Číselné charakteristiky znaků

Doposud jsme se zabývali funkcionálními charakteristikami znaků, jako jsou $p(x,y)$, $p_1(x)$, $p_2(y)$, $F(x)$, $f(x,y)$, $f_1(x)$, $f_2(y)$, které nesou úplnou informaci o rozložení četností. Nyní zavedeme číselné charakteristiky, které nás informují o některých rysech tohoto rozložení četností: o poloze (úrovni) hodnot znaku, o jejich variabilitě (rozptýlení), o těsnosti závislosti dvou znaků a pod. Pro různé typy znaků se používají různé číselné charakteristiky, proto se nejdřív seznámíme s jednotlivými typy znaků.

Typy znaků (třídění podle stupně kvantifikace)

Nominální znak: připouští obsahovou interpretaci pouze u relace rovnosti =. O dvou variantách nominálního znaku lze pouze konstatovat, že jsou buď stejné nebo různé. Čísla, která přiřadíme jednotlivým variantám znaku, nereprezentují skutečnou hodnotu použitých čísel, ale jsou pouhým označením variant znaku.
Příklady nominálních znaků: lékařská diagnóza, typ profese, barva očí, rodinný stav, národnost, …

Ordinální znak: připouští obsahovou interpretaci nejen u relace rovnosti =, ale též u relace uspořádání <. Můžeme tedy konstatovat, že varianta $x_{[j]}$ je větší (dokonalejší, silnější, vhodnější) než varianta $x_{[k]}$.
Příklad ordinálního znaku: školní klasifikace vyjadřuje menší nebo větší znalosti zkoušených žáků – jedničkář je lepší než dvojkař, ale intervaly mezi známkami nemají obsahovou interpretaci. Nelze tvrdit, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem.
Další příklady: Různá bodování ve sportovních a uměleckých soutěžích, posuzování různých rysů sociálního chování, posuzování stavu pacientů, hodnocení postojů respondentů k různým otázkám, …

Intervalový znak: kromě relací rovnosti = a uspořádání < umožňuje obsahovou interpretaci také u operace rozdílu -, tj. stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v extenzitě zkoumané vlastnosti.
Příklad intervalového znaku: teplota měřená ve stupních Celsia. Např. naměříme-li ve čtyřech po sobě jdoucích dnech polední teploty 0, 2, 4, 6 ºC, znamená to, že každým dnem stouply teploty o 2 ºC. Nelze však říci, že z druhého na třetí den vzrostla teplota dvojnásobně, kdežto ze třetího na čtvrtý den pouze jeden a půl krát.
Další příklady: kalendářní systémy, směr větru, inteligenční kvocient, …
Společný znak intervalových znaků: nula byla stanovena uměle, pouhou konvencí.

Poměrový znak: kromě relací rovnosti = a uspořádání < umožňuje obsahovou interpretaci také u operací rozdílu - a podílu /, tj. stejný poměr mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný podíl v extenzitě zkoumané vlastnosti.
Příklad poměrového znaku: délka předmětu měřená v cm. Má-li jeden předmět délku 8 cm a druhý 16 cm, má smysl prohlásit, že druhý předmět je dvakrát delší než první předmět.
Další příklady: počet dětí v rodině, výška kapesného v Kč, hmotnost osoby, …
Společný znak poměrových znaků: Poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.

Mimo uvedenou klasifikaci stojí alternativní znaky, které nabývají jen dvou hodnot, např. 0,1, což znamená absenci a prezenci nějakého jevu. Například 0 bude znamenat neúspěch, l úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

Číselné charakteristiky nominálních znaků

Charakteristika polohy: modus – nejčetnější varianta resp. střed nejčetnějšího třídicího intervalu.

Charakteristika variability: mutabilita $M = \frac{n^2 - \sum_{j=1}^{r} n_j^2}{n(n-1)}$, nabývá hodnot z intervalu [0, 1].

Jsou–li všechny hodnoty znaku stejné, pak M = 0. Jsou-li všechny hodnoty znaku navzájem různé, pak M = 1.

Příklad na stanovení modu a výpočet mutability:
20 náhodně vybraných osob mělo odpovědět na otázku, který z pěti výrobků (označíme je A, B, C, D, E) preferují. Výsledky máme v tabulce:

Výrobek	A	B	C	D	E
Četnost odpovědí	3	5	3	6	3

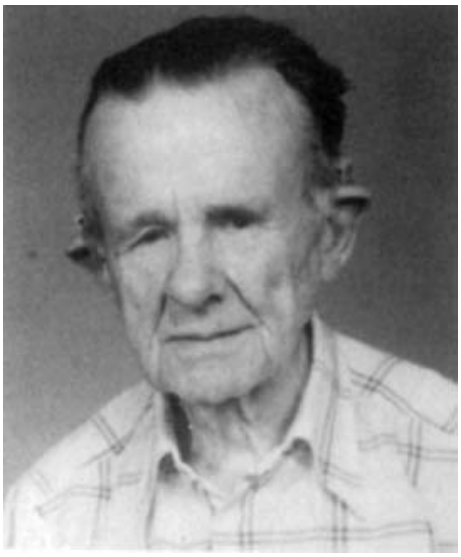
Stanovte modus a vypočtěte mutabilitu.

Řešení:
Modus = D

Mutabilita: $M = \frac{n^2 - \sum_{j=1}^{r} n_j^2}{n(n-1)} = \frac{20^2 - \left(3^2 + 5^2 + 3^2 + 6^2 + 3^2\right)}{20 \cdot 19} = 0{,}821$

Vidíme, že daný datový soubor vykazuje dosti vysokou míru proměnlivosti.

Charakteristika těsnosti závislosti dvou nominálních znaků: Cramérův koeficient kontingence.



Carl Harald Cramér (1893 – 1985): Švédský matematik

Nechť znak X nabývá variant $x_{[1]}, ..., x_{[r]}$ a znak Y nabývá variant $y_{[1]}, ..., y_{[s]}$.

Máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$. Zjistíme absolutní četnosti n_{jk} dvojice variant $(x_{[j]},y_{[k]})$, j = 1, …, r, k = 1, …, s a uspořádáme je do kontingenční tabulky:

x \ y	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	n_{11}	...	n_{1s}	$n_{1.}$
⋮	...	...	...	...
$x_{[r]}$	n_{r1}	...	n_{rs}	$n_{r.}$
$n_{.k}$	$n_{.1}$	...	$n_{.s}$	n

Vypočteme tzv. teoretické četnosti $\frac{n_{j.}n_{.k}}{n}$ a s jejich pomocí pak statistiku

$K = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$. Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r,s\}$.

Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je 1, tím je těsnější závislost mezi X a Y, čím blíže je 0, tím je tato závislost volnější.

Význam hodnot Cramérova koeficientu:
mezi 0 až 0,1 … zanedbatelná závislost,
mezi 0,1 až 0,3 … slabá závislost,
mezi 0,3 až 0,7 … střední závislost,
mezi 0,7 až 1 … silná závislost.

Příklad na výpočet Cramérova koeficientu:
686 náhodně vybraných osob bylo dotázáno, zda vlastní auto (znak X, varianty 1 – ano, 2 – ne) a zda jsou ochotny používat MHD (znak Y, varianty 1 – ano, 2 – ne). Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y		$n_{j.}$
	ano	ne	
ano	56	312	368
ne	283	35	318
$n_{.k}$	339	347	686

Vypočtěte a interpretujte Cramérův koeficient.

Řešení: Nejprve vypočteme teoretické četnosti:

$$\frac{n_{1.}n_{.1}}{n}=\frac{368\cdot 339}{686}=181{,}8542,\ \frac{n_{1.}n_{.2}}{n}=\frac{368\cdot 347}{686}=186{,}1458,$$

$$\frac{n_{2.}n_{.1}}{n}=\frac{318\cdot 339}{686}=157{,}1458,\ \frac{n_{2.}n_{.2}}{n}=\frac{318\cdot 347}{686}=160{,}8542$$

Nyní dosadíme do vzorce pro výpočet statistiky K:

$$K=\frac{(56-181{,}8542)^2}{181{,}8542}+\frac{(312-186{,}1458)^2}{1861485}+\frac{(283-157{,}1458)^2}{157{,}1458}+\frac{(35-160{,}8542)^2}{160{,}8542}=371{,}456$$

Nakonec vypočteme Cramérův koeficient:

$$V=\sqrt{\frac{371{,}456}{686\cdot 1}}=0{,}7358$$

Hodnota Cramérova koeficientu svědčí o tom, že mezi znaky X a Y existuje silná závislost.

Číselné charakteristiky ordinálních znaků

Charakteristika polohy: α-kvantil. Je-li $\alpha\in(0;1)$, pak α-kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1 – α všech dat. Pro výpočet α-kvantilu slouží algoritmus:

$$n\alpha=\begin{cases}\text{celé číslo } c\Rightarrow x_\alpha=\dfrac{x_{(c)}+x_{(c+1)}}{2}\\ \text{necelé číslo}\Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c\Rightarrow x_\alpha=x_{(c)}\end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – medián, $x_{0,25}$ – dolní kvartil, $x_{0,75}$ – horní kvartil, $x_{0,1}$, ..., $x_{0,9}$ – decily, $x_{0,01}$, ..., $x_{0,99}$ – percentily.

Charakteristika variability: kvartilová odchylka: $q = x_{0,75} - x_{0,25}$.

Příklad na výpočet kvantilů:
U 50 žáků 7. ročníku jedné základní školy byly na pololetním vysvědčení zjištěny známky z matematiky:

známka	1	2	3	4	5
četnost známky	9	15	20	4	2

Určete medián, 1. a 9. decil a kvartilovou odchylku.

Řešení:
Pro snadnější výpočet tabulku doplníme ještě o absolutní kumulativní četnosti:

známka	1	2	3	4	5
n_j	9	15	20	4	2
N_j	9	24	44	48	50

Rozsah souboru n = 50

α	nα	c	x_α
0,50	50.0,5=25	25	$\frac{x_{(25)}+x_{(26)}}{2}=\frac{3+3}{2}=3$
0,10	50.0,1 = 5	5	$\frac{x_{(5)}+x_{(6)}}{2}=\frac{1+1}{2}=1$
0,90	50.0,9 = 45	45	$\frac{x_{(45)}+x_{(46)}}{2}=\frac{4+4}{2}=4$
0,25	50.0,25 = 12,5	13	$x_{(13)}=2$
0,75	50.0,75 = 37,5	38	$x_{(38)}=3$

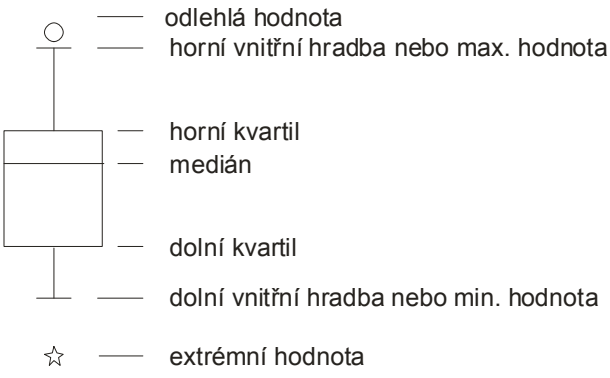
Kvartilová odchylka: q = 3 – 2 = 1.

Interpretace např. dolního kvartilu: V souboru 50 žáků je aspoň čtvrtina takových, kteří mají z matematiky jedničku nebo dvojku.

Grafické znázornění ordinálních dat pomocí krabicového diagramu

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot.

Způsob konstrukce

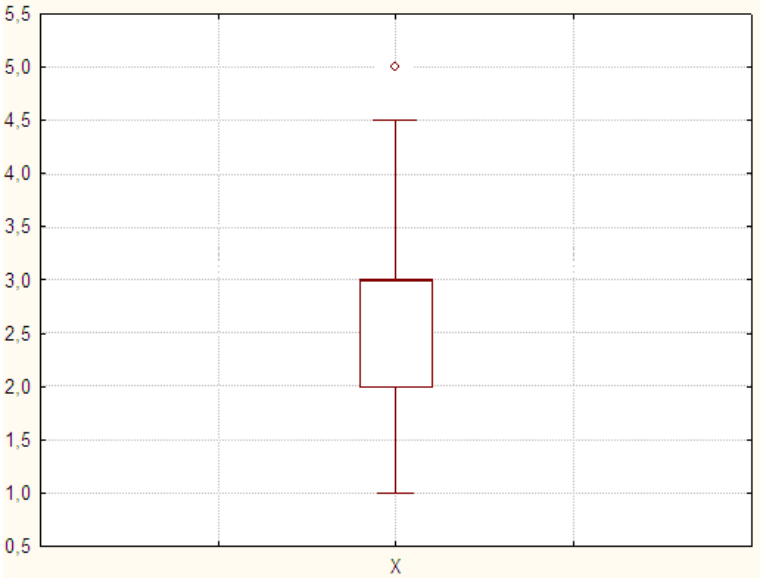


Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1,5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1,5q)$.
Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

Příklad na konstrukci krabicového diagramu
Pro datový soubor známek z matematiky 50 žáků 7. ročníku ZŠ sestrojte krabicový diagram

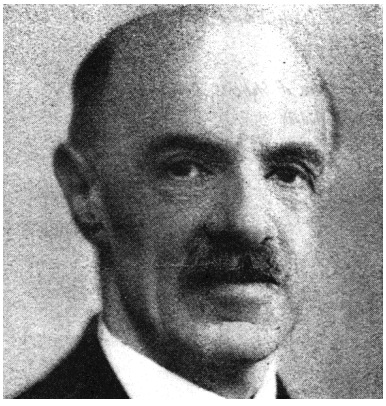
Řešení:

Již jsme spočítali medián $x_{0,50} = 3$, dolní kvartil $x_{0,25} = 2$, horní kvartil $x_{0,75} = 3$, kvartilová odchylka $q = 3 - 2 = 1$. Dále vypočítáme
dolní vnitřní hradba: $x_{0,25} - 1{,}5q = 2 - 1{,}5.1 = 0{,}5$,
horní vnitřní hradba: $x_{0,75} + 1{,}5q = 3 + 1{,}5.1 = 4{,}5$,
dolní vnější hradba: $x_{0,25} - 3q = 2 - 3.1 = -1$,
horní vnější hradba: $x_{0,75} + 3q = 3 + 3.1 = 6$.
Nakonec sestrojíme krabicový diagram.



Vidíme, že medián splyne s horním kvartilem, soubor známek tedy nemá symetrické rozložení četností. Vyskytuje se zde odlehlá hodnota 5, extrémní hodnoty nikoliv.

Charakteristika těsnosti závislosti dvou ordinálních znaků: Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik

Nejprve je nutné vysvětlit pojem pořadí čísla v posloupnosti čísel.
Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.

a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.
b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

Příklad na stanovení pořadí
a) Jsou dána čísla 9, 4, 5, 7, 3, 1.
b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.
Stanovte pořadí těchto čísel.

Řešení
ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,5	2,5	2,5	2,5	5,5	5,5	7	8,5	8,5	10

Předpokládejme, že máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \cdots & \cdots \\ x_n & y_n \end{pmatrix}$. Označíme R_i pořadí hodnoty x_i a Q_i pořadí hodnoty y_i, i = 1, ..., n.

Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$.

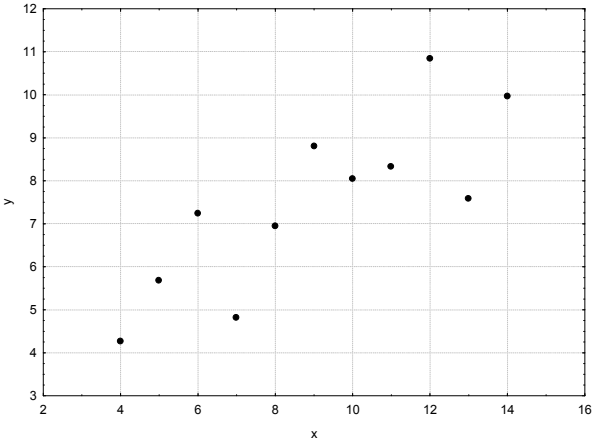
Vlastnosti Spearmanova koeficientu pořadové korelace:
Koeficient nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi znaky X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi znaky X a Y.
Je-li $r_S = 1$ resp. $r_S = -1$, pak dvojice (x_i, y_i) leží na nějaké vzestupné resp. klesající funkci.
Hodnoty r_S se nezmění, když provedeme vzestupnou transformaci původních dat.
Hodnoty r_S se vynásobí -1, když provedeme sestupnou transformaci původních dat.
Koeficient je symetrický.
Koeficient je rezistentní vůči odlehlým hodnotám.

Význam absolutní hodnoty Spearmanova koeficientu:

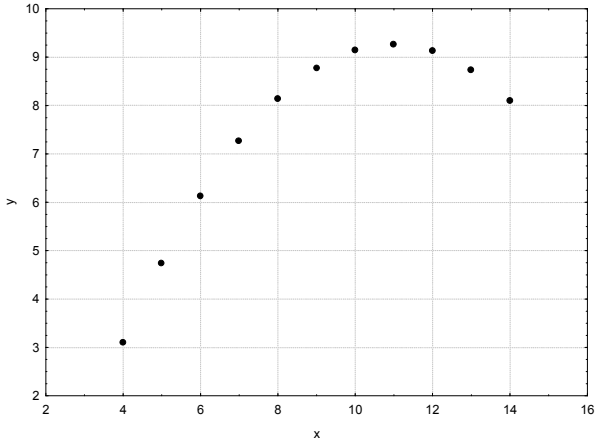
mezi 0 až 0,1 … zanedbatelná pořadová závislost,
mezi 0,1 až 0,3 … slabá pořadová závislost,
mezi 0,3 až 0,7 … střední pořadová závislost,
mezi 0,7 až 1 … silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace

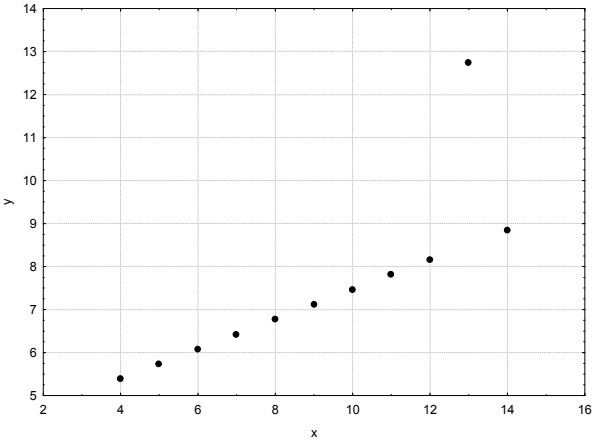
$r_S = 0{,}82$



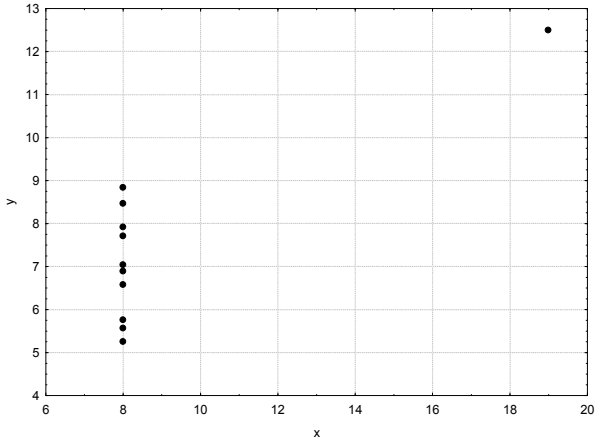
$r_S = 0{,}69$



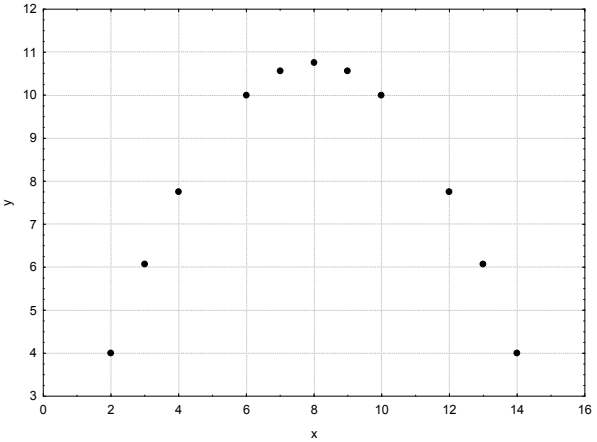
$r_S = 0{,}99$



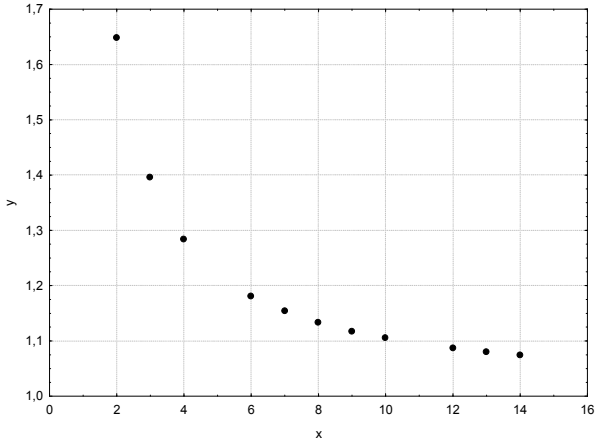
$r_S = 0{,}5$



$r_S = 0$



$r_S = -1$



Příklad na výpočet Spearmanova koeficientu pořadové korelace:
Je dán dvourozměrný datový soubor

$$\begin{pmatrix} 2,5 & 13,4 \\ 3,4 & 15,2 \\ 1,3 & 11,8 \\ 5,8 & 13,1 \\ 3,6 & 14,5 \end{pmatrix}$$

Vypočtěte Spearmanův koeficient pořadové korelace.

Řešení:

x_i	2,5	3,4	1,3	5,8	3,6
y_i	13,4	15,2	11,8	13,1	14,5
R_i	2	3	1	5	4
Q_i	3	5	1	2	4
$(R_i-Q_i)^2$	1	4	0	9	0

$$r_S = 1 - \frac{6}{n\left(n^2-1\right)}\sum_{i=1}^{n}\left(R_i - Q_i\right)^2 = 1 - \frac{6}{5\cdot 24}\left(1+4+0+9+0\right) = 1 - \frac{6\cdot 14}{5\cdot 24} = 0,3$$

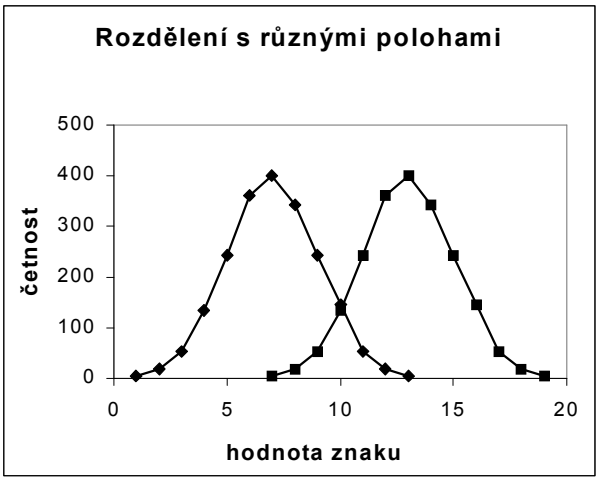
Znamená to, že mezi znaky X a Y existuje slabá přímá pořadová závislost.

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

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

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: $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - m^2$

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:

$$s_1^{\,2} = \frac{1}{3}\left[(7-8)^2 + (8-8)^2 + (9-8)^2\right] = \frac{1+0+1}{3} = \frac{2}{3} = 0,\overline{6}$$

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

Výpočet pomocí výpočetního vzorce:

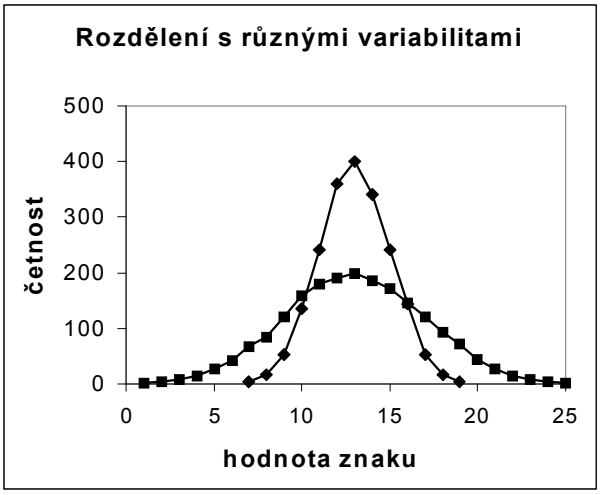
$$s_1^{\,2} = \frac{1}{3}\left(7^2 + 8^2 + 9^2\right) - 8^2 = \frac{49+64+81}{3} - 64 = \frac{194}{3} - 64 = \frac{194-192}{3} = \frac{2}{3} = 0,\overline{6}$$

$$s_2^{\,2} = \frac{1}{3}\left(1^2 + 10^2 + 13^2\right) - 8^2 = \frac{1+100+169}{3} - 64 = \frac{270}{3} - 64 = \frac{270-192}{3} = \frac{78}{3} = 26$$

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

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 více než 5 v obou směrech, většina čísel leží tedy mezi 3 a 13.

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



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

$$\frac{1}{n}\sum_{i=1}^{n}[(x_i - m) - 0]^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = s^2$$

Rozptyl standardizovaných hodnot je 1, protože

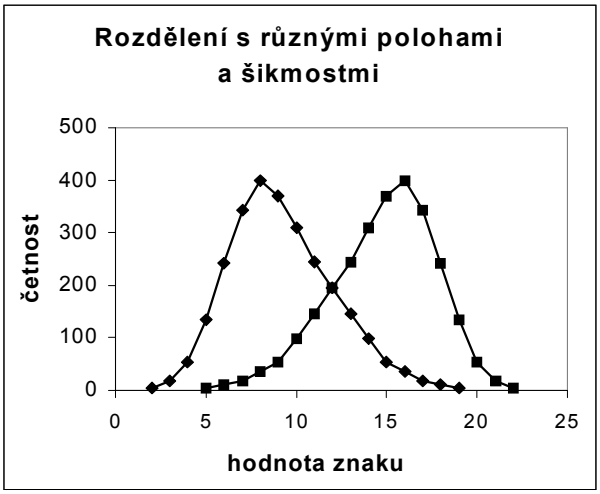
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s} - 0\right)^2 = \frac{1}{s^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{s^2}{s^2} = 1$$

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 $\alpha_3 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^3}{\sqrt{s^3}}$

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, $\alpha_3 < 0$.

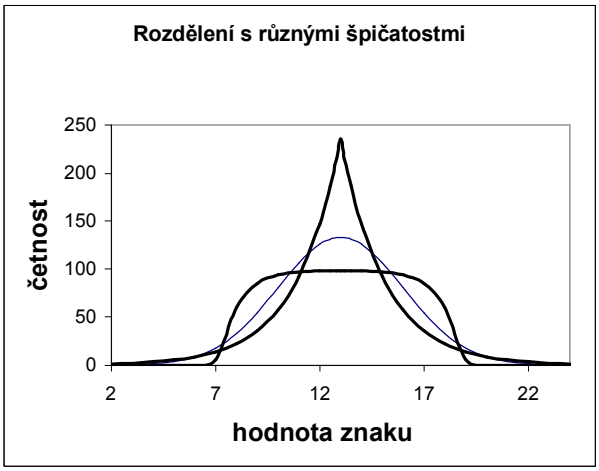
Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem a šikmostí



Charakteristika koncentrace dat kolem průměru: špičatost

$$\alpha_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^4}{\sqrt{s^4}} - 3$$

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

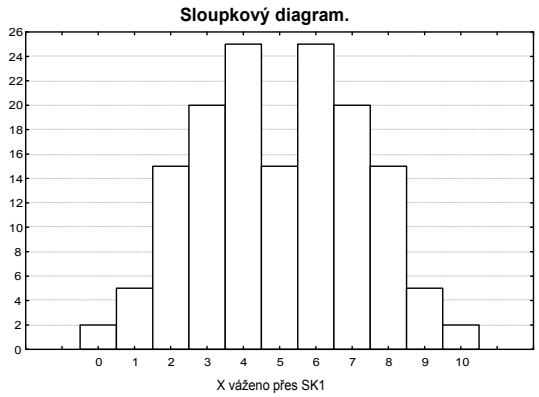
Variable	m	s^2	alfa3	alfa4
X	5	5	0	-0,759

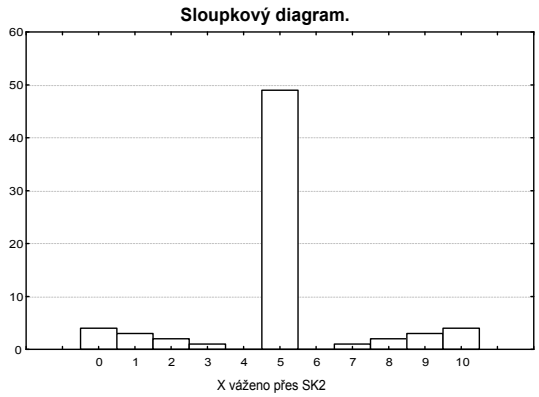
2. skupina

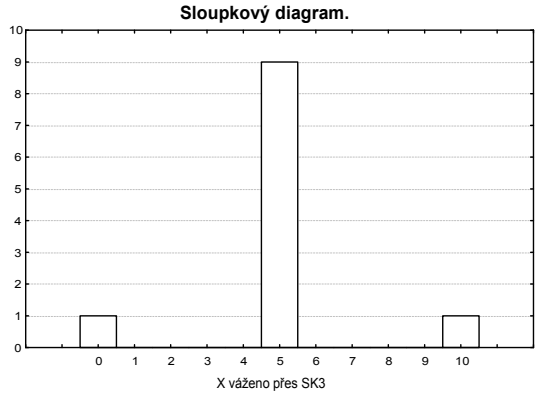
Variable	m	s^2	alfa3	alfa4
X	5	5	0	1,291

3. skupina

Variable	m	s^2	alfa3	alfa4
X	5	5	0	5,000







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

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2).$$

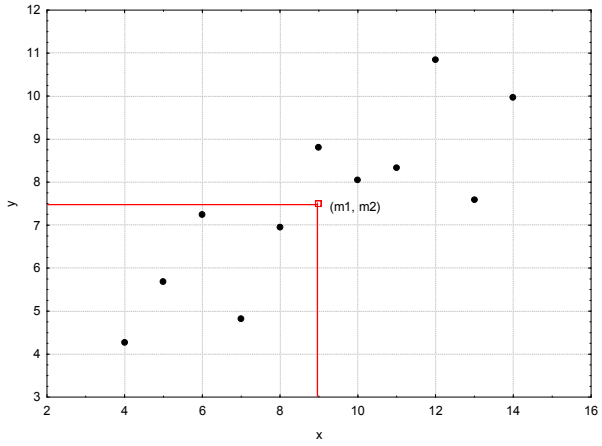
Kovariance je průměrem součinů centrovaných hodnot.
Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s nadprůměrnými (podprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot $x_i - m_1$ a $y_i - m_2$ vesměs kladné a jejich průměr (tj. kovariance) rovněž. Znamená to, že mezi znaky 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.

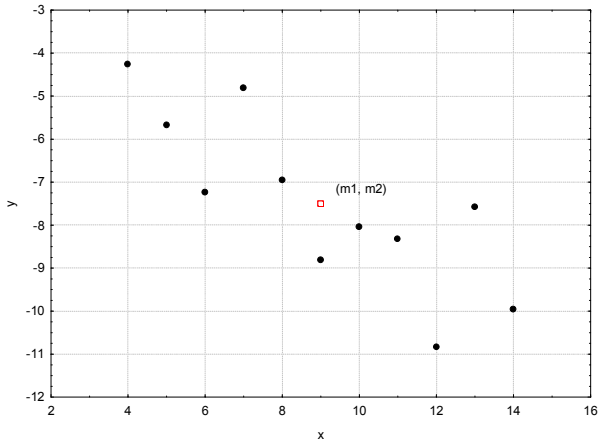
Je-li kovariance nulová, pak řekneme, že znaky X, Y jsou nekorelované a znamená to, že mezi nimi neexistuje žádná lineární závislost.

Znázornění významu kovariance

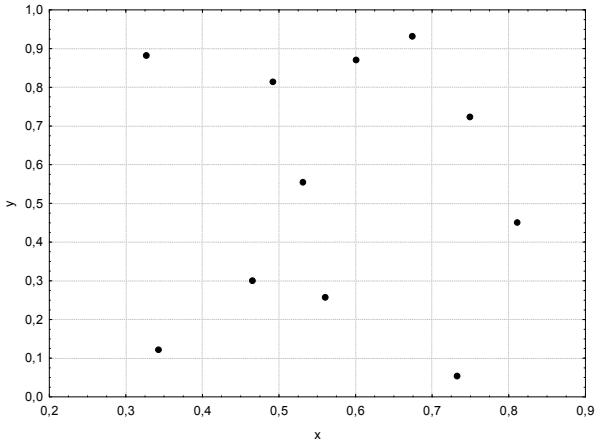
$s_{12} = 5{,}5$



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



$s_{12} = 0$



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

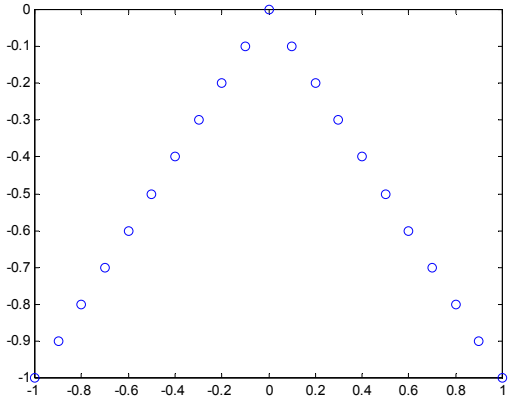
Charakteristika těsnosti závislosti dvou intervalových znaků: Pearsonův koeficient korelace

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 $r_S = 1$ resp. $r_S = -1$, pak dvojice (x_i, y_i) leží na nějaké vzestupné 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:

$$\sum_{i=1}^{8} x_i = 450, \sum_{i=1}^{8} y_i = 400, \sum_{i=1}^{8} x_i^2 = 26684, \sum_{i=1}^{8} y_i^2 = 20836, \sum_{i=1}^{8} x_i y_i = 23214$$

Řešení:
Vypočteme aritmetické průměry a rozptyly:

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

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_1^2 = \frac{1}{8}26684 - 56{,}25^2 = 171{,}4385, s_1 = 13{,}0934$$

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - m_2^2 = \frac{1}{8}20836 - 50^2 = 104{,}5, s_2 = 10{,}2225$$

Dále vypočteme kovarianci:

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2 = \frac{1}{8}23214 - 56{,}25 \cdot 50 = 89{,}25$$

Dosadíme do vzorce pro výpočet koeficientu korelace:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{89{,}26}{13{,}0934 \cdot 10{,}2225} = 0{,}6668$$

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: $m = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}$

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

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

Při použití vážených číselných charakteristik u intervalového rozložení četnosti si musíme uvědomit, že výpočty jsou přesné jen tehdy, souhlasí-li průměry v jednotlivých třídicích intervalech se středy těchto intervalů, resp. vykompenzují-li se vzájemně chyby, které vzniknou v důsledku odchylek středů intervalů od průměru v těchto intervalech. Oba tyto případy jsou však vzácné a většinou se dopustíme určité chyby.

Příklad na výpočet vážených číselných charakteristik
Je dán datový soubor 12 1,1 6,3 3,9 11 5,8 2,5 8 4,1 2 9,5 6,6 1,7 3,4 4,9 3 10,3 2,2 5,4 15,5. Stanovíme třídicí intervaly $\langle 1,2\rangle, \langle 2,4\rangle, \langle 4,7\rangle, \langle 7,11\rangle, \langle 11,16\rangle$.
Vypočtěte vážený průměr a vážený rozptyl.
Řešení:
Sestavíme tabulku rozložení četností:

$\langle u_j, u_{j+1}\rangle$	$x_{[j]}$	d_j	n_j
$\langle 1,2\rangle$	1,5	1	3
$\langle 2,4\rangle$	3	2	5
$\langle 4,7\rangle$	5,5	3	6
$\langle 7,11\rangle$	9	4	4
$\langle 11,16\rangle$	13,5	5	2

Vážený průměr:

$$m = \frac{1}{n}\sum_{j-1}^{r} n_j x_{[j]} = \frac{1}{20}\left(3\cdot 1{,}5 + 5\cdot 3 + 6\cdot 5{,}5 + 4\cdot 9 + 2\cdot 13{,}5\right) = \frac{1}{20}1155 = 5{,}775$$

Vážený rozptyl:

$$s^2 = \frac{1}{n}\sum_{j=1}^{r} n_j x_{[j]}^{\ 2} - m^2 = \frac{1}{20}\left(3\cdot 1{,}5^2 + 3\cdot 3^2 + 6\cdot 5{,}5^2 + 4\cdot 9^2 + 2\cdot 13{,}5^2\right) - 5{,}775^2 = 11{,}84$$

$$s = \sqrt{11{,}84} = 3{,}44$$

Pro srovnání: průměr vypočítaný z původního datového souboru je 5,96, rozptyl 14,85 a směrodatná odchylka 3,85.

Číselné charakteristiky poměrových znaků

Charakteristika polohy: aritmetický průměr. Jsou-li všechny hodnoty znaku kladné, lze definovat geometrický průměr $\sqrt[n]{x_1 \cdot \ldots \cdot x_n}$.

Příklad na výpočet geometrického průměru:
Rolník měl obdélníkový pozemek o stranách 80 m a 20 m. Rozoráním mezí získal pozemek čtvercový o stejné ploše. Jaká je strana čtverce?
Řešení: Strana čtverce bude geometrickým průměrem stran obdélníka, tedy $\sqrt{80\cdot 20} = \sqrt{1600} = 40$ m.

Charakteristiky variability: stejně jako u intervalových znaků používáme rozptyl a směrodatnou odchylku. Navíc definujeme koeficient variace $cv = \frac{s}{m}$. Často se vyjadřuje v procentech. Používá se zvláště tehdy, chceme-li porovnat variabilitu několika datových souborů.

Příklad na výpočet koeficientu variace:
Mezi místy A a B jezdí tramvaj a autobus. V době ranní špičky byla 6x použita tramvaj a 5x autobus. Naměřené časy cestování (v minutách) jsou pro tramvaj 32, 39, 42, 37, 34, 38 a pro autobus 30, 34, 28, 26, 32. Posuďte variabilitu časů cestování tramvají a autobusem pomocí koeficientů variace.
Řešení:
Vypočteme průměrné časy cestování: $m_1 = 37$, $m_2 = 30$. Dále vypočteme rozptyly a směrodatné odchylky: $s_1^2 = 10{,}67$, $s_2^2 = 8{,}33$, $s_1 = 3{,}27$, $s_2 = 2{,}89$. Po dosazení do vzorce pro koeficient variace dostaneme: $cv_1 = \frac{3{,}27}{37} \cdot 100\% = 8{,}8\%$,

$cv_2 = \frac{2{,}89}{30} \cdot 100\% = 9{,}6\%$

Vidíme, že poněkud vyšší variabilitu mají časy cestování autobusem.

Charakteristika společné variability dvou poměrových znaků: kovariance.

Charakteristika těsnosti lineární závislosti dvou poměrových znaků: Pearsonův koeficient korelace.