STATISTIKA

© Ing. Tomáš Löster, Ph.D.; 2015

Kontaktní údaje:

- ***e-mail:*** **tomas.loster@vse.cz** nebo **losterto@vse.cz**

Zdroje dat a software:

- **Statistický software**: SAS
  Statistica
  Statgraphics
  SPSS
  S-Plus
  MS Excel (stat. funkce, analýza dat)
- **zdroje dat:** **http://www.czso.cz/** (ČSÚ)
  http://www.cnb.cz/ (ČNB)
  http://www.mpsv.cz/ (MPSV)
  http://archive.ics.uci.edu/ml/datasets.html
  atd.

Úvod

- ***statistika***
  - statistické údaje o hromadných jevech
  - praktická činnost
  - vědní disciplína

- **základní pojmy**
  - ***statistický soubor***

    = množina prvků (jednotek), které mají alespoň jednu společnou vlastnost

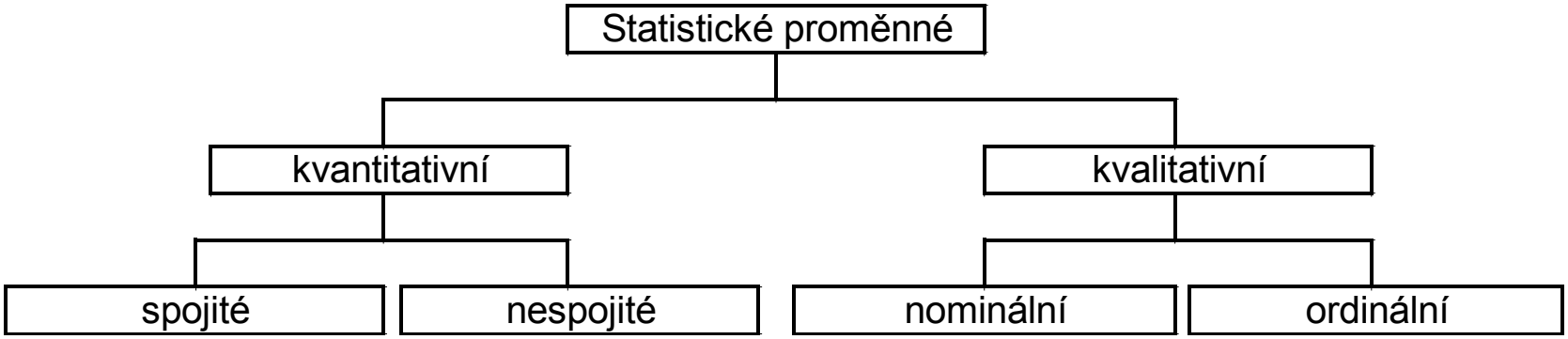
    - ***základní soubor*** (obsahuje všechny existující statistické jednotky)
    - ***výběrový soubor*** (obsahuje pouze část statistických jednotek)
  - ***statistická jednotka***

    = prvek statistického souboru, který má různé vlastnosti
  - ***statistický znak***

    = vlastnost statistické jednotky

    - ***identifikační znak*** (společná vlastnost pro všechny jednotky souboru)
    - ***statistická proměnná*** (nabývá různých obměn u jednotlivých jednotek v souboru)

- ***statistické proměnné***
  - ***slovní (kvalitativní)***
    - ***nominální (názvové)*** – jednotlivé varianty proměnné nelze seřadit podle pořadí)
    - ***ordinální (pořadové)*** – (jednotlivé varianty lze seřadit od nejnižší do nejvyšší obměny)
  - ***číselné (kvantitativní, numerické)***
    - ***spojité*** – nabývají hodnot z konečného nebo nekonečného intervalu)
    - ***nespojité (diskrétní)*** – nabývají hodnot malého počtu jednoznačně izolovaných hodnot


Statistické proměnné
kvantitativní
kvalitativní
spojité
nespojité
nominální
ordinální

Popisná statistika

- **tabulka rozdělení četností**
- **číselné charakteristiky**
- **grafy**

- **TABULKA ROZDĚLENÍ ČETNOSTÍ**
  - četnost
    - absolutní četnost $\sum_{i=1}^{k} n_i = n$
    - relativní četnost $p_i = \frac{n_i}{n}$
  - kumulativní četnost
    - kumulativní absolutní četnost $N_i = \sum_{j=1}^{i} n_j$
    - kumulativní relativní četnost $P_i = \sum_{j=1}^{i} p_j$

- ***tabulka rozdělení četností pro nespojitý znak***

  - x_i (sledovaný znak) = počet automobilů firmy

x_i	n_i	p_i	N_i	P_i
0	5	0,05	5	0,05
1	15	0,15	20	0,20
2	56	0,56	76	0,76
3	14	0,14	90	0,90
4	10	0,10	100	1,00
Σ	**100**	**1,00**	**-**	**-**

- ***tabulka rozdělení četností pro spojitý znak***
  - pravidlo pro stanovení počtu intervalů $k = \sqrt{n}$
  - pravidlo pro stanovení šíře intervalů $\frac{x_{\max} - x_{\min}}{k}$
    - x_i (sledovaný znak) = měsíční obrat firmy (v mil. Kč)

x_i	n_i	p_i	N_i	P_i
0-20	5	0,05	5	0,05
21-40	15	0,15	20	0,20
41-60	56	0,56	76	0,76
61-80	14	0,14	90	0,90
81-100	10	0,10	100	1,00
Σ	**100**	**1,00**	**-**	**-**

- **ČÍSELNÉ CHARAKTERISTIKY**
  - ***charakteristiky polohy***
    - střední hodnoty
    - kvantily
  - ***charakteristiky variability***
    - absolutní variability
    - relativní variability
  - použití charakteristik prostých a vážených v závislosti na způsobu zadání vstupních údajů

- **STŘEDNÍ HODNOTY**
  - ***aritmetický průměr***
    - prostý AP $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
    - vážený AP $\bar{x} = \frac{\sum_{i=1}^{k} x_i \, n_i}{\sum_{i=1}^{k} n_i}$ $\bar{x} = \sum_{i=1}^{k} x_i \, p_i$ $\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i \, n_i}{\sum_{i=1}^{k} n_i}$ $\bar{x} = \frac{\sum_{i=1}^{k} x_i^{STŘ} n_i}{\sum_{i=1}^{k} n_i}$
    - vlastnosti AP $\overline{x+k} = \bar{x} + k$

      $\bar{k} = k$

      $\overline{x \cdot k} = \bar{x} \cdot k$

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

- ***harmonický průměr***
  - prostý HP $$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$
  - vážený HP $$\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \qquad \bar{x}_H = \frac{1}{\sum_{i=1}^{k} \frac{p_i}{x_i}}$$
  - použití HP

- ***geometrický průměr***
  - prostý GP $\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$
  - vážený GP $\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$
  - použití GP
- ***modus*** $\hat{x}$

- **KVANTILY**
  - ***p*-procentní kvantil** $\tilde{x}_p$
    - určení **pořadí** jednotky $n \cdot \frac{p}{100} < z_p < n \cdot \frac{p}{100} + 1$
  - **pojmenované kvantily**
    - ***kvartily* (25%, 50% a 75% kvantily)**
    - ***decily* (10%, 20%, ..., 90% kvantily)**
    - ***percentily* (1%, 2%, ..., 99% kvantily)**
    - **další**

- **MÍRY ABSOLUTNÍ VARIABILITY**
  - ***rozpětí***
    - ***variační rozpětí*** $R = x_{\max} - x_{\min}$
    - ***kvartilové rozpětí*** $R_Q = \tilde{x}_{75} - \tilde{x}_{25}$
    - ***decilové rozpětí*** $R_D = \tilde{x}_{90} - \tilde{x}_{10}$
    - ***percentilové rozpětí*** $R_P = \tilde{x}_{99} - \tilde{x}_{1}$

- ***průměrná absolutní odchylka***

  - prostý tvar $$d_x = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n}$$

  - vážený tvar $$d_x = \frac{\sum_{i=1}^{k} \left| x_i - \bar{x} \right| n_i}{\sum_{i=1}^{k} n_i} \qquad d_x = \sum_{i=1}^{k} \left| x_i - \bar{x} \right| p_i$$

- ***rozptyl***
  - prostý tvar $$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$
  - vážený tvar $$s_x^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{\sum_{i=1}^{k} n_i} \qquad s_x^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 p_i$$
  - vlastnosti rozptylu $$s_{x+k}^2 = s_x^2$$

$$s_k^2 = 0$$

$$s_{x \cdot k}^2 = k^2 \cdot s_x^2$$

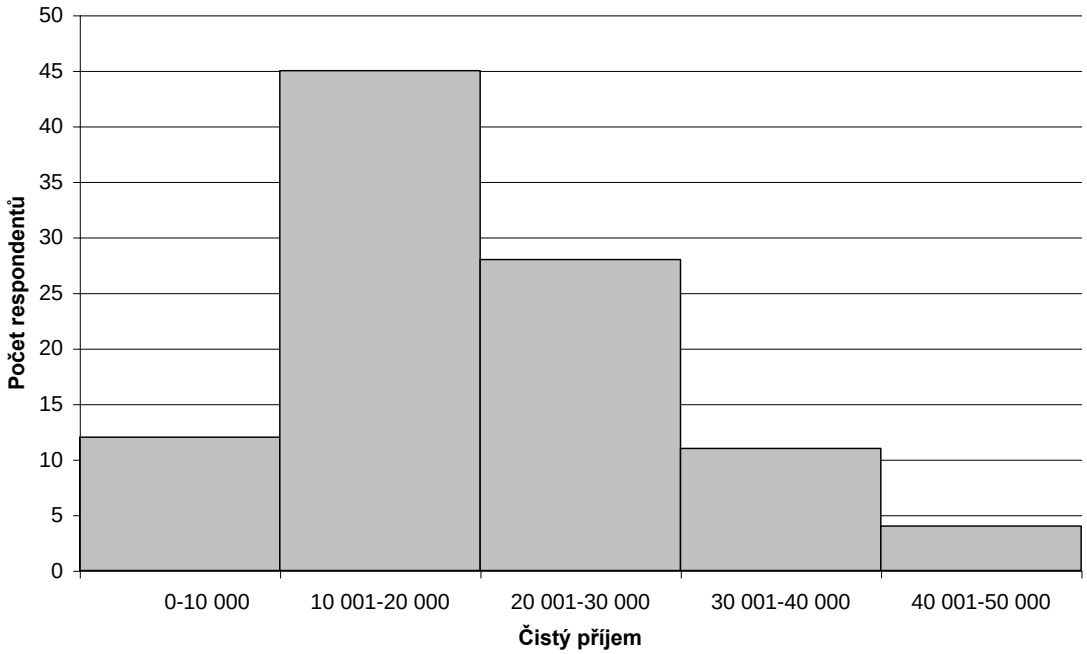
$$s_{x \pm y}^2 = s_x^2 + s_y^2$$

- ***směrodatná odchylka*** $s_x = \sqrt{s_x^2}$

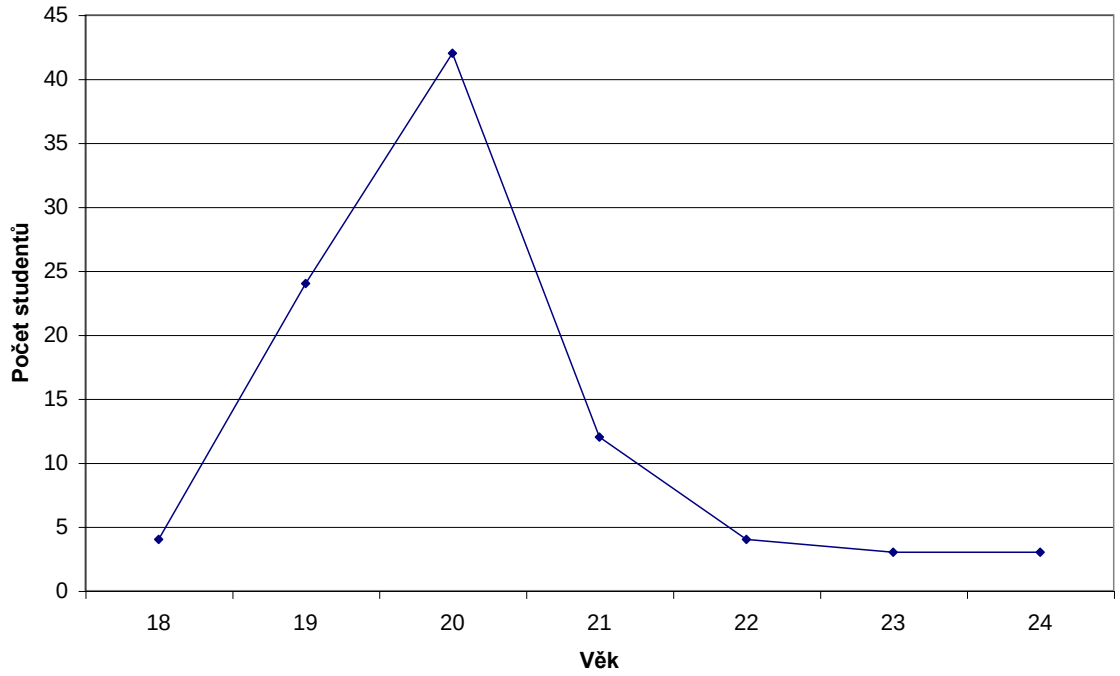
- **MÍRY RELATIVNÍ VARIABILITY**

  - ***variační koeficient*** $v_x = \frac{s_x}{\bar{x}}$

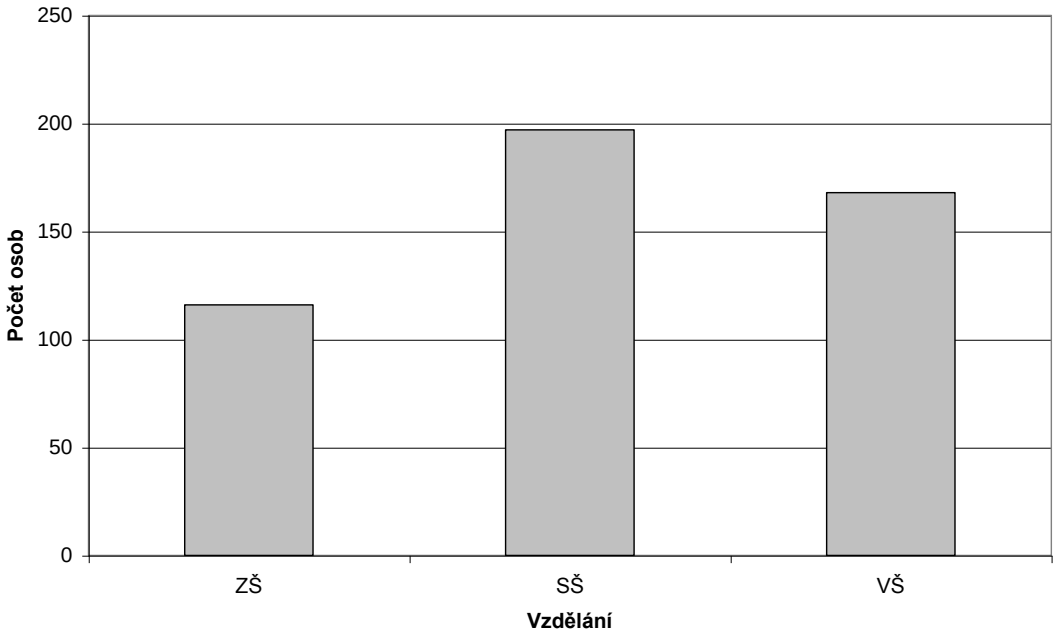
- **GRAFY**
  - ***Histogram četností***



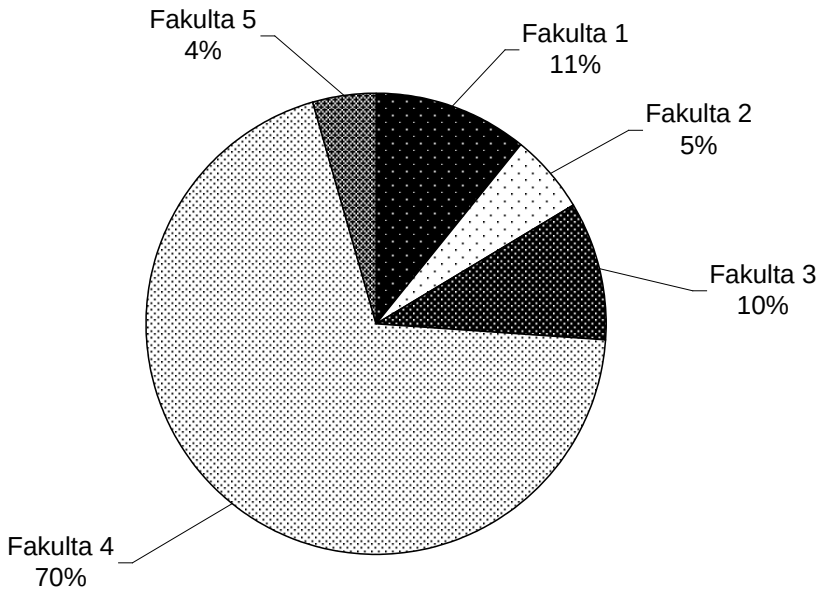
- ***Polygon četností***



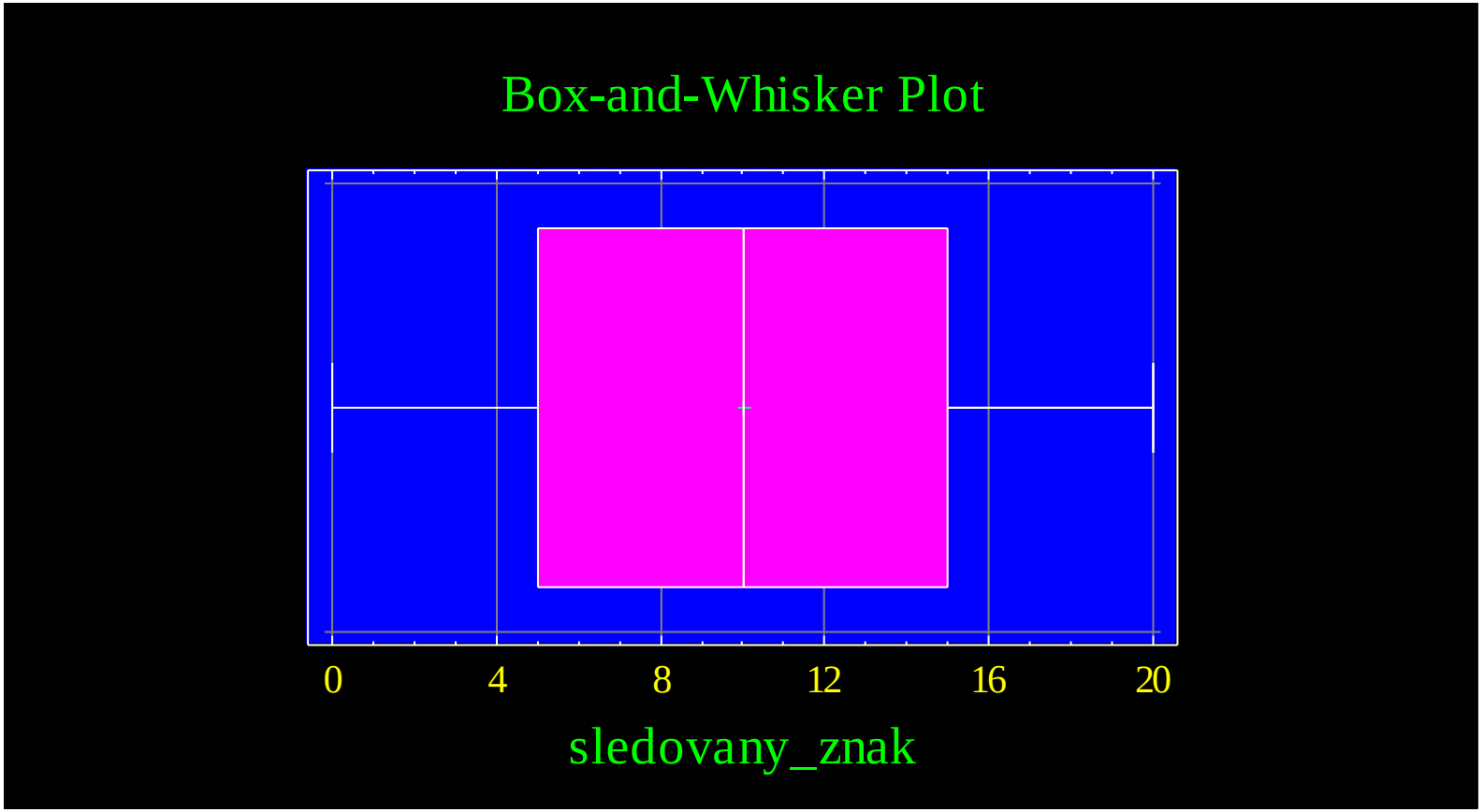
- ***Sloupcový graf***



- ***Koláčový graf***



- ***Krabičkový (BOX) graf***



- ***Příklad č.1***

Na základě hodnot v prvním sloupci vypočítejte zástupce všech skupin popisných charakteristik a výsledné hodnoty interpretujte. Zároveň porovnejte oba soubory hodnot.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$x_i^{(\text{uspoř})}$
1	-2	4	1
2	-1	1	2
4	1	1	2
2	-1	1	4
6	3	9	6
15	**-**	**16**	**15**

n	5		
$\bar{x}$	3		
s^2_x	3,20		
s_x	1,79		
V_x	0,60		
z_{50}	2,5	3,5	
x_{50}	2		
R	5		

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$x_i^{(\text{uspoř})}$
-5	-8	64	-5
2	-1	1	2
4	1	1	2
2	-1	1	4
12	9	81	12
15	**-**	**148**	**15**

n	5		
$\bar{x}$	3		
s^2_x	29,60		
s_x	5,44		
V_x	1,81		
z_{50}	2,5	3,5	
x_{50}	2		
R	17		

- ***Příklad č.2***

Na základě hodnot zadaných hodnot v prvním a druhém sloupci vypočítejte zástupce charakteristik polohy a variability.

x_i	n_i	$x_{i*}n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$	N_i
0	5	0	-2,09	4,37	21,84	5
1	15	15	-1,09	1,19	17,82	20
2	56	112	-0,09	0,01	0,45	76
3	14	42	0,91	0,83	11,59	90
4	10	40	1,91	3,65	36,48	100
Σ	**100**	**209**	**-**	**-**	**88,19**	**-**

n	100	
$\bar{x}$	2,09	
s^2_x	0,88	
s_x	0,94	
V_x	0,45	
z_{50}	50	51
x_{50}	2	
R	4	

- ***Příklad č.3***

Řidič jel z města A do města B rychlostí 90 km/h a z města B do města C rychlostí 110 km/h. Vypočítejte průměrnou rychlost řidiče na celé trase, tj. z města A do města C, pokud víme, že vzdálenost měst A-B a B-C je stejná.

Řešení:

$$\bar{x}_H = \frac{2}{\frac{1}{90} + \frac{1}{110}} = 99\text{km/h}$$

- ***Příklad č.4***

  Na základě tabulky, která obsahuje tempa růstu ceny výrobku A (koeficient růstu je definován jako podíl dvou sousedních cen, kde v čitateli je hodnota daného roku a ve jmenovateli hodnota předchozího roku), stanovte průměrné tempo růstu cen za celé sledované období.

Řešení:

2000	2001	2002	2003	2004	2005	2006
-	1,10	1,20	0,93	1,17	1,15	1,20

Πx_i 1,9821

$\bar{x}_G$ 1,1208

- ***Příklad č.5***

  Hodnota HDP (měřeného v mld. Kč, v běžných cenách) vzrostla od roku 1999 do roku 2007 z 1466,5 na 3231,6. Určete, jaký byl průměrný relativní meziroční přírůstek hodnoty HDP.

Řešení:

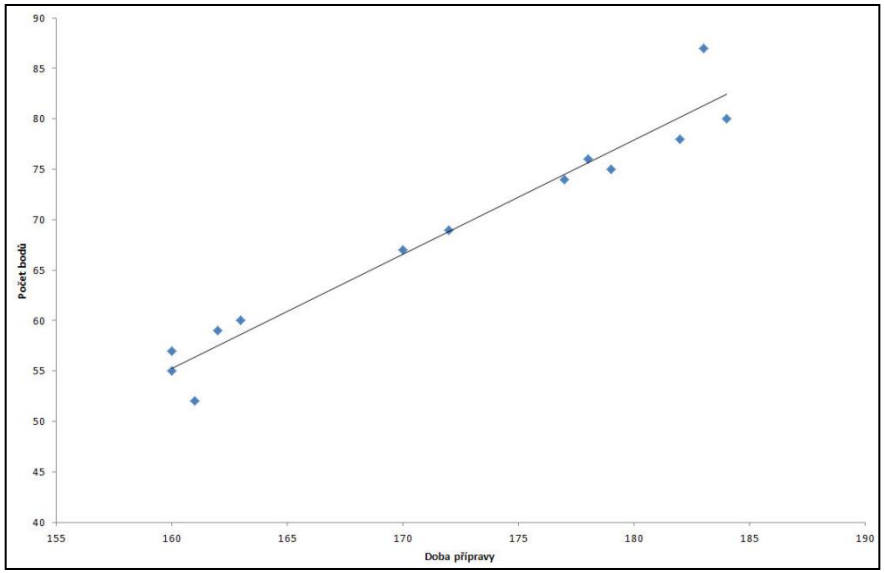
$$\bar{x}_G = \sqrt[8]{\frac{3231,6}{1466,5}} = 1,104$$

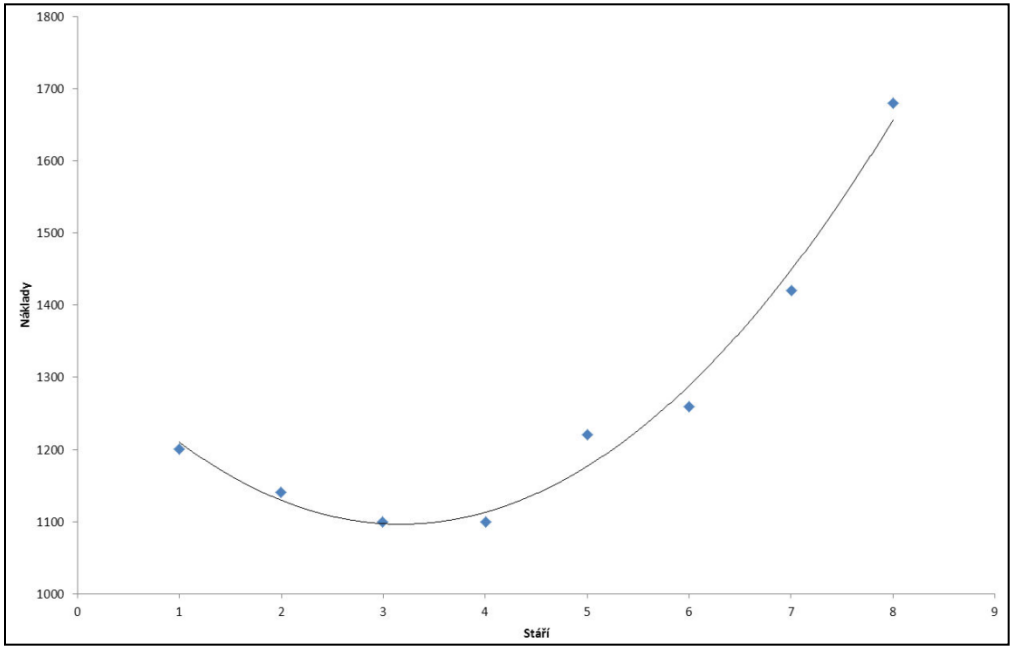
ZKOUMÁNÍ ZÁVISLOSTI MEZI KVANTITAVNÍMI PROMĚNNÝMI

- *druh závislosti*
  - **pevná (funkční)** => změně jednoho znaku jednoznačně odpovídá změna druhého znaku (podle funkčního vztahu)
  - **volná (statistická)** => změnám jedné veličiny odpovídají změny druhé veličiny tak, že změna jedné proměnné zvýší pravděpodobnost změny druhé proměnné.
- *směr závislosti*
  - **jednostranné závislosti** => regresní analýza
  - **oboustranné (vzájemné) závislosti** => korelační analýza

Bodový diagram

- slouží ke grafickému zobrazení závislosti





- z grafu je možné usuzovat na:

  => průběh závislosti: (lineární, nelineární)

  => intenzitu závislosti: (podle kolísání hodnot kolem křivky)

- **REGRESNÍ ANALÝZA**

  => Slouží k popisu statistických závislostí

  => **Cíl**: pomocí hodnot jedné či více proměnných X_i (kvantitativní) odhadovat hodnoty proměnné Y (kvantitativní).

  - **Jednoduchá regresní analýza**

    => Y … vysvětlovaná (závislá) proměnná

    => X … vysvětlující (nezávislá) proměnná

    $$y = f(x)$$

  - **Vícenásobná regresní analýza**

    => Y … vysvětlovaná (závislá) proměnná

    => $X_1, X_2, \ldots X_k$, vysvětlující (nezávislé) proměnné

    $$y = f(x_1, x_2, ..., x_k)$$

- základní typy regresních funkcí

=> **lineární** z hlediska regresních parametrů

regresní přímka: $Y = \beta_0 + \beta_1 x$,

regresní rovina: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$,

regresní nadrovina: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K$,

regresní hyperbola: $Y = \beta_0 + \beta_1 \frac{1}{x}$,

regresní logaritmická funkce: $Y = \beta_0 + \beta_1 \ln x$,

regresní parabola: $Y = \beta_0 + \beta_1 x + \beta_2 x^2$.

=> **nelineární** z hlediska regresních parametrů (**převoditelné**)

regresní mocninná funkce: $Y = \beta_0 x^{\beta_1}$,

regresní exponenciální funkce: $Y = \beta_0 \beta_1^{\,x}$.

- odhad parametru β_1

$$b_1 = b_{yx} = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

- odhad parametru β_0

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} = \bar{y} - b_1\bar{x}$$

- korelační koeficient

$$r_{yx} = r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

$$r_{xy} \in \langle -1;1 \rangle$$

- výpočet R^2 z hodnoty korelačního koeficientu

$$R^2 = r_{yx}^2$$

• ***Příklad***

U 13 náhodně vybraných studentů byla pomocí experimentu zjišťována doba přípravy na určitý test (v minutách) a počet dosažených bodů. Pomocí regresní přímky vyjádřete závislost počtu bodů na době přípravy studenta. Zhodnoťte kvalitu regresního modelu, vyjádřete intenzitu závislosti počtu bodů na době přípravy. Odhadněte střední hodnotu počtu bodů studenta, který se připravoval 182 minut.

Doba př.	160	160	162	163	161	170	172	177	179	178	182	184	183
Počet b.	57	55	59	60	52	67	69	74	75	76	78	80	87

i	x_i	y_i	$x_i y_i$	x_i^2
1	160	57	9 120	25 600
2	160	55	8 800	25 600
3	162	59	9 558	26 244
4	163	60	9 780	26 569
5	161	52	8 372	25 921
6	170	67	11 390	28 900
7	172	69	11 868	29 584
8	177	74	13 098	31 329
9	179	75	13 425	32 041
10	178	76	13 528	31 684
11	182	78	14 196	33 124
12	184	80	14 720	33 856
13	183	87	15 921	33 489
Σ	2 231	889	153 776	383 941

$$b_1 = \frac{13 \cdot 153776 - 2231 \cdot 889}{13 \cdot 383941 - 2231^2} = 1{,}13387$$

$$b_0 = \frac{889}{13} - 1{,}13387 \cdot \frac{2231}{13} = -126{,}204$$

$$Y = -126{,}204 + 1{,}13387 \cdot x$$

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
1	160	57	55,2152	3,1855	129,6095
2	160	55	55,2152	0,0463	179,1479
3	162	59	57,4829	2,3015	88,0710
4	163	60	58,6168	1,9132	70,3018
5	161	52	56,3491	18,9144	268,4556
6	170	67	66,5539	0,1990	1,9172
7	172	69	68,8216	0,0318	0,3787
8	177	74	74,4910	0,2411	31,5325
9	179	75	76,7587	3,0931	43,7633
10	178	76	75,6249	0,1407	57,9941
11	182	78	80,1603	4,6671	92,4556
12	184	80	82,4281	5,8956	134,9172
13	183	87	81,2942	32,5560	346,5325
Σ	2 231	889	-	73,1853	1 445,0769

$$Y = -126{,}204 + 1{,}13387 \cdot x$$

$$S_y = 1445{,}0769$$

$$S_{y,R} = 73{,}1853$$

$$S_{y,T} = 1445{,}0769 - 73{,}1853 = 1371{,}8994$$

$$R^2 = \frac{1371{,}8994}{1445{,}0769} = 1 - \frac{73{,}1853}{1445{,}0769} = 0{,}9494$$

$$r_{yx} = \sqrt{R^2} = \sqrt{0{,}9494} = 0{,}9743$$

ZÁKLADNÍ MÍRY DYNAMIKY ČASOVÝCH ŘAD

- první diference (absolutní přírůstek)

$$\Delta y_t = y_t - y_{t-1}$$

- průměrná první diference (průměrný absolutní přírůstek)

$$\overline{\Delta} = \frac{\sum_{t=2}^{n} \Delta y_t}{n-1} = \frac{(y_2 - y_1) + (y_3 - y_2) + \ldots + (y_n - y_{n-1})}{n-1} = \frac{y_n - y_1}{n-1}$$

- koeficient růstu

$$k_t = \frac{y_t}{y_{t-1}}$$

- průměrný koeficient růstu

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

- relativní přírůstek

$$\delta_t = k_t - 1$$

- průměrný relativní přírůstek

$$\bar{\delta} = \bar{k} - 1$$

- ***Příklad č.* 1**

Stanovte základní míry dynamiky sledované časové řady za celé sledované období, tj. absolutní, relativní, průměrný absolutní a průměrný relativní přírůstek.

t	y_t
2002	10
2003	12
2004	13
2005	9
2006	16

- **řešení:**

t	y_t	Δ	$\overline{\Delta}$	k_t	$\bar{k}$	δ_t	$\bar{\delta}$
2002	10	-	1,50	-	1,125	-	12,5%
2003	12	2		1,200		20,0%	
2004	13	1		1,083		8,3%	
2005	9	-4		0,692		-30,8%	
2006	16	7		1,778		77,8%	