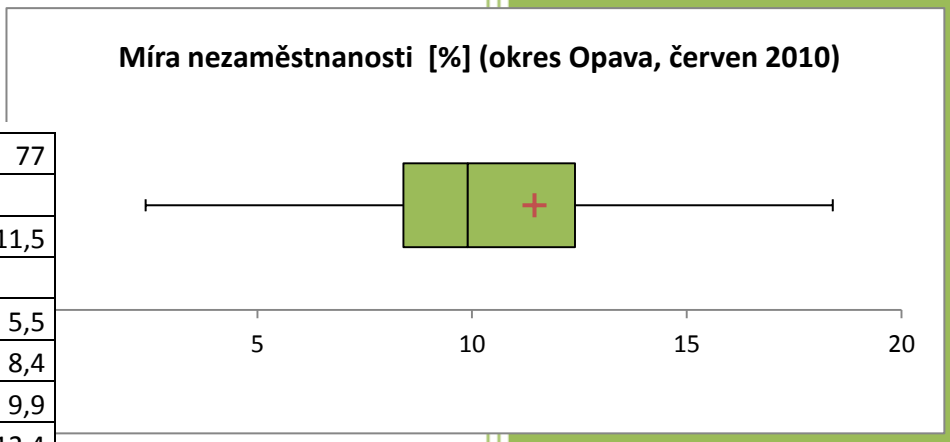
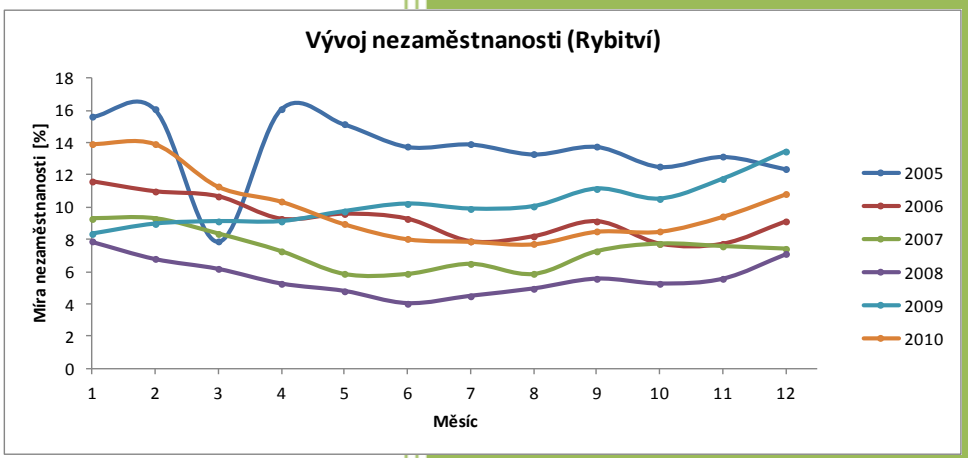
2011

Průzkumová analýza jednorozměrných dat (Teorie)

Rozsah	77
Průměr	11,5
Minimum	5,5
Dolní kvartil	8,4
Medián	9,9
Horní kvartil	12,4
Maximum	34,4
Rozptyl	28,3
Směrodatná odchylka	5,3
Variační koeficient	46%
Šikmost	2,5
Špičatost	7,5





Martina Litschmannová

VŠB-TU Ostrava

Obsah

1 Průzkumová analýza jednorozměrných dat

Co najdete v této kapitole?

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

Původním posláním statistiky bylo zjišťování údajů o populaci na základě výběrového souboru. Pod pojmem **populace** přitom rozumějme množinu všech prvků, které sledujeme při statistickém výzkumu. Populace (základní soubor) bývá zadána buď výčtem prvků, nebo vymezením některých jejich společných vlastností. Například:

Provádíme-li stat. výzkum týkající se měsíčních příjmů žen ve věkové kategorie „nad 50 let“, populaci tvoří všechny ženy, které mají více než 50 let. Zkoumáme-li stav nezaměstnanosti v Severomoravském kraji, budeme za populaci považovat všechny správní celky (obce) v tomto kraji.

Vzhledem k tomu, že **rozsah** (počet prvků) **populace (*N*)** je obvykle vysoký, získáváme informace o populaci prostřednictvím statistického výzkumu. Nejběžnějším druhem statistického výzkumu je tzv. **výběrové šetření**, při němž je statistik pouze pasivním pozorovatelem – do průběhu šetření zasahuje co nejméně (ideálně vůbec). Zkoumaná část populace se nazývá **výběr**, popř. výběrový soubor. Počet prvků ve výběru (**rozsah výběru**) označujeme ***n***. Otázkou je jak stanovit takový výběr, aby byl skutečně reprezentativní, tj. aby charakteristiky 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 pouze v domovech důchodců, popř. na schůzích mladých konzervativců. Existuje několik způsobu jak výběr provést, přičemž nejčastěji volíme **náhodný výběr**, v němž každý prvek populace má stejnou šanci být zařazen do výběru.

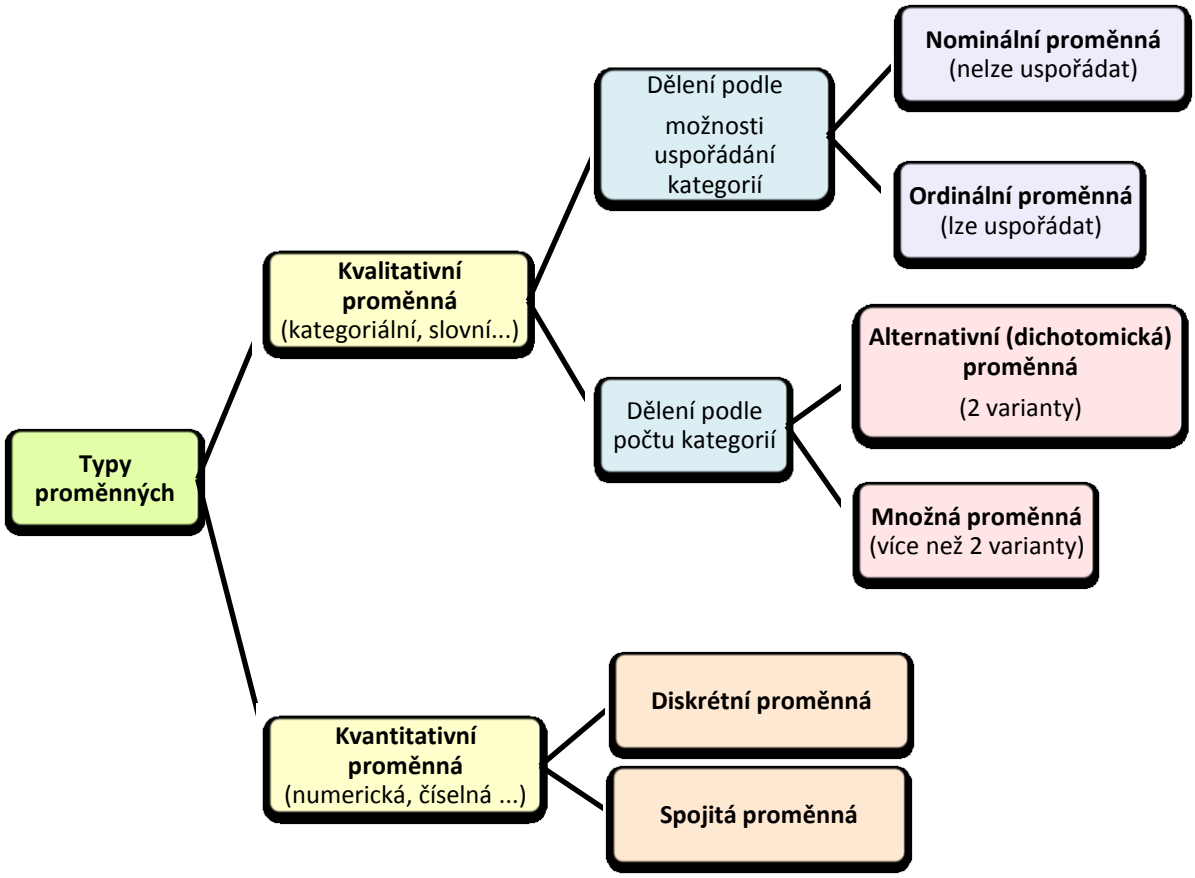
Je zřejmé, že výběrové šetření nemůže být nikdy tak přesné jako průzkum celé populace. Proč jej tedy preferujeme? Jmenujme tři nejdůležitější důvody.

- Úspora času a finančních prostředků (zejména u rozsáhlé populace)
- **Minimalizace ztrát v důsledku destruktivního 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)
- **Nedostupnost celé populace** (při srovnávání působení faktorů okolí a dědičných znaků poskytují nejlepší informace jednovaječná dvojčata – jak je všechna najít a přesvědčit ke spolupráci?)

Při statistickém zpracování dat pak nejprve analyzujeme informace z příslušného výběru a následně se snažíme o přenášení závěrů z výběru na celou populaci, což je nazýváno **statistická indukce.** Metody statistické indukce ponechme v tuto chvíli statistikům a

zaměřme se na to, jak získat základní poznatky z výběru. Oblast statistiky zabývající se analýzou výběru bývá nazývána **průzkumová analýza dat**, popř. **explorační analýza**, zkráceně EDA.

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 obrázku 1.1.



Obr. 1.1: Demonstrace základních typů proměnných

- **Proměnná kategoriální** (kvalitativní, 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, XL), velikost obce)

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

  - **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 (např. měsíční příjem v tis. Kč, věk v letech, …).

  - **Proměnné spojité** nabývající libovolných hodnot z $\mathbb{R}$ (poznámka: $\mathbb{R}$ označujeme množinu reálných čísel) nebo z nějaké podmnožiny $\mathbb{R}$ (např. průměrný měsíční příjem, ...)

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 kategoriálních 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

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

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

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

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.

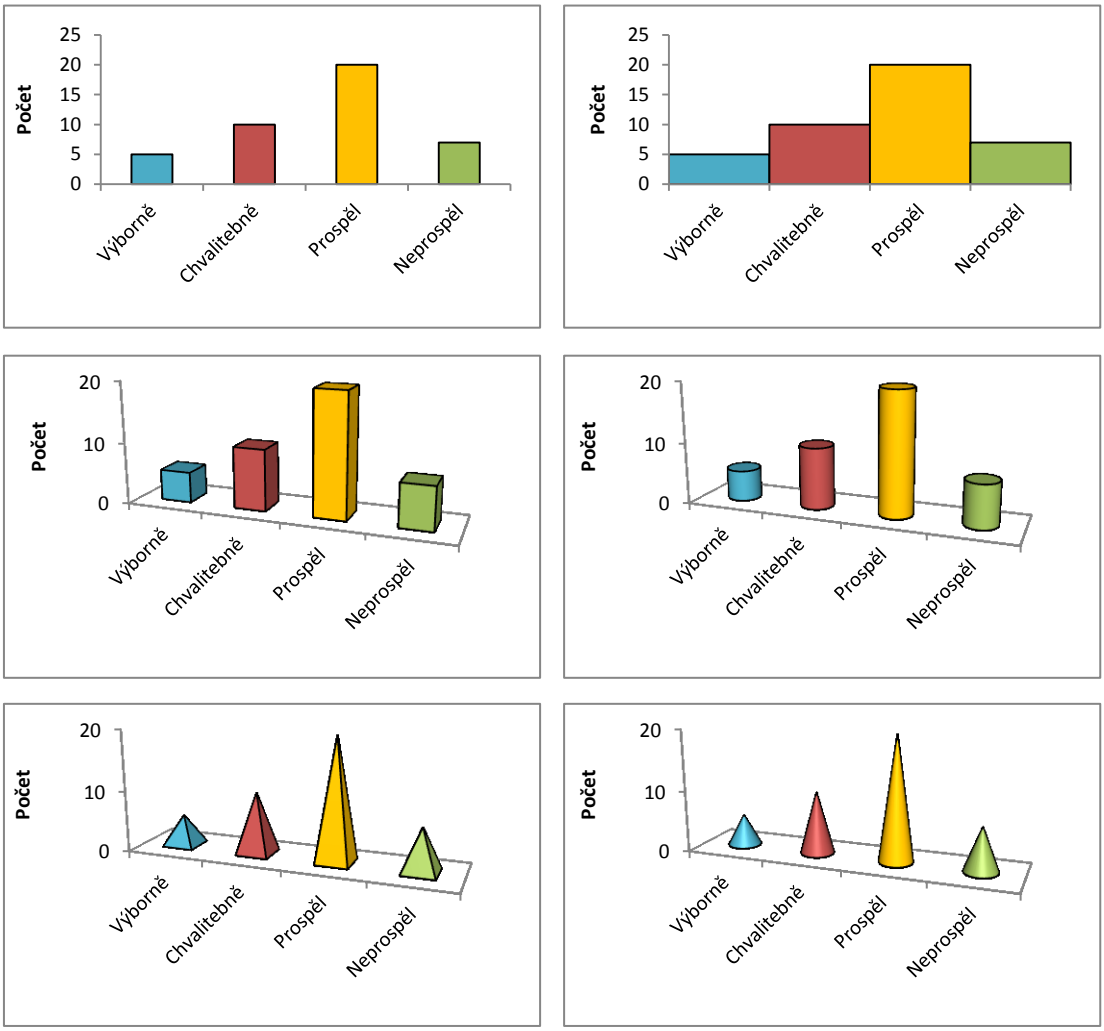
Modus tedy můžeme chápat jako typického reprezentanta souboru. V případě, že se ve statistickém souboru vyskytuje více variant s maximální četnosti, modus neurčujeme.

Grafické znázornění nominální proměnné

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

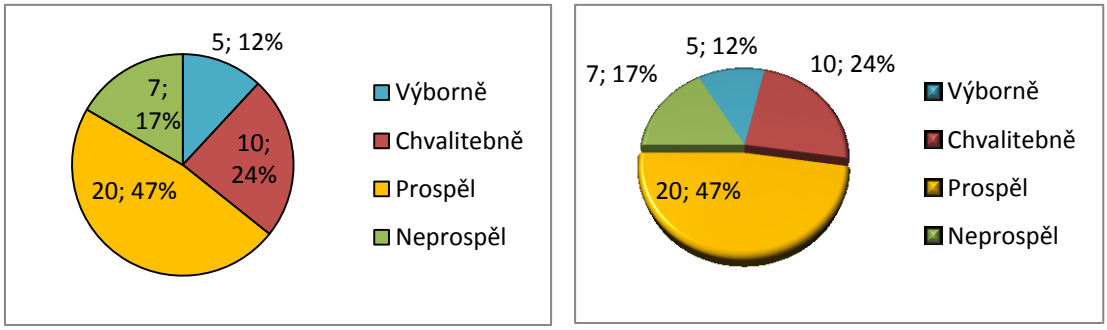
- **Sloupcový graf** (angl. bar chart),
- **Výsečový graf** (také koláčový graf, angl. pie chart).

Sloupcový graf 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 sloupcových grafů

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.)

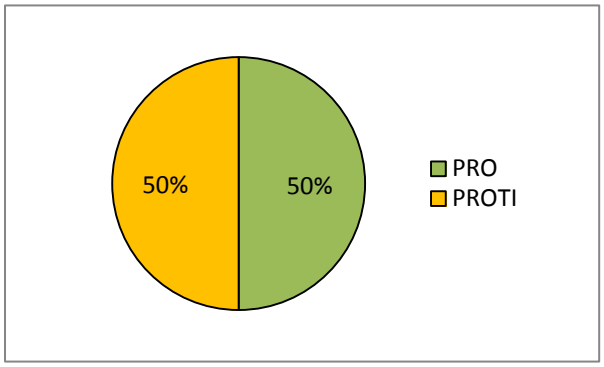


Obr. 1.3: Ukázky výsečových grafů

POZOR!!! V případě výsečového grafu si dejte zvláštní pozor na popis grafu. Jednotlivé výseče nestačí označit relativními četnostmi bez uvedení četnosti absolutních, popř. bez

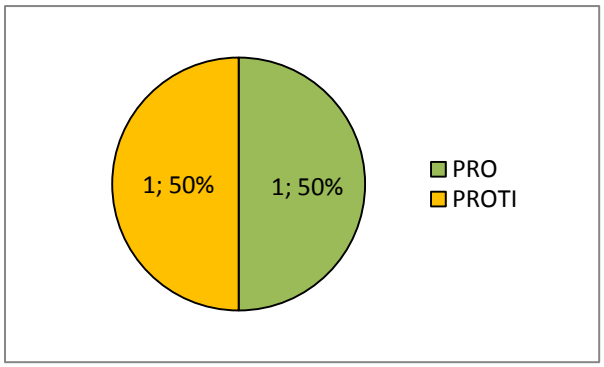
uvedení celkového počtu pozorování, to by mohlo vést k matení (ať už záměrnému nebo nechtěnému) toho, komu je graf určen. Zamyslete se nad následující ukázkou.

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 dotazováni 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.

1.1.2 Ordinální proměnná

Ordinální proměnná, stejně jako nominální proměnná, 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 nominální proměnné (četnost, relativní četnost, modus + sloupcový graf, výsečový graf), rozšířené o další dvě charakteristiky (kumulativní četnost, kumulativní relativní četnost), které berou v úvahu uspořádání ordinální proměnné.

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

Uvažte např. proměnnou "velikost obce", která nabývá variant: "pod 500 obyvatel", "500-1000 obyvatel", "1001-2000 obyvatel", "nad 2000 obyvatel", pak např. kumulativní četnost pro variantu "1001-2000 obyvatel" bude rovna počtu obcí, které mají nejvýše 2000 obyvatel.

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** 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:

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

Kumulativní relativní četnost se často uvádí v procentech. Pak $F_i = \frac{m_i}{n} \cdot 100$.

Obdobně jako pro nominální proměnné, můžeme i pro 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.

Tab. 1.2: 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	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	-----	-----

Grafické znázornění ordinální proměnné

Pro základní zobrazení ordinální proměnné můžete použít sloupcový graf, popř. výsečový graf, stejně jako u proměnné nominální.

1.2 Statistické charakteristiky kvantitativních proměnných

Pro popis kvantitativní proměnné můžeme v případě, že proměnnou kategorizujeme (tj. rozdělíme do intervalů) použít většinu statistických charakteristik užívaných pro popis proměnné ordinální (četnost, relativní četnost, kumulativní četnost, kumulativní relativní četnost). Tímto postupem se však ochuzujeme o velkou část informace, kterou bychom z výběru mohli získat. Vhodnější je použít dvě skupiny charakteristik:

- **Míry polohy** určující 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

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, average)

  Jeho hodnotu získáme pomocí známého vztahu

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

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

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

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

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

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.

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

- **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.

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

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

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

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

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 a kvantily.

- **Modus**

  Pro diskrétní proměnnou definujeme **modus** jako hodnotu nejčetnější varianty proměnné (podobně jako u kvalitativní proměnné). U spojité proměnné je definice módu složitější a pro praktické účely ji nejspíš nebudete potřebovat.

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** – $x_{0,1}; x_{0,2}; \ldots ; x_{0,9}$

    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.

Pro určení kvantilů můžete používat například MS Excel, algoritmus pro jejich výpočet nebude v tomto materiálu prezentován.

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.)

Prostřednictvím kvantilů je definována i další statistika kvantitativní proměnné – interkvartilové rozpětí.

- **Interkvartilové rozpětí IQR**

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

  $$IQR = x_{0,75} - x_{0,25}.$$

  Hodnota *IQR* jako taková nenese pro vás žádnou užitečnou informaci. Jde o pomocnou proměnnou, která nám pomůže při identifikaci odlehlých pozorování, o níž se zmíníme později.

1.2.2 Míry variability

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 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.

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 $Kč^2$. 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%.

$$V_x = \frac{s}{\bar{x}}, \qquad \text{popř. } V_x = \frac{s}{\bar{x}} \cdot 100\ [\%]$$

  Při praktickém hodnocení považujeme variabilitu dat za přiměřenou, mají-li variační koeficient nižší než 0,5 (50%).

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.

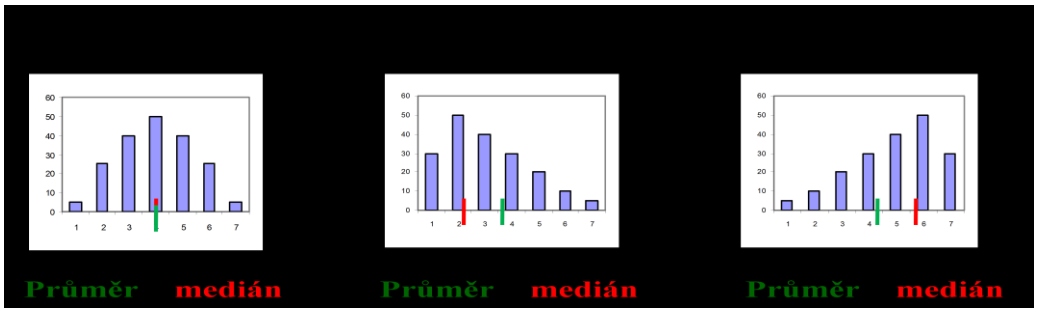
- **Výběrová šikmost *a*** (angl. skewness)

  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



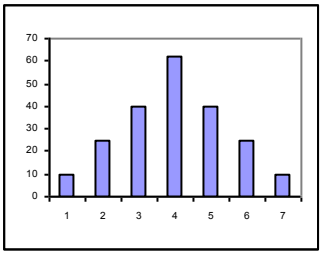
- **Výběrová špičatost *b*** (angl. kurtosis)

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

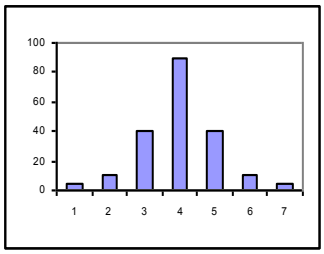
$$b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}.$$

A jak interpretujeme výběrovou špičatost?

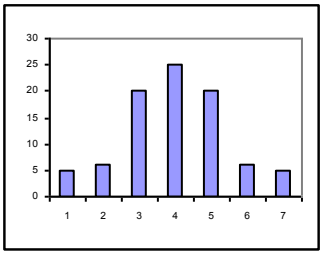
$b = 0$ ... špičatost odpovídá normálnímu rozdělení (bude definováno později)
$b > 0$ ... špičaté rozdělení proměnné
$b < 0$ ... ploché rozdělení proměnné



b=0



b>0



b<0

1.2.3 Odlehlá pozorování

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 dva z nich.

- **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:

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

- **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:

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

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

V konkrétním případě můžete pro identifikaci odlehlých pozorování zvolit libovolné z těchto dvou pravidel.

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.

Co dělat, když v datech identifikujeme odlehlá pozorování?
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.

1.2.4 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. 1.2: 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. 1.3: Příklady správného zápisu číselných charakteristik

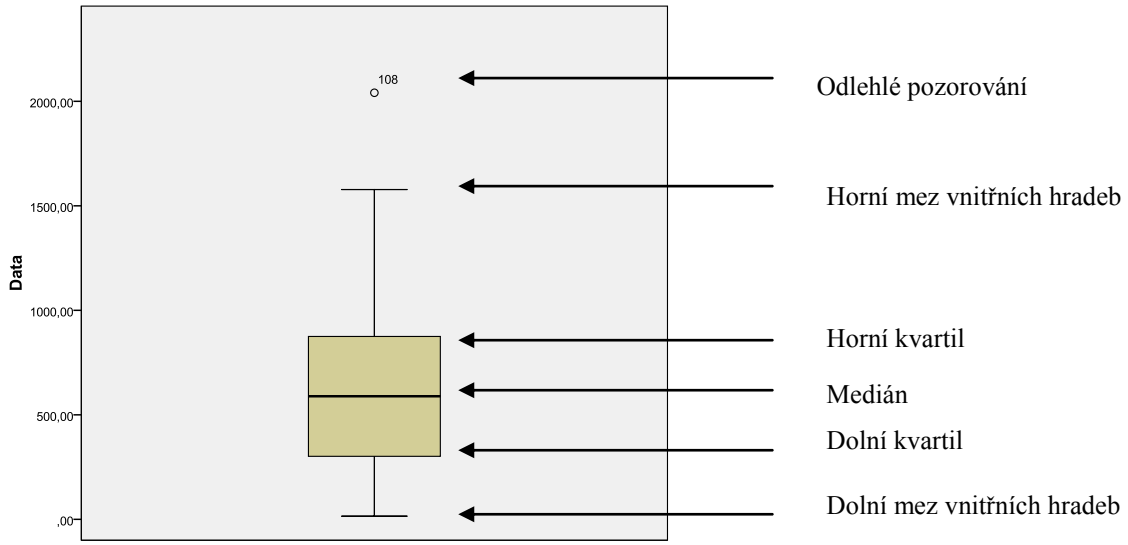
	Délka [m]	Váha [kg]	Teplota [^{0}C]
Průměr	2,26	128	14 600
Medián	2,68	118	13 700
Směrodatná odchylka	0,78	24	1 200

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.

1.2.5 Grafické znázornění kvantitativní 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 níže uvedeném obrázku.



Obr. 1.6: Krabicový graf

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.

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.

Poznámka: *Z popisu krabicového grafu je zřejmé, že jeho konstrukci začínáme zakreslením odlehlých pozorování a až poté vyznačujeme ostatní číselné charakteristiky proměnné (vnitřní hradby a kvartily).*

- **Histogram** (angl. histogram)
  Histogram představuje grafické zobrazení intervalového členění kvantitativní proměnné. Umožňuje získat dobrou představu o struktuře dat.

  Postup při konstrukci histogramu:
  1. Seřaďte data vzestupně, tj. od nejmenší po nejvyšší hodnotu.
  2. Určete minimální a maximální hodnotu v souboru (MIN(x) a MAX(x))
  3. Určete variační rozpětí R, kde $R = MAX(x) - MIN(x)$.
  4. Určete počet tříd histogramu, tj. počet jeho sloupců.
     Počet tříd k můžete určit buď intuitivně, nebo pomocí níže uvedených vztahů.

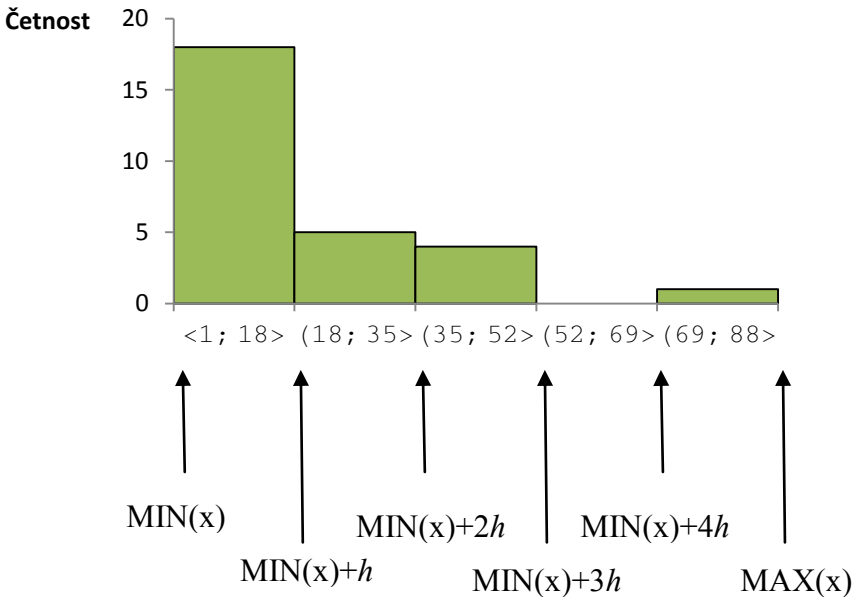
n (rozsah výběru)	k (doporučený počet tříd histogramu)
$n > 100$	$k \cong 10 \cdot \log n$
$40 < n \leq 100$	$k \cong 2\sqrt{n}$
$n \leq 40$	$k \cong 1 + 1{,}4426 \cdot \ln n$

  5. Vypočtěte šířku tříd h.
     $$h \cong R/k$$
  6. Určete meze jednotlivých tříd (viz Obr. 1.7).
  7. Určete četnosti dat ve stanovených třídách a zakreslete histogram.

  Jak byl aplikován výše uvedený postup při konstrukci histogramu na Obr. 1.7**?**
  - Histogram na uvedeném obrázku byl zkonstruován pro soubor, který obsahoval 28 měření.
  - Jejich nejmenší hodnota MIN(x) byla 1, nejvyšší hodnota MAX(x) = 88, tj. variační rozpětí $R = 88 - 1 = 87$.
  - Počet tříd byl navržen intuitivně k = 5. (Použili-li bychom doporučený počet tříd, bylo by $k \cong 1 + 1{,}4426 \cdot \ln(28) = 5{,}81 \Rightarrow k = 6$.)
  - Následně byla určena šířka tříd: $h \cong \frac{87}{5} = 17{,}4 \Rightarrow h = 17$.
  - V dalším kroku byly určeny meze jednotlivých tříd (viz Obr. 1.7) a určeny počty pozorování v jednotlivých třídách.

Interval	Četnost
<1; 10>	4
(10; 20>	15
(20; 30>	3
(30; 40>	2
(40; 88>	4

  - Posledním krokem pak je vykreslení histogramu.



Obr. 1.7: Histogram numerické proměnné

2 Analýza závislostí

Co najdete v této kapitole?

- explorační analýza kontingenčních tabulek,
- explorační analýza závislosti spojitých proměnných,
- explorační analýza závislosti ordinálních veličin.

V praxi většinou u statistických jednotek (pozorovaných osob nebo jiných objektů) zjišťujeme současně celou řadu znaků. Například:

- spotřeba, objem motoru, hmotnost a zrychlení automobilů,
- výše mzdy, velikost IQ, hmotnost a výška mužů,
- školní prospěch a pocit deprese u dětí, apod.

Jednotlivé znaky pak můžeme analyzovat metodami, s nimiž jsme se seznámili v předchozích kapitolách. Většinou však jednotlivé znaky nestudujeme jako takové, zajímají nás především jejich vazby k jiným znakům. Například nás může zajímat, zda existuje závislost mezi spotřebou automobilu a jeho hmotností, výši mzdy a velikostí IQ, pocitem deprese u dětí a školním prospěchu.

V případě, že znak X působí na znak Y, avšak znak Y již nepůsobí zpětně na znak X, mluvíme o **jednostranné závislosti**. Příkladem jednostranné závislosti může být vztah mezi typem absolvované střední školy a (bodovým) výsledkem přijímací zkoušky z matematiky nebo vztah mezi výškou a váhou.

Pokud v analyzovaném vztahu nelze jednoznačně určit příčinu a důsledek, tzn. pokud znak X ovlivňuje znak Y a znak Y zpětně působí na znak X, hovoříme o **závislosti oboustranné**. (Například: vztah mezi výdaji domácností na oblečení a na potraviny.) V této kapitole se seznámíme se základními metodami analýzy oboustranné závislosti – vymezíme si metody pro analýzu síly vazeb mezi dvojicemi znaků, tj. metody pro analýzu síly závislostí dvojic náhodných veličin.

Výběr vhodné metody závisí na typu analyzovaných veličin. V Tab. 2.1 jsou uvedeny jednotlivé metody analýzy závislostí pro různé typy dat.

Tab. 2.1: Metody analýzy oboustranné závislostí

		Typ znaku Y		
		kategoriální	diskrétní	spojitá
Typ znaku X	kategoriální	analýza závislosti v kontingenčních tabulkách,		
	diskrétní		analýza závislosti ordinálních znaků	
	spojitá			analýza závislosti v normálním rozdělení

2.1 Analýza závislostí v kontingenčních tabulkách

2.1.1 Motivační příklad

Analýzou dat v kontingenční tabulce nás provede následující příklad.

Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Šetření se účastnilo 100 pracovníků z Prahy a 200 pracovníků z venkova. Výsledky šetření jsou v následující tabulce.

místo/stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
Praha	10	25	50	15
Venkov	20	10	130	40

Výsledky šetření analyzujte.

2.1.2 Základní pojmy

Výsledky šetření jsou uvedeny v tzv. kontingenční tabulce. **Kontingenční tabulka** vzniká setříděním prvků výběru podle variant dvou kategoriálních znaků, např. znaku X a znaku Y. Nechť znak X nabývá variant $x_{[1]}, \dots, x_{[r]}$ a znak Y nabývá variant $y_{[1]}, \dots, y_{[s]}$. V kontingenční tabulce jsou uspořádány absolutní četnosti n_{ij} dvojice variant $(x_{[i]}, y_{[j]})$, přičemž názvy jednotlivých variant znaků X a Y jsou uvedeny v hlavičce tabulky.

Tab. 2.2: Schéma kontingenční tabulky

$X \backslash Y$	$y_{[1]}$	$y_{[2]}$	$\cdots$	$y_{[s]}$
$x_{[1]}$	n_{11}	n_{12}	$\cdots$	n_{1s}
$x_{[2]}$	n_{21}	n_{22}	$\cdots$	n_{2s}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$x_{[r]}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}

Pokud lze mezi analyzovanými znaky X a Y pozorovat kauzalitu (příčinnou souvislost), volíme označení X pro nezávislý znak a označení Y pro znak závislý. (*Všimněte si, že v motivačním příkladu jsme jako znak X, tj. znak jehož varianty jsou identifikátory řádků, zvolili umístění podniku…*)

Kontingenční tabulku často rozšiřujeme o další zajímavé číselné charakteristiky, jejichž výpočet pro data z motivačního příkladu můžete sledovat v Tab. 2.4.

- **Marginální četnosti**, které udávají celkové četnosti jednotlivých variant znaku X, resp. znaku Y. Marginální četnosti označujeme

  $n_{i\cdot}$ … součet všech četností v i-té řádce,
  $n_{\cdot j}$ … součet všech četností v j-tém sloupci

a zapisujeme je na okraj kontingenční tabulky (viz Tab. 2.3).

Tab. 2.3: Schéma rozšířené kontingenční tabulky

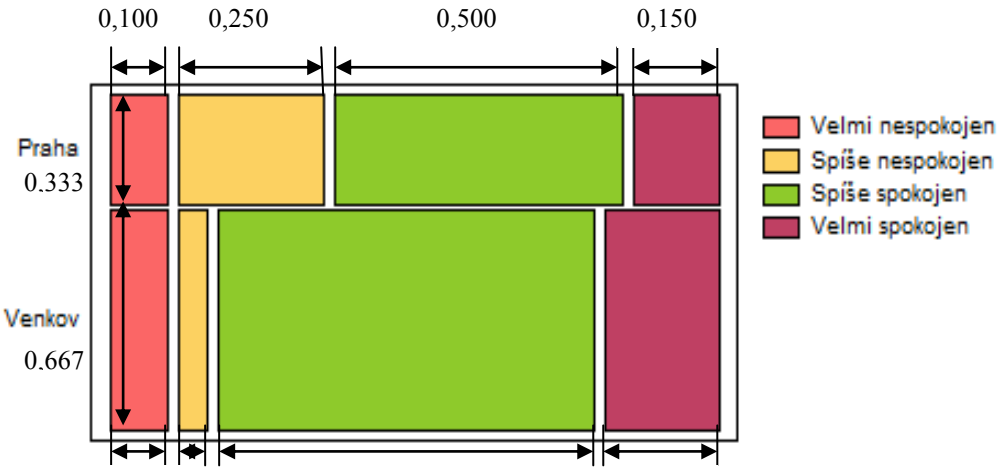
$X\backslash Y$	$y_{[1]}$	$y_{[2]}$	⋯	$y_{[s]}$	Celkem
$x_{[1]}$	n_{11}	n_{12}	⋯	n_{1s}	$n_{1\cdot}$
$x_{[2]}$	n_{21}	n_{22}	⋯	n_{2s}	$n_{2\cdot}$
⋮	⋮	⋮	⋯	⋮	⋮
$x_{[r]}$	n_{r1}	n_{r2}	⋯	n_{rs}	$n_{r\cdot}$
Celkem	$n_{\cdot 1}$	$n_{\cdot 2}$	⋯	$n_{\cdot s}$	n

- **Celkový rozsah výběru** n
- **Relativní četnosti**, které pro každé pole rozšířené kontingenční tabulky určíme jako podíl příslušné absolutní četnosti a celkového rozsahu výběru n. (Např.: Z celkového počtu 300 respondentů bylo 5,0% velmi spokojených respondentů zaměstnaných v Praze.)
- **Řádkové rel. četnosti**, které udávají relativní četnosti znaku Y za předpokladu, že znak X nabývá určité varianty. Určujeme je jako podíl příslušné absolutní četnosti a marginální četnosti v odpovídajícím řádku. (Např.: Ze všech v Praze zaměstnaných respondentů bylo 10,0% velmi nespokojených.)
- **Sloupcové rel. četnosti**, které udávají relativní četnosti znaku X za předpokladu, že znak Y nabývá určité varianty. Určujeme je jako podíl příslušné absolutní četnosti a marginální četnosti v odpovídajícím sloupci. (Např. Ze všech velmi spokojených respondentů je 20, 0% zaměstnaných na venkově.)

Tab. 2.4: Rozšířená kontingenční tabulka pro data z motivačního příkladu (pozorované četnosti, celkový rozsah výběru, marginální četnosti, relativní četnosti, řádkové rel. četnosti, sloupcové rel. četnosti)

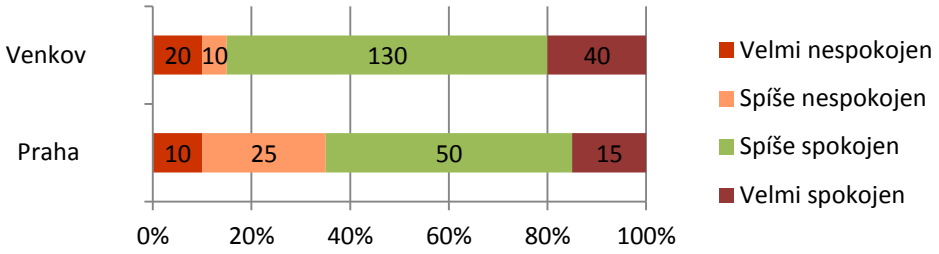
místo/stupeň spokojenosti	velmi nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen	celkem
Praha	10 0,033 (10/300) 0,100 (10/100) 0,333 (10/30)	25 0,083 (25/300) 0,250 (25/100) 0,714 (25/35)	50 0,167 (50/300) 0,500 (50/100) 0,278 (50/180)	15 0,050 (15/300) 0,150 (15/100) 0,273 (15/55)	100 0,333 (100/300)
venkov	20 0,067 (20/300) 0,100 (20/200) 0,667 (20/30)	10 0,033 (10/300) 0,050 (10/200) 0,286 (10/35)	130 0,433 (130/300) 0,650 (130/200) 0,722 (130/180)	40 0,133 (40/300) 0,200 (40/200) 0,727 (40/55)	200 0,067 (200/300)
celkem	30 0,100 (30/300)	35 0,117 (35/300)	180 0,600 (180/300)	55 0,183 (55/300)	**300**

Grafickou obdobou kontingenční tabulky je **mozaikový graf**. Mozaikový graf se skládá z r řad obdélníků, přičemž r je počet variant (nezávislého) znaku X. (V našem případě r=2.) Každá řada obsahuje s obdélníků, přičemž s je počet variant (závislého) znaku Y. (V našem případě s=4.) Výšky jednotlivých řad obdélníků odpovídají příslušným marginálním relativním četnostem. Šířky obdélníků v jednotlivých řadách odpovídají příslušným řádkovým relativním četnostem (viz Obr. 2.1: Mozaikový graf pro data z motivačního příkladu).



Obr. 2.1: Mozaikový graf pro data z motivačního příkladu

Pokud by byl mozaikový graf v tomto případě tvořen svislými pruhy (jednotlivé obdélníky stejných barev by měly stejné šířky), znamenalo by to, že sledované znaky jsou nezávislé. Čím je mozaikový graf členitější, tím silnější závislost mezi znaky X a Y lze předpokládat. Dle obr. 10.1 lze předpokládat, že spokojenost v práci závisí na umístění závodu. (*Podívejte se znovu na Obr. 2.1 a zvažte, jaký následek by mělo sloučení variant „spíše nespokojen“ a „spíše spokojen“.*)



Obr. 2.2: 100% skládaný pruhový graf

Obdobou mozaikového grafu je **100% skládaný pruhový graf** (např. MS Excel). Od mozaikového grafu se tento graf liší tím, že šířky všech řádků jsou stejné, tzn. že tento typ grafu nezohledňuje řádkové marginální relativní četnosti.

Kromě mozaikového grafu se pro prezentaci dat zapsaných v kontingenční tabulce používají **shlukový**, popř. **kumulativní sloupcový graf.**

2.2 Úvod do korelační analýzy

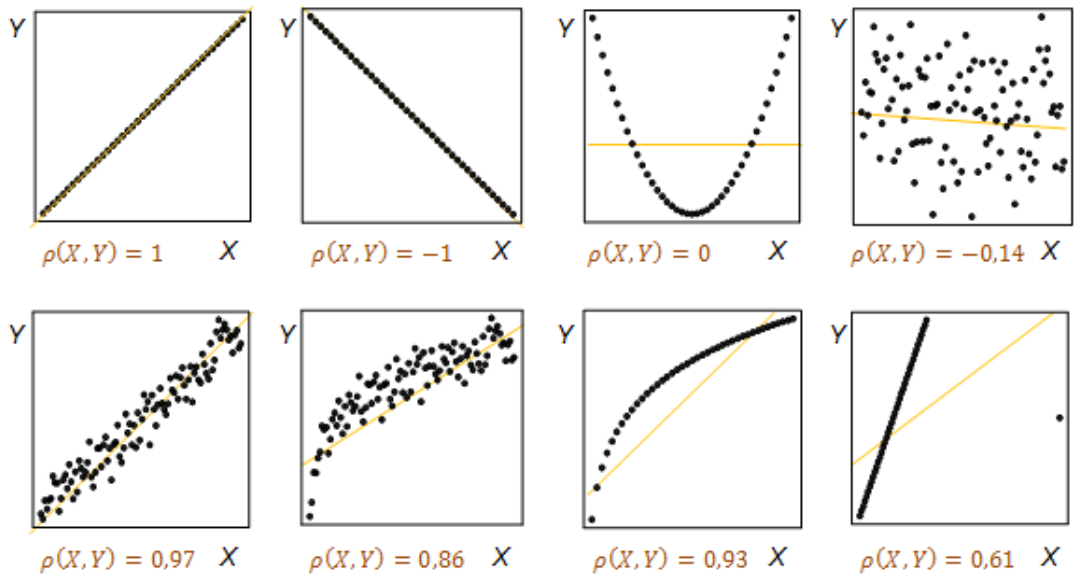
Pro měření závislosti dvou kvantitativních (číselných) proměnných se v praxi používá statistika nazývána jednoduchý korelační koeficient, dále jen korelační koeficient. Korelační koeficient $\rho(X,Y)$ je mírou **lineární** závislosti dvou náhodných veličin X, Y.

***Poznámka**: Pojem náhodná veličina je základním pojmem z teorie pravděpodobnosti. Můžete si představit, že jde o obecné označení výsledku náhodného pokusu, který nabývá číselných*

hodnot. Příklady náhodných veličin: počet ekonomicky aktivních obyvatel v obci (obec vybíráme náhodně), počet nezaměstnaných žen v obci, roční příjem ekonomicky aktivního obyvatele ČR, ...)

Vlastnosti korelačního koeficientu

- $-1 \leq \rho(X,Y) \leq 1$, tj. korelační koeficient je číslo mezi -1 a 1 (včetně).
- $\rho(X,Y) = \rho(Y,X)$, tj. *„při výpočtu korelačního koeficientu nezáleží na tom, kterou náhodnou veličinu označíme X a kterou Y“.*
- $\rho(X,X) = 1$, viz Obr. 2.3.
- jsou-li *X*, *Y* nezávislé náhodné veličiny, pak $\rho(X,Y) = 0$,
- POZOR!!! je-li $\rho(X,Y) = 0$, říkáme, že *X*, *Y* jsou **nekorelované**[1] náhodné veličiny,
- je-li $\rho(X,Y) > 0$, říkáme, že *X*, *Y* jsou **pozitivně korelované** (s rostoucím *X* roste *Y*),
- je-li $\rho(X,Y) < 0$, říkáme, že *X*, *Y* jsou **negativně korelované** (s rostoucím *X* klesá *Y*).



Obr. 2.3: Grafická prezentace souvislosti mezi ρ(*X*,*Y*) a závislosti náhodných veličin *X*, *Y*

Nevýhodou korelačního koeficientu je, že jej většinou nedokážeme přesně určit (proto zde není uveden ani definiční vztah pro ρ). Jedná se o tzv. populační charakteristiku, k jejímuž výpočtu potřebujeme znát informace o analyzovaných náhodných veličinách (tj. jejich pravděpodobnostní popis, popř. hodnoty všech jejích realizací.) Nic však není ztraceno! Přestože neumíme ρ určit přesně, dokážeme jej dobře odhadnout. Následující kapitoly ukazují, jak na to.

[1] **POZOR!!!** Jsou-li *X*, *Y* nekorelované náhodné veličiny, nemusí to znamenat, že jsou nezávislé!!! Jsou-li *X*, *Y* nekorelované náhodné veličiny, víme pouze to, že mezi *X*, *Y* neexistuje **lineární** závislost (viz Obr. 2.3).

Pokud používáme korelační koeficient, je třeba mít na paměti, že tento koeficient je pouze mírou lineární závislosti proměnných. "Pěkný" korelační koeficient (hodnota blízká jedné nebo minus jedné) ještě vůbec neznamená, že srovnávané proměnné dávají "pěkně" závislé výsledky. Znamená to pouze silnou LINEÁRNÍ závislost mezi proměnnými. "Špatný" (malý v absolutní hodnotě) korelační koeficient vůbec neznamená, že závislost je málo silná. Může (ale nemusí!) jít např. o silnou nelineární závislost, např. kvadratickou.

Uvedeme-li korelační koeficient pouze jako číslo a nedoplníme-li jej bodovým grafem (tzv. rozptylogramem, angl. scatter plot), můžeme z jeho velikosti získat naprosto mylnou představu o intenzitě analyzované závislosti (viz Obr. 2.3).

2.2.1 Pearsonův koeficient korelace

Mějme výběr $(X_1; Y_1), \dots, (X_n; Y_n)$ z nějakého dvourozměrného rozdělení. V případě, že X, Y jsou spojité náhodné veličiny s normálním rozdělením, je vhodným odhadem korelačního koeficientu tzv. Pearsonův koeficient korelace. (***Poznámka**: Opět narážíme na nedefinovaný pojem. Pokud byste měli zájem o jeho vysvětlení, můžete se podívat například do učebnice [1]. V rámci práce s MS Excel Vám bude poskytnut výpočetní applet, který uvedený předpoklad testuje.*) Označme

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}.$$

Pak je Pearsonův korelační koeficient definován jako

$$r = \begin{cases} \dfrac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}}, & S_X^2, S_Y^2 \neq 0, \\ 0 & jinak. \end{cases}$$

2.3 Analýza závislosti diskrétních proměnných

V předcházející kapitole jsme viděli, že možnost použití Pearsonova korelačního koeficientu r je vázána na splnění předpokladu, že výběr pochází z dvourozměrného normálního rozdělení. Při porušení tohoto předpokladu, resp. v případě, že chceme analyzovat závislost dvou diskrétních proměnných, můžeme použít například **Spearmanův koeficient korelace.**

2.3.1 Spearmanův korelační koeficient

Mějme náhodný výběr $(X_1; Y_1), \dots, (X_n; Y_n)$ z dvourozměrného rozdělení. Nechť $R_{X_1}, \dots, R_{X_n}$ jsou pořadí veličin $X_1, \dots, X_n$ a nechť $R_{Y_1}, \dots, R_{Y_n}$ jsou pořadí veličin $Y_1, \dots, Y_n$.

Kdyby s rostoucími hodnotami X_i vzrůstaly i hodnoty Y_i, byla by zřejmě pořadí obou veličin shodná, tj. $R_{X_i} = R_{Y_i}$ pro $i = 1, \dots, n$. Jestliže s rostoucími hodnotami X_i klesají hodnoty Y_i,

jsou pořadí obou veličin právě opačná. Při nezávislosti veličin *X* a *Y* jsou pořadí zpřeházená zcela náhodně. Spearmanův korelační koeficient r_S se proto definuje pomocí diferencí pořadí $\left(R_{X_i} - R_{Y_i}\right)$ jako

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \left(R_{X_1} - R_{Y_1}\right)^2.$$

Při shodném pořadí nabývá koeficient r_S maximální hodnoty 1, při opačném pořadí minimální hodnoty -1. V ostatních případech je $-1 < r_S < 1$. Je-li hodnota Spearmanova korelačního koeficientu $r_S = 0$, pořadí veličin *X* a *Y* jsou náhodně zpřeházená, a mezi sledovanými veličinami tedy není závislost.

Pokud se v náhodných výběrech, z nichž je r_S počítán, vyskytuje mnoho shod (tj. stejně velkých pozorování), doporučuje se používat **korigovaný Spearmanův korelační koeficient** $r_{S_{korig}}$. Označme t_X počty stejně velkých *X*-ových hodnot. (Je-li mezi pozorovanými hodnotami náhodné veličiny *X* několik skupin stejně velkých pozorování, pak t_X jsou rozsahy těchto skupin.) Podobně definujme t_Y. Pak

$$r_{S_{korig}} = 1 - \frac{6}{n^3 - n - T_X - T_Y} \sum_{i=1}^{n} \left(R_{X_1} - R_{Y_1}\right)^2,$$

kde $T_X = \frac{1}{2}\sum(t_X^3 - t_X)$, $T_Y = \frac{1}{2}\sum(t_Y^3 - t_Y)$.

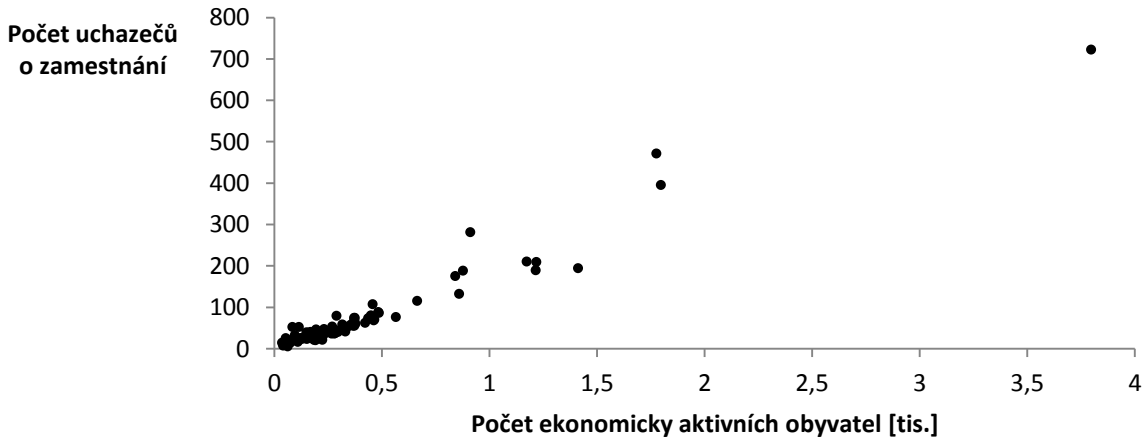
Poznámka: *Nesprávně použitý Pearsonův výběrový korelační koeficient by ukazoval na mnohem těsnější závislost (silnější korelaci) mezi X a Y.*

3 Velmi stručný úvod do lineární regrese

Co najdete v této kapitole?

- základní pojmy regresní analýzy,
- velmi zjednodušený návod jak vyhodnotit kvalitu odhadu regresní funkce.

Často chceme prozkoumat vztah mezi dvěma veličinami, kde jedna z nich, tzv. **nezávisle proměnná *x***, má ovlivňovat druhou, tzv. **závisle proměnnou *Y***. Předpokládá se, že obě veličiny jsou spojité. Prvním krokem ve zkoumání by mělo být zakreslení dat do bodového grafu, tzv. **korelačního pole** a ověření toho, zda mezi veličinami skutečně existuje předpokládaná závislost, tzv. **regrese**.



Obr. 3.1: Korelační pole

Nejjednodušší formou regrese je **jednoduchá lineární regrese**, která předpokládá lineární závislost mezi dvěma veličinami.

Rovnici regresní přímky zapisujeme ve tvaru: $Y_i = \beta_0 + \beta_1 \cdot x_i + e_i$

Odhad regresní přímky nazýváme **vyrovnávací přímka** a zapisujeme ji například ve tvaru:

$$\hat{Y}_i = b_0 + b_1 x_i.$$

Chyby odhadu, tj. hodnoty $e_i = \hat{Y}_i - Y_i$, nazýváme **rezidua**. Pokud jsou splněny podmínky lineárního regresního modelu, můžeme koeficienty regresní přímky odhadovat **metodou nejmenších čtverců** (pro odhad koeficientu používáme v praxi statisticky software, popř. tabulkový procesor MS Excel).

Podmínky lineárního regresního modelu $Y_i = \beta_0 + \beta_1 x_i + e_i$ jsou tyto:

- $E(e_i) = 0$ pro každé i=1,2,…,n, tj. střední hodnota náhodné složky je nulová.
- $D(e_i) = \sigma^2$ pro každé *i=1,2,...,n,* tj. *r*ozptyl náhodné složky je konstantní.

- $Cov(e_i, e_j) = 0$ pro každé $i \neq j$, kde $i, j = 1,2,\ldots,n$, tj. kovariance náhodné složky je nulová.
- Náhodné složky e_i mají pro $i = 1,2,\ldots,n$ normální rozdělení.
- Regresní parametry β_i mohou nabývat libovolných hodnot.
- Regresní model je lineární v parametrech.

Podmínky lineárníhu regresního modelu je nutno v rámci regresní analýzy ověřit. Za minimální způsob ověření je považována analýza reziduí. Znázorníme-li rezidua pomocí grafu, v němž na horizontální osu vynášíme pozorované hodnoty závislé veličiny a na vertikální osu hodnoty reziduí, měli bychom ověřit, zda:

- Rezidua jsou rovnoměrně rozložena kolem nuly.
- Histogram reziduí je symetrický, jeho tvar odpovídá přibližně Gaussově křivce.
- Rozptyl reziduí je konstantní, tj. rezidua se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími odhadovanými hodnotami $\hat{Y}_i$.
- Graf reziduí nevykazuje funkční závislost. Autokorelace se na tomto grafu projeví tak, že se rezidua systematicky snižují nebo zvyšují, resp. můžeme mezi reziduí a předpovídanými hodnotami pozorovat nelineární závislost.

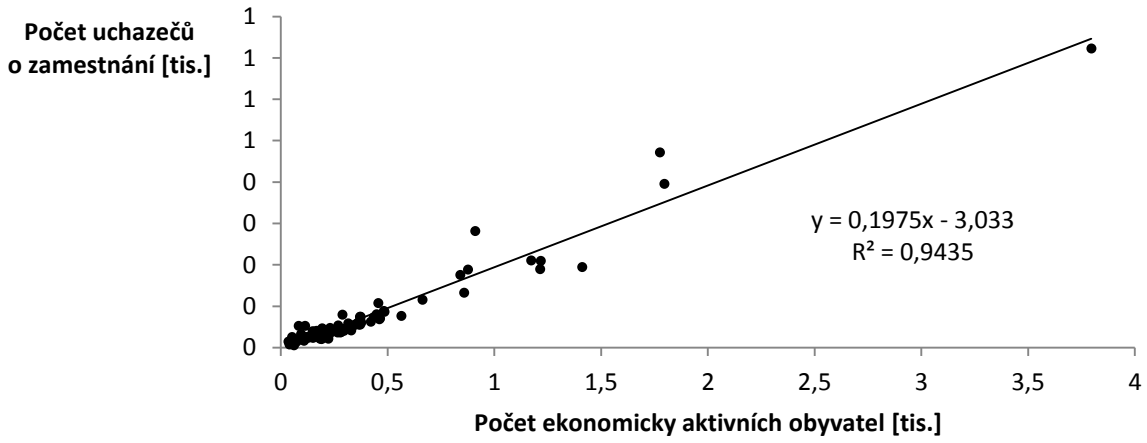
Kvalitu regresního modelu udává **index determinace R^2**. Přesněji řečeno udává kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

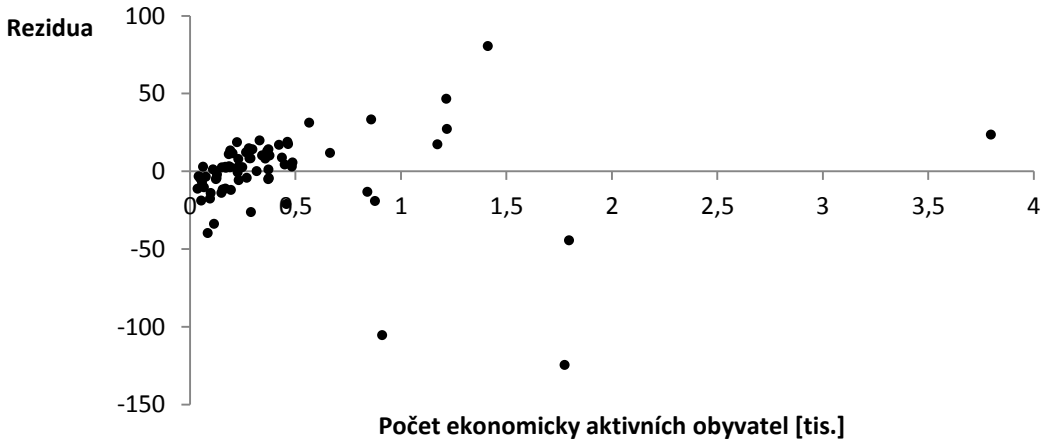
kde y_i jsou pozorované hodnoty vysvětlované proměnné, $\hat{Y}_i$ jsou odhadované hodnoty vysvětlované proměnné a $\bar{y}$ je průměr pozorovaných hodnot vysvětlované proměnné.

Tento index nabývá hodnot od nuly do jedné (teoreticky i včetně těchto krajních mezí), přičemž hodnoty blízké nule značí špatnou kvalitu regresního modelu, hodnoty blízké jedné značí dobrou kvalitu regresního modelu. Udává se většinou v procentech. Je-li $R^2 = 1$, pak regresní model vysvětluje závislost vysvětlované proměnné na vysvětlující proměnné úplně (tzv. dokonalá lineární závislost). Naopak, je-li $R^2 = 0$, pak model nevysvětluje nic. (*Jinými slovy řečeno: Čím je hodnota R^2 vyšší, tím kvalitnější model máme. V praxi považujeme za kvalitní model, jehož R^2 je větší než 0,8.*)

POZOR! *Vyjde-li nízká hodnota indexu determinace, nemusí to ještě znamenat nízký stupeň závislosti mezi proměnnými, ale může to signalizovat chybnou volbu typu regresní funkce.*



Obr. 3.2: Odhad regresní přímky, včetně indexu determinace



Obr. 3.3: Graf reziduí

Všimněte si navržené regresní přímky, indexu determinace (Obr. 3.2) a grafu reziduí (Obr. 3.3) pro vzorová data. Přestože model dle Obr. 3.2 vypadá velmi dobře (index determinace je mnohem vyšší než 0,8), na grafu reziduí vidíme, že modelu nelze důvěřovat (velmi výrazně se mění rozptyl reziduí, rozložení reziduí kolem nuly není rovnoměrné).

Regresní model nám umožňuje provádět rovněž **extrapolaci**, tj. odhad závisle proměnné pro hodnoty nezávisle proměnné ležící mimo interval naměřených hodnot. Extrapolace je vždy spojena s rizikem, že regresní model mimo interval naměřených hodnot pozbývá platnosti.

Tuto kapitolu uzavřeme citací: „Lépe je znát několik užitečných pravidel, než nastudovat mnoho neužitečných věcí.“ (Seneca, volně dle Ing Pavla Blažíčka, IV. zjazd Slovenskej spoľočnosti klinickej biochémie, Stará Ľubovňa, květen 2000) a uvedením několika základních pravidel, která byste při použití regresní analýzy měli dodržovat.

- Závěry plynoucí z našich výsledků platí pouze pro rozsah hodnot, pro které byl model navržen. Jakákoliv extrapolace je přinejmenším ošidná.

- Na data se vždy nejprve "podíváme" pomocí korelačního pole. Z korelačního pole usuzujeme, zda nejsou přítomny tzv. **vlivné** resp. vychýlené **body**. Bod, který je silně vychýlený ve směru pouze jedné ze souřadnic, často nazýváme odlehlý (outlier). Bod, který je vychýlený ve směru obou souřadnic, označujeme často jako extrém. Terminologie není ustálená. Vlivné body mohou mít silný vliv na odhadovanou regresní funkci.

- Problém odlehlých bodů bývá často řešen tím, že jsou z výběrového souboru vyloučeny a to na základě odhadu (jsou patrné už na výše zmíněném korelačním poli). To může být účelné v případě, že máme dostatečné množství dat. Nikdy bychom však neměli vlivný bod vyloučit, aniž bychom vysvětlili příčinu jeho vzniku nebo se přesvědčili, že se jedná o artefakt (např. hrubá chyba).

- Dále je třeba, aby každá proměnná měla v ideálním případě normální (Gaussovo) anebo v praxi alespoň symetrické rozdělení dat. Při troše zkušenosti to poznáme už z korelačního pole eventuelně z empirické hustoty (histogramu) příslušné proměnné.

4 Explorační analýza časových řad

Co najdete v této kapitole?

- základní pojmy související s popisem časových řad,
- možnosti grafického zobrazení časových řad,
- způsoby průměrování časových řad,
- vyhlazování časových řad klouzavými průměry,
- základní míry dynamiky časových řad.

Časová řada je numerická proměnná, jejíž hodnoty podstatně závisí na čase, v němž byly získány (posloupnost chronologicky uspořádaných pozorování). Časové okamžiky, kdy byla data získána, jsou od sebe většinou stejně vzdáleny. Jde například o

- počty nezaměstnaných v jednotlivých měsících,
- počty automobilových nehod na Barandovském mostě v jednotlivých měsících,
- denní produkce mléka Veselé krávy.

Popis pomocí popisných statistik (krabicové grafy) – poskytuje dobrou představu o vlastnostech časové řady jako jednoho celku dat, ale neposkytuje informace o jejím časovém vývoji (roční průměrná mzda).

4.1 Základní pojmy

Časové řady lze klasifikovat podle různých hledisek, např.: podle charakteru dat, jejichž hodnoty tvoří časovou řadu

- **časové řady intervalové** - data závisí na délce intervalu, který je sledován (např. měsíční výroba cementu v ČR)
- **časové řady okamžikové** - data se vztahují k určitému okamžiku (počet nezaměstnaných v ČR v jednotlivých měsících)

podle periodicity, s jakou jsou data sledována

- **časové řady údajů ročních**
- **časové řady krátkodobé**

podle druhu sledovaných dat

- **časové řady absolutních ukazatelů** – např. počet obsloužených klientů za měsíc
- **časové řady odvozených charakteristik** – např. časová řada kumulativní (kumulativní časové řady, které vznikají postupným načítáním (kumulováním) jednotlivých hodnot (u

okamžikových časových řad nemají smysl, neboť výše jejich hodnot nezávisí na daném časovém intervalu, např. aktuální počet obsloužených klientů od začátku roku)

4.1.1 Očištění časové řady o důsledky kalendářních variací

Chceme-li porovnávat jednotlivé hodnoty u intervalových krátkodobých časových řad, musí se tyto hodnoty vztahovat ke stejně dlouhým časovým intervalům.

Očištění na měsíce

- standardní měsíc o délce 30 dnů – údaj za každý měsíc se vydělí počtem dnů v měsíci a vynásobí se 30, součet měsíčních údajů za rok potom odpovídá „roku“ o délce 360 dní
- standardní měsíc o délce 365/12 dnů – součet měsíčních údajů za rok odpovídá délce roku 365 dní

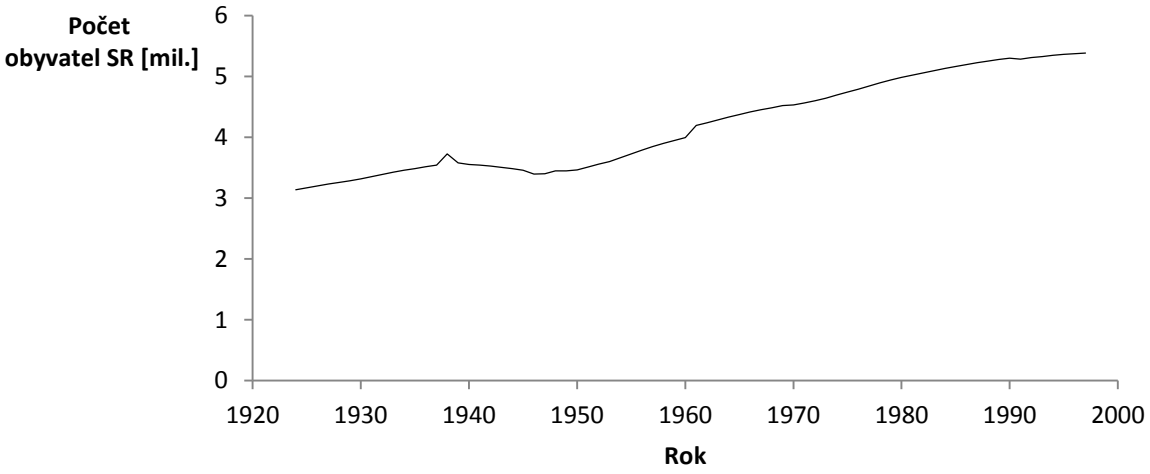
očištění na pracovní dny – provádí se obdobně jako očištění na měsíce

4.2 Grafická analýza časových řad

Jedním ze základních prostředků prezentace časových řad je jejich graf. Nejčastěji se graficky znázorňují původní hodnoty časové řady, nebo řady kumulativní. Často se ale časové řady zobrazují tak, aby více vynikly jejich charakteristické vlastnosti a rysy. K tomu slouží speciální typy grafů.

4.2.1 Spojnicový graf jedné časové řady

Základní informace pro analýzu časových řad získáme ze spojnicových grafů. Jejich princip spočívá v zakreslení jednotlivých hodnot časové řady do souřadnicového systému. Na osu horizontální se vynáší časová proměnná a na osu vertikální hodnoty časové řady nebo její funkce.



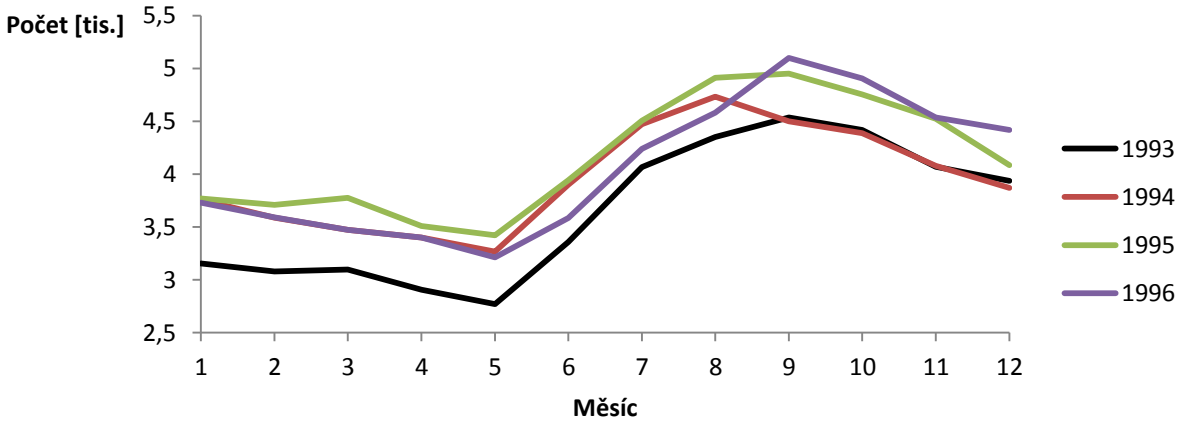
Obr. 4.1: Vývoj počtu obyvatel SR v letech 1924-1997

4.2.2 Spojnicový graf dvou a více časových řad

Do spojnicového grafu můžeme zakreslit i více časových řad. V případě, že zobrazujeme např. dvě časové řady lišící se měřítkem, je možné použít kromě levé i pravou vertikální osu.

4.2.3 Graf ročních hodnot sezónních časových řad

Speciálním případem spojnicového grafu dvou a více časových řad je graf ročních hodnot sezónních časových řad. Tento graf zobrazuje hodnoty časové řady uspořádané podle roků a tak charakterizuje, jak se v jednotlivých letech liší úroveň hodnot v daných sezónách za celou časovou řadu.



Obr. 4.2: Vývoj počtu nezaměstnaných absolventů gymnázií v SR

4.3 Popisné charakteristiky časových řad

4.3.1 Průměrování časových řad

Při práci s časovými řadami je někdy důležité zjistit jejich průměrné hodnoty.

- **intervalové řady** – výpočet se provádí pomocí aritmetického průměru.
- **okamžikové řady** – výpočet se provádí v případě stejných vzdálenosti mezi jednotlivými okamžiky pomocí prostého chronologického průměru, v případě nestejných vzdálenosti pomocí váženého chronologického průměru

  Nechť časovým okamžikům $t_1, t_2, \ldots, t_n$ odpovídají hodnoty časové řady $y_1, y_2 \ \ldots, y_n$:

  Prostý chronologický průměr: $$\frac{\frac{y_1}{2} + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1}$$

  Vážený chronologický průměr:

$$\frac{\frac{y_1 + y_2}{2}(t_2 - t_1) + \frac{y_2 + y_3}{2}(t_3 - t_2) + \cdots + \frac{y_{n-1} + y_n}{2}(t_n - t_{n-1})}{t_n - t_{n-1}}$$

4.3.2 Míry dynamiky

Kromě průměrů nás mnohdy zajímají základní míry dynamiky časových řad, které umožňují charakterizovat základní rysy jejich "chování". Charakteristiky, které si dále uvedeme

vyžadují stejnou délku časových intervalů v intervalových časových řadách nebo stejné vzdálenosti mezi okamžiky zjišťování v okamžikových časových řadách.

- **Absolutní přírůstky**
  Nejjednodušší mírou dynamiky je absolutní přírůstek, který nám říká „o kolik“ se změnila časová řada mezi jednotlivými okamžiky.

$$\Delta^{(1)} y_t = y_t - y_{t-1}, \quad t = 2,3,\dots,n$$

- **Průměrný absolutní přírůstek**
  nám říká „o kolik“ se průměrně změnila časová řada za období mezi dvěma měřeními během sledovaného období.

$$\bar{\Delta} = \frac{1}{n-1}\sum_{t=2}^{n} \Delta^{(1)} y_t = \frac{y_n - y_1}{n-1}$$

- **Koeficienty růstu**
  Koeficienty růstu udávají „kolikrát“ se změnila časová řada mezi jednotlivými okamžiky.

$$k_t = \frac{y_t}{y_{t-1}} \qquad t = 2,3,\dots,n$$

- **Průměrný koeficient růstu**
  nám říká „kolikrát“ se průměrně změnila časová řada za období mezi dvěma měřeními během sledovaného období. Vzhledem k tomu, že průměrujeme poměrové proměnné, musíme pro jeho výpočet použít geometrický průměr.

$$\bar{k} = \sqrt[n-1]{k_2 k_3 \dots k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

- **Meziroční koeficienty růstu**
  jsou podíly hodnot časové řady ve stejných obdobích (sezónách) v po sobě jdoucích letech. V případě čtvrtletní časové řady má meziroční koeficient růstu tvar

$$k_t = \frac{y_t}{y_{t-4}}, \qquad t = 5,6,\dots,n.$$

- **Relativní přírůstky [%]**
  Chceme-li vědět „o kolik procent“ se změnila časová řada mezi jednotlivými okamžiky, použijeme relativní přírůstky.

$$\delta_t = \frac{\Delta^{(1)} y_t}{y_{t-1}} \cdot 100 = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100 = (k_t - 1) \cdot 100, \qquad t = 2,3,\dots,n$$

- **Průměrný relativní přírůstek [%]**

udávající „o kolik %“ se průměrně změnila časová řada za období mezi dvěma měřeními během sledovaného období pak jednoduše vypočteme dle vztahu

$$\bar{\delta}_t = \left(\bar{k} - 1\right) \cdot 100 \ [\%].$$

4.4 Dekompozice časových řad

Časovou řadu lze rozložit na součet (nebo součin) několika složek, z nichž každá bude podstatně jednodušší a bude mít jasnou interpretaci. Těmito složkami jsou:

- Trend D_t
- Sezónní složka S_t
- Cyklická složka C_t
- Náhodná složka E_t

Trend
- Odráží dlouhodobý vývoj (obvykle růst nebo pokles, ale obecně nemusí být tato složka monotónní) daného procesu.
- Co je dlouhodobé? (růst teploty během dne z pohledu lyžaře, zemědělce, meteorologa).

Sezónní složka
- Odráží periodické změny, které se mohou v dané řadě projevovat, a jejich perioda je svázána s kalendářem (mají periodu jednu hodinu, jeden den, týden, měsíc, rok, století…).
- Velký význam v ekonomii, meteorologii…

Cyklická složka
- Odráží periodické změny, které se mohou v dané řadě projevovat, a jejichž perioda neodpovídá délce nějaké kalendářní jednotky.
- V technických vědách se sezónní složka obvykle neuvažuje a všechny periodické jevy se zahrnují do cyklické složky.

Náhodná (reziduální) složka
- Zbývá v časové řadě po odstranění trendu, sezónních a cyklických složek.
- Je tvořena náhodnými fluktuacemi, které nemají žádný systematický charakter.

Nadále budeme pracovat vždy s náhodným procesem a s některou jeho realizací. Pro odlišení budeme časové řady označovat velkými písmeny (např. X_t) a jejich konkrétní realizace malými písmeny (např. x_t).
Znalost každé jednotlivé složky nám umožní například lepší odhad vývoje daného procesu do budoucna (predikci).

Budeme-li se snažit rozložit časovou řadu $\{X_t\}$ na součet složek, budeme předpokládat, že ho lze zapsat ve tvaru:

$$X_t = D_t + S_t + C_t + E_t$$

Tomuto způsobu rozkladu časové řady říkáme aditivní rozklad a používáme jej v případě, že se variabilita hodnot časové řady se v průběhu času příliš nemění.

Mění-li se variabilita hodnot časové řady v čase výrazně, používáme tzv. multiplikativní model.

$$X_t = D_t \cdot S_t \cdot C_t \cdot E_t$$

4.4.1 Metody hledání trendu

Pro hledání trendu se používají buď regresní metody (předpokládáme, že $X_t = D_t + E_t$), nebo adaptivní přístupy, které dokážou reagovat na změny v charakteru trendu. Mezi nejznámější adaptivní přístupy patří metoda klouzavých průměrů.

- **Klouzavé průměry**

Chceme-li z časové řady odstranit šum vznikající působením náhodných vlivů, lze použít metodu klouzavých průměrů spočívající v tom, že se řada původních pozorování nahradí řadou vypočtených klouzavých průměrů. Tato metoda je adaptivní, tzn. dokáže pracovat s časovou řadou, která v čase mění svůj charakter a nelze ji proto popsat jedinou křivkou.

Klouzavý průměr je určitou lineární kombinací *2p+1* členů původní řady. Čím větší je délka klouzavého průměru, tím větší je „vyhlazení" časové řady. V případě, že zvolená délka klouzavého průměru je „lichá", získáme jejich hodnoty jako obyčejné aritmetické průměry dané délky (prosté klouzavé průměry).

Prosté klouzavé průměry

Úseky časové řady o délce *2p+1* vyrovnáme tak, že je nahradíme prostým aritmetickým průměrem.

$$\overline{y}_t = \frac{1}{2p+1}\sum_{i=-p}^{p} y_{t+i} = \frac{y_{t-p} + y_{t-p+1} + \ldots + y_{t+p-1} + y_{t+p}}{2p+1} \qquad t = p+1, p+2, \ldots, n-p$$

p hodnot na začátku a p hodnot na konci časové řady zůstává nevyrovnáno.

Sudá délka klouzavých průměrů se volí jen velmi zřídka. V případě, že délku klouzavých průměrů zvolíme rovnu *2p*, používáme tzv. centrovaných klouzavých průměrů.

Centrované klouzavé průměry

$$y_t = \frac{1}{4p}\left(y_{t-p} + 2y_{t-p+1} + \cdots + 2y_{t+p-1} + y_{t+p}\right) \quad t = p+1, \ldots, n-p$$

Myšlenka tohoto vyrovnání je prostá. K tomu, aby vyrovnaná hodnota odpovídala danému období, potřebujeme lichý počet členů v klouzavém průměru. Ten získáme nejlépe tak, že místo prvního členu v klouzavém průměru vezmeme průměr první a poslední hodnoty dané periody.Tedy klouzavé průměry mají tvar

$$y_t = \frac{\frac{y_{t-p} + y_{t+p}}{2} + y_{t-p+1} + \cdots + y_{t+p+1}}{2p} = \frac{1}{4p}\left(y_{t-p} + 2y_{t-p+1} + \cdots + 2y_{t+p-1} + y_{t+p}\right).$$

U neperiodických časových řad se nejčastěji používají průměry délky 3, 5, 7. Pokud chceme z naší časové řady odstranit sezónní vliv (např. kolísání hodnot během týdne, měsíce…), využíváme klouzavých průměrů s délkou rovnou délce sezónního období.

Příklad: Časová řada (viz. tab.) udává roční objemy vývozu piv (v mil. l.) z ČSFR v letech 1980 až 1991. Vyrovnejte časovou řadu 3-člennými a 4-člennými klouzavými průměry.

Řešení:

Rok	y_t	3-členné klouzavé průměry	5-členné klouzavé průměry
1980	215		
1981	219	218,667	
1982	222	225,333	221,125
1983	235	219,667	218,000
1984	202	214,667	212,125
1985	207	198,667	203,875
1986	187	199,333	196,500
1987	204	188,333	188,625
1988	174	183,333	186,000
1989	172	182,333	196,250
1990	201	215	
1991	272		

3-členné klouz. průměry: $2p + 1 = 3 \Rightarrow p = 1 \Rightarrow$ 1 hodnota na začátku a 1 hodnota na konci bude vynechána.

3-členný klouzavý průměr v roce 1981 $= \dfrac{215 + 219 + 222}{3} = 218{,}667$

3-členný klouzavý průměr v roce 1982 $= \dfrac{219 + 222 + 235}{3} = 225{,}333$

4-členné klouz. průměry: $2p = 4 \Rightarrow p = 2 \Rightarrow$ 2 hodnoty na začátku a 2 hodnoty na konci budou vynechány.

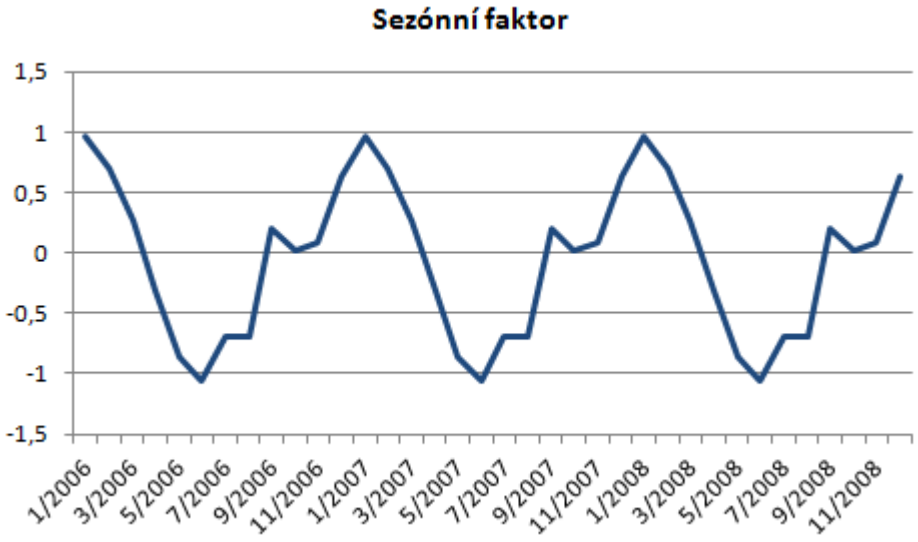
4-členný klouzavý průměr v roce 1982 $= \dfrac{\frac{215}{2} + 219 + 222 + 235 + \frac{202}{2}}{4} = 221{,}125$

atd.

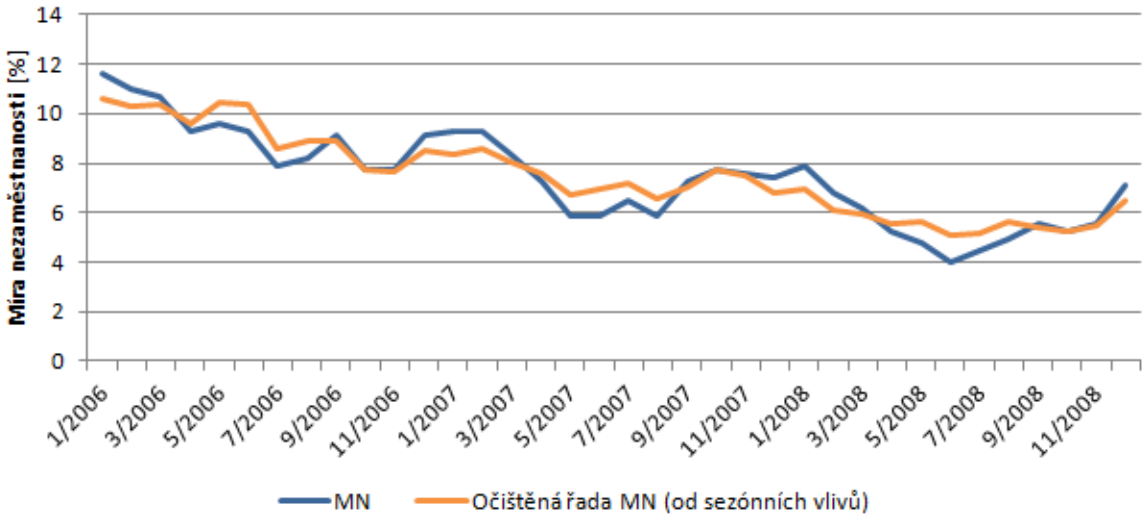
4.4.2 Očištění časové řady od sezónních vlivů

Pro očištění časové řady musíme nejdříve stanovit sezónní faktor v aditivní formě [2]. **Sezónní faktor** stanovíme pomocí **odchylky** časové řady a centrovaných klouzavých průměru o délce rovné periodě časové řady, nejčastěji o délce 12). (viz [2], příklad 6.4)

Sezónní faktor pro určitý měsíc pak určíme jako průměrnou měsíční odchylku, tj. lednový sezónní faktor se určí jako průměr všech lednových odchylek. (Všimněte si, že sezónní faktor je pro všechny roky stejný – viz obrázek.)



Časovou řadu očištěnou od sezónní složky získáme tak, že sezónní faktor odečteme od původní časové řady. Takto očištěná časová řada se pak používá pro další statistické vyhodnocení (regresní analýza, modelování časových řad, atd. [2]).



Literatura

[1] LITSCHMANNOVÁ, M. (2011), *Úvod do statistiky*, skripta - pilotní verze , dostupné z: http://mi21.vsb.cz/modul/uvod-do-statistiky

[2] HANČLOVÁ, J., TVRDÝ, L., Úvod do analýzy časových řad, dostupné z: http://gis.vsb.cz/pan/cz/skoleni.html