1 EXPLORATORNÍ ANALÝZA PROMĚNNÝCH

Čas ke studiu kapitoly: 120 minut

Cíl: Po prostudování této kapitoly budete umět použít

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

Výklad:

Původním posláním statistiky bylo zjišťování údajů o populaci na základě výběrového souboru. Pod pojmem **populace** přitom rozumějte souhrn všech existujících prvků, které sledujeme při statistickém výzkumu. Například:

1. *Provádíme-li stat. výzkum týkající se výšky 15-ti letých dívek, populaci tvoří všechny dívky, které mají právě 15 let.*
2. *Zkoumáme-li pevnost lan L50 vyrobených firmou LANOS, budeme za populaci považovat všechna lana L50 vyrobena firmou LANOS*

Vzhledem k tomu, že rozsah (počet prvků) populace je obvykle vysoký, provádí se většinou tzv. **výběrová šetření**, kdy se namísto celé populace zkoumá pouze její část. Zkoumaná část populace se nazývá **výběr**, popř. výběrový soubor. Otázkou je jak stanovit takový výběr, aby byl skutečně reprezentativní, tj. aby parametry výběru (např. průměr) dostatečně přesně reprezentovaly parametry populace. Jen si zkuste představit k jakým výsledkům bychom došli při předvolebním průzkumu prováděném na vzorku voličů, který bychom získali v domovech důchodců, popř. na schůzích mladých konzervativců.

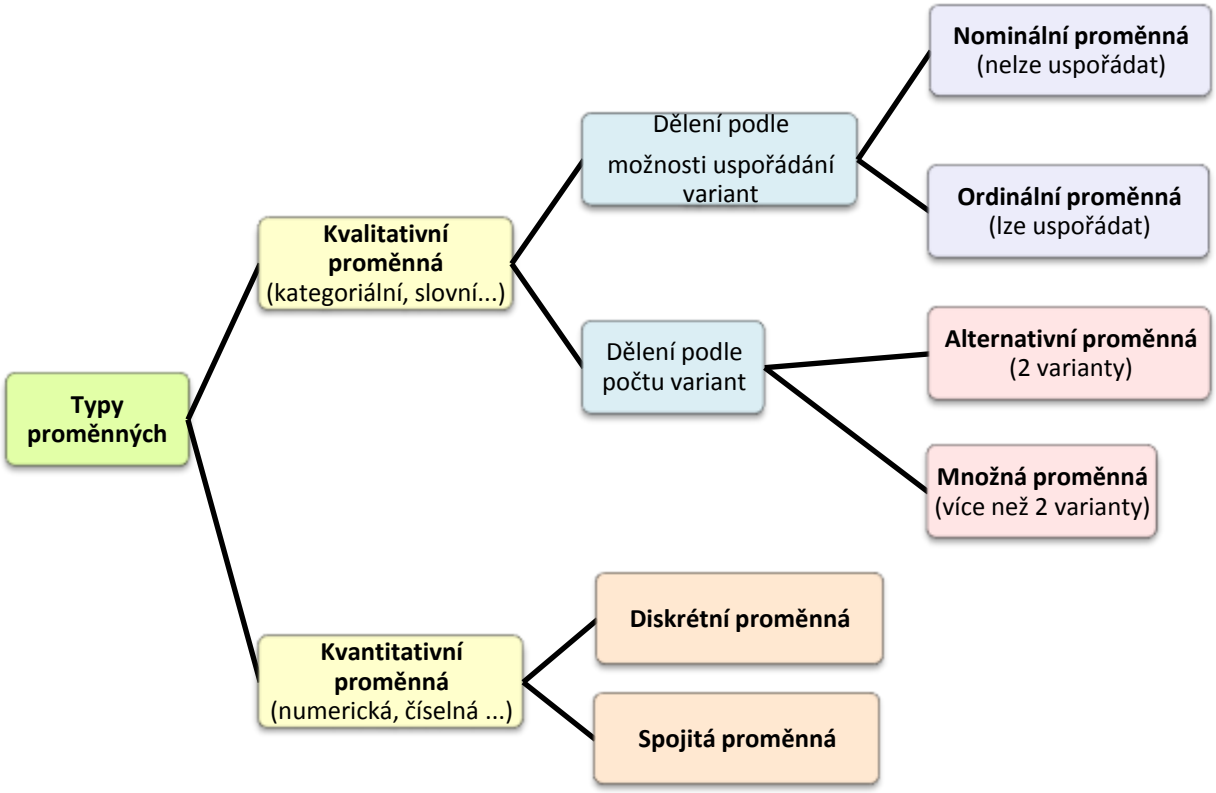
Existuje několik způsobu jak výběr provést. Abychom se vyvarovali opomenutí některých prvků populace, zvolíme tzv. **náhodný výběr,** v němž každý prvek populace má stejnou šanci být zařazen do výběru.

Je zřejmé, že výběrové šetření nemůže být nikdy tak přesné jako průzkum celé populace. Proč jej tedy preferujeme?

1. Úspora času a finančních prostředků (zejména u rozsáhlé populace)
2. Destruktivní testování (některé testy – pevnost lan, životnost zářivek, obsah cholesterolu v krvi, atd. – vedou k destrukci zkoumaných prvků; zamyslete se sami k čemu by vedlo testování celé populace)
3. Nedostupnost celé populace (při srovnávání působení faktorů okolí a dědičných znaků poskytují nejlepší informace identická dvojčata –jak je všechna sehnat a přesvědčit ke spolupráci?)

Nyní tedy víte, že statistikové dokáži popsat celou populaci na základě poznatků z výběru, proto přejdeme k základním výběrovým šetřením neboli k exploratorní analýze (exploratory data analysis – EDA). Údajům, které u souboru sledujeme budeme říkat **proměnné** a jejich jednotlivým hodnotám **varianty proměnné. Exploratorní (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:



- **Proměnná kvalitativní** – její varianty jsou vyjádřeny slovně a podle vztahu mezi jednotlivými hodnotami se dělí na dvě základní podskupiny:

    - **Proměnná nominální (jmenná)** – nabývá rovnocenných variant; nelze je 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, XL))

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

    - **Proměnná alternativní** – nabývá pouze dvou různých variant (např. pohlaví...)

    - **Proměnná množná** – nabývá více než dvou různých variant (např. vzdělání, jméno, barva očí...)

- **Proměnná kvantitativní –** je vyjádřena číselně a dělí se na:

    - **Proměnná diskrétní** – nabývá konečného nebo spočetného množství variant (např. známka z matematiky)

- **Proměnná diskrétní konečná –** nabývá konečného počtu variant (např. známka z matematiky)

- **Proměnná diskrétní spočetná –** nabývá spočetného množství variant (např. věk v letech, výška v centimetrech, váha v kilogramech...)

▪ **Proměnná spojitá** - nabývá libovolné hodnoty z $\Re$ nebo z nějaké podmnožiny $\Re$ (např. výška, hmotnost, vzdálenost měst...)

Průvodce studiem:

Tak, 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 "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. A jdeme na to!

Výklad:

1.1 Statistické charakteristiky kvalitativních proměnných

V tuto chvíli již víte, že kvalitativní (slovní) proměnná má dva základní typy – nominální a ordinální.

1.1.1 Nominální proměnná

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

- **Četnost n_i** (absolutní četnost, 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 četnost označíme n_1, n_2, …, 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 danou variantou, použijeme pro popis proměnné relativní četnost.

- **Relativní četnost p_i** (relative frequency)

  je definována jako:

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

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

  $$p_1 + p_2 + \ldots + p_k = \sum_{i=1}^{k} p_i = 1$$

Při zpracování kvalitativní proměnné je vhodné četnosti i relativní četnosti uspořádat do tzv. **tabulky rozdělení četnosti** (frequency table):

TABULKA ROZDĚLENÍ ČETNOSTI		
Hodnoty x_i	**Absolutní četnost**	**Relativní četnost**
	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.

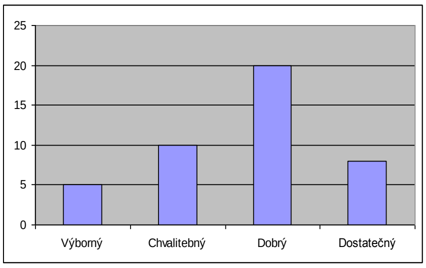
  Modus tedy můžeme chápat jako typického reprezentanta souboru. V případě, že se ve statistickém souboru vyskytuje více variant s maximální četnosti, modus neurčujeme.

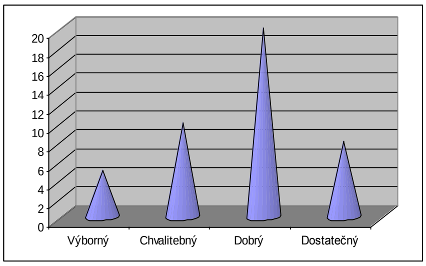
1.1.2 Grafické znázornění kvalitativní proměnné

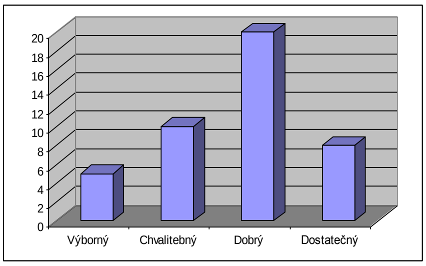
Pro větší názornost analýzy proměnných se ve statistice často užívají **grafy**. Pro nominální proměnnou jsou to tyto dva typy:

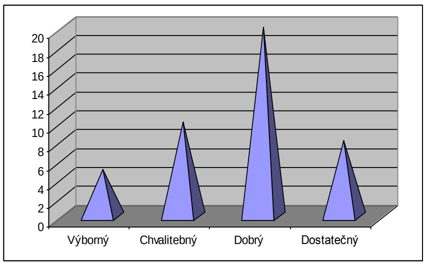
- **Histogram** (sloupcový graf, bar chart)
- **Výsečový graf** (koláčový graf, pie chart)

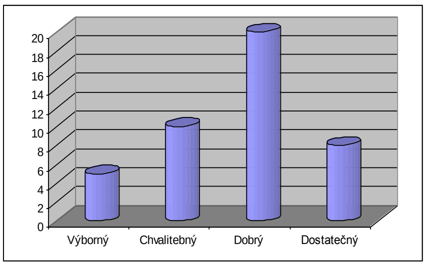
Histogram je klasickým grafem, v němž na jednu osu vynášíme varianty proměnné a na druhou osu jejich četnosti. Jednotlivé hodnoty četnosti jsou pak zobrazeny jako sloupce (obdélníky, popř. úsečky, hranoly, kužely...)

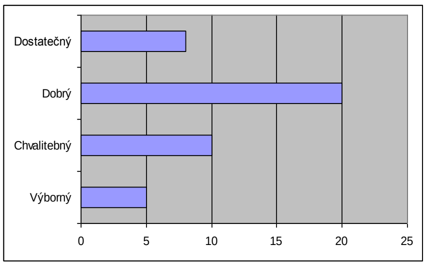




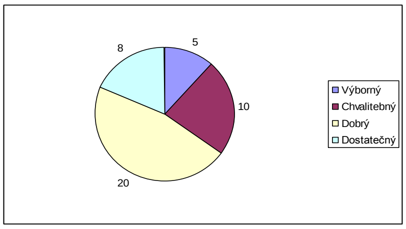


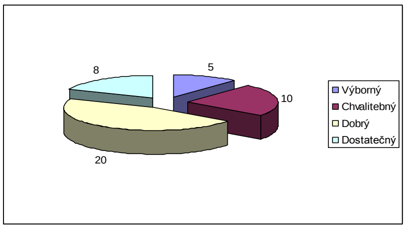


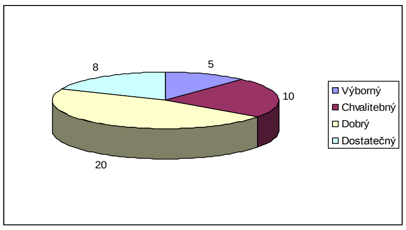


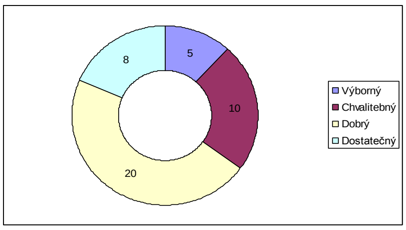


Výsečový graf prezentuje relativní četnosti jednotlivých variant proměnné, přičemž jednotlivé relativní četnosti jsou úměrně reprezentovány plochami příslušných kruhových výsečí. (Změnou kruhu na elipsu dojde k trojrozměrnému efektu.)



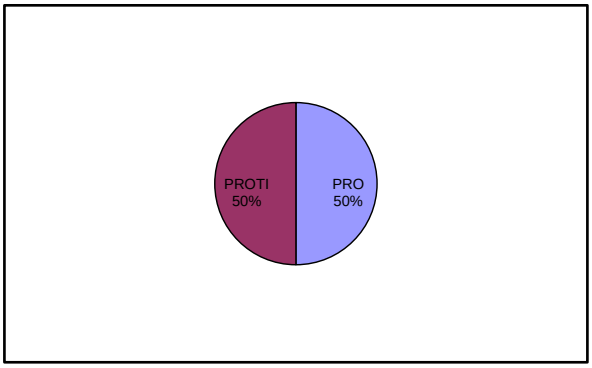




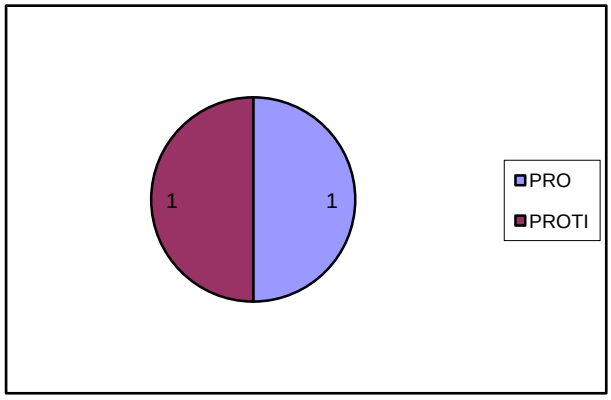


POZOR!!! V případě výsečového grafu si dejte zvláštní pozor na popis grafu. Jednotlivé výseče nestačí označit relativními četnostmi bez uvedení četnosti absolutních, popř. bez uvedení celkového počtu pozorování, to by mohlo vést k matení (ať už záměrnému nebo nechtěnému) toho, jemuž je graf určen. Zamyslete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



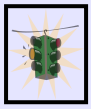
Co vy na to? Zajímavé výsledky, že? A věřte, nevěřte – pravdivé. A teď graf doplníme tak, jak jsme Vám to doporučili:



Co si myslíte nyní? Z druhého grafu je patrné, že byli dotazováni dva lidé – jeden byl pro a jeden proti. Jaká je vypovídací schopnost takovéto ankety? Jaký je nyní Váš názor na prezentované výsledky? A závěr? Vy 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 či zda je to jeho záměr.

Průvodce studiem:

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



Řešený příklad:

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

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

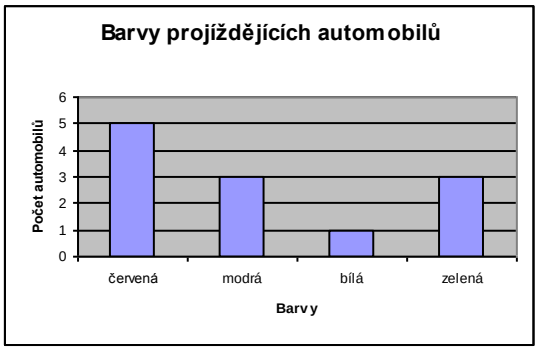
Řešení:

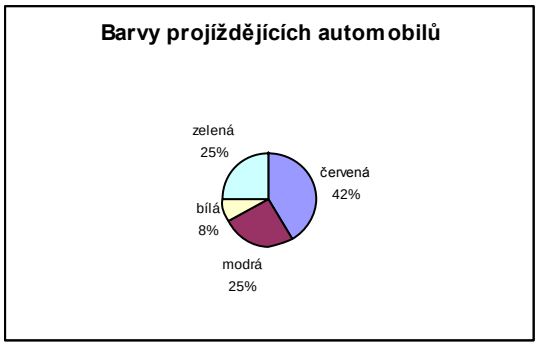
Je zřejmé, že se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že barvy automobilů nemá smysl seřazovat ani porovnávat, můžeme konstatovat, že se jedná o proměnnou nominální.

Pro její popis tedy zvolíme tabulku četností, určíme modus a barvu projíždějících automobilů znázorníme prostřednictvím histogramu a výsečového grafu.

TABULKA ROZDĚLENÍ ČETNOSTI		
Barvy projíždějících automobilů	**Absolutní četnost** n_i	**Relativní četnost** 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

Modus = červená (tj. v zaznamenaném vzorku se vyskytlo nejvíce červených automobilů)





Celkem bylo sledováno 12 automobilů

Výklad:

1.1.3 Ordinální proměnná

Dále budeme pokračovat popisem ordinální proměnné. Ordinální proměnná, stejně jako nominální, nabývá v rámci souboru různých slovních variant, avšak tyto varianty jsou seřaditelné, tj. můžeme určit, která je "menší" a která je "větší".

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

- **Kumulativní četnost m_i**

  definujeme jako počet hodnot proměnné, které nabývají varianty nižší nebo rovné i-té variantě.

  Uvažte např. proměnnou "známka ze statistiky", která nabývá variant: "výborný", "velmi dobrý", "dobrý", "neprospěl", pak např. kumulativní četnost pro variantu "dobrý" bude rovna počtu studentů, kteří ze statistiky získali známku "dobrý" 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 – n.

  $$m_k = n$$

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

- **Kumulativní relativní četnost F_i**

  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:

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

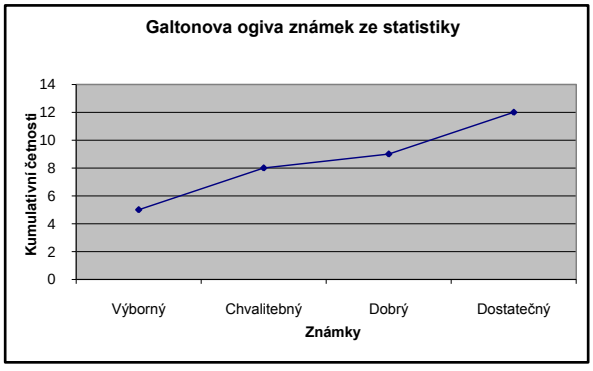
Obdobně jako u nominální proměnné, můžeme i u ordinální proměnné prezentovat statistické charakteristiky pomocí tabulky rozdělení četnosti. Ta obsahuje ve srovnání s tabulkou rozdělení četností pro nominální proměnnou navíc hodnoty kumulativních a kumulativních relativních četností.

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

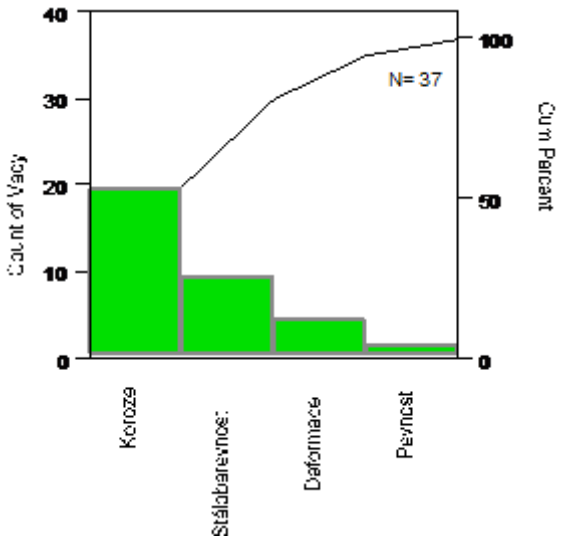
1.1.4 Grafické znázornění ordinální proměnné

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

Polygon kumulativních četností (Galtonova ogiva, S křivka) je spojnicovým grafem, v němž se na vodorovnou osu vynáší jednotlivé varianty proměnné v pořadí od "nejmenší" do "největší" a na svislou osu příslušné hodnoty kumulativních četností.
Všimněte si, směrnice (sklon) polygonu kumulativních četností je tím nižší, čím nižší je četnost jednotlivých variant.



Paretův graf je v technických disciplínách často užívaným grafem tvořeným spojením histogramu a polygonu kumulativních četností, v němž se na vodorovnou osu vynáší jednotlivé varianty proměnné v pořadí "od té s největším po tu s nejmenším významem".



Průvodce studiem:

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

Řešený příklad:

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:

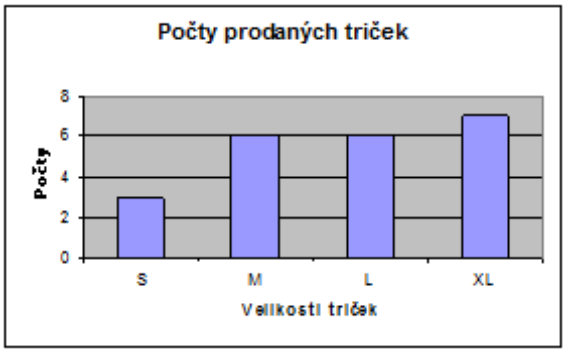
ada) Zřejmě se jedná o kvalitativní (slovní) proměnnou a vzhledem k tomu, že velikosti triček lze seřadit, jde o proměnnou ordinální. Pro její popis proto použijeme tabulku četností pro ordinální proměnnou, v níž varianty velikosti triček budou seřazeny od nejmenší po největší (S, M. L, XL) a modus.

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

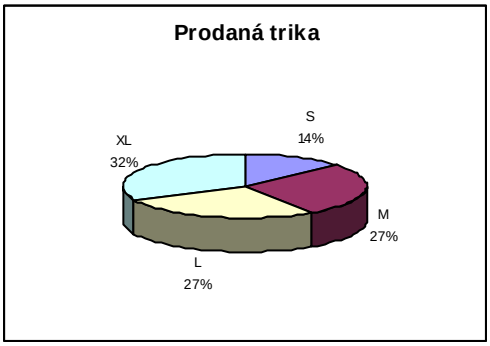
Modus = XL (nejvíce lidí si koupilo tričko velikosti XL)

Grafický výstup bude tvořit histogram, výsečový graf a polygon kumulativních četností (jelikož se nejedná o technická data, Paretův graf vytvářet nebudeme).

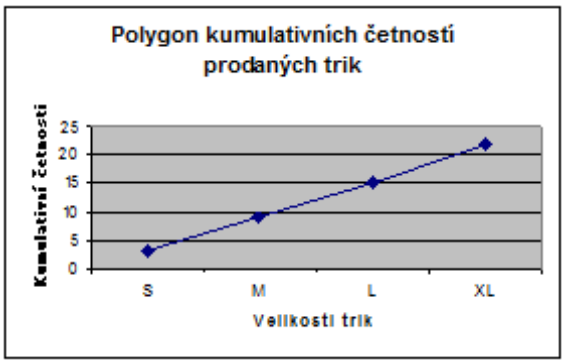
Grafický výstup:



Histogram



Celkem bylo prodáno 22 triček



Galtonova ogiva, S-křivka

adb) 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.

Výklad:

1.2 Statistické charakteristiky kvantitativních proměnných

Pro popis kvantitativní proměnné můžeme použít většinu statistických charakteristik užívaných pro popis proměnné ordinální (četnost, relativní četnost, kumulativní četnost, kumulativní relativní četnost), což doplníme dalšími dvěmi skupinami charakteristik:

- **míry polohy** – ty určují typické rozložení hodnot proměnné (jejich rozmístění na číselné ose)

a

- **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** $\bar{x}$

Jeho hodnotu získáme pomocí známého vztahu:

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

kde: x_i ... jednotlivé hodnoty proměnné
n ... rozsah výběrového souboru (počet hodnot proměnné)

Poměrně známé jsou i **vlastnosti aritmetického průměru**:

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

   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. $\forall (a \in \Re): \left(\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} \Rightarrow \frac{\sum_{i=!}^{n} (a + x_i)}{n} = a + \bar{x} \right)$

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

3. $\forall (b \in \Re): \left(\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} \Rightarrow \frac{\sum_{i=!}^{n} (bx_i)}{n} = b\bar{x} \right)$

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

Přestože to tak na první pohled vypadá, aritmetický průměr není vždy pro výpočet průměru výběrového souboru nejvhodnější. Pracujeme-li, například, s proměnnou představující relativní změny (růstové indexy, cenové indexy...), používáme tzv. geometrický průměr. Pro výpočet průměru v případech, kdy proměnná má charakter části z celku (úlohy o společné práci...), používáme průměr harmonický.

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 dokáží 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).

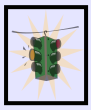
  Naproti tomu **u spojité proměnné** považujeme za modus $\hat{x}$ hodnotu kolem níž je největší koncentrace hodnot proměnné.
  Pro určení této hodnoty využijeme **shorth**, což je nejkratší interval, v němž leží alespoň 50% hodnot proměnné (v případě výběru o rozsahu $n = 2k\ (k \in \mathrm{N})$ (sudý počet hodnot), leží v shorthu k hodnot – což je 50% (n/2) hodnot proměnné, v případě výběru o rozsahu $n = 2k+1\ (k \in \mathrm{N})$ (lichý počet hodnot), leží v shorthu $k+1$ hodnot - což je o ½ více než je 50% hodnot proměnné (n/2+½)).
  Modus pak definujeme jako střed shorthu.

  Z předcházejících definic vyplývá, že délka shorthu (horní mez – dolní mez) je jednoznačně dána, to však neplatí 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 procvičit na jednoduchém příkladu, který Vám snad případné nejasnosti ozřejmí.



Řešený příklad:

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 (zdůvodnění snad není nutné):

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

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 (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: n/2+½ = 11/2 +1/2 = 12/2 = 6).

A další postup?

- Proměnnou seřadíme
- Určíme délky všech 6-ti členných intervalů, v nichž $x_i < x_{i+1} < \ldots < x_{i+5}$
- 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 shorthu:

$$\hat{x} = \frac{34 + 43}{2} = 38{,}5$$

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

Výklad:

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

- **Výběrové kvantily**

  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) definován 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 $\mathbf{x_p}$.

  V praxi se nejčastěji setkáváme s těmito kvantily:

- **Kvartily**

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

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

  Horní kvartil $x_{0,75}$ = 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 stejně četné části.

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

  Decily dělí výběrový soubor na 10 stejně četných části.

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

  Percentily dělí výběrový soubor na 100 stejně četných části.

- **Minimum x_{min} a Maximum x_{max}**

  $x_{\min} = x_0$, tj. 0% hodnot je menších než minimum
  $x_{\max} = x_1$, tj. 100% hodnot je menších než maximum

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:

   $z_p = np + 0{,}5$

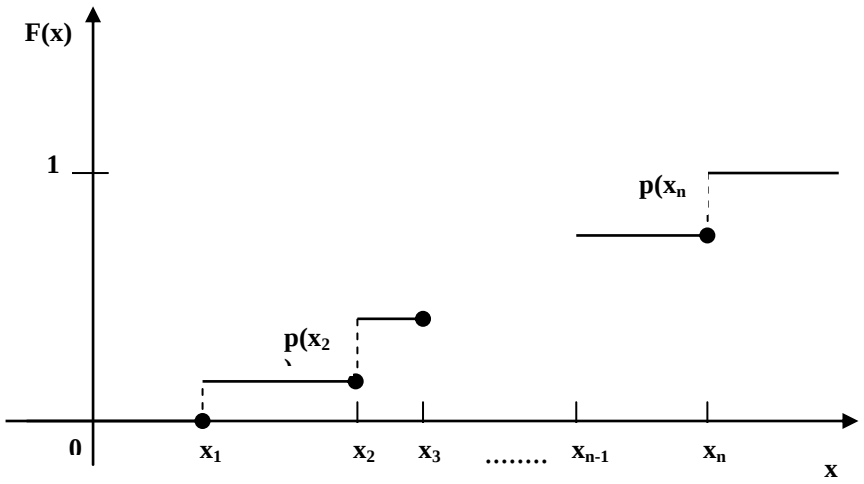
   Není-li z_p celé číslo, pak daný kvantil určíme jako průměr prvků s pořadím $[z_p]$ a $[z_p]+1$. (**Pozn.**: $[a]$ značíme celou část čísla a.)

Za zmínku zajisté stojí i **vztah mezi kvantily a kumulativní relativní četnosti**. Zřejmě lze říci, že hodnota p udává kumulativní relativní četnost kvantilu x_p, tj. relativní četnost těch hodnot proměnné, které jsou menší než kvantil x_p. Kvantil a kumulativní relativní 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:

$$F(x)=\begin{cases} 0 & pro\ x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & pro\ x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & pro\ x_n < x \end{cases}$$



  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**

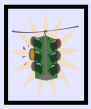
  Název MAD je zkratkou anglické definice – **m**edian **a**bsolute **d**eviation from the median, čili česky: medián absolutních odchylek od mediánu

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:

Moc teorie? Abyste se ujistili, že nic není tak černé jak to vypadá, zkuste pokračovat v předcházejícím řešeném příkladu.



Řešený příklad:

Pro data z předcházejícího příkladu určete:

a) všechny kvartily,
b) interkvartilové rozpětí
c) MAD
d) zakreslete empirickou distribuční funkci

Řešení:

ada) Naším úkolem je určit dolní kvartil $x_{0,25}$; medián $x_{0,5}$ a horní kvartil $x_{0,75}$. Budeme-li dodržovat postup doporučený pro určování kvantilů, znamená to – data seřadit a přiřadit jim pořadí. Splnění prvních dvou bodů postupu ukazuje následující tabulka:

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

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

Dolní kvartil $x_{0,25}$: $p = 0{,}25;\ n = 11 \Rightarrow z_p = 11.0{,}25 + 0{,}5 = 3{,}25$,

Dolní kvartil je tedy průměrem prvků s pořadím 3 a 4 - $x_{0,25} = \frac{27+34}{2} = 30{,}5$ 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 $x_{0,5}$: $p = 0{,}5;\ n = 11 \Rightarrow z_p = 11.0{,}5 + 0{,}5 = 6 \Rightarrow x_{0,5} = 35$

Tj. polovina hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 35 let (50% z nich má 35 let a více).

Horní kvartil $x_{0,75}$: $p = 0{,}75;\ n = 11 \Rightarrow z_p = 11.0{,}75 + 0{,}5 = 8{,}75$

Horní kvartil je tedy průměrem prvků s pořadím 8 a 9 - $x_{0,75} = \frac{42+43}{2} = 42{,}5$ let.

Tj. 75% hudebníků vystupujících na přehlídce dechových orchestrů je mladších než 42,5 let (25% z nich má 42,5 let a více).

adb) **Interkvartilové rozpětí IQR**:

IQR = $x_{0,75} - x_{0,25} = 42{,}5 - 30{,}5 = 12$

adc) **MAD**

Chceme-li určit tuto statistiku, budeme postupovat přesně podle toho co nám říká definice (medián absolutních odchylek od mediánu), tudíž dodržíme výše uvedený postup, jehož aplikaci vám ukazuje následující tabulka.

$x_{0,5} = 35$

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 **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	$1 = \lvert 34 - 35 \rvert$	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	$47 = \lvert 82 - 35 \rvert$	47

$$MAD = M_{0,5}$$
$$p = 0{,}5;\ n = 11 \Rightarrow z_p = 11.0{,}5 + 0{,}5 = 6 \Rightarrow M_{0,5} = 8$$

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

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

$$F(x) = \begin{cases} 0 & pro\ x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & pro\ x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & pro\ x_n < x \end{cases}$$

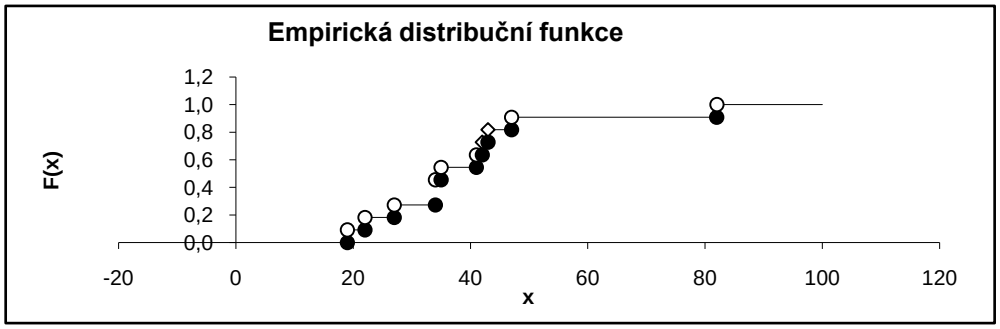
- do tabulky si zapíšeme seřazené hodnoty proměnné, jejich četnosti, relativní četnosti a z nich odvodíme empirickou distribuční funkci:

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

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.

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



Průvodce studiem:

Zvládli jste to? Gratuluji. Pokud jste s příkladem měli nějaké problémy, doporučuji Vám, abyste si pasáž o kvantilech a empirické distribuční funkci znovu důkladně prostudovali – není to naposled, co o nich slyšíte.

Výklad:

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 celou proměnnou.

Následující tři 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**

  je nejrozšířenější mírou variability výběrového souboru. Určujeme jej podle vztahu:

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

  tzn. 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 konstanty je roven nule,

     neboli: jsou-li všechny hodnoty proměnné stejné, má soubor nulovou rozptýlenost

2. $$\forall a \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge \left(y_i = a + x_i \right) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = s^2$$

   neboli: přičteme-li ke všem hodnotám proměnné libovolnou konstantu, výběrový rozptyl proměnné se nezmění

3. $$\forall b \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge \left(y_i = bx_i \right) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = b^2 s^2$$

   neboli: 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 rozměr této charakteristiky je druhou mocninou rozměru proměnné. (Např. je-li proměnnou denní tržba uvedena v Kč, bude výběrový rozptyl této proměnné vyjádřen v $Kč^2$.) Tento nedostatek odstraňuje další míra variability, a tou je:

- **Výběrová směrodatná odchylka s**

  je definována prostě jako kladná odmocnina výběrového rozptylu:

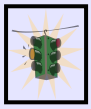
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

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

- **Variační koeficient V_x**

  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%.

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



Řešený příklad:

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 byly následující:

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

Porovnejte obě technologie pomocí základních charakteristik exploratorní statistiky (průměru a rozptylu, popř. směrodatné odchylky).

Řešení:

- Nejprve se pokusíme porovnat obě technologie pouze za pomocí průměru:

Průměr pro starou technologii:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{475 + 436 + \ldots + 426}{5} = 463{,}0 \quad [^oC]$$

Průměr pro novou technologii:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{485 + 390 + \ldots + 488}{5} = 468{,}6 \quad [^oC]$$

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

A co na to míry variability?

Stará technologie:

Výběrový rozptyl:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(475 - 463{,}0)^2 + (436 - 463{,}0)^2 + \ldots + (426 - 463{,}0)^2}{5-1} = 916{,}3 \quad [^oC^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s_x^2} = \sqrt{916{,}3} = 30{,}3 \quad [^oC\ ]$$

Nová technologie:

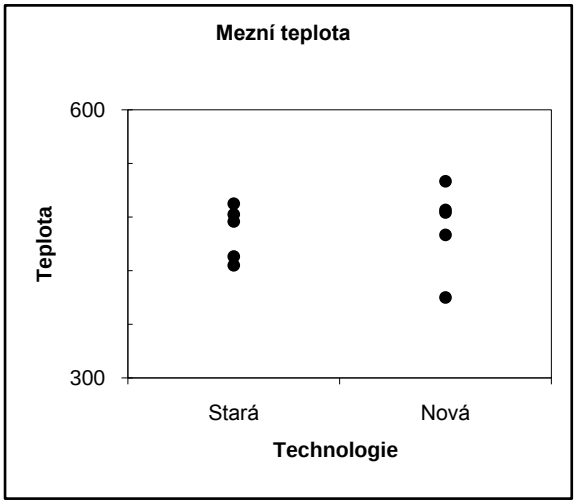
Výběrový rozptyl:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{(485-468{,}6)^2 + (390-468{,}6)^2 + \ldots + (488-468{,}6)^2}{5-1} = 2384{,}4 \quad [^oC^2]$$

Výběrová směrodatná odchylka:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{s_y^2} = \sqrt{2384{,}4} = 48{,}8 \quad [^oC\ ]$$

Tady pozor. 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.



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 exploratorní analýzy, statistika nám nabízí exaktnější metody pro rozhodnutí takovýchto případů (testování hypotéz), s nimiž se seznámíte později.

Výklad:

A nyní se vrátíme k exploratorní statistice jako takové. Vzpomínáte si ještě na zmínku o odlehlých pozorováních? Dozvěděli jste se, že jako odlehlá pozorování označujeme ty hodnoty proměnné, které se mimořádně liší od ostatních hodnot a tím ovlivňují např. reprezentativnost průměru. Nyní se dozvíte jak se tyto hodnoty identifikují.

- **Identifikace odlehlých pozorování (outliers)**

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

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

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

  2. **z-souřadnice:** 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:

     $$z - souř._i = \frac{x_i - \bar{x}}{s}$$
     $$\left(\left|z - souř._i\right| > 3\right) \Rightarrow x_i \text{ je odlehlým pozorováním}$$

  3. **$x_{0,5}$-souřadnice:** 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 1,483.MAD. Tedy:

     $$\text{mediánová souř.}_i = \frac{x_i - x_{0,5}}{1{,}483.MAD}$$
     $$\left(\left|\text{mediánová souř.}_i\right| > 3\right) \Rightarrow x_i \text{ je odlehlým pozorováním}$$

  V konkrétním případě si můžete pro identifikaci odlehlých pozorování zvolit libovolné z těchto tří pravidel. Za zmínku stojí snad jen to, že z-souřadnice je "méně přísná" k odlehlým pozorováním než mediánová souřadnice. To je způsobeno tím, 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. Využívá je však velká část statistických programů.

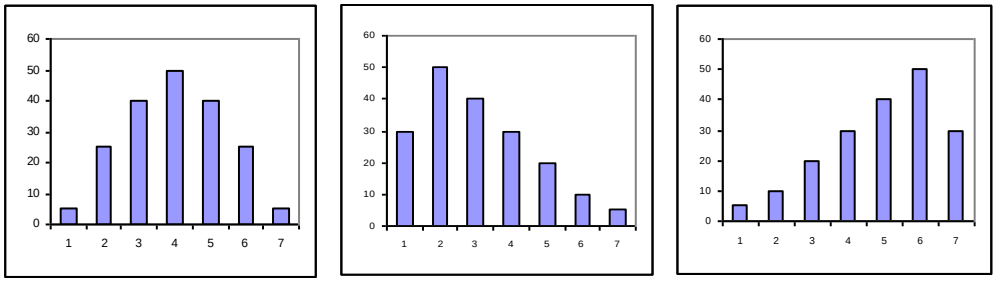
- **Výběrová šikmost (skewness) a**

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



a=0 a>0 a<0

Souvislost mezi šikmostí a charakteristikami polohy

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

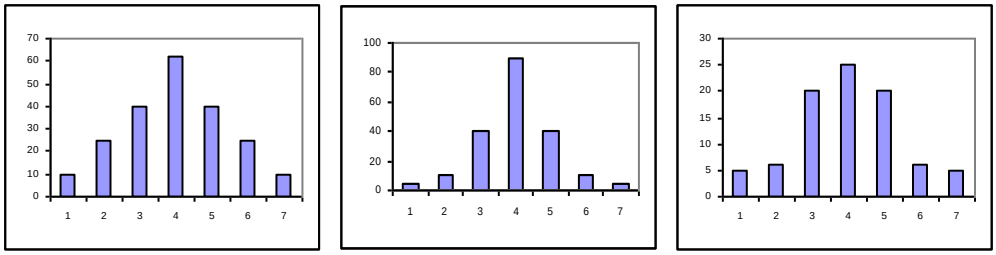
- **Výběrová špičatost (kurtosis) b**

  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 interpretujeme 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

Průvodce studiem:

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

Výklad:

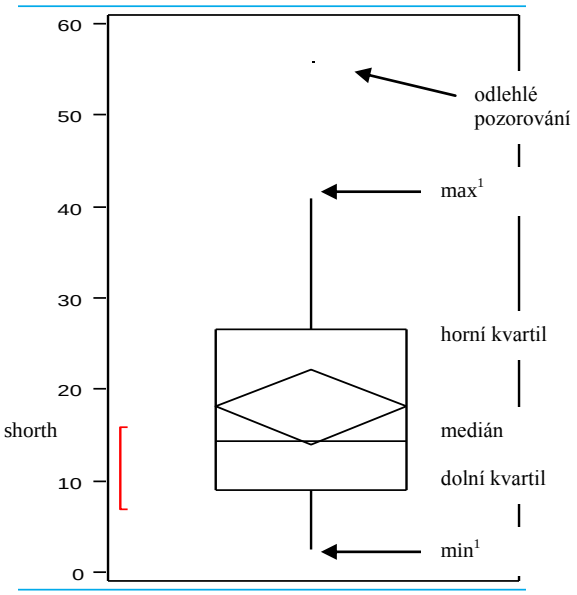
1.2.2 Grafické znázornění kvantitativní proměnné

- **Krabicový graf (Box plot)**

Krabicový graf se ve statistice využívá od roku 1977, kdy jej poprvé prezentoval statistik 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 výše uvedeném obrázku.

Odlehlá pozorování jsou znázorněna jako izolované body, konec horního (popř. konec dolního) vousu představují maximum max[1] (popř. minimum min[1]) proměnné po vyloučení odlehlých pozorování, "víko" krabice udává horní kvartil, "dno" dolní kvartil, vodorovná úsečka uvnitř krabice označuje medián. Svorka vně krabice ukazuje shorth.

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.



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).

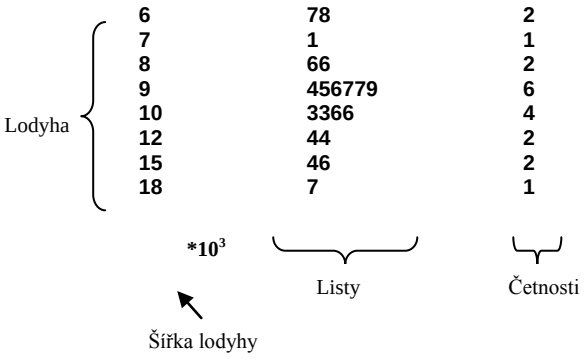
- **Číslicový histogram (Stem and leaf plot, Lodyha s listy...)**

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

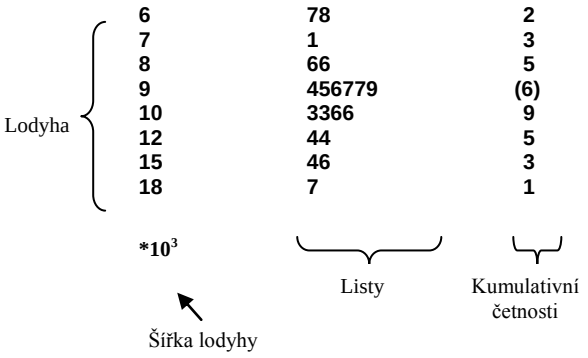
A vy nyní stojíte před problémem – jak tato data znázornit. Pokud se nad touto otázkou trochu zamyslíme, zjistíme, že pro naší informaci nejsou tak důležité koruny ani desetikoruny rozdílu. V tomto případě se nám jedná přinejmenším o stokoruny. Co kdybychom tedy informaci o "nedůležitých" řádech zanedbali a znázornili setříděná data pouze na základě vyšších řádů? My jsme se rozhodli, že důležitý řád jsou pro nás



stovky. 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.
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 prezentující číslicový histogram pro náš případ. Např. první řádek reprezentuje dvě hodnoty – (6.7 a 6.8)*10^3 Kč, tj. 6700 Kč a 6800 Kč (koruny a desetikoruny jsme zanedbali), šestý řádek reprezentuje také dvě hodnoty – (12.4 a 12.4)*10^3 Kč, tj. dvě osoby s průměrným měsíčním příjmem 12400 Kč, atd. – už je to jasnější, dokázali byste tento graf sestrojit sami?

Existují různé modifikace tohoto grafu. 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 snad jen jednu radu – dlouhé lodyhy s krátkými listy a krátké lodyhy s dlouhými listy svědčí o nevhodné volbě měřítka.



0	66788999999	11
1	000022558	9

*10^4

 **Shrnutí:**

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
- Paterův graf
- Polygon kumulativních četností (Galtonova ogiva)

Kvantitativní - Numerická proměnná

Míry polohy

- Průměr $\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n}$
- Modus (střed shorthu)
- Kvantity (dolní kvartil, medián, horní kvartil, …)

Míry variability

- Interkvartilové rozpětí $IQR = x_{0,75} - x_{0,25}$
- Výběrový rozptyl $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
- Výběrová směrodatná odchylka $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
- Variační koeficient $V_x = \frac{s}{\bar{x}}$
- Výběrová šikmost $\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$
- Výběrová špičatost $\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$

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 $z - souř._i = \frac{x_i - \bar{x}}{s}$
- Mediánová souřadnice $mediánová\ souř._i = \frac{x_i - x_{0,5}}{1{,}483.MAD}$

Grafické zobrazení numerické proměnné:

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

Otázky

1. Čím se zabývá exploratorní statistika?

2. Charakterizujte základní typy proměnných.

3. Které statistické charakteristiky mohou obsahovat tabulky četnosti (pro který typ proměnné)?

4. Definujte základní statistiky popisující kvalitativní proměnnou.

5. Co jsou to odlehlá pozorování a jak je identifikujeme?

6. Na výskyt odlehlých pozorování ve výběru je citlivý:
   a) Medián
   b) Aritmetický průměr
   c) Horní kvartil

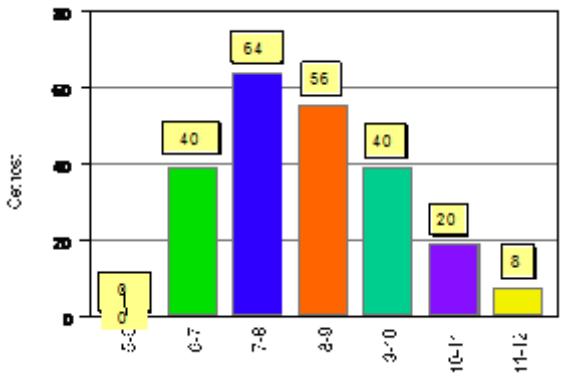
7. Definujte základní míry variability.

8. Co je to empirická distribuční funkce?

9. Jaké jsou možnosti grafické prezentace kvalitativní (kvantitativní) proměnné?

Úlohy k řešení

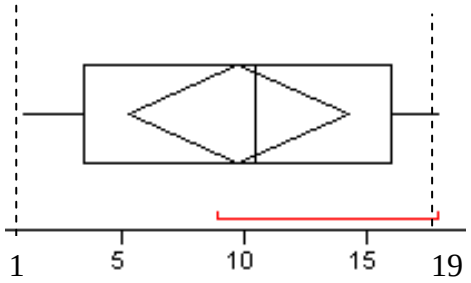
1. Následující histogram zobrazuje platy zaměstnanců (v tis. Kč) jedné akciové společnosti.



Které z následujících výroků jsou určitě **chybné,** popř. neověřitelné?

a) Modus platů je třída od 7 do 8tis. Kč
b) Celkový počet zaměstnanců firmy (zahrnutých do průzkumu) je 250
c) Průměrný plat činí 7 977,- Kč

2. Tento krabicový graf vypovídá o výdělcích (v tis. Kč,-) studentů během letních prázdnin.



Označte výroky, které zjevně **neodpovídají** zobrazené skutečnosti.

a) Student si vydělal maximálně 19 tis. Kč,-
b) Interkvartilové rozpětí výdělků činí zhruba 10 tis. Kč,-
c) Polovina studentů si vydělala méně než cca. 11 tis. Kč,-
d) Nejkratší interval, v němž leží alespoň 50% výdělků (Shorth), je cca (5;15) tis. Kč,-

3. Následující graf Stem & leaf zobrazuje roční úhrn srážek (v mm) na Lysé hoře v letech 1966 – 1996.

4	73 86	2
5	15 27 52 53 61	7
6	05 09 23 30 33 33 41 60 64 65 72 98	(12)
7	05 14 25 41 48 59 98	11
8	09 32 37	4
9	10	1

Multiply by 10^2

Označte výroky, které zjevně **neodpovídají** zobrazené skutečnosti.

a) Údaje ve třetím sloupci udávají kumulativní četnosti (při kumulaci shora a zdola, hodnota ve třetím řádku udává absolutní četnost)
b) Medián ročních úhrnu srážek činí 668mm.
c) V roce 1994 byl roční úhrn srážek na Lysé hoře 832mm.
d) V roce 1966 byl zaznamenán nejnižší roční úhrn srážek na Lysé hoře.

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

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

5. Následující data představují dobu čekání [min] zákazníka na obsluhu. Zakreslete box plot a graf stem and leaf.

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

6. 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:

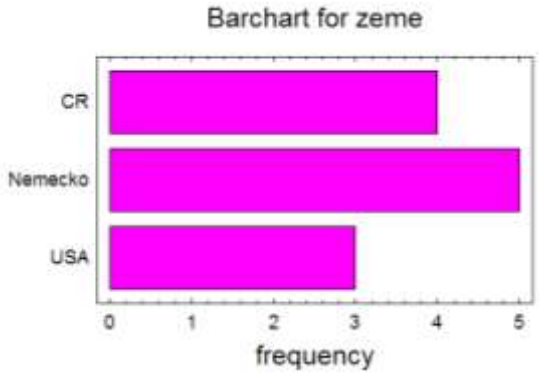
1. b), c)

2. b), d)

3. b), c), d)

4.

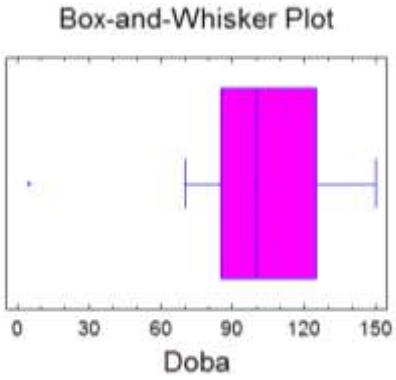
Class	Value	Frequency	Relative Frequency
1	CR	4	0,3333
2	Nemecko	5	0,4167
3	USA	3	0,2500



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

5.

```
Count = 12
Average = 99,5833
Variance = 1447,54
Standard deviation = 38,0465
Minimum = 5,0
Maximum = 150,0
Range = 145,0
Stnd. skewness = -1,81656
Stnd. kurtosis = 1,99356
```



Stem and leaf

0	5	1
7	0	2
8	0	3
9	0	4
10	000	7
11	0	(1)
12	0	4
13	0	3
14	0	2
15	0	1

*10

6.

```
Count (počet) = 77
Average (průměr)= 4,35065
Variance(rozptyl) = 4,49385
Standard deviation(směrodatná odchylka) = 2,11987
Minimum = 1,0
Maximum = 9,0
Range (rozpětí)= 8,0
Stnd. skewness (standardizovaná šikmost)= 1,12981
Stnd. kurtosis (standardizovaná špičatost)= -1,24037
```

$x_{0,25} = 3$; $x_{0,75} = 6$; IQR = 3