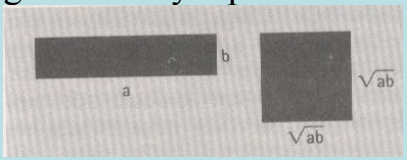
Číselné charakteristiky intervalových a poměrových znaků

Připomenutí: Intervalový znak umožňuje obsahovou interpretaci u operace rozdílu, poměrový znak i u operace podílu.

Charakteristika polohy: aritmetický průměr (arithmetic mean or mean) $m = \frac{1}{n}\sum_{i=1}^{n} x_i$.

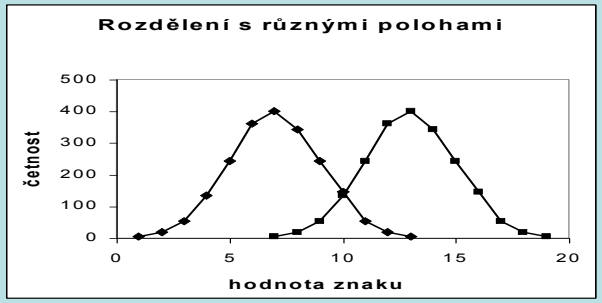
U poměrových znaků, které nabývají pouze kladných hodnot, lze použít geometrický průměr (geometric mean) $g = \sqrt{x_1 \cdot \ldots \cdot x_n}$. Vyskytuje se tam, kde má věcný význam součin hodnot znaku. Je zřejmé, že jde o aritmetický průměr logaritmů hodnot $x_1, \ldots, x_n$. Přitom geometrický průměr hodnot $x_1, \ldots, x_n$ je vždy menší nebo roven aritmetickému průměru těchto hodnot ($g \leq m$) a rovnosti je dosaženo právě tehdy, jsou-li všechny hodnoty znaku X stejné.

Příklad použití geometrického průměru: Máme-li obdélník a čtverec o stejných plochách, pak strana čtverce je geometrickým průměrem stran obdélníku.



Pomocí průměru zavedeme i-tou centrovanou hodnotu $x_i - m$ (podle znaménka poznáme, zda i-tá hodnota je podprůměrná či nadprůměrná).

Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem:



Příklad na výpočet geometrického průměru: Růst cen za měsíce červen, červenec a srpen roku 2010 byl postupně 1,2 %, 1,9 % a 1,9 %. Vypočtěte průměrný růst cen.

Řešení: $g = \sqrt{1{,}2 \cdot 1{,}9 \cdot 1{,}9} = {,}63$

Průměrný růst cen je přibližně 1,63 %. Znamená to, že výsledná cena by taková byla i v případě, že by růst cen byl konstantní, každý měsíc o 1,63 %.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech případech a jedné proměnné X. Do X zapíšeme hodnoty 1,2 1,9 1,9.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – Proměnné X – OK – Detailní výsledky – zaškrtneme Geom. Průměr a všechny ostatní volby odškrtneme – Výpočet.

Proměnná	Geometrický Průměr
X	1,630157

Vlastnosti aritmetického průměru

- Aritmetický průměr si lze představit jako těžiště dat – součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot – oba součty jsou v rovnováze.

- Průměr centrovaných hodnot je nulový, protože $\frac{1}{n}\sum_{i=} \quad - n = \frac{1}{n}\sum_{i=} \quad - \frac{1}{n}\sum_{i=} \quad = - \frac{1}{n} \cdot n \cdot m = 0$.

- Výraz $\sum_{i=1}^{n} x_i - a$ (tzv. kvadratická odchylka) nabývá svého minima pro a = m. Uvedený výraz charakterizuje celkovou chybu, které se dopustíme, když datový soubor nahradíme jedinou hodnotou a. Tato chyba je tedy nejmenší, když datový soubor nahradíme aritmetickým průměrem, přičemž za míru chyby považujeme kvadratickou odchylku.

- Pokud každou hodnotu x_i podrobíme lineární transformaci $y_i = a + bx_i$, pak průměr transformovaných hodnot je roven lineární transformaci původního průměru, tj. $m_2 = a + bm_1$.

- Mají-li znaky X, Y průměry m_1, m_2, pak znak Z = X + Y má průměr $m_1 + m_2$.

- Aritmetický průměr je silně ovlivněn extrémními hodnotami.

- Aritmetický průměr je vhodné použít, pokud je rozložení dat přibližně symetrické.

Příklad na vlastnosti aritmetického průměru:

U skupiny 20 pracovníků v určité dílně byly zjišťovány měsíční mzdy. Průměr mezd činil 15 500 Kč. Určete průměr mezd, jestliže mzdy všech pracovníků se zvýší

a) o 300 Kč, b) 1,1 krát, c) o 20%.

Řešení:

Označme m_1 průměr hodnot $x_1, \ldots, x_n$ a m_2 průměr hodnot $y_1, \ldots, y_n$, přičemž $y_i = a + bx_i$, $i = 1, \ldots, n$. Pak $m_2 = a + bm_1$.

ad a) $m_2 = 300 + m_1 = 15\,800$

Průměr se zvýšil o 300 Kč na 15 800 Kč.

ad b) $m_2 = 1{,}1.m_1 = 17\,050$

Průměr se zvýšil na 17 050 Kč.

ad c) $m_2 = 1{,}2.m_1 = 18\,600$

Průměr se zvýšil na 18 600 Kč.

Charakteristiky variability intervalových a poměrových znaků

Variační rozpětí (range) $R = x_{(n)} - x_{(1)}$ (nevýhoda – bere v úvahu pouze nejmenší a největší hodnotu datového souboru),

rozptyl (variance) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ (nevýhoda – vychází ve druhých mocninách jednotek, v nichž byl měřen znak X)

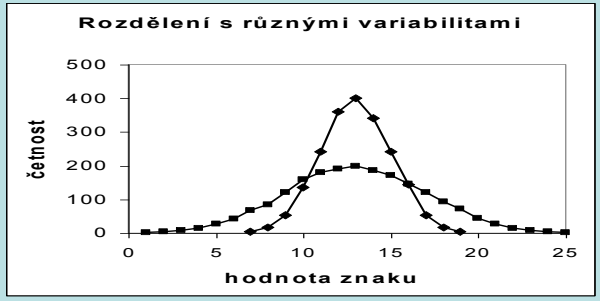
směrodatná odchylka (standard deviation) $s = \sqrt{s^2}$.

Pomocí směrodatné odchylky zavedeme i-tou standardizovanou hodnotu $\frac{x_i - m}{s}$ (vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru).

U poměrových znaků se jako charakteristika variability používá též:

koeficient variace (coefficient of variation) $cv = \frac{s}{m}$ (často se udává v procentech a udává, kolika procent průměru dosahuje směrodatná odchylka),

Znázornění rozložení četností dvou datových souborů, které se liší rozptylem:



Příklad na výpočet charakteristik variability:

Kurzy akcií společnosti AAA Auto Group v průběhu 23 dní v měsíci srpnu 2010 byly následující: 17,75; 17,74; 17,85; 17,59; 17,92; 17,98; 18,39; 18,25; 18,30; 18,00; 18,15; 18,15; 18,22; 18,40; 18,25; 17,95; 18,25; 18,23; 17,95; 17,90; 17,80; 17,87; 17,87. Vypočtěte charakteristiky variability.

Řešení:

Nejprve vypočítáme variační rozpětí: $R = x_{(n)} - x_{(1)} = 8,4 - 7,59 = 0,81$.

Před výpočtem dalších charakteristik variability musíme získat aritmetický průměr: $m = \frac{1}{23}(7,75 + 7,74 + \ldots + 7,87) = 8,033$.

Rozptyl: $s^2 = \frac{1}{n}\sum_{i=} \; - n^2 = \frac{1}{23}(7,75^2 + 7,74^2 + \ldots + 7,87^2) - 8,033^2 = 0,049$

Směrodatná odchylka: $s = \sqrt{s^2} = \sqrt{0,049} = 0,2213$

Koeficient variace: $\frac{s}{m}100\% = \frac{0,2213}{18,033}100\% = 2,23\%$

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné X a 23 případech. Do proměnné X zapíšeme zjištěné kurzy akcií. Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr, Rozptyl, Rozpětí – Výpočet.

Systém STATISTICA počítá rozptyl podle vzorce $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$, proto výsledek musíme vynásobit $\frac{n-}{n}$.

Ve výstupní tabulce přidáme za proměnnou Rozptyl tři nové proměnné nazvané rozptyl, směr. odch. a koef. variace. Do Dlouhého jména proměnné rozptyl napíšeme =v3*22/23, do Dlouhého jména proměnné směr. odch. napíšeme =sqrt(v4) a do Dlouhého jména proměnné koef. variace napíšeme =100*v5/v1.

Proměnná	Průměr	Rozpětí	Rozptyl	rozptyl =v3*22/2	směr. odch. =sqrt(v4)	koef. variace =100*v5/v1
x	18,03304	0,810000	0,051231	0,049004	0,221367976	1,22756858

Vlastnosti rozptylu:

- Rozptyl je nulový pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladný.

- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť $\frac{1}{n}\sum_{i=}$ [illegible] $= s^2$.

- Rozptyl standardizovaných hodnot je 1, protože $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s} - 0\right)^2 = \frac{1}{s^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{s^2}{s^2} = 1.$

- Rozptyl se zpravidla počítá podle vzorce $s^2 = \frac{1}{n}\sum_{i=}$ [illegible] $- m^2$.

- Pokud každou hodnotu x_i podrobíme lineární transformaci $y_i = a + bx_i$, pak rozptyl transformovaných hodnot je roven původnímu rozptylu vynásobenému b^2, tj. $s_2^2 = b^2 s_1^2$.

- Rozptyl je stejně jako průměr silně ovlivněn extrémními hodnotami.

- Rozptyl se nehodí jako charakteristika variability, je-li rozložení dat nesymetrické.

Příklad na využití vlastností rozptylu:

U skupiny 20 pracovníků v určité dílně byly zjišťovány měsíční mzdy. Směrodatná odchylka výše mezd činila 900 Kč.

Určete směrodatnou odchylku výše mezd, jestliže mzdy všech pracovníků se zvýší

a) o 300 Kč, b) 1,1 krát, c) o 20%.

Řešení:

Označme s_1 směrodatnou odchylku hodnot $x_1, \ldots, x_n$ a s_2 směrodatnou odchylku hodnot $y_1, \ldots, y_n$, přičemž $y_i = a + bx_i$, $i = 1, \ldots, n$. Pak $s_2 = bs_1$.

ad a) $s_2 = 1.s_1 = 900$

Směrodatná odchylka zůstala stejná.

ad b) $s_2 = 1{,}1s_1 = 1{,}1.900 = 990$

Směrodatná odchylka se zvýšila na 990 Kč.

ad c) $s_2 = 1{,}2s_1 = 1{,}2.900 = 1080$

Směrodatná odchylka se zvýšila na 1080 Kč.

Vážené číselné charakteristiky polohy a variability

Známe-li absolutní četnosti $n_1, \ldots, n_r$ či relativní četnosti $p_1, \ldots, p_r$ variant $x_{[1]}, \ldots, x_{[r]}$, můžeme spočítat

vážený průměr (weighted mean) $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]} = \sum_{j=1}^{r} p_j x_{[j]}$,

vážený rozptyl (weighted variance) $s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j (x_{[j]} - m)^2 = \sum_{j=1}^{r} p_j (x_{[j]} - m)^2$ (výpočetní vzorec:

$s^2 = \frac{1}{n}\sum_{j=} x_{[j]}^2 - n^2 = \sum_{j=} x_{[j]}^2 - n^2$)

Příklad na vážené číselné charakteristiky:
U 35 zaměstnanců byl zjištěn počet odpracovaných hodin za měsíc.

Počet odpracovaných hodin	184	185	186	187	188	189
Počet zaměstnanců	4	6	7	6	7	5

Vypočtěte průměr, směrodatnou odchylku a koeficient variace počtu odpracovaných hodin.

Řešení:
Hodnot je celkem 35, nikoliv 6 (častá chyba!)

Vážený průměr: $m = \frac{1}{n}\sum_{j=} x_{[j]} = \frac{1}{35}(4\cdot 184 + 6\cdot 185 + 7\cdot 186 + 6\cdot 187 + 7\cdot 188 + 5\cdot 189) = 186{,}6$

Vážený rozptyl: $s^2 = \frac{1}{n}\sum_{j=} x_{[j]}^2 - m^2 = \frac{1}{35}(4\cdot 184^2 + 6\cdot 185^2 + 7\cdot 186^2 + 6\cdot 187^2 + 7\cdot 188^2 + 5\cdot 189^2) - 186{,}6^2 = 2{,}5257$

Vážená směrodatná odchylka: $s = \sqrt{s^2} = \sqrt{2{,}5257} = 1{,}59\,\text{h} = 1\text{h } 35\text{ min}$

Koeficient variace: $\frac{s}{m}100\,\% = \frac{1{,}59}{186{,}6}100\,\% = 0{,}85\,\%$

Vidíme, že zaměstnanci odpracovali za měsíc v průměru 186,6 h, přičemž směrodatná odchylka dosahuje 0,85 % průměrné odpracované doby.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o šesti případech a dvou proměnných X a četnost. Zapíšeme zjištěné údaje

Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – klikneme na ikonu závaží – Proměnná vah četnost – OK – Stav Zapnuto – OK - Detailní výsledky – vybereme Průměr, Rozptyl – Výpočet.

Ve výstupní tabulce přidáme za proměnnou Rozptyl tři nové proměnné nazvané rozptyl, směr. odch. a koef. variace. Do Dlouhého jména proměnné rozptyl napíšeme =v*34/35, do Dlouhého jména proměnné směr. odch. napíšeme =sqrt(v3) a do Dlouhého jména proměnné koef. variace napíšeme =100*v4/v1.

Proměnná	Průměr	Rozptyl	rozptyl =v2*34/3	smer. odch. =sqrt(v3)	koef. variace =100*v4/v1
X	186,6000	2,600000	2,525714	1,5892496	0,851687888

Převod desetinných částí hodiny na minuty můžeme provést např. pomocí aplikace na adrese http://www.prevody-jednotek.cz/.

Počáteční a centrální momenty

Aritmetický průměr a rozptyl jsou speciální případy momentů. Zavedeme

k-tý počáteční moment $m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, $k = 1, 2, \ldots$,

k-tý centrální moment $m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k$, $k = 1, 2, \ldots$

Pomocí 3. a 4. centrálního momentu se definuje šikmost a špičatost.

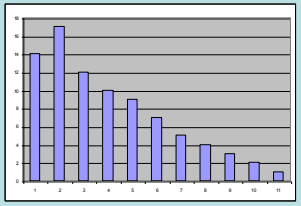
Šikmost (skewness): $\alpha_3 = \frac{m_3}{s^3}$ - měří nesouměrnost rozložení četností kolem průměru.

Je-li rozložení dat symetrické kolem aritmetického průměru (symmetrical distribution), pak $\alpha_3 = 0$.

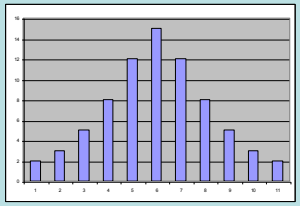
Má-li rozložení dat prodloužený pravý konec, jde o kladně zešikmené rozložení, $\alpha_3 > 0$.

Má-li rozložení dar prodloužený levý konec, jde o záporně zešikmené rozložení, $\alpha_3 < 0$.

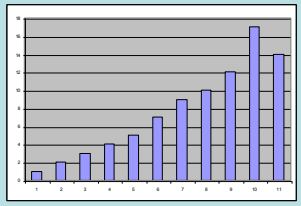
Příklad kladně sešikmeného rozložení



Příklad symetrického rozložení



Příklad záporně sešikmeného rozložení



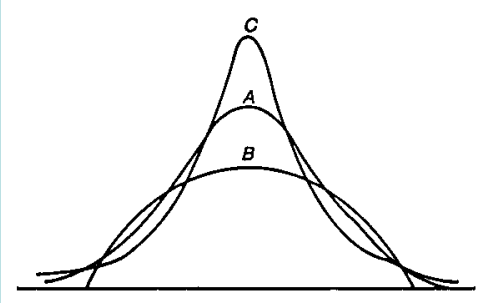
Špičatost (kurtosis): $\alpha = \frac{m_4}{s^4} - 3$ - měří koncentraci rozložení četností kolem průměru.

Je-li rozložení dat normální (mesokurtic), pak $\alpha_4 = 0$.

Je-li rozložení dat strmé (leptokurtic), pak $\alpha_4 > 0$.

Je-li rozložení dat ploché (platykurtic), pak $\alpha_4 < 0$.

Znázornění rozložení četností tří datových souborů, které se liší špičatostí



A … normální rozložení

B … ploché rozložení

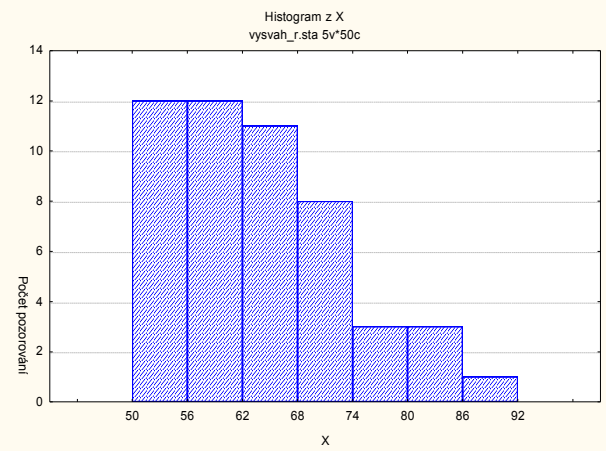
C … strmé rozložení

Příklad na výpočet šikmosti a špičatosti pomocí systému STATISTICA:
Datový soubor vysvah.sta obsahuje v proměnné X údaje o hmotnosti 50 náhodně vybraných studentů. Vypočtěte šikmost a špičatost znaku X.

Řešení:
Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme pouze Šikmost, Špičatost – Výpočet.

Proměnná	Šikmost	Špičatost
X	0,713596	-0,037538

Vidíme, že rozložení hmotností je kladně sešikmené a je poněkud plošší než normální rozložení. Asymetrie rozložení je patrná z histogramu:



Charakteristika společné variability dvou intervalových znaků: kovariance

Předpokládejme, že máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \cdots & \cdots \\ x_n & y_n \end{pmatrix}$. Označme m_1, m_2 průměry znaků X, Y a s_1, s_2 směrodatné odchylky znaků X, Y. Zavedeme kovarianci (covariance) jako charakteristiku společné variability znaků X, Y kolem jejich průměrů $s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2)$.

Kovariance je průměrem součinů centrovaných hodnot.

Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s nadprůměrnými (podprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot $x_i - m_1$ a $y_i - m_2$ vesměs kladné a jejich průměr (tj. kovariance) rovněž. Znamená to, že mezi znaky X, Y existuje určitý stupeň přímé lineární závislosti. Říkáme, že znaky X, Y jsou kladně korelované.

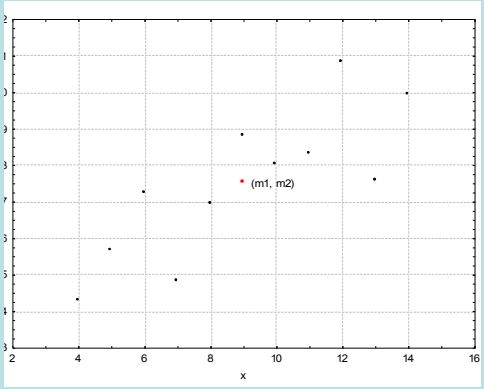
Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s podprůměrnými (nadprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot vesměs záporné a jejich průměr rovněž. Znamená to, že mezi znaky X a Y existuje určitý stupeň nepřímé lineární závislosti. Říkáme, že znaky X, Y jsou záporně korelované.

Je-li kovariance nulová, pak řekneme, že znaky X, Y jsou nekorelované a znamená to, že mezi nimi neexistuje žádná lineární závislost.

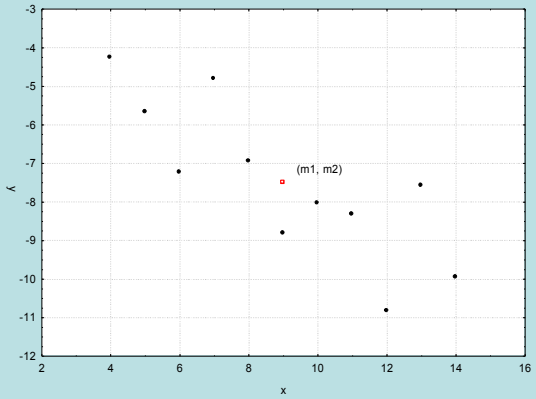
Pro výpočet kovariance používáme vzorec: $s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2$.

Znázornění významu kovariance

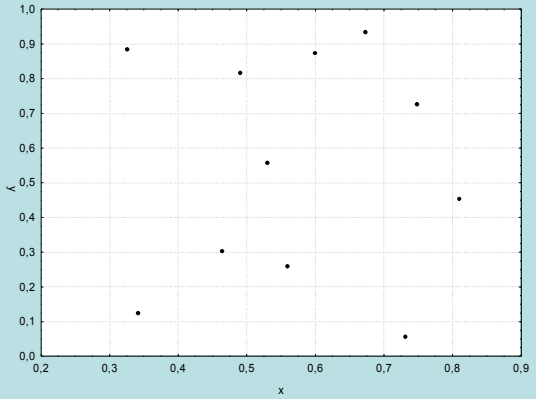
= 5,5



s_{12} = -5,5



s_{12} = 0



Vážená kovariance

Má-li znak X r variant $x_{[1]}, ..., x_{[r]}$ a znak Y s variant $y_{[1]}, ..., y_{[r]}$ a známe-li simultánní absolutní resp. relativní četnosti n_{jk} resp. p_{jk} dvojic variant $(x_{[j]}, y_{[k]})$, j = 1, …, r, k = 1, …, s, můžeme spočítat váženou kovarianci

$$s_{12} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk}\left(x_{[j]} - m_1\right)\left(y_{[k]} - m_2\right) = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} n_{jk}x_{[j]}y_{[k]} - m_1 m_2 = \sum_{j=1}^{r}\sum_{k=1}^{s} p_{jk}\left(x_{[j]} - m_1\right)\left(y_{[k]} - m_2\right) = \sum_{j=1}^{r}\sum_{k=1}^{s} p_{jk}x_{[j]}y_{[k]} - m_1 m_2$$

Vlastnosti kovariance

- Kovariance je nulová, právě když aspoň jeden ze znaků X, Y má všechny hodnoty stejné.

- Nechť m_1, m_2 jsou aritmetické průměry, s_1^2, s_2^2 rozptyly a s_{12} kovariance znaků X, Y. Pak znak $U = X + Y$ má aritmetický průměr $m_3 = m_1 + m_2$ a rozptyl $s_3^2 = s_1^2 + s_2^2 + s_{12}$.

- Nechť s_{12} je kovariance a m_1, m_2 jsou aritmetické průměry znaků X, Y. Pak znaky $U = a + bX, V = c + dY$ mají kovarianci $s_{34} = bds_{12}$

Příklad na vlastnosti kovariance: Hodnoty znaků X a Y mají postupně aritmetické průměry 3 a –1, rozptyly 2 a 3 a jejich kovariance je rovna 4. Vypočtěte aritmetický průměr a rozptyl hodnot znaku Z = X + Y.

Řešení:

$m_1 = 3, m_2 = -1, s_1^2 = 2, s_2^2 = 3, s_{12} = 4$

$m_3 = m_1 + m_2 = 3 - 1 = 2$

$s_3^2 = s_1^2 + s_2^2 + 2\, s_{12} = 2 + 3 + 2.4 = 13$

Příklad: Pro datový soubor obsahující údaje o příjmu manžela (znak X) a příjmu manželky (znak Y) vypočtěte kovarianci znaků X, Y.

příjem manžela	příjem manželky	příjem manžela	příjem manželky	příjem manžela	příjem manželky
16210	13710	31760	30250	24420	14640
30310	27960	38620	21980	15460	12800
33900	24930	27030	25410	37600	24200
40580	36720	43670	37540	42190	28650
19070	12940	45270	30580	15960	14500
29800	25810	39210	25470	18650	20210
26000	24590	14470	10550	26020	30150
37500	34810	23630	14820	23570	18840
21950	18860	15840	16340	20630	12760
19020	21530	25720	18700	31450	26840
17460	19870	17290	11560	19950	17960
13840	14320	18900	12080	16840	20900
29200	21200	47920	35620	16790	15740
14400	17300	29740	31420	26930	23980
15340	11930	13930	15790	46090	27960
23400	13220	25920	12870	22020	17400
18780	12760	21770	15980	31230	13580
33290	27140	17670	14320	20320	18490
31890	36970	19880	14800	19960	20500
18990	15470	14880	12680	36550	24360

Řešení:

$m_1 = 23485$, $m_2 = 20804{,}33$, $s_1 = 9398{,}96$, $s_2 = 7566{,}14$

$$s_{12} = \frac{1}{n}\sum_{i=} y_i - n_1 m_2 = \frac{1}{60}(6210 \cdot 13710 + 0310 \cdot 27960 + \ldots + 6550 \cdot 24360 - 23485 \cdot 20804{,}33 = 5773499$$

Výpočet pomocí systému STATISTICA:
Otevřeme datový soubor prijmy.sta.
Výpočet kovariance: Statistiky – Vícenásobná regrese - Proměnné Nezávislá X, Závislá Y – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky - Kovariance.

Proměnná	X	Y
X	88340482	56718812
Y	56718812	57246442

Vysvětlení: Na hlavní diagonále jsou rozptyly proměnných X, Y, mimo hlavní diagonálu je kovariance. STATISTICA však ve vzorci pro výpočet kovariance nepoužívá 1/n, ale 1/(n-1).
Získanou kovarianci přepočítáme: k výstupní tabulce přidáme novou proměnnou, kterou vložíme za proměnnou v2. Do jejího Dlouhého jména napíšeme =v2*59/60. Dostaneme tabulku:

Proměnná	X	Y	NProm =v2*59/60
X	88340482	56718812	55773498,9
Y	56718812	57246442	56292334,6

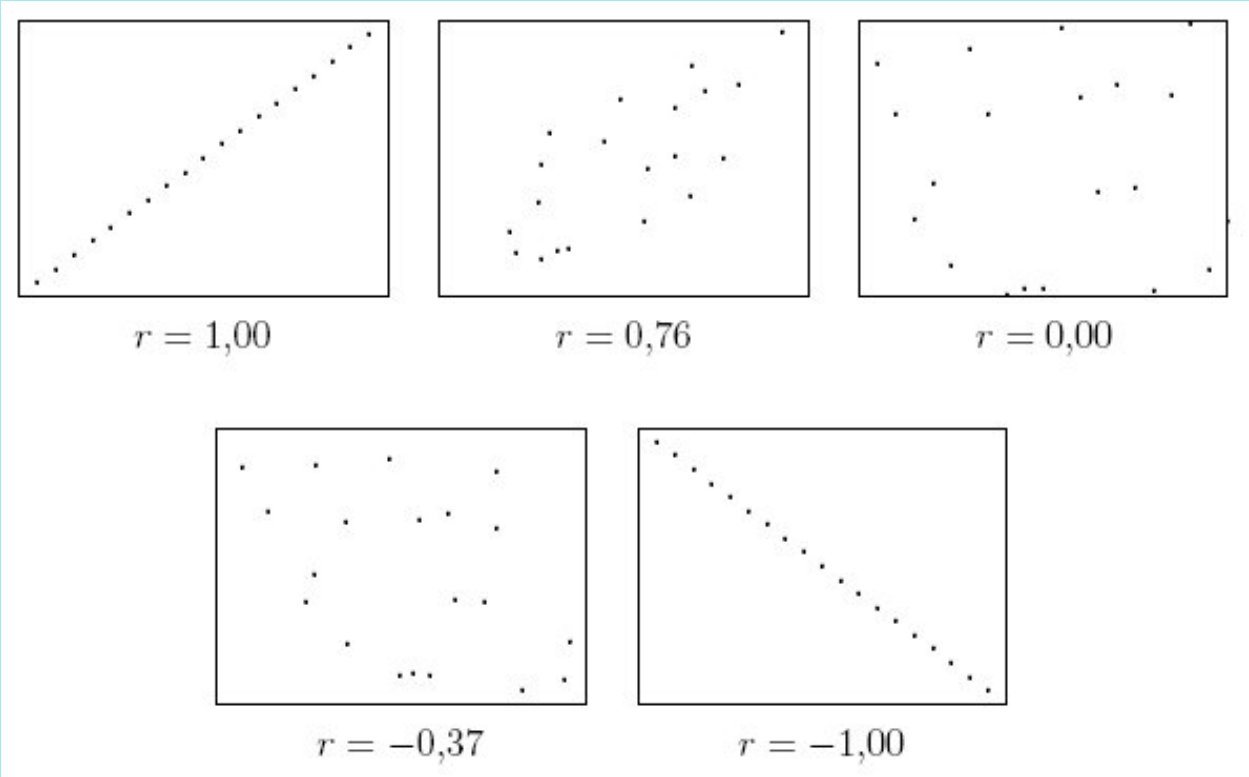
Na prvním řádku této nové proměnné najdeme kovarianci 55 773 499.

Charakteristika těsnosti závislosti dvou intervalových či poměrových znaků: Pearsonův koeficient korelace

Jsou-li směrodatné odchylky s_1, s_2 nenulové, pak definujeme Pearsonův koeficient korelace (Pearson correlation coefficient) znaků X, Y vzorcem: $r_{12} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - m_1}{s_1} \frac{y_i - m_2}{s_2}$. Je to průměr součinů standardizovaných hodnot. Počítá se podle vzorce

$$r_{12} = \frac{s_{12}}{s_1 s_2}.$$

Ilustrace různých hodnot koeficientu korelace



Příklad: Pro datový soubor obsahující údaje o příjmu manžela (znak X) a příjmu manželky (znak Y) vypočtěte koeficient korelace znaků X, Y. Přitom již víme, že $s_1 = 9\,398{,}96$, $s_2 = 7\,566{,}14$, $s_{12} = 55\,773\,499$.

Řešení:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{55773499}{9398{,}96 \cdot 7566} = 0{,}7976$$

Koeficient korelace svědčí o tom, že mezi oběma znaky existuje silná přímá lineární závislost – čím je vyšší příjem manžela, tím je vyšší příjem manželky a čím je nižší příjem manžela, tím je nižší příjem manželky.

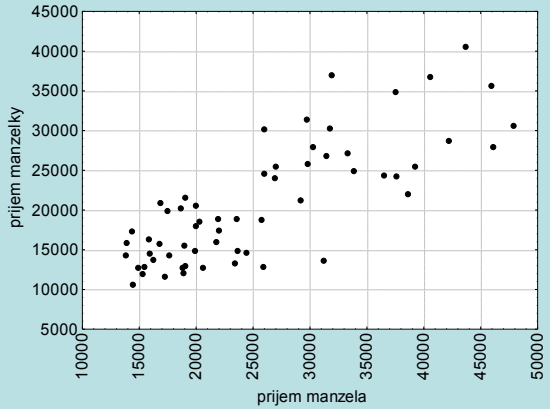
Výpočet pomocí systému STATISTICA:

Otevřeme datový soubor prijmy.sta.

Výpočet koeficientu korelace: Statistiky – Vícenásobná regrese - Proměnné Nezávislá X, Závislá Y – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Korelace

Proměnná	X	Y
X	1,000000	0,797578
Y	0,797578	1,000000

Dvourozměrný tečkový diagram



Vlastnosti Pearsonova koeficientu korelace:

Pro koeficient korelace platí $-1 \leq r_{12} \leq 1$ a rovnosti je dosaženo právě když mezi hodnotami $x_1, ..., x_n$ a $y_1, ..., y_n$ existuje úplná lineární závislost, tj. existují konstanty a, b tak, že $y_i = a + bx_i$, i = 1, ..., n, přičemž znaménko + platí pro $b > 0$, znaménko – pro $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)
Tedy čím je r_{12} bližší 1, tím je silnější přímá lineární závislost mezi znaky X a Y, čím je bližší –1, tím je silnější nepřímá lineární závislost mezi X a Y.

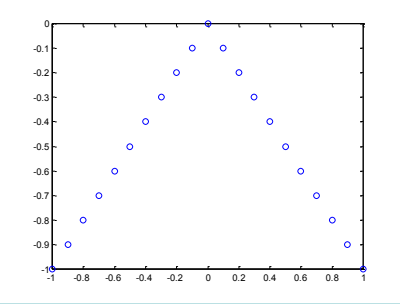
Je-li $r_{12} = 1$ resp. $r_{12} = -1$, pak dvojice (x_i, y_i) leží na nějaké rostoucí resp. klesající přímce.

Hodnoty r_{12} se nezmění, když u x-ových a y-ových hodnot současně provedeme vzestupnou resp. sestupnou lineární transformaci.

Hodnoty r_{12} se vynásobí -1, když u x-ových hodnot provedeme vzestupnou (resp. sestupnou) a u y-ových hodnot sestupnou (resp. vzestupnou) lineární transformaci.

Koeficient je symetrický, tj. $r_{12} = r_{21}$.

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu znaků X a Y. Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.



Vysvětlení významu Pearsonova korelačního koeficientu:

Máme 4 dvourozměrné datové soubory

X	Y_1
4	4,260
5	5,680
6	7,240
7	4,820
8	6,950
9	8,810
10	8,040
11	8,330
12	10,840
13	7,580
14	9,960

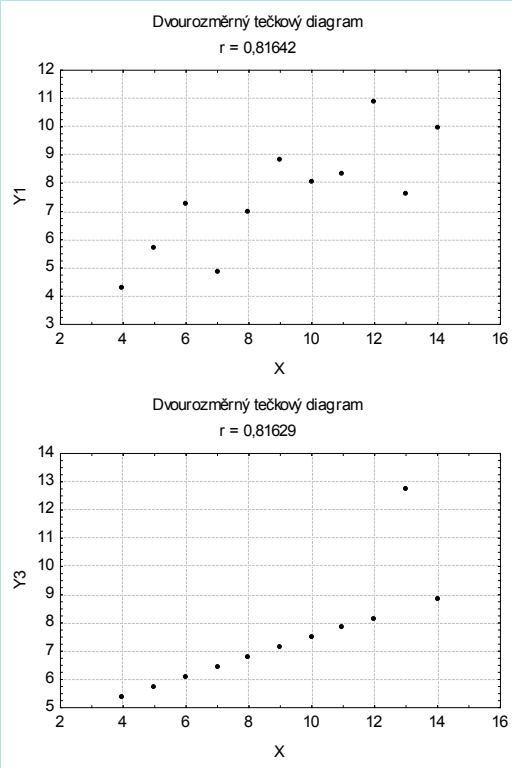
X	Y_2
4	3,100
5	4,740
6	6,130
7	7,260
8	8,140
9	8,770
10	9,140
11	9,260
12	9,130
13	8,740
14	8,100

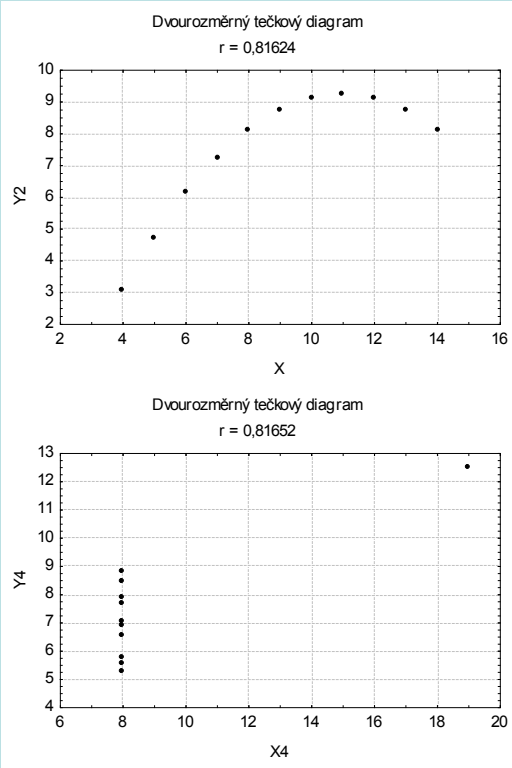
X	Y_3
4	5,390
5	5,730
6	6,080
7	6,420
8	6,770
9	7,110
10	7,460
11	7,810
12	8,150
13	12,740
14	8,840

X_4	Y_4
8	6,580
8	5,760
8	7,710
8	8,840
8	8,470
8	7,040
8	5,250
8	5,560
8	7,910
8	6,890
19	12,500

Pro každou z dvojic proměnných (X,Y_1), (X,Y_2), (X,Y_3), (X_4,Y_4) vypočtěte Pearsonův korelační koeficient a nakreslete dvourozměrný tečkový diagram. Pro které dvojice proměnných se hodí Pearsonův korelační koeficient jako vhodná míra těsnosti lineární závislosti?

Pro všechny dvojice proměnných vyjde korelační koeficient roven 0,816, zdálo by se tedy, že ve všech čtyřech případech existuje mezi proměnnými silná přímá lineární závislost. Oprávněnost této domněnky ověříme pomocí dvourozměrných tečkových diagramů.





Při pohledu na dvourozměrné tečkové diagramy je zřejmé, že pouze v prvním případě je použití Pearsonona korelačního koeficientu oprávněné.

Příklad na výpočet vážených číselných charakteristik

Z dvourozměrného datového souboru rozsahu 27, v němž znak X má varianty 1, 2, 3 a znak Y má rovněž varianty 1, 2, 3, byly určeny simultánní absolutní četnosti: $n_{11} = 5$, $n_{12} = 1$, $n_{13} = 3$, $n_{21} = 4$, $n_{22} = 3$, $n_{23} = 4$, $n_{31} = 2$, $n_{32} = 3$, $n_{33} = 2$.

Vypočtěte a interpretujte koeficient korelace znaků X a Y.

Řešení:

Kontingenční tabulka simultánních absolutních četností:

x	y			$n_{j.}$
	1	2	3	
1	5	1	3	9
2	4	3	4	11
3	2	3	2	7
$n_{.k}$	11	7	9	27

Nejprve vypočteme vážené průměry: $m_1 = \frac{1}{27}(\cdot 9 + \cdot 11 + \cdot 7) = \frac{52}{27} = ,926$, $m_2 = \frac{1}{27}(\cdot 11 + \cdot 7 + \cdot 9) = \frac{52}{27} = ,926$.

Dále spočítáme vážené rozptyly:

$s_1^2 = \frac{1}{27}(^2 \cdot 9 + ^2 11 + ^2 \cdot 7) - \left(\frac{52}{27}\right)^2 = \frac{116}{27} - \frac{2704}{729} = \frac{428}{729}$, $s_1 = 0{,}766$

$s_2^2 = \frac{1}{27}(^2 \cdot 11 + ^2 \cdot 7 + ^2 \cdot 9) - \left(\frac{52}{27}\right)^2 = \frac{120}{27} - \frac{2704}{729} = \frac{536}{729}$, $s_2 = 0{,}857$

Následuje výpočet vážené kovariance:

$s_{12} = \frac{1}{27}(\cdot 1 \cdot 5 + \cdot 2 \cdot 1 + \cdot 3 \cdot 3 + \cdot 1 \cdot 4 + \cdot 2 \cdot 3 + \cdot 3 \cdot 4 + \cdot 1 \cdot 2 + \cdot 2 \cdot 3 + \cdot 3 \cdot 2) - \frac{52}{27} \cdot \frac{52}{27}$

$= \frac{102}{27} - \frac{2704}{729} = \frac{2754 - 2704}{729} = \frac{50}{729} = ,0685871$

Dosadíme do vzorce pro výpočet koeficientu korelace: $r_{12} = \frac{\frac{50}{729}}{\sqrt{\frac{428}{729} \cdot \frac{536}{729}}} = ,10439$.

Mezi znaky X a Y existuje velmi slabá přímá lineární závislost.

Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 9 případech a 3 proměnných X, Y, četnost. Do proměnné X napíšeme 1 1 1 2 2 2 3 3 3, do proměnné Y napíšeme 1 2 3 1 2 3 1 2 3 a do proměnné četnost napíšeme 5 1 3 4 3 4 2 3 2.

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 2 seznamy – 1. seznam X, 2. seznam Y – OK- klikneme na ikonu závaží – zaškrtneme Stav zapnuto – Proměnná vah četnost - OK - OK – Výpočet

Ve výstupní tabulce zvětšíme počet desetinných míst.

Proměnná	Y
X	0,1044