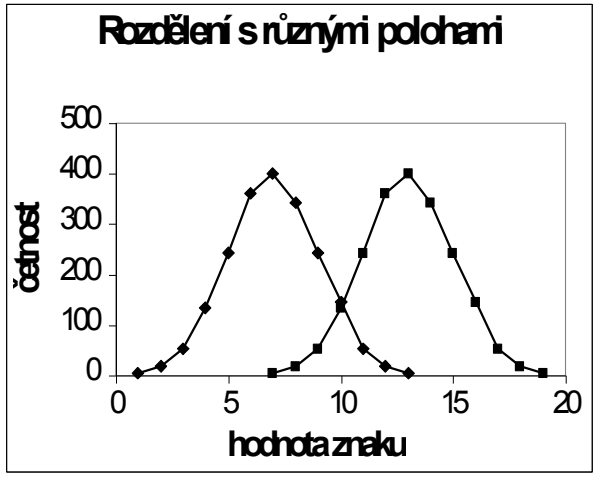
Číselné charakteristiky intervalových znaků

Charakteristika polohy: aritmetický průměr je součet hodnot dělený jejich počtem: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$. Pomocí průměru zavedeme i-tou centrovanou hodnotu $x_i - m$ (podle znaménka poznáme, zda i-tá hodnota je podprůměrná či nadprůměrná).

Příklad: (na výpočet aritmetického průměru)
Je dán datový soubor (2 8 9 10 1 0 5). Vypočtěte jeho průměr.
Řešení:

$$m = \frac{2 + 8 + 9 + 10 + 1 + 0 + 5}{7} = \frac{35}{7} = 5$$

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



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

- Výraz $\sum_{i=1}^{n}(x_i - a)^2$ (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.

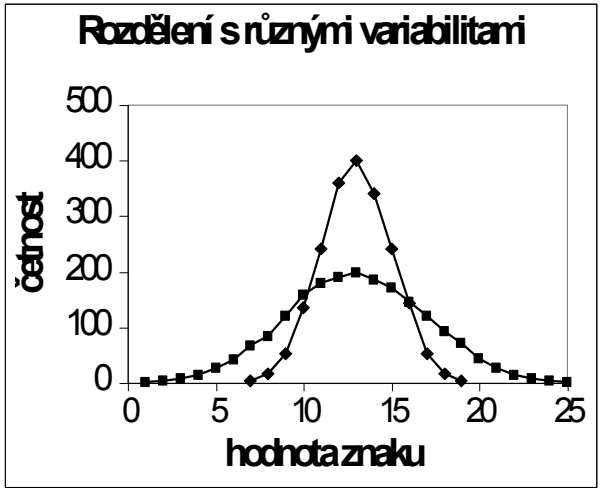
- Aritmetický průměr je silně ovlivněn extrémními hodnotami.

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

Charakteristika variability: rozptyl je průměrná kvadratická odchylka hodnot od jejich aritmetického průměru $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$. Kladná odmocnina z rozptylu se nazývá směrodatná odchylka $s = \sqrt{s^2}$. 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).

Výpočetní tvar vzorce pro rozptyl: $s^2 = \frac{1}{n}\sum_{i=1}^{n}$.

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



Příklad na výpočet rozptylu a směrodatné odchylky:
Jsou dány dva datové soubory, a to (7 8 9) a (1 10 13). V obou případech vypočtěte rozptyl a směrodatnou odchylku.

Řešení:
Pro první datový soubor je průměr $m_1 = 8$, pro druhý datový soubor je průměr m_2 také 8.
Výpočet pomocí definičního vzorce:

$$s_1^2 = \frac{(7-8)^2 + (8-8)^2 + (9-8)^2}{3} = \frac{1+0+1}{3} = \frac{2}{3}$$

$$s_2^2 = \frac{(1-8)^2 + (10-8)^2 + (13-8)^2}{3} = \frac{49+4+25}{3} = \frac{78}{3} = 26$$

Výpočet pomocí výpočetního vzorce:

$$s_1^2 = \frac{7^2 + 8^2 + 9^2}{3} - 8^2 = \frac{194}{3} - 64 = \frac{194-192}{3} = \frac{2}{3}$$

$$s_2^2 = \frac{1^2 + 10^2 + 13^2}{3} - 8^2 = \frac{270}{3} - 64 = 90 - 64 = 26$$

$$s_1 = \sqrt{\frac{2}{3}} = 0{,}82,\ s_2 = \sqrt{26} = 5{,}1$$

Interpretace směrodatné odchylky pro první soubor: většina čísel se odchyluje od průměru 8 o méně než 1 v obou směrech, většina čísel leží tedy mezi 7 a 9.
Interpretace směrodatné odchylky pro druhý soubor: většina čísel se odchyluje od průměru 8 o méně než 5 v obou směrech, většina čísel leží tedy mezi 3 a 13.

Vlastnosti rozptylu a směrodatné odchylky:

- Směrodatná odchylka je nulová pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladná.

- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - 0)^2 =$

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

- Směrodatná odchylka je stejně jako průměr silně ovlivněna extrémními hodnotami.

- Směrodatná odchylka se nehodí jako charakteristika variability, je-li rozložení dat zešikmené.

Charakteristika nesymetrie dat: šikmost $\alpha_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{s^2}^3}$

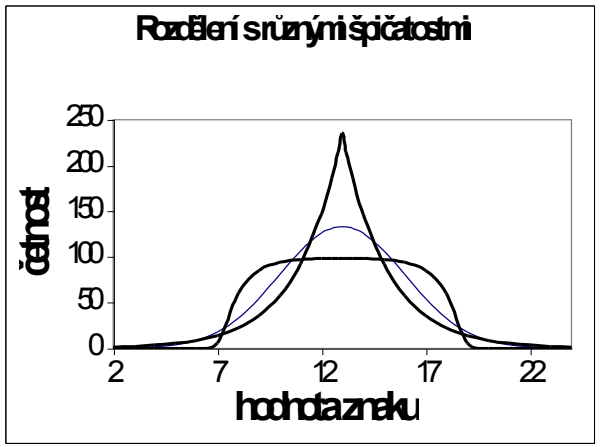
Je-li rozložení dat symetrické kolem aritmetického průměru, pak $\alpha_3 = 0$.
Má-li rozložení dat prodloužený pravý konec, jde o kladně zešikmené rozložení, $\alpha_3 > 0$.
Má-li rozložení dar prodloužený levý konec, jde o záporně zešikmené rozložení, $\alpha_3 < 0$.

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

Charakteristika koncentrace dat kolem průměru: špičatost $\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sqrt{s}}\right)^4 - 3$

Je-li rozložení dat normální (Gaussovo), pak $\alpha_4 = 0$.
Je-li rozložení dat strmé, pak $\alpha_4 > 0$.
Je-li rozložení dat ploché, pak $\alpha_4 < 0$.

Znázornění rozložení četností dvou datových souborů, které se liší špičatostí



Příklad na ilustraci významu špičatosti

Tři skupiny studentů o počtech 149, 69 a 11 odpovídaly při testu na 10 otázek. Znak X je počet správně zodpovězených otázek. Známe absolutní četnosti znaku X ve všech třech skupinách.

č. sk.	X										
	0	1	2	3	4	5	6	7	8	9	10
1	2	5	15	20	25	15	25	20	15	5	2
2	4	3	2	1	0	49	0	1	2	3	4
3	1	0	0	0	0	9	0	0	0	0	1

Vypočtěte průměr, rozptyl, šikmost a špičatost počtu správně zodpovězených otázek ve všech třech skupinách.
Nakreslete sloupkové diagramy absolutních četností.

Řešení:

1. skupina

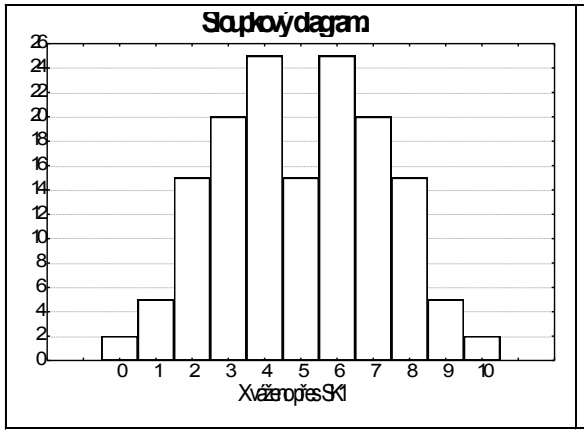
Variable	m	s²	alfa	alfa
X	5	5	0	-0,7

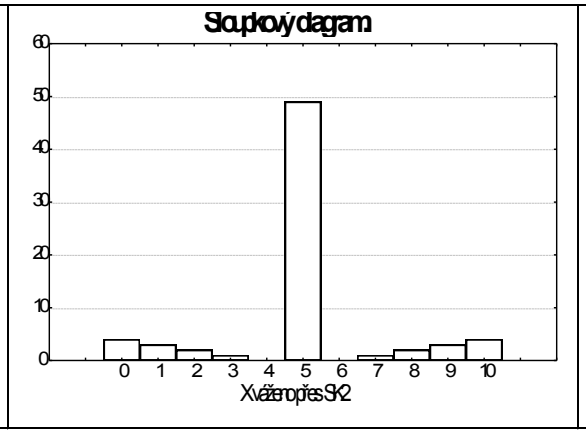
2. skupina

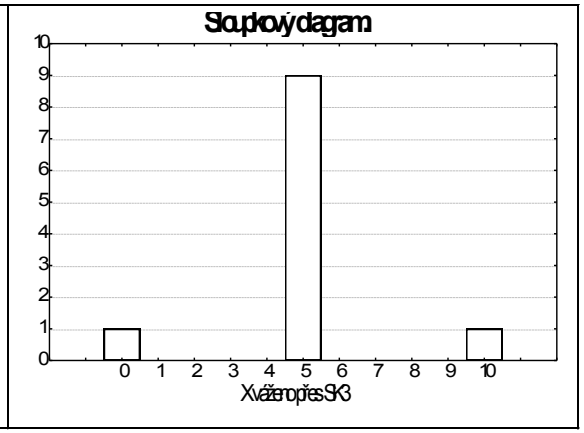
Variable	m	s²	alfa	alfa
X	5	5	0	1,2

3. skupina

Variable	m	s²	alfa	alfa
X	5	5	0	5,0







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 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.

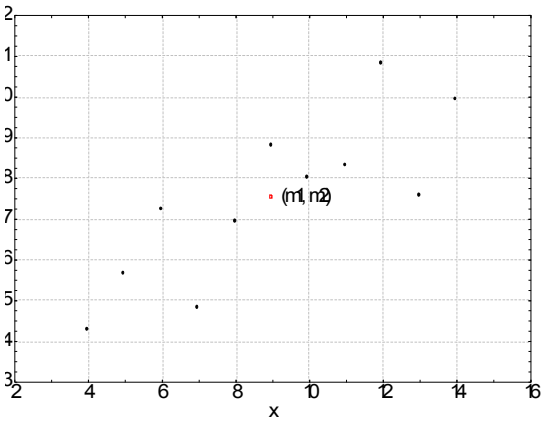
Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s podprůměrnými (nadprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot vesměs záporné a jejich průměr rovněž. Znamená to, že mezi znaky X a Y existuje určitý stupeň nepřímé lineární závislosti. Říkáme, že znaky X, Y jsou záporně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

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

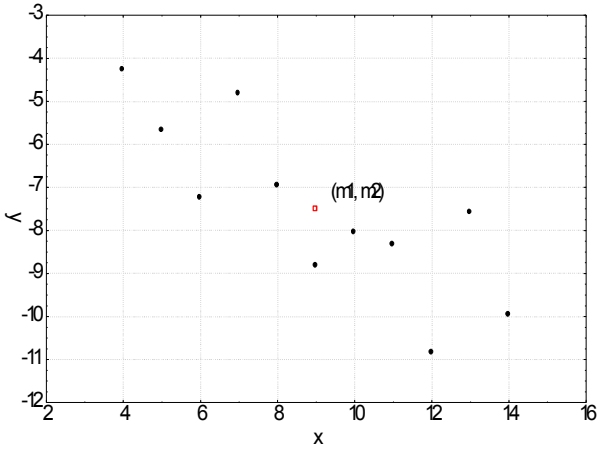
Pro výpočet kovariance používáme vzorec: $s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2$.

Znázornění významu kovariance

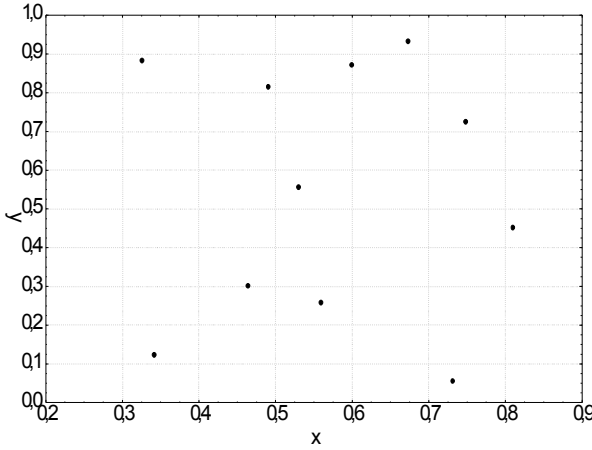
$s_{12} = 5{,}5$



$s_{12} = -5{,}5$



$s_{12} = 0$



Charakteristika těsnosti závislosti dvou intervalových znaků: Pearsonův koeficient korelace

Jsou-li směrodatné odchylky s_1, s_2 nenulové, pak definujeme Pearsonův koeficient korelace 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 $r_{12} = \frac{s_{12}}{s_1 s_2}$.

Vlastnosti Pearsonova koeficientu korelace:

Koeficient nabývá hodnot mezi –1 a 1. Čím je 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.

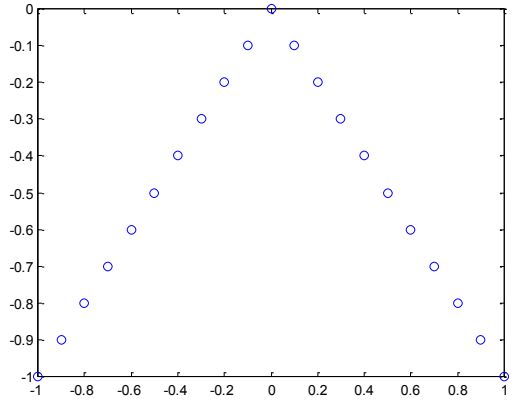
Je-li $r_{12} = 1$ resp. $r_{12} = -1$, pak dvojice (x_i, y_i) leží na nějaké rostoucí resp. klesající přímce.

Hodnoty r_{12} se nezmění, když provedeme vzestupnou lineární transformaci původních dat.

Hodnoty r_{12} se vynásobí -1, když provedeme sestupnou lineární transformaci původních dat.

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.



Příklad na výpočet Pearsonova koeficientu korelace
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Vypočtěte a interpretujte koeficient korelace. Pro usnadnění výpočtů máte k dispozici tyto součty:

$$\sum_{i=1}^{8} x_i = 450, \quad \sum_{i=1}^{8} y_i = 400, \quad \sum_{i=1}^{8} x_i^2 = 26668, \quad \sum_{i=1}^{8} y_i^2 = 20816, \quad \sum_{i=1}^{8} x_i y_i = 23232$$

Řešení:
Vypočteme aritmetické průměry a rozptyly:

$$m_1 = \frac{450}{8} = 56{,}25 \qquad m_2 = \frac{400}{8} = 50$$

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_1^2 = \frac{26668}{8} - 56{,}25^2 = 3334{,}5 - 3163{,}0625 = 171{,}4375 \qquad s_1 = 13{,}093$$

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - m_2^2 = \frac{20816}{8} - 50^2 = 2602 - 2500 = 102 \qquad s_2 = 10{,}0995$$

Dále vypočteme kovarianci:

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2 = \frac{23232}{8} - 56{,}25 \cdot 50 = 2904 - 2812{,}5 = 91{,}5$$

Dosadíme do vzorce pro výpočet koeficientu korelace:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{91{,}5}{13{,}093 \cdot 10{,}0995} = 0{,}6926$$

Lze tedy soudit, že mezi výsledky obou testů existuje středně silná přímá lineární závislost.

Vážené číselné charakteristiky

Pokud nemáme k dispozici původní datový soubor, ale jenom tabulku rozložení četností (resp. kontingenční tabulku), můžeme vypočítat tzv. vážené číselné charakteristiky.

Vážený aritmetický průměr: $\bar{x} = \frac{1}{n} \sum_{j=1}^{r} n_j x_{[j]}$

Vážený rozptyl: $s^2 = \frac{1}{n} \sum_{j=1}^{r} n_j (x_{[j]} - \bar{x})^2 = \frac{1}{n} \sum_{j=1}^{r} n_j x_{[j]}^2 - \bar{x}^2$

Vážená kovariance: $s_{12} = \frac{1}{n} \sum_{j=1}^{r} \sum_{k=1}^{s} n_{jk} (x_{[j]} - \bar{x}_1)(y_{[k]} - \bar{x}_2) = \frac{1}{n} \sum_{j=1}^{r} \sum_{k=1}^{s} n_{jk} x_{[j]} y_{[k]} - \bar{x}_1 \bar{x}_2$

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: $n_{11} = 5$, $n_{12} = 1$, $n_{13} = 3$, $n_{21} = 4$, $n_{22} = 3$, $n_{23} = 4$, $n_{31} = 2$, $n_{32} = 3$, $n_{33} = 2$.

a) Vypočtěte průměry a směrodatné odchylky znaků X a Y.

b) 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

ad a) $m_1 = \frac{1}{7}(\cdot + \cdot + \cdot) = \frac{\cdot}{7} = \cdot,\ m_2 = \frac{1}{7}(\cdot + \cdot + \cdot) = \frac{\cdot}{7} = \cdot$

$s_1^2 = \frac{1}{7}(\cdot^2 \cdot + \cdot^2 + \cdot^2 \cdot) - (\cdot)^2 = \frac{\cdot}{7} - \cdot = \cdot$, $s_1 = 0{,}766$

$s_2^2 = \frac{1}{7}(\cdot^2 \cdot + \cdot^2 + \cdot^2 \cdot) - (\cdot)^2 = \frac{\cdot}{7} - \cdot = \cdot$, $s_2 = 0{,}857$

ad b)

$s_{12} = \frac{1}{7}(\cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot + \cdot\cdot) - \cdot\cdot = \frac{\cdot}{7} - \cdot = \frac{\cdot}{7} = \cdot$

$r_{12} = \frac{\frac{50}{729}}{\sqrt{72 \cdot 729}} = \cdot$

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

Pro poměrové znaky používáme jako charakteristiku variability koeficient variace $\frac{s}{\bar{x}}$. Je to bezrozměrné číslo, které se často vyjadřuje v procentech. Umožňuje porovnat variabilitu několika znaků.
Jsou-li všechny hodnoty poměrového znaku kladné, pak jako charakteristiku polohy lze užít geometrický průměr $\sqrt[n]{x_1 \times \cdots \times}$.