Popisná statistika - zavedení pojmů

Soubor individuálních údajů o objektech nazýváme **základní soubor** nebo také **populace**. Zkoumané objekty jsou tzv. **statistické jednotky** a sledujeme u nich vytypované vlastnosti - **statistické znaky** (veličiny, parametry atd.), které nabývají pozorovatelných **hodnot** (**úrovní**).

Podstatou statistických metod je, že informace o základním souboru nezjišťujeme u všech jeho jednotek, ale jen u některých, které získáme tzv. **výběrem**. Vedou nás k tomu různá omezení, např. dosažitelnost všech jednotek, velký rozsah základního souboru, způsob získávání informací (zkoušky životnosti, ověření opotřebení atd.), náklady na statistické sledování a další. Počet vybraných jednotek se nazývá **rozsah** výběru. Dle rozsahu dělíme výběry na **malé** (obvykle do 30 až 50) a **velké** (řádově stovky, tisíce i více). Toto dělení je relativní a závisí na okolnostech statistického sledování. Výběr by měl být **reprezentativní** (poskytovat informace bez omezení) a **homogenní** (bez vlivu dalších různých faktorů). To však často nelze v plné míře verifikovatelně zajistit, a proto obvykle vybíráme statistické jednotky do výběru **náhodně**, ovšem s rizikem, že výběr může poskytnout více či méně zkreslené informace o základním souboru. Podle způsobu provedení rozlišujeme výběry:

- **bez opakování** (každá jednotka může být vybrána nejvýše jednou);
- **s opakováním** (každá jednotka může být vybrána vícekrát);
- **záměrný** (vybíráme typické jednotky);
- **oblastní** (základní soubor rozdělíme na podmnožiny a z nich provedeme části výběru);
- **systematický** nebo **mechanický** (vybíráme vždy několikátou jednotku co do pořadí při realizaci výběru).

Hodnoty znaku, pozorované či zjištěné na statistických jednotkách z výběru o rozsahu n, tvoří **statistický soubor s rozsahem** n. Pro jednorozměrný znak X získáme **jednorozměrný statistický soubor** $(x_1, ..., x_n)$, kde x_i je **pozorovaná hodnota** znaku X u $i-$té statistické jednotky,$i = 1,\ldots, n$. Analogicky pro dvourozměrný znak $(X,\ Y)$ obdržíme **dvourozměrný statistický soubor** $((x_1, y_1), ..., (x_n, y_n))$ apod.

1 Jednorozměrný statistický soubor s kvantitativním znakem

Neroztříděný statistický soubor – získaný statistický soubor $x_1, \ldots x_n)$.

Rozsah statistického souboru – počet prvků: n.

Uspořádaný statistický soubor – $\left(x_{(1)}, \ldots, x_{(n)}\right)$, kde $x_{(i)} \leq x_{(i+1)}$ pro všechny indexy i.

Variační obor – interval $\left\langle x_{(1)}; x_{(n)}\right\rangle$

Rozpětí statistického souboru – délka variačního oboru: $x_{(n)} - x_{(1)}$

Při velkém rozsahu statistického souboru nebo z důvodu dalšího zpracování původní soubor **roztřídíme** a dále již můžeme pracovat s tímto roztříděným statistickým souborem. Tříděním už zároveň získáváme první údaje o statistickém souboru.

Roztříděný statistický soubor získáme pokrytím variačního oboru systémem disjunktních intervalů (obvykle zleva otevřených a zprava uzavřených), tzv. **tříd** o počtu m, které mají obvykle stejnou **délku** h.

Počet tříd m volíme obvykle přibližně $1 + 3,3 \log n$ (pro statistický soubor symetrického charakteru) anebo $\sqrt{n}$ až $2\sqrt{n}$ (pro statistický soubor asymetrického charakteru).

Délka třídy - $h \approx \frac{x_{(n)} - x_{(1)}}{m}$.

Každá třída $\langle x_j,\ x_{j+1}\rangle$je reprezentována uspořádanou dvojicí $\left(x_j^*\ ,\ f_j\right)$, kde x_j^* je **reprezentant** j-té třídy a f_j je **absolutní četnost** j-té třídy $j = 1, ..., m$.

repreyentant j-té třídy - často se nahrazuje **středem** j-té třídy $x_j^* = \frac{x_{(j)} + x_{(j+1)}}{2} \quad j = 1, ..., m$. Při určování délky třídy bereme ohled na požadavek, aby střed třídy x_j^* byl zaokrouhlené číslo. U diskrétního znaku volíme obvykle za středy tříd přímo hodnoty, kterých tento znak může nabývat.

Absolutní četnost j-té třídy f_j - počet prvků x_i původního neroztříděného statistického souboru, které leží v j-té třídě $(x_i \in \langle x_j,\ x_{j+1}\rangle)$.Platí $\sum_{j=1}^{m} f_j = n$.

Relativní četnost j-té třídy $\frac{f_j}{n}$.Uvádí se též v %. Platí$\sum_{j=1}^{m} \frac{f_j}{n} = 1$.

Kumulativní absolutní četnost $F_j = \sum_{k=1}^{j} f_k$.

Kumulativní relativní četnost $\frac{F_j}{n}$.

Roztříděný statistický soubor zapisujeme do tzv. **četnostní tabulky** pro různé typy četností, např. pro absolutní četnosti, viz Tabulka **??**.

x_j^*	x_1^*	...	x_m^*
f_j	f_1	...	f_m

Tabulka 1: Četnostní tabulka

Pro jednorozměrný roztříděný statistický soubor se v případě spojitého znaku X užívají nejčastěji následující dva typy grafů:

1. **Histogram** je soustava obdélníků v kartézské souřadné soustavě, jejichž základny jsou třídy a výšky jsou četnosti tříd (absolutní, relativní, kumulativní atd.)

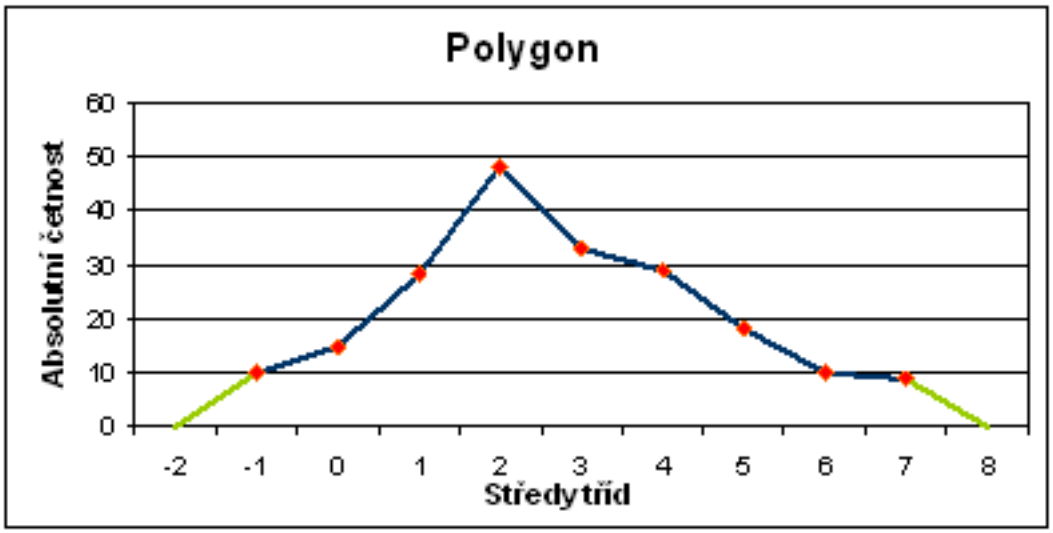
2. **Polygon** je lomená čára v kartézské souřadné soustavě spojující body, jejichž x-ová souřadnice je střed třídy, příp. horní hranice třídy pro kumulativní četnosti a y-ová souřadnice je četnost třídy.

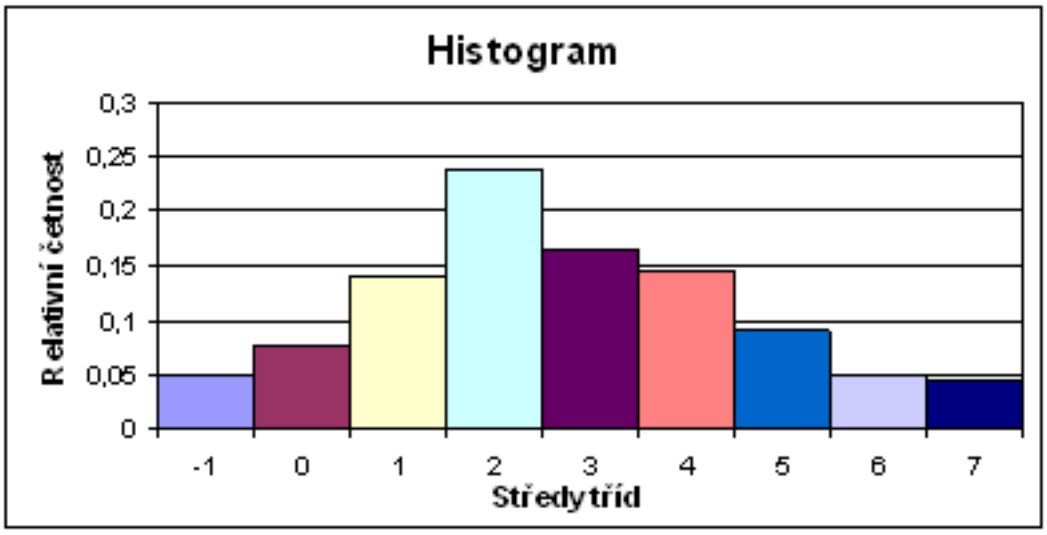
1. Příklad Znázorněte pomocí histogarmu a polygonu informace z Tabulky **??**.

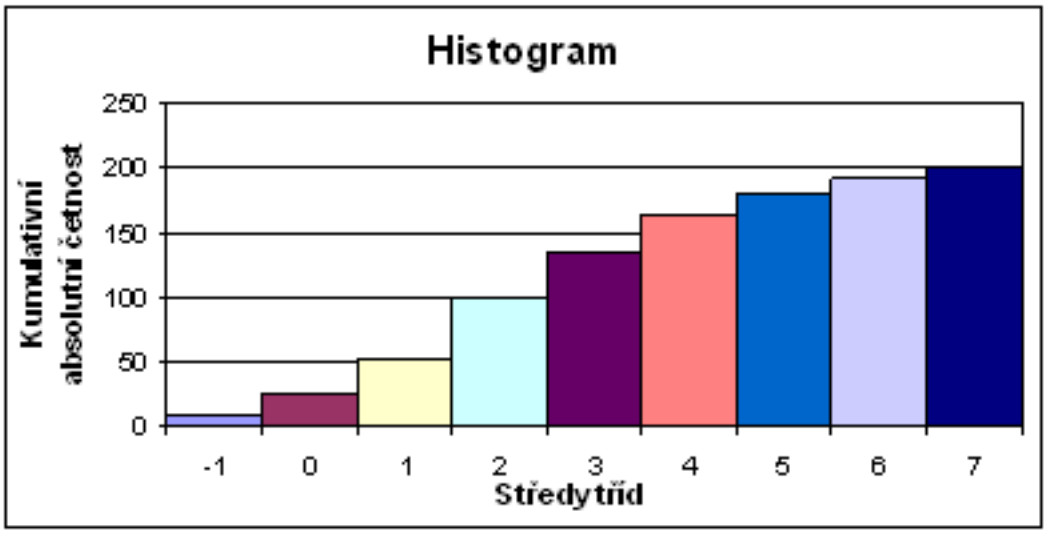
Střed třídy	-1	0	1	2	3	4	5	6	7
Absolutní četnost	10	15	28	48	33	29	18	10	9
Relativní četnost	0,05	0,075	0,14	0,24	0,165	0,145	0,09	0,05	0,045
Kumulativní absolutní četnost	10	25	53	101	134	163	181	191	200
Kumulativní relativní četnost	0,05	0,125	0,265	0,505	0,67	0,815	0,905	0,955	1

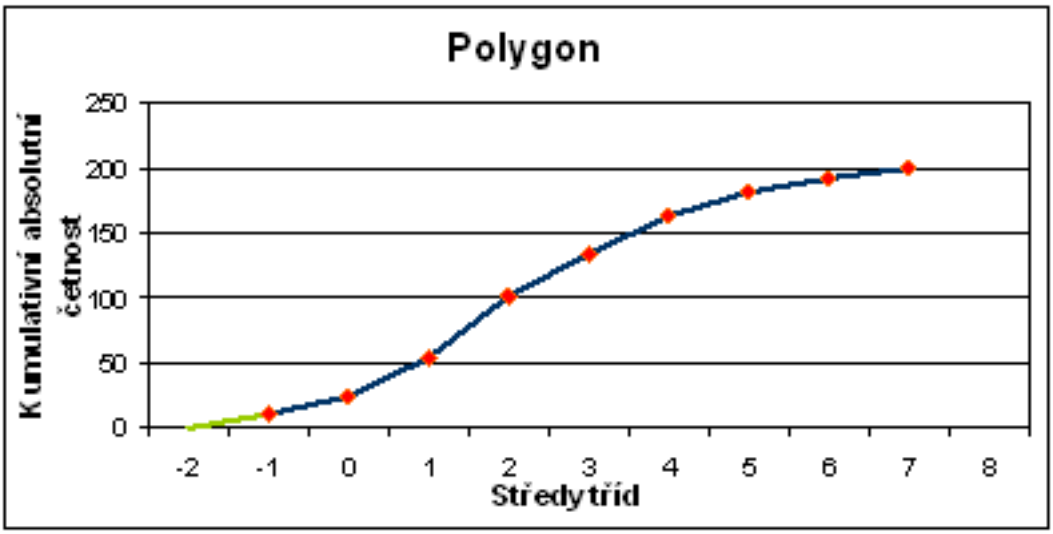
Tabulka 2: Četnostní tabulka k Příkladu **??**

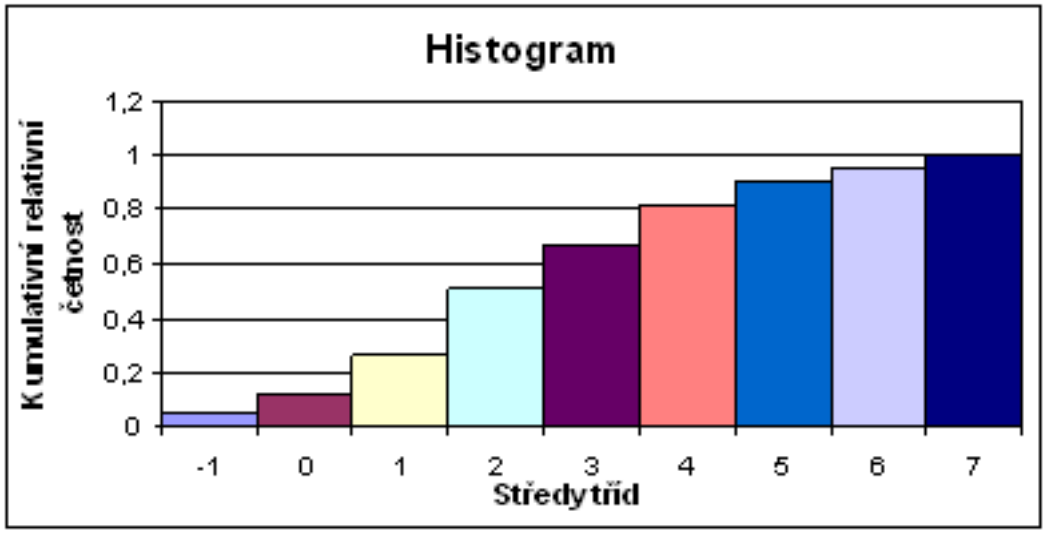
Řešení Řešení je vidět na Obrázcích **??**–**??**.




Histogram
Relativní četnost
0,3
0,25
0,2
0,15
0,1
0,05
0
-1
0
1
2
3
4
5
6
7
Středy tříd


Histogram
Kumulativní absolutní četnost
250
200
150
100
50
0
-1
0
1
2
3
4
5
6
7
Středy tříd


Polygon
Kumulativní absolutní četnost
250
200
150
100
50
0
-2
-1
0
1
2
3
4
5
6
7
8
Středy tříd



Významné vlastnosti statistického souboru vyjadřují v koncentrované formě jeho následující **číselné** (**empirické**) **charakteristiky**. Jde zejména o **charakteristiky polohy**, **proměnlivosti a souměrnosti**.

1.1 Základní charakteristiky polohy

Základní charakteristiky polohy statistického souboru jsou:

1. **Aritmetický průměr**

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

   $\bar{x} = \frac{1}{n} \sum_{j=1}^{m} f_j x_j^*$ pro roztříděný soubor.

   Někdy se užívá též **vážený aritmetický průměr**

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

   kde $w_i \geq 0$ jsou **váhy** (vhodně stanovená reálná čísla, z nichž aspoň jedno je nenulové) hodnot x_i, které vyjadřují jejich význam, např. přesnost.

2. **Medián** pro neroztříděný statistický soubor

   $$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{pro lichá } n\,, \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & \text{pro sudá } n\,. \end{cases}$$

   Medián rozděluje statistický soubor na "dolní polovinu" a "horní polovinu" hodnot x_i. Jde o **robustní** charakteristiku, která je oproti aritmetickému průměru málo citlivá na extrémně odchýlené hodnoty. Pro roztříděný soubor se k výpočtu mediánu užívá vhodná aproximace.

3. **Modus** $\hat{x}$ je číslo, v jehož okolí je nejvíce hodnot x_i, resp. je to střed x_j^* třídy s největší absolutní četností f_j. Modus má tytéž vlastnosti jako aritmetický průměr i medián a dle potřeby se počítá vhodnou aproximací (např. pro roztříděný soubor).

1.2 Základní charakteristiky proměnlivosti (variability)

Základní charakteristiky proměnlivosti (variability) statistického souboru jsou:

1. **Rozptyl** (**disperze**, **variance**)

   $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$ pro neroztříděný soubor,

   $s^2 = \frac{1}{n}\sum_{j=1}^{m} f_j\left(x_j^* - \bar{x}\right)^2 = \left(\frac{1}{n}\sum_{j=1}^{m} f_j x_j^{*\,2}\right) - \bar{x}^2$ pro roztříděný soubor.

   Dle potřeby a také pro zdůraznění znaku X někdy píšeme $s^2(x)$ apod.

   Větší proměnlivosti znaku X odpovídá větší rozptyl a naopak. Při výpočtech se také užívá jiný vzorec pro rozptyl, když výraz $\frac{1}{n}$ zaměníme výrazem $\frac{1}{n-1}$. Takto vypočtený rozptyl je roven číslu $\frac{n}{n-1}s^2 > s^2$ (pro $s^2 \neq 0$). Zdůvodnění výrazu $\frac{1}{n-1}$ plyne z požadavků uvedených v kapitole 6 a 7.

2. **Směrodatná odchylka** $s = \sqrt{s^2}$.

   Dle potřeby také píšeme $s(x)$.

   Větší proměnlivosti znaku X odpovídá větší směrodatná odchylka a naopak.

2 Dvourozměrný statistický soubor s kvantitavními znaky

Při popisování objektů nemusíme zjišťovat pouze jeden údaj. Můžeme zjistit více informací o objektu, které přeneseme do tabulky (např. jeden řádek tabulky popisuje jeden objekt). Tím dostáváme vícerozměrný statistický soubor. V dalším popisu se omezíme na dvojrozměrný statistický soubor a hlavně na vztah mezi znaky. Vyšetřování vícerozměrného statistického souboru je analogické.

Neroztříděný statistický soubor – $((x_1, y_1), ..., (x_n, y_n))$ s rozsahem n lze zapsat například do Tabulky **??**. Každý sloupec je jednorozměrný statistický soubor: $(x_1, ..., x_n)$, $(y_1, ..., y_n)$. Zpracováním

x_1	y_1
x_2	y_2
...	...
x_n	y_n

Tabulka 3: Neroztříděný statistický soubor – $((x_1, y_1), ..., (x_n, y_n))$ s rozsahem n

těchto souborů získáme jejich číselné charakteristiky $\bar{x}$, $\bar{y}$, $s^2(x)$, $s^2(y)$ atd.

Rozsah statistického souboru – počet prvků: n.

Roztříděný dvourozměrný statistický soubor získáme roztříděním jednorozměrných statistických souborů $(x_1, \ldots, x_n)$ a $(y_1, \ldots, y_n)$, přičemž oba roztříděné soubory mohou mít různé počty tříd i jejich délky. Předpokládejme, že soubor $(x_1, \ldots, x_n)$ byl roztříděn na m_1 tříd a soubor $(y_1, \ldots, y_n)$ byl roztříděn na m_2 tříd. Dostaneme tak dvourozměrné třídy se **středy** a **absolutními četnostmi.**

Středy tříd – $\left(x_j^*, y_k^*\right)$

Absolutní četnost – f_{jk}, $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$.

Relativní četnost – $\frac{f_{jk}}{n}$,$j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$.

Kumulativní absolutní četnost – F_{jk}, $F_{jk} = \sum_{r=1}^{j}\sum_{s=1}^{k} f_{rs}$, $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$.

Kumulativní relativní četnost – $\frac{F_{jk}}{n}$, $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$.

Marginální (**okrajové**) **četnosti** – f_{xj} a f_{yk}

$$f_{xj} = \sum_{k=1}^{m_2} f_{jk}, \quad j = 1, \ldots, m_1$$

$$f_{yk} = \sum_{j=1}^{m_1} f_{jk}, \quad k = 1, \ldots, m_2$$

Platí :

$$\sum_{j=1}^{m_1} f_{xj} = \sum_{k=1}^{m_2} f_{yk} = \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk} = n.$$

Přehledný zápis těchto četností je ve formě **četnostní tabulky**. Následující Tabulka **??** je pro absolutní četnosti a marginální četnosti.

$y_k^*\ x_j^*$	y_1^*	$\dots$	$y_{m_2}^*$	f_{xj}
x_1^*	f_{11}	$\dots$	$f_{1\,m_2}$	$f_{x\,1}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{m_1}^*$	$f_{m_1 1}$	$\dots$	$f_{m_1\,m_2}$	$f_{x\,m_1}$
f_{yk}	$f_{y\,1}$	$\dots$	$f_{y\,m_2}$	n

Tabulka 4: Absolutní četnosti a marginální četnosti

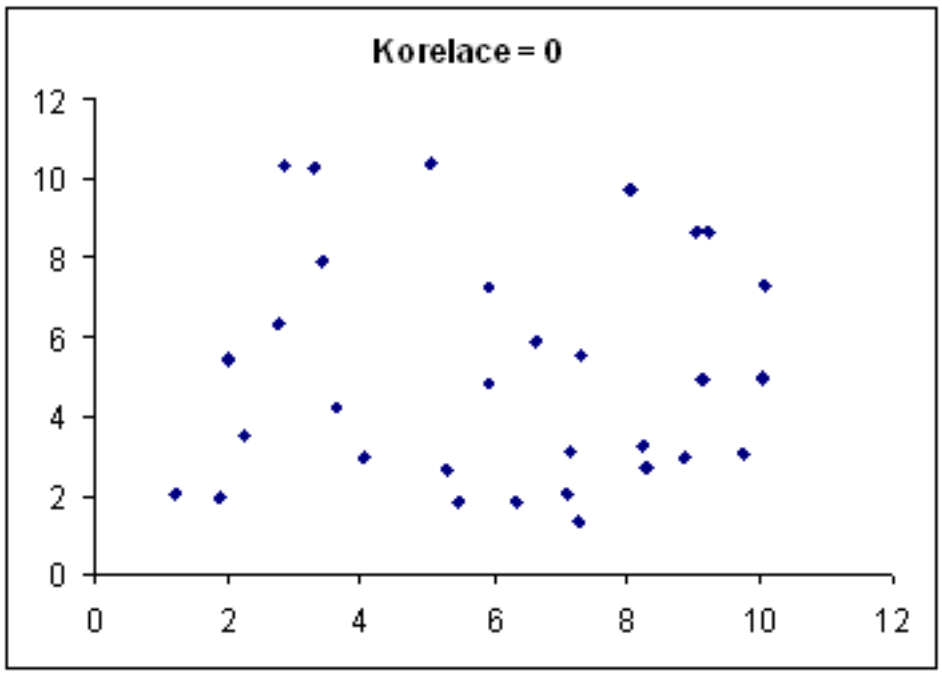
Pro roztříděné jednorozměrné statistické soubory $\left(x_j^*\ ,\ f_{xj}\right)$, $j = 1, \dots, m_1$, a $\left(y_k^*\ ,\ f_{yk}\right)$, $k = 1, \dots, m_2$, obdržíme jejich číselné charakteristiky $\bar{x}$, $\bar{y}$, $s^2(x)$, $s^2(y)$ atd.

Koeficient korelace (**korelační koeficient**) – r určuje míru lineární závislosti znaků X a Y

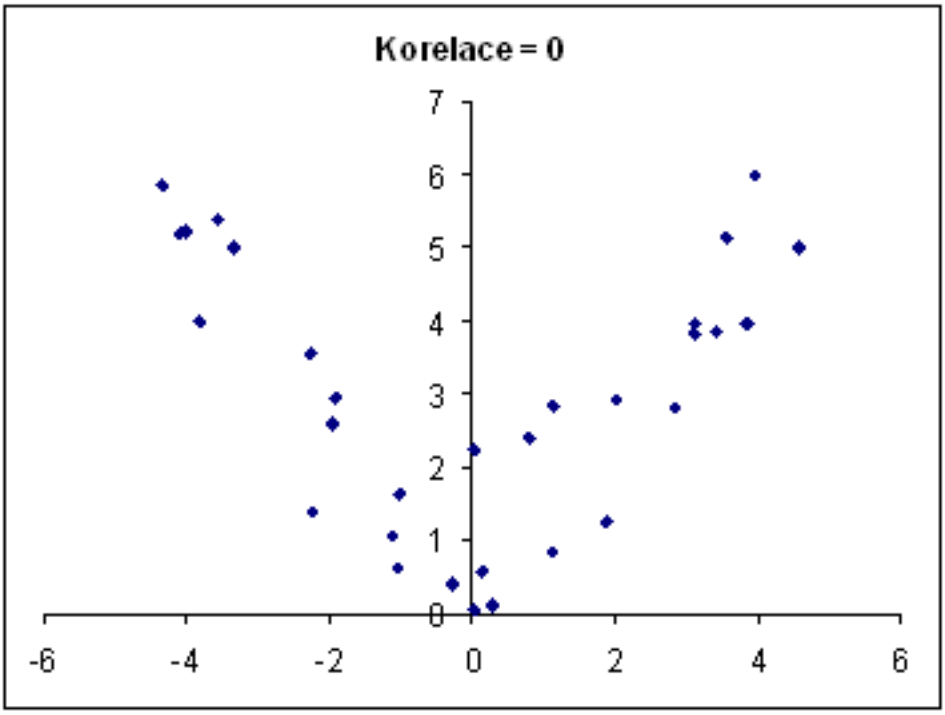
$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s(x)s(y)} \qquad \text{pro neroztříděný soubor,}$$

$$r = \frac{\frac{1}{n}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk}\left(x_j^* - \bar{x}\right)\left(y_k^* - \bar{y}\right)}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} f_{jk} x_j^* y_k^* - \bar{x}\bar{y}}{s(x)s(y)}, \qquad \text{pro roztříděný soubor,}$$

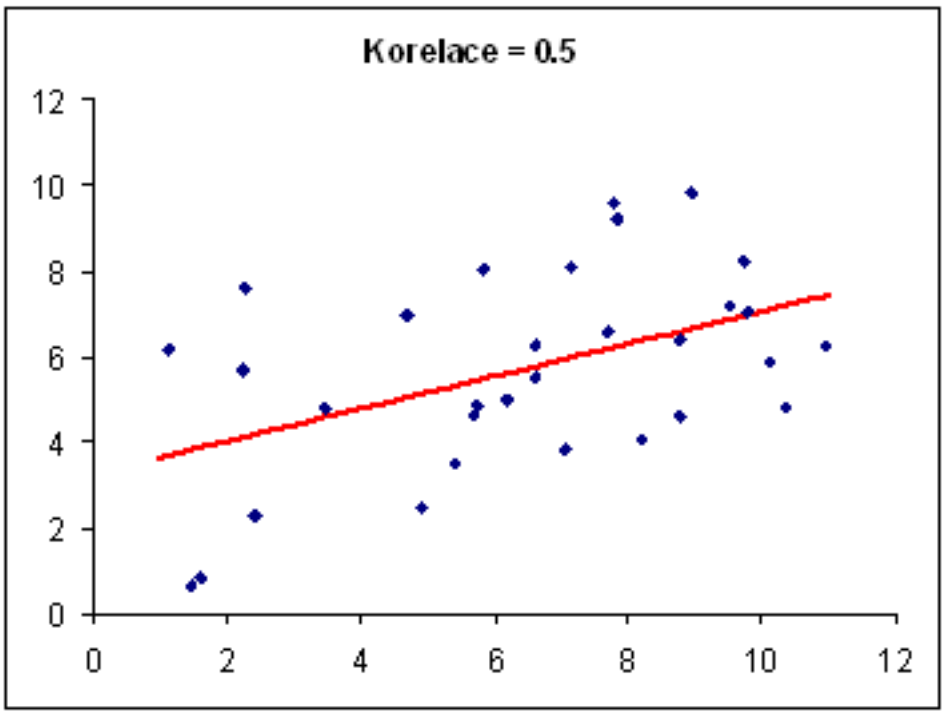
přičemž čitatelé ve všech zlomcích vyjadřují tzv. **kovarianci**, kterou značíme **cov**. Někdy pro zdůraznění znaků X, Y píšeme $r(x,\, y)$, resp. **cov**$(x,\, y)$. Koeficient korelace r je pouze mírou lineární závislosti mezi znaky X a Y. Čím je jeho hodnota bližší 1 anebo -1, tím je závislost bližší lineární závislosti a body (x_i, y_i) bližší přímce. Jeho kladná (záporná) hodnota odpovídá celkově rostoucí (klesající) závislosti mezi X a Y. Hodnota blízká 0 vyjadřuje, že závislost není lineární popřípadě znaky X, Ymohou být nezávislé. Pro grafické vyjádření dvourozměrného neroztříděného statistického souboru se užívá **rozptylový graf**. Na Obrázcích **??-??** jsou rovněž uvedeny pro ilustraci hodnoty koeficientu korelace.



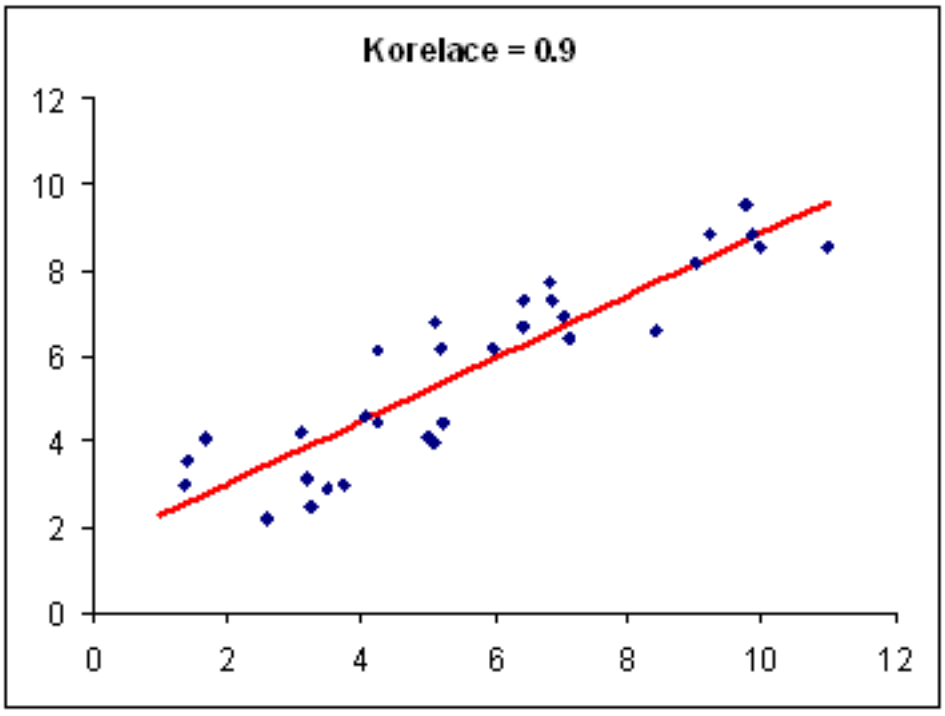
Obrázek 1:



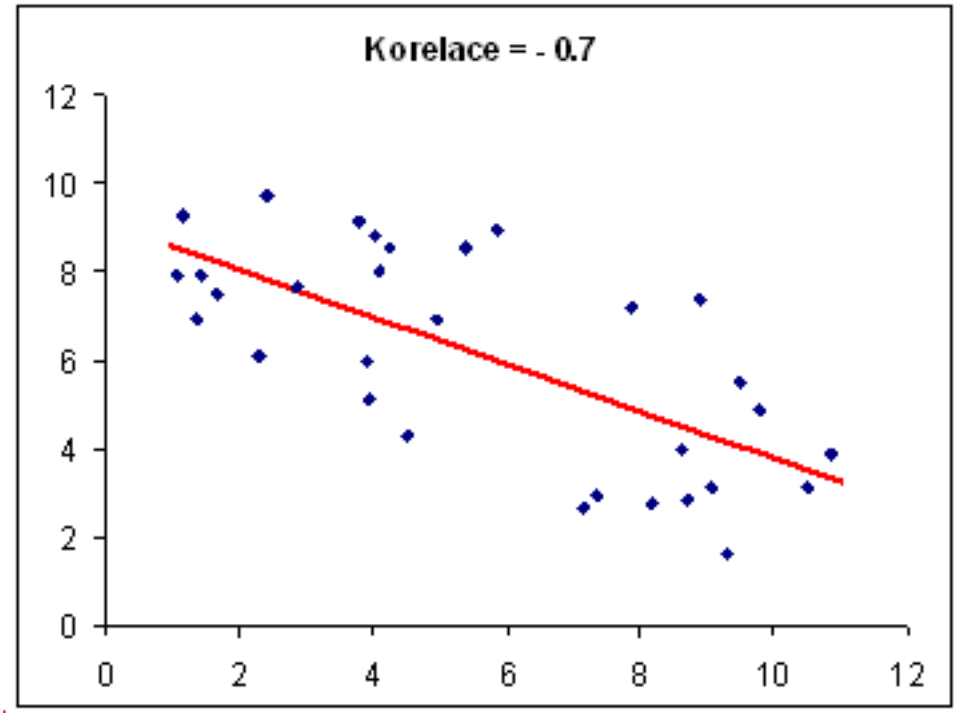
Obrázek 2:



Obrázek 3:



Obrázek 4:



Obrázek 5:

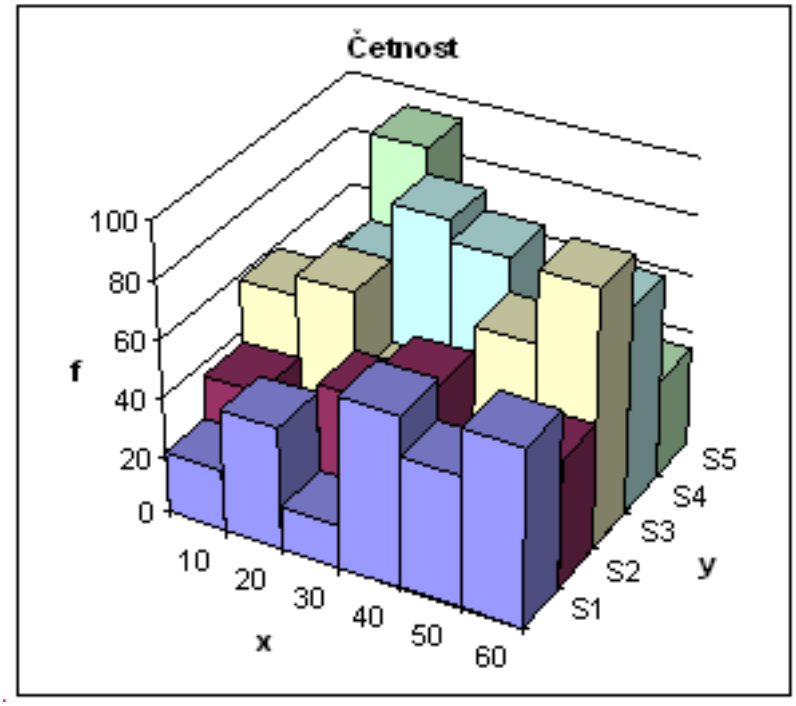
Pro grafické vyjádření dvourozměrného roztříděného statistický souboru se užívá třírozměrný **histogram** (viz Obrázky **??**, resp. **??**), případně třírozměrný **sloupcový graf** pro diskrétní znaky X, Y dané Tabulkou **??**, resp. **??**.

Četnost	10	20	30	40	50	60
1	21	42	15	59	45	61
2	36	17	45	56	25	44
3	58	65	36	41	65	89
4	12	63	84	77	47	71
5	43	93	62	43	32	34

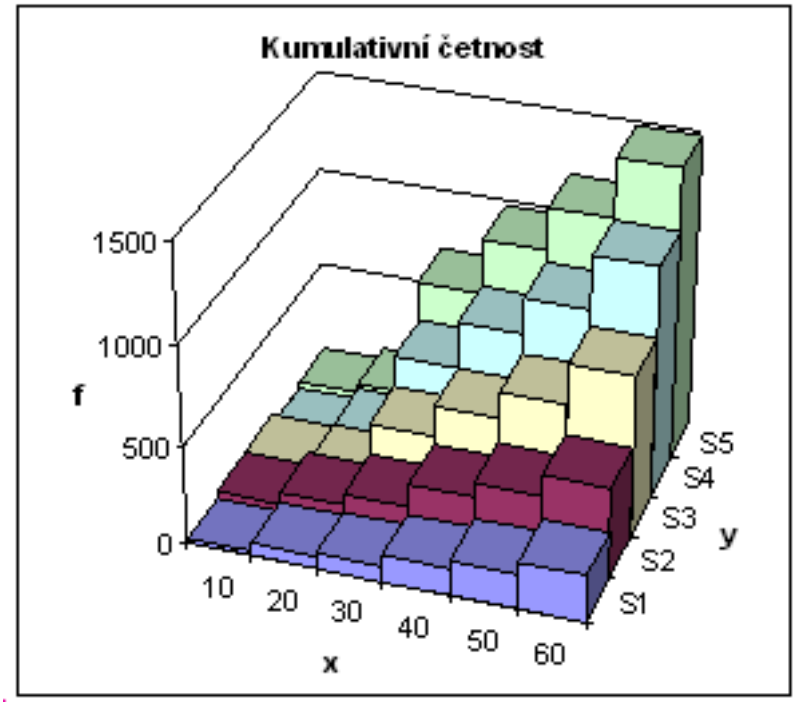
Tabulka 5:

Kumulativní četnost	10	20	30	40	50	60
1	21	63	78	137	182	243
2	57	116	176	291	361	466
3	58	65	335	491	626	820
4	12	63	494	727	909	1174
5	43	93	692	968	1182	1481

Tabulka 6:



Obrázek 6: Třírozměrný histogram k Tabulce **??**



Obrázek 7: Třírozměrný histogram k Tabulce **??**

3 Statistické soubory s kvalitativními znaky

Jednorozměrný statistický soubor s kvalitativním znakem $(x_1, \ldots, x_n)$ s rozsahem n vyjadřujeme pomocí **četnostní tabulky**, kde x_j^* jsou možné slovní hodnoty znaku X a $f_j$jsou četnosti těchto hodnot v původním souboru, $j = 1, \ldots, m$. Číselné charakteristiky se až na výjimky (variabilitu) nepoužívají. Ke grafickému vyjádření souboru slouží **sloupcový graf**, **koláčový graf** apod.

Dvourozměrný statistický soubor s kvalitativními znaky $((x_1, y_1), \ldots, (x_n, y_n))$ s rozsahem n vyjadřujeme pomocí **četnostní tabulky** podobně jako pro kvantitativní znaky, kde $\left(x_j^*, y_k^*\right)$ jsou dvojice možných slovních hodnot dvourozměrného kvalitativního znaku (X, Y) a f_{jk} jsou četnosti těchto hodnot v původním souboru pro $j = 1, \ldots, m_1$ a $k = 1, \ldots, m_2$. Z číselných charakteristik se užívají především různé míry závislosti znaků X a Y. Ke grafickému vyjádření souboru slouží třírozměrný **sloupcový graf** podobný třírozměrnému sloupcovému grafu pro dvourozměrný diskrétní kvantitativní znak.