petr, řidič zkušebního automobilu, jel z Ostravy do Olomouce rychlostí 70 km/h. Zpět jel rychlostí 90 km/h. Jaká byla průměrná rychlost zkušebního automobilu na trase Ostrava – Olomouc – Ostrava?

4. V jistém supermarketu byla ve stejné chvíli na 8 pokladnách měřena doba, během které pokladní ověří platnost platební karty zákazníka v bance. U pěti zákazníků trvalo ověření 2 minuty, u zbývajících tří to byly 3 minuty. Určete průměrnou dobu potřebnou k ověření platnosti karty.

5. Nákladní automobil jel z města A do města B rychlostí 40 km/h, z města B do města C rychlostí 50 km/h a z města C do města D rychlostí 60 km/h. Vypočítejte průměrnou rychlost, které dosáhl automobil na celé trase, víte-li, že:
   a) vzdálenost všech úseků je stejná – 5 km.
   b) Vzdálenost z A do B je 15% trasy a vzdálenost z C do D je 60% trasy.

6. Cena jedné akcie energetické společnosti vzrostla na burze XY v období od 13. do 15. března téhož roku z 952,50 Kč na 982,00 Kč. Jaký byl průměrný relativní přírůstek ceny této akcie?

7. Při sledování proměnné x byl určen aritmetický průměr 110 a rozptyl 800. Dodatečně byly zjištěny chyby u dvou údajů. Místo 85 mělo být správně 95 a místo 120 má být 150. Ostatních 18 údajů bylo správných. Opravte vypočítané charakteristiky (průměr a rozptyl).

8. Ze čtyřiceti hodnot byl vypočítán aritmetický průměr 7,50 a rozptyl 2,25. Při kontrole bylo zjištěno, že chybí dvě hodnoty proměnné – 3,8 a 7. Opravte uvedené charakteristiky.

9. V důsledku výstavby satelitního městečka poklesl průměrný věk obyvatel vesnice o 19%, rozptyl věku vzrostl o 21%. Jak se změnil variační koeficient?

10. Ze známých dat byl určen rozptyl měsíčních mezd 250 000 $K\check{c}^2$. Určete směrodatnou odchylku mezd, zvýší-li se všechny měsíční mzdy
    a) o 150,- Kč
    b) 1,2 krát
    c) o 4%.

11. Máme n údajů o měření teploty ve ^oC. Průměrná teplota je 20^oC a rozptyl je 10^oC^2. Určete

    a) průměrnou teplotu ve stupních Fahrenheita (^oF),

    b) rozptyl teploty ve stupních Fahrenheita (^oF),

    c) variační koeficienty teploty ve stupních Celsia (^oC) a ve stupních Fahrenheita (^oF). (Vztah pro převod stupňů Celsia na stupně Fahrenheita:$T_{o_F} = 1,8 \cdot T_{o_C+32}$)

12. Následující data představují zemi výroby automobilu. Data vyhodnoťte (četnost, rel. četnost, resp. kum. četnost a rel. kum. četnost, modus) a graficky znázorněte (histogram, výsečový graf).

USA	USA	Německo
ČR	Německo	Německo
Německo	ČR	ČR
ČR	USA	Německo

13. Následující data představují dobu čekání v minutách zákazníka na obsluhu. Zakreslete krabicový graf a číslicový histogram.

120	80	100	90
150	5	140	130
100	70	110	100

14. Při dopravním průzkumu byla sledována vytíženost vjezdu do určité křižovatky. Student provádějící průzkum si vždy při naskočení zeleného světla zapsal počet aut, čekajících ve frontě u semaforu. Jeho zapsané výsledky jsou:

    3 1 5 3 2 3 5 7 1 2 8 8 1 6 1 8 5 5 8 5 4 7 2 5 6 3 4 2 8 4 4 5 5 4 3 3 4 9 6 2 1 5 2 3 5 3 5 7 2 5 8 2 4 2 4 3 5 6 4 6 9 3 2 1 2 6 3 5 3 5 3 7 6 3 7 5 6

    Nakreslete krabicový graf, empirickou distribuční funkci a vypočtěte následující výběrové statistiky: průměr, výběrová směrodatná odchylka a interkvartilové rozpětí.

Řešení

Test 1c, 2a, 3c, 4d, 5b, 6a, 7a, 8d, 9d, 10a, 11a, 12a, 13c, 14b, 15d, 16d, pravdivá tvrzení – 17a, 17c a 17e, 18b, pravdivá tvrzení – 19b a 19c, 20d, nepravdivé, resp. neověřitelné výroky – 21b (Median je 130,- Kč.), 21d (Přispívající studenti věnovali na humanitární účely částky od 10,- Kč do 350,- Kč.)

Úlohy k řešení

1. 1,27 kg

2. 1,11 %

3. 78,8 km/h (harmonický průměr)

4. 2,3 min (vážený harmonický průměr)

5. a) 48,7 km/h
   b) 53,3 km/h

6. 1,54%

7. $\bar{x} = 112, s^2 = 854$

8. $\bar{x} = 7,40, s^2 = 2,46$

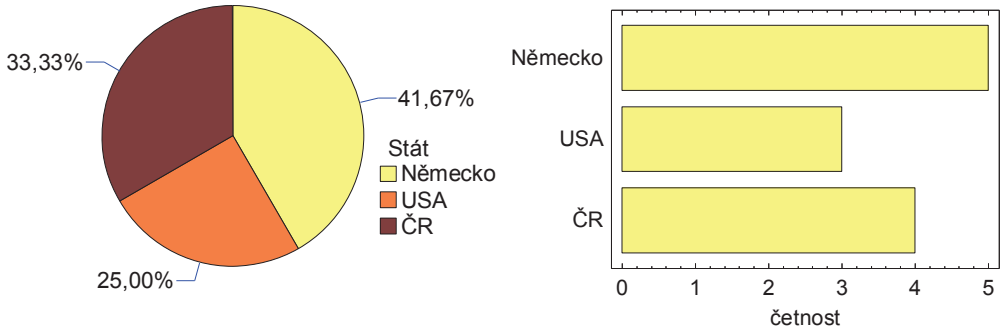
9. Vzrostl o 35,8%.

10. a) 500
    b) 600
    c) 520

11. a) $68^o F$
    b) $32^o F$
    c) $\mathrm{V}_{o_C} = 15,8\%$ $\mathrm{V}_{o_F} = 8,4\%$

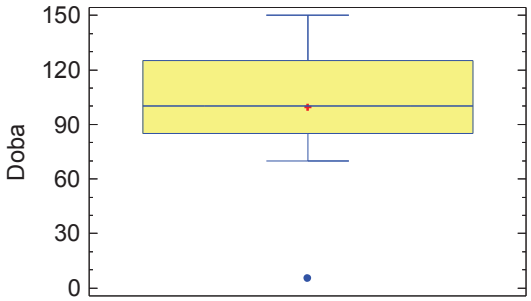
12. Kumulativní četnost a kumulativní relativní četnost nemá v tomto případě smysl. Modem, tj. zemí, v níž bylo vyrobeno nejvíce automobilů, je Německo.

Class	Value	Frequency	Relative Frequency
1	CR	4	0,3333
2	Nemecko	5	0,4167
3	USA	3	0,2500



13.

Average = 100
Median = 100
Variance = 1448
Standard deviation = 38
Minimum = 5,0
Maximum = 150,0
Lower quartile = 85
Upper quartile = 125
Stnd. skewness = -1,0
Stnd. kurtosis = 2,0
Coeff. of variation = 38,2%



Stem-and-Leaf Display for Doba: unit = 10,0 1 | 2 **represents 120,0**

LO| 5, 0
1 0 |
1 0 |
1 0 |
2 0 | 7
4 0 | 89
(4) 1 | 0001
4 1 | 23
2 1 | 45

14.

Count = 77
Average = 4,4
Median = 4,0
Variance = 4,5
Standard deviation = 2,1
Minimum = 1,0
Maximum = 9,0
Range = 8,0
Lower quartile = 3,0
Upper quartile = 6,0
Stnd. skewness = 1,1
Stnd. kurtosis = -1,2
Coeff. of variation = 48,7%