Číselné charakteristiky znaků

Doposud jsme se zabývali funkcionálními charakteristikami znaků, jako jsou
empirická distribuční funkce $F(x)$,
simultánní četnostní funkce $p(x,y)$,
marginální četnostní funkce $p_1(x)$, $p_2(y)$,
simultánní hustots četnosti $f(x,y)$,
marginální hustoty četnosti $f_1(x)$, $f_2(y)$,
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.

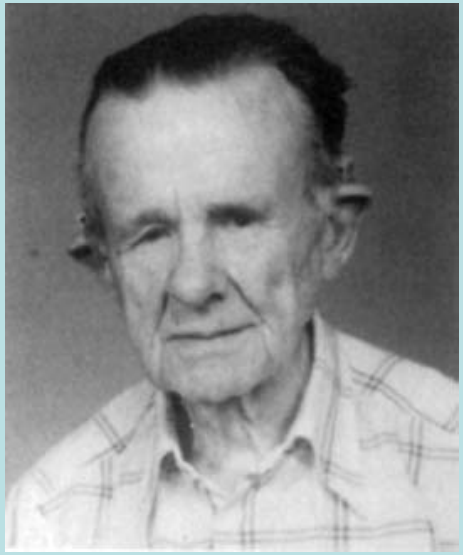
Příklad na stanovení modu
20 náhodně vybraných osob mělo odpovědět na otázku, který z pěti výrobků (označíme je A, B, C, D, E) preferují. Výsledky máme v tabulce:

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

Stanovte modus.

Řešení:
Modus = D

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:

	y	$y_{[1]}$	…	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		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:
800 náhodně vybraných osob bylo dotázáno na věk (znak X, varianty 1 – nejvýše 29, 2 – od 29 do 49, 3 – nad 49) a zda jsou ochotny volit v parlamentních volbách (znak Y, varianty 1 – ano, 2 – neví, 3 – ne). Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y			$n_{j.}$
	ano	nevím	ne	
nejvýše 29	128	21	27	176
od 29 do 49	223	58	39	320
nad 49	198	73	33	304
$n_{.k}$	549	152	99	800

Vypočtěte a interpretujte Cramérův koeficient.
Řešení: Nejprve vypočteme teoretické četnosti:

$$\frac{n_1 n_1}{n} = \frac{176 \cdot 549}{800} = 120{,}78 \qquad \frac{n_1 n_2}{n} = \frac{176 \cdot 152}{800} = 33{,}44 \qquad \frac{n_1 n_3}{n} = \frac{176 \cdot 99}{800} = 21{,}78$$

$$\frac{n_2 n_1}{n} = \frac{320 \cdot 549}{800} = 219{,}60 \qquad \frac{n_2 n_2}{n} = \frac{320 \cdot 152}{800} = 60{,}80 \qquad \frac{n_2 n_3}{n} = \frac{320 \cdot 99}{800} = 39{,}60$$

$$\frac{n_3 n_1}{n} = \frac{304 \cdot 549}{800} = 208{,}62 \qquad \frac{n_3 n_2}{n} = \frac{304 \cdot 152}{800} = 57{,}76 \qquad \frac{n_3 n_3}{n} = \frac{304 \cdot 99}{800} = 37{,}62$$

Nyní dosadíme do vzorce pro výpočet statistiky K:

$$K = \frac{(128 - 120{,}78)^2}{120{,}78} + \frac{(21 - 33{,}44)^2}{33{,}44} + \cdots + \frac{(33 - 37{,}62)^2}{37{,}62} = 16{,}3026$$

Nakonec vypočteme Cramérův koeficient:

$$V = \sqrt{\frac{16{,}3026}{800 \cdot 2}} = 0{,}1427$$

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

Číselné charakteristiky ordinálních znaků

Charakteristika polohy: α-kvantil. Je-li $\alpha \in$ ⟨, 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:

$n\alpha =$ ⟨ celé číslo $\Rightarrow x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}$; necelé číslo zaokrouhlíme na nejbližší celé číslo $\Rightarrow x_{(c)}$

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	x_{25}
0,10	50.0,1 = 5	5	x_5
0,90	50.0,9 = 45	45	x_{45}
0,25	50.0,25 = 12,5	13	$x_{(13)} = 2$
0,75	50.0,75 = 37,5	38	$x_{(38)} = 3$

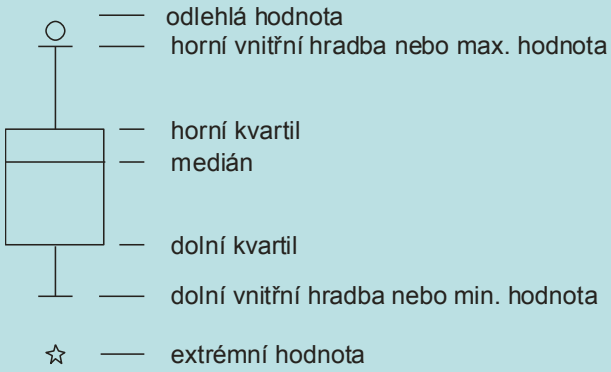
Kvartilová odchylka: q = 3 – 2 = 1.

Interpretace např. dolního kvartilu: V souboru žáků je aspoň čtvrtina takových, kteří mají z matematiky jedničku nebo dvojku (neboli v souboru 50 žáků jsou aspoň tři čtvrtiny takových, kteří mají z matematiky dvojku či horší známku).

Grafické znázornění ordinálních dat pomocí krabicového diagramu

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

Způsob konstrukce

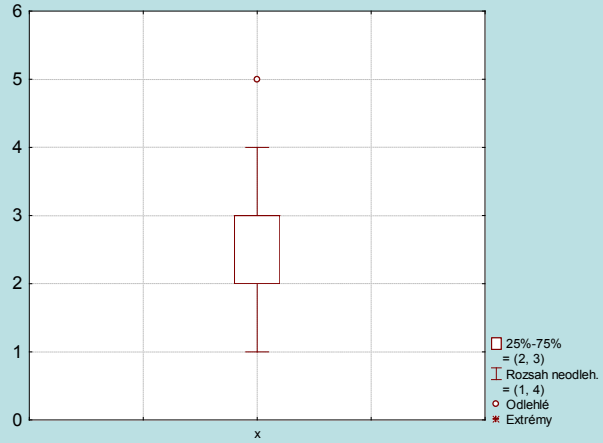


Odlehlá hodnota leží mezi vnějšími a vnitřními hradbami, tj. v intervalu $(x_{0,75} + 1{,}5q, x_{0,75} + 3q)$ či v intervalu $(x_{0,25} - 3q, x_{0,25} - 1{,}5q)$.
Extrémní hodnota leží za vnějšími hradbami, tj. v intervalu $(x_{0,75} + 3q, \infty)$ či v intervalu $(-\infty, x_{0,25} - 3q)$.

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

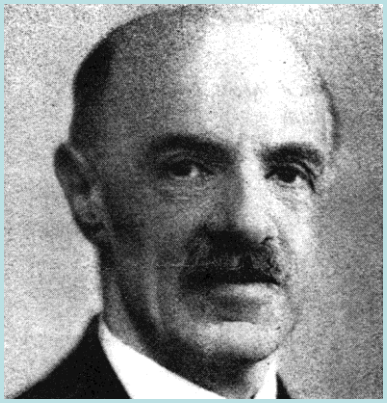
známka	1	2	3	4	5
n_j	9	15	20	4	2
N_j	9	24	44	48	50

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	2,25	2,25	2,25	2,25	5,5	5,5	7	8,5	8,5	10

Vzorec pro výpočet Spearmanova koeficientu:

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

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,

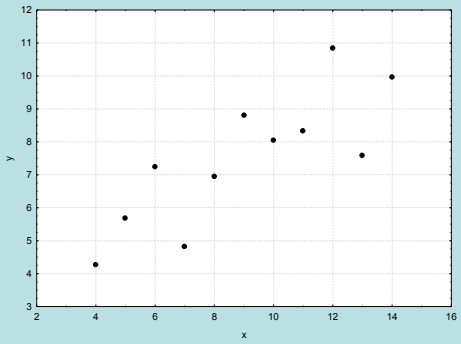
mezi 0,1 až 0,3 … slabá pořadová závislost,

mezi 0,3 až 0,7 … střední pořadová závislost,

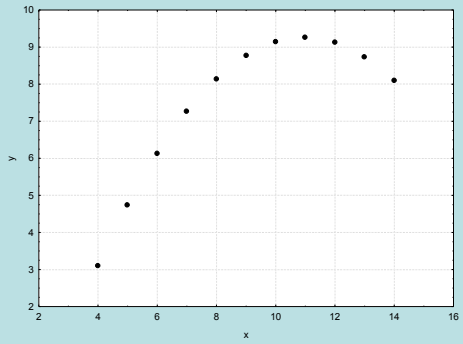
mezi 0,7 až 1 … silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace:

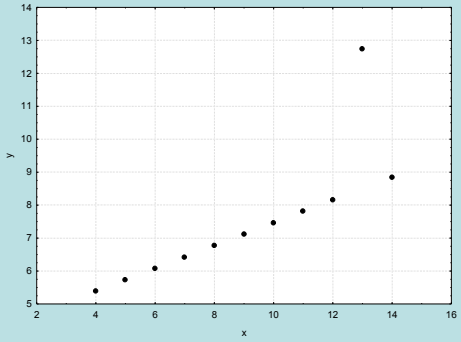
$r_S = 0{,}82$



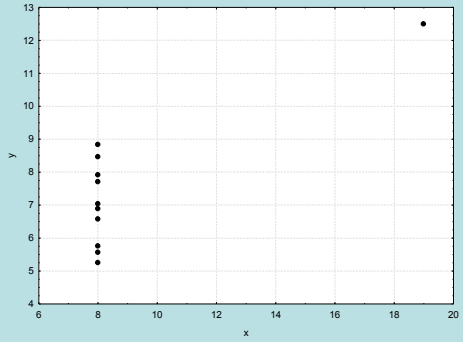
$r_S = 0{,}69$



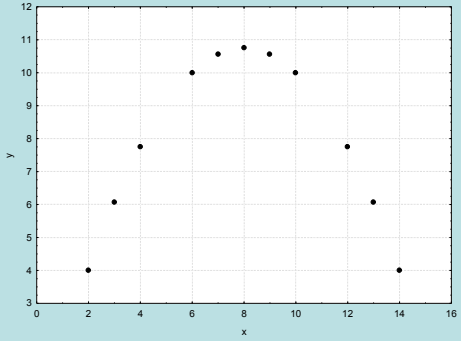
$r_S = 0{,}99$



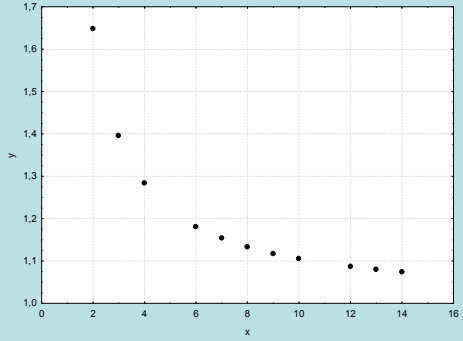
$r_S = 0{,}5$



$r_S = 0$



$r_S = -1$



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

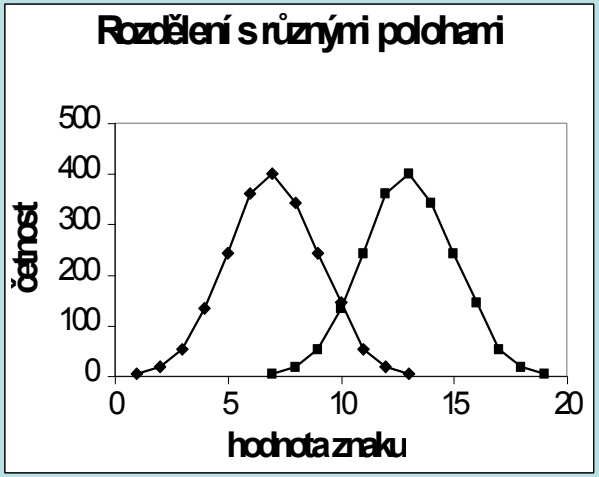
$$r_S = 1 - \frac{6\sum_{i=1}^{n}(R_i-Q_i)^2}{n^3-n} = 1 - \frac{6 \cdot 14}{5^3-5} = 1 - \frac{84}{120} = 0,3$$

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).

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



Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte aritmetické průměry znaků X, Y.

154	178		83	98		73	76		
133	164		106	111		77	85		
58	75		92	104		47	61		
145	161		85	103		68	85		
94	107		112	118		137	142		
113	141		98	102		44	68		
86	97		103	108		92	116		
121	127		99	119		141	157		
119	138		104	128		155	189		
112	125		107	118		136	155		
85	97		98	140		82	81		
41	72		97	115		136	163		
96	113		105	101		72	79		
45	89		71	93		66	81		
99	109		39	69		42	61		
51	95		122	147		113	123		
101	114		33	52		42	85		
160	169		78	117		133	147		
87	101		114	137		153	179		
88	139		125	149		85	91		

Řešení:

$m_1 = \frac{15 + 3 + \cdots + 9}{60} = 9,$ $m_2 = \frac{17 + 5 + \cdots + 1}{60} = 1,4$

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.

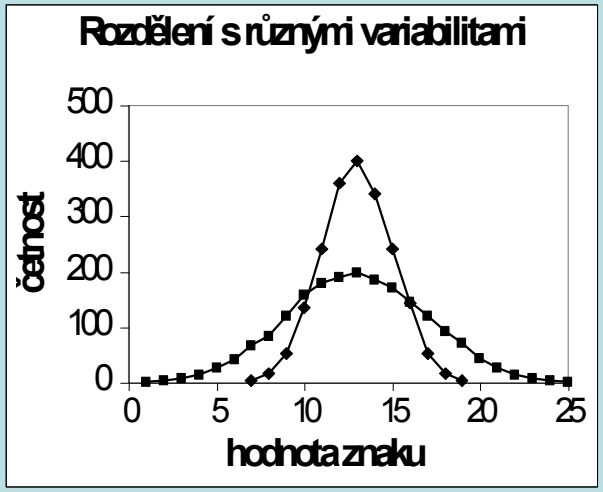
- Aritmetický průměr je silně ovlivněn extrémními hodnotami.

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

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



Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte rozptyly a směrodatné odchylky znaků X, Y. Přitom již víme, že $m_1 = 95{,}5$ a $m_2 = 114{,}4$.

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

Řešení:

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{80}\left(154^2 + 133^2 + \cdots + 88^2 - 80 \cdot 95{,}5^2\right) = 1054{,}20 \qquad s_1 = \sqrt{1054{,}20} = 32{,}47$$

$$s_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{80}\left(178^2 + 164^2 + \cdots + 139^2 - 80 \cdot 114{,}4^2\right) = 1052{,}71 \qquad s_2 = \sqrt{1052{,}71} = 32{,}45$$

Vlastnosti rozptylu

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

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

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

Rozptyl je stejně jako průměr silně ovlivněn vybočujícími hodnotami.

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

Kromě již uvedeného tvaru pro rozptyl je pro výpočty praktičtější tvar: $s^2 = \frac{1}{n}\sum_{i=1}^{n}$.

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

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

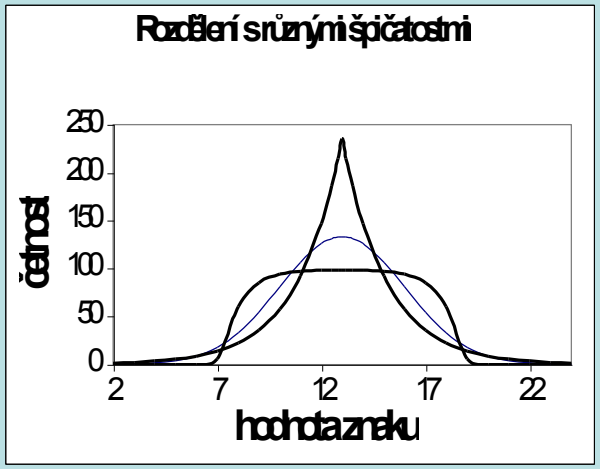
Informaci o koncentraci dat kolem průměru přináší špičatost $\alpha_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$

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

Je-li rozložení dat strmější než normální rozložení, pak $\alpha_4 > 0$.

Je-li rozložení dat plošší než normální rozložení, pak $\alpha_4 < 0$.

Znázornění rozložení četností dvou datových souborů, které se liší špičatostí



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.

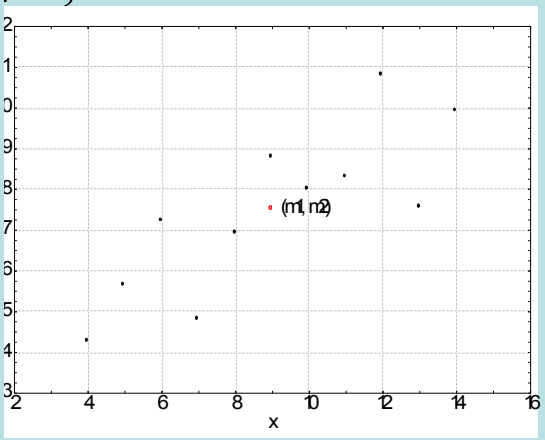
Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s podprůměrnými (nadprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot vesměs záporné a jejich průměr rovněž. Znamená to, že mezi znaky X a Y existuje určitý stupeň nepřímé lineární závislosti. Říkáme, že znaky X, Y jsou záporně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

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

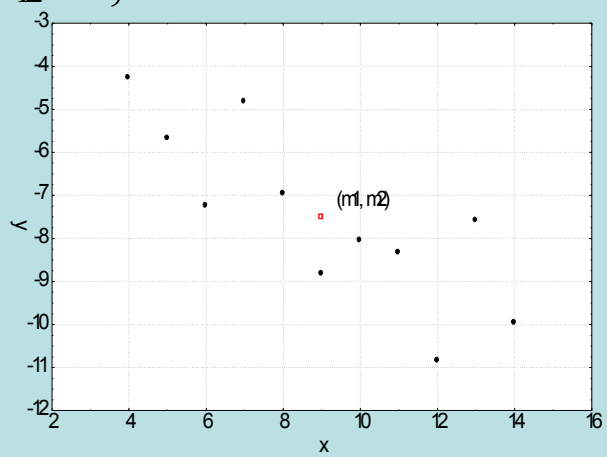
Pro výpočet kovariance používáme vzorec: $s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2$.

Znázornění významu kovariance

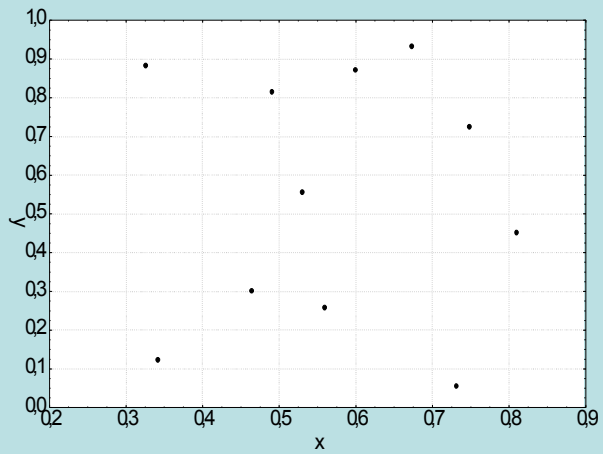
$s_{12} = 5,5$



$s_{12} = -5,5$



$s_{12} = 0$



Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte kovarianci znaků X, Y. Přitom již víme, že $m_1 = 95{,}5$, $m_2 = 114{,}4$, $s_1 = 32{,}4$, $s_2 = 32{,}5$

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

Řešení:

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_1 m_2 = \frac{1}{}(15 \cdot 17 + 13 \cdot 16 + \cdots + 91 \cdot 85) - 95{,}5 \cdot 114{,}4 = 3{,}5$$

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

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte koeficient korelace znaků X, Y. Přitom již víme, že $m_1 = 95,5$, $m_2 = 114,4$, $s_1 =32,4$, $s_2 = 32,5$, $s_{12} = 985,76$.

Řešení:

$$r_{12}=\frac{s_{12}}{s_1 s_2}=\frac{985,76}{32,4\cdot 32,5}=0,93$$

Koeficient korelace svědčí o tom, že mezi oběma znaky existuje velmi silná přímá lineární závislost – čím je vyšší mez plasticity, tím je vyšší mez pevnosti a čím je nižší mez plasticity, tím je nižší mez pevnosti.

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.

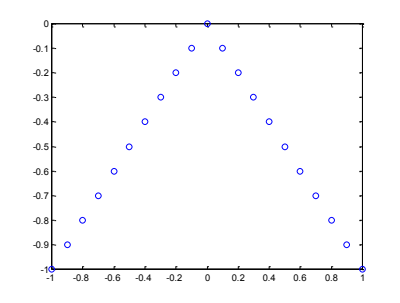
Je-li $r_{12} = 1$ resp. $r_{12} = -1$, pak dvojice (x_i, y_i) leží na nějaké rostoucí resp. klesající přímce.

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.



Početní pravidla pro číselné charakteristiky

Nechť m_1 je aritmetický průměr a s_1^2 rozptyl znaku X. Pak znak $Y = a + bX$ má aritmetický průměr $m_2 = a + bm_1$ a rozptyl $s_2^2 = b^2 s_1^2$.

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 $s_3^2 = s_1^2 + s_2^2 + 2s_{12}$.

Nechť s_{12} je kovariance znaků X, Y 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:

a) Znak X má aritmetický průměr 2 a rozptyl 3. Najděte aritmetický průměr a rozptyl znaku $Y = -1 + 3X$.

b) Znaky X a Y mají aritmetické průměry 3 a 2, rozptyly 2 a 3, kovarianci 1,5. Vypočtěte aritmetický průměr a rozptyl znaku $Z = 5X - 4Y$.

c) Součet rozptylů dvou znaků je 120, součin 1000 a rozptyl jejich součtů je 100. Vypočtěte koeficient korelace těchto znaků.

Řešení:

ad a) $m_2 = -1 + 3m_1 = -1 + 3 \times 2 = 5$, $s_2^2 = 3^2 \times s_1^2 = 9 \times 3 = 27$.

ad b) $m_3 = 5m_1 - 4m_2 = 5 \times 3 - 4 \times 2 = 7$, $s_3^2 = 5^2 \times s_1^2 + (-4)^2 \times s_2^2 + 2 \times 5 \times (-4) \times s_{12} = 25 \times 2 + 16 \times 3 - 40 \times 1{,}5 = 38$.

ad c) $s_1^2 + s_2^2 = 120$, $s_1^2 \times s_2^2 = 1000$, $s_{1+2}^2 = 100 = s_1^2 + s_2^2 + 2s_{12}$ $\Rightarrow$ $s_{12} =$ [illegible]

$r_{12} = \frac{s_{12}}{s_1 \times}$ = [illegible]

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)

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

$$s_1^2 = \frac{1}{7}(\cdot + + \cdot) - ^2 = \frac{116270442}{7} - 29 = 2, \quad s_1 = 0{,}766$$

$$s_2^2 = \frac{1}{7}(\cdot + \cdot + \cdot) - ^2 = \frac{120270453(}{7} - 29 = 2, \quad s_2 = 0{,}857$$

ad b)

$$s_{12} = \frac{1}{7}(\cdot \cdot + \cdot + \cdot + \cdot + \cdot + \cdot + \cdot + \cdot + \cdot) - \cdot = \frac{52}{}$$

$$= \frac{10227042752}{7} - 29 = \frac{70450}{} - 29 = \frac{0685871}{} - 29 = \frac{729}{} = 2$$

$$r_{12} = \frac{50}{\sqrt{72 \cdot 729}} = \frac{729}{42853(} = 04$$

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. Udává, jakým násobkem průměru je směrodatná odchylka. Umožňuje porovnat variabilitu několika znaků.

Jsou-li všechny hodnoty poměrového znaku kladné, pak jako charakteristiku polohy lze užít geometrický průměr $\sqrt[n]{x_1 \dots x_n}$. Geometrický průměr je vhodný tehdy, má-li smysl počítat součin pozorovaných hodnot, např. chceme-li charakterizovat vývoj prodeje určitého zboží pomocí řetězových indexů, pak vhod-nou charakteristikou souboru získaných indexů je právě geometrický průměr.

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtěte koeficienty variace znaků X, Y. Přitom již víme, že $m_1 = 95{,}5$, $m_2 = 114{,}4$, $s_1 = 32{,}4$, $s_2 = 32{,}5$

Řešení:

$$c_{V1} = \frac{s_1}{\bar{x}_1} = \frac{32{,}4}{95{,}5} = 0{,}339 \qquad c_{V2} = \frac{s_2}{\bar{x}_2} = \frac{32{,}5}{114{,}4} = 0{,}284$$

Vidíme, že mez plasticity oceli má poněkud vyšší variabilitu než mez pevnosti oceli.