



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 hustoty 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

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, nerepresentují 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 – jedničkař je lepší než dvojkař, ale intervaly mezi známkami nemají obsahovou interpretaci. Nelze tvrdit, že rozdíl ve znalostech mezi jedničkařem 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, ...



Typy znaků

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 °C, znamená to, že každým dnem stouply teploty o 2 °C. 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, 1 úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.



Typy znaků II

➤ Demografické znaky:

- Klienta (věk, pohlaví, rodinný stav, počet dětí, druh bydlení, kraj/okres trvalého bydliště...)
- Prodejního místa (kraj/okres, typ, prodejní plocha,...)
- Prodejce (věk, pohlaví, kraj/okres trvalého bydliště...)

➤ Behaviourální znaky:

- Klienta („stáří“ klienta, doposud splacená jistina, dlužná jistina, počet dní po splatnosti,...)
- Prodejního místa („stáří“ prodejny, počet uzavřených smluv, objem uzavřených smluv, podíl nesplácených úvěrů,...)
- Prodejce (počet uzavřených smluv, objem uzavřených smluv, podíl nesplácených úvěrů ...)

➤ Produktové znaky:

- Výše úvěru, délka smlouvy, akontace, RPSN,...



Číselné charakteristiky nominálních znaků

Charakteristika polohy: **modus** – nejčetnější varianta resp. střed nejčetnějšího třídícího 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

Označení: \hat{x}



Cramérův koeficient

Charakteristika těsnosti závislosti dvou nominálních znaků: Cramérův koeficient kontingence.



Carl Harald Cramér (1893 – 1985): Švédský matematik

Cramérův koeficient

Nechť znak X nabývá variant $x_{[1]}, \dots, x_{[r]}$ a znak Y nabývá variant $y_{[1]}, \dots, y_{[s]}$. 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 $(x_{[j]}, y_{[k]})$, $j = 1, \dots, r$, $k = 1, \dots, s$ a uspořádáme je do kontingenční tabulky:

	Y	$y_{[1]}$	\dots	$y_{[s]}$	$n_{j\cdot}$
X	n_{jk}				
$x_{[1]}$	n_{11}	\dots	n_{1s}	$n_{1\cdot}$	
\vdots	\dots	\dots	\dots	\dots	
$x_{[r]}$	n_{r1}	\dots	n_{rs}	$n_{r\cdot}$	
$n_{\cdot k}$	$n_{\cdot 1}$	\dots	$n_{\cdot s}$	n	

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

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

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.



Cramérův koeficient

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

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

Vypočtete a interpretujte Cramérův koeficient.

Řešení: Nejprve vypočteme teoretické četnosti:

$$\frac{n_{1,n_1}}{n} = \frac{368 \cdot 339}{686} = 181,8542, \quad \frac{n_{1,n_2}}{n} = \frac{368 \cdot 347}{686} = 186,1458,$$

$$\frac{n_{2,n_1}}{n} = \frac{318 \cdot 339}{686} = 157,1458, \quad \frac{n_{2,n_2}}{n} = \frac{318 \cdot 347}{686} = 160,8542$$

Nyní dosadíme do vzorce pro výpočet statistiky K:

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

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 $1 - \alpha$ všech dat. Pro výpočet α -kvantilu slouží algoritmus:

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

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

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

Příklad

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 v ýpočet 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	$n\alpha$	c	x_α
0,50	$50 \cdot 0,5 = 25$	25	$\frac{x_{(25)} + x_{(26)}}{2} = \frac{3+3}{2} = 3$
0,10	$50 \cdot 0,1 = 5$	5	$\frac{x_{(5)} + x_{(6)}}{2} = \frac{1+1}{2} = 1$
0,90	$50 \cdot 0,9 = 45$	45	$\frac{x_{(45)} + x_{(46)}}{2} = \frac{4+4}{2} = 4$
0,25	$50 \cdot 0,25 = 12,5$	13	$x_{(13)} = 2$
0,75	$50 \cdot 0,75 = 37,5$	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 u (neboli v souboru 50 žáků jsou aspoň tři čtvrtiny takových, kteří mají z matematiky dvojku či horší známku).

Příklad

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4	4	5	6	8	8	12	12	13	14	14	14	18	19

$$x_{0,25}$$

$$x_{0,5}$$

$$x_{0,75}$$

	n_j	N_j	p_j	F_j
1	1	1	0,07	0,07
4	2	3	0,13	0,20
5	1	4	0,07	0,27
6	1	5	0,07	0,33
8	2	7	0,13	0,47
12	2	9	0,13	0,60
13	1	10	0,07	0,67
14	3	13	0,20	0,87
18	1	14	0,07	0,93
19	1	15	0,07	1,00
Součet	15	x	1,00	x

$$x_{0,25} = 5$$

$$x_{0,5} = \tilde{x} = 12$$

$$x_{0,75} = 14$$

$$\hat{x} = 14$$

$$q = x_{0,75} - x_{0,25} = 14 - 5 = 9$$

$x_{0,25}$ je tedy hodnota, u které F_j poprvé překročí 0,25.

!!! Pokud ale $F_j = \alpha$ pro nějaké $x_{[j]}$, $x_\alpha = (x_{[j]} + x_{[j+1]})/2$

$$\rightarrow x_{0,2} = (4+5)/2 = 4,5$$



Modus a kvantily pro intervalově tříděná data

$$\hat{x} = d_m + \frac{n_m - n_{m-1}}{2n_m - n_{m-1} - n_{m+1}} \cdot h$$

d_m je dolní mez modální třídy,

n_m, n_{m-1}, n_{m+1} je četnost modální, předcházející a následující třídy,

h je šířka třídy

$$x_P = d_P + \frac{P - F_{P-1}}{p_P} \cdot h$$

d_P je dolní mez třídy obsahující příslušný P -kvantil,

p_P je relativní četnost této třídy,

F_{P-1} je kumulativní relativní četnost předcházející třídy,

h je šířka třídy



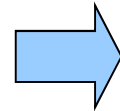
Příklad

Určete modus a medián.

x_i	N_i
méně než 15>	22
(15;20>	34
(20;25>	72
(25;30>	102
(30;35>	127
více než 35	135

Příklad

x_i	N_i
méně než 15>	22
(15;20>	34
(20;25>	72
(25;30>	102
(30;35>	127
více než 35	135



x_i	n_i	p_i	N_i	F_i
12,5	22	0,16	22	0,16
17,5	12	0,09	34	0,25
22,5	38	0,28	72	0,53
27,5	30	0,22	102	0,76
32,5	25	0,19	127	0,94
37,5	8	0,06	135	1,00
Součet	135	1,00	x	x

$$\hat{x} = 20 + \frac{38 - 12}{2 \cdot 38 - 12 - 30} \cdot 5$$

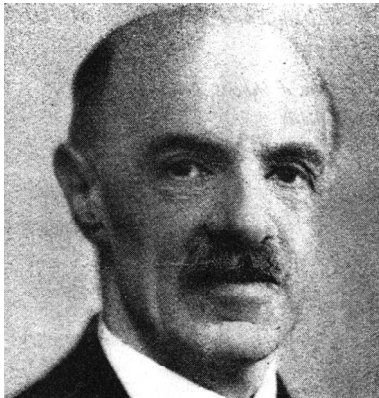
$$= 23,82$$

$$\tilde{x} = 20 + \frac{0,5 - 0,25}{0,28} \cdot 5$$

$$= 24,46$$

Spearmanův koeficient

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, \dots, 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, \dots, 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

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



Spearmanův koeficient

Vzorec pro výpočet Spearmanova koeficientu:

Předpokládejme, že máme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$. Označíme R_i pořadí

hodnoty x_i a Q_i pořadí hodnoty y_i , $i = 1, \dots, n$.

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.



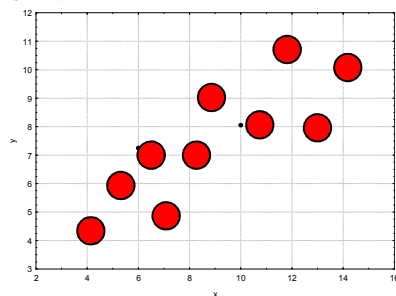
Spearmanův koeficient

Význam absolutní hodnoty Spearmanova koeficientu:

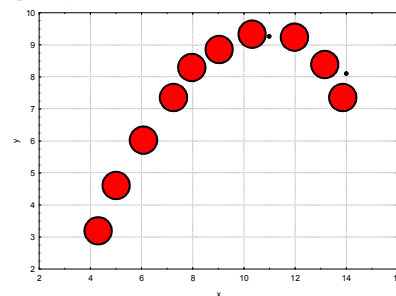
mezi 0 až 0,1 ... zanedbatelná pořadová závislost,
mezi 0,1 až 0,3 ... slabá pořadová závislost,
mezi 0,3 až 0,7 ... střední pořadová závislost,
mezi 0,7 až 1 ... silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace

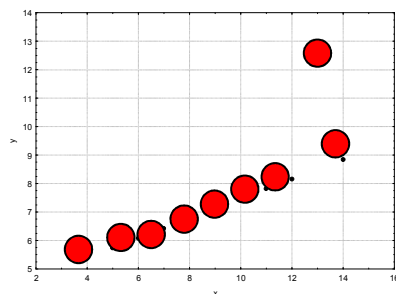
$r_s = 0,82$



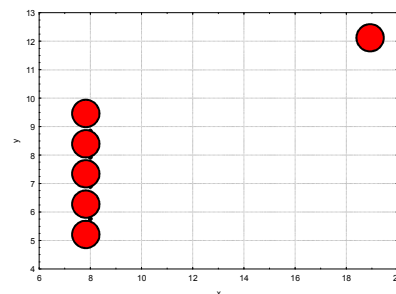
$r_s = 0,69$



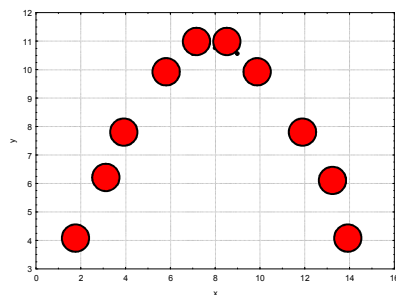
$r_s = 0,99$



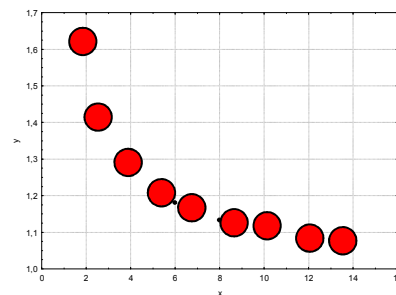
$r_s = 0,5$



$r_s = 0$



$r_s = -1$





Příklad

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

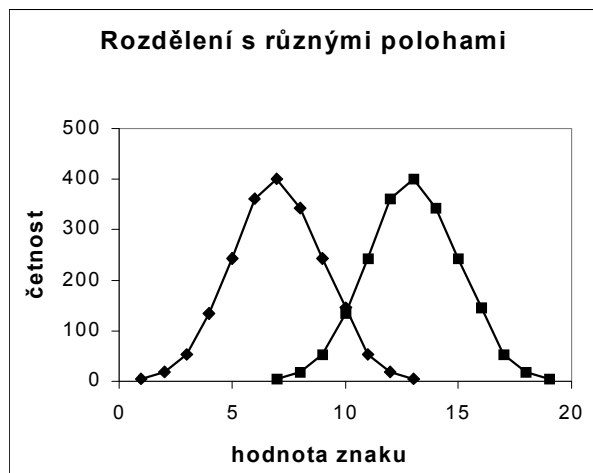
$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - Q_i)^2 = 1 - \frac{6}{5 \cdot 24} (1 + 4 + 0 + 9 + 0) = 1 - \frac{6 \cdot 14}{5 \cdot 24} = 0,3$$

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).

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



Často se aritmetický průměr označuje \bar{x} :

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$



Příklad

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtete 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:

$$m_1 = \frac{154 + 133 + \dots + 85}{60} = 95,9, \quad m_2 = \frac{178 + 164 + \dots + 91}{60} = 114,4$$



Aritmetický průměr

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.

Rozptyl, směrodatná odchylka

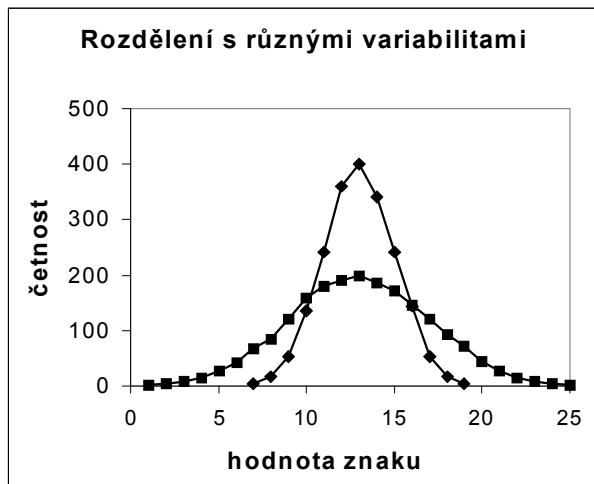
Charakteristika variability: rozptyl je průměrná kvadratická odchylka hodnot od jejich aritmetického průměru

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

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





Příklad

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
80	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:

$$s_1^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - m_1^2 = \frac{1}{60} (154^2 + 133^2 + \dots + 85^2) - 95,5^2 = 1052,40, s_1 = \sqrt{1052,40} = 32,4$$

$$s_2^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - m_2^2 = \frac{1}{60} (178^2 + 164^2 + \dots + 91^2) - 114,4^2 = 1057,21, s_2 = \sqrt{1057,21} = 32,5$$



Rozptyl, směrodatná odchylka - vlastnosti

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 $\frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - m}{s} - 0 \right)^2 = \frac{1}{s^2} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = \frac{s^2}{s^2} = 1$
- Rozptyl či směrodatná odchylka jsou stejně jako průměr silně ovlivněny extrémními hodnotami.
- Rozptyl či směrodatná odchylka se nehodí jako charakteristiky variability, je-li rozložení dat nesymetrické.

Šikmost

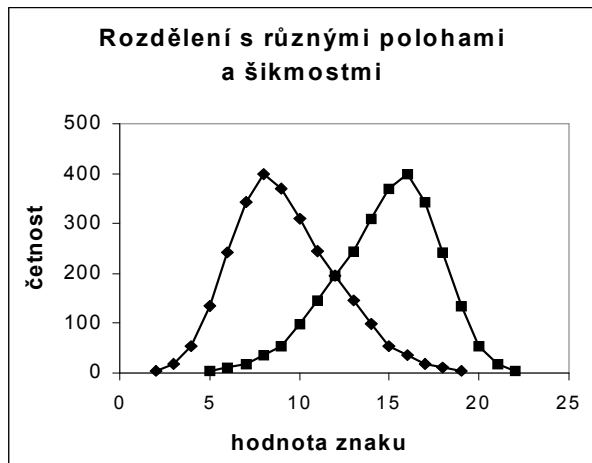
Charakteristika nesymetrie dat: šikmost $\alpha_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m)^3}{\sqrt{s^3}}$

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 dat prodloužený levý konec, jde o **záporně zešikmené rozložení**, $\alpha_3 < 0$.

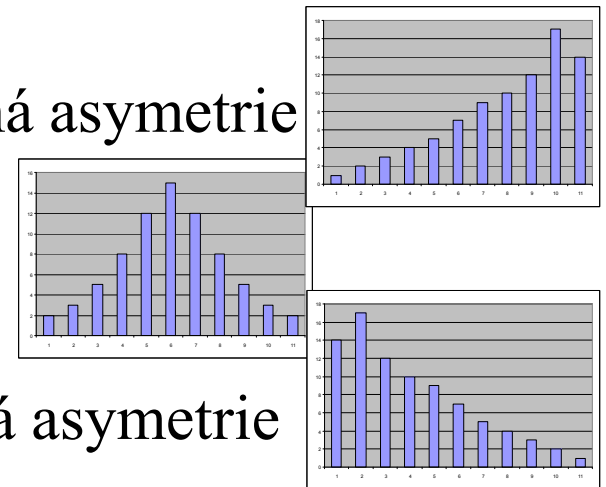
Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem a šikmostí



$\alpha_3 < 0$: Pravostranná asymetrie

$\alpha_3 = 0$: Symetrie

$\alpha_3 > 0$: Levostranná asymetrie



Špičatost

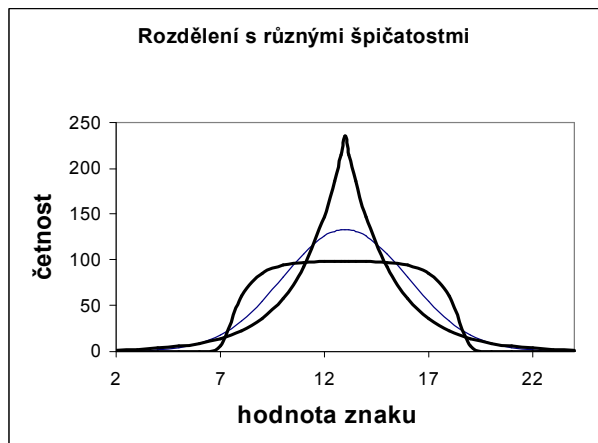
Charakteristika koncentrace dat kolem průměru: špičatost $\alpha_4 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^4 - 3\sigma^4$

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 $\alpha_4 < 0$.

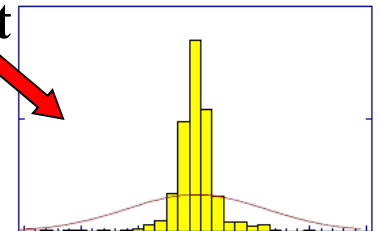
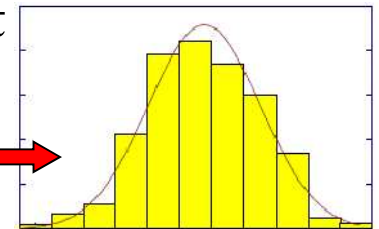
Znázornění rozložení četností dvou datových souborů, které se liší špičatostí



$\alpha_4 < 0$: Podnormální špičatost

$\alpha_4 = 0$: Normální špičatost

$\alpha_4 > 0$: Nadnormální špičatost





Kovariance

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

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

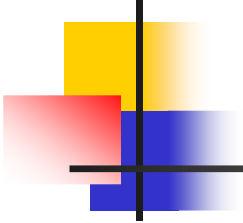
Kovariance je průměrem součinů centrovaných hodnot.

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.

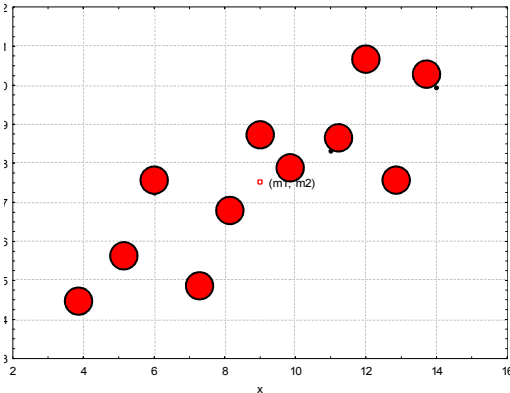
Pro výpočet kovariance používáme vzorec: $s_{12} = \frac{1}{n} \sum_{i=1}^n x_i y_i - m_1 m_2$.



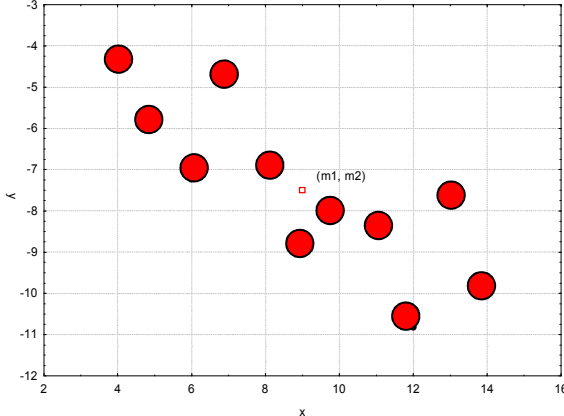
Kovariance

Znázornění významu kovariance

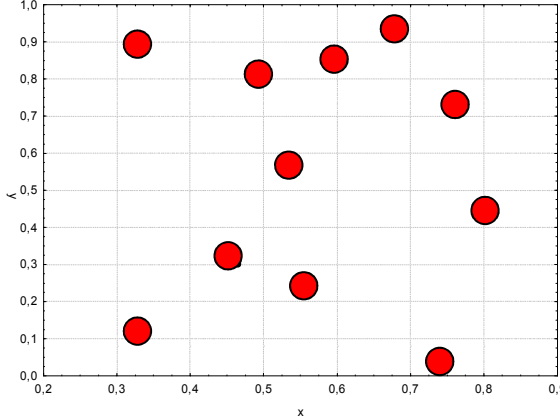
$s_{12} = 5,5$



$s_{12} = -5,5$



$s_{12} = 0$





Příklad

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:

$$s_{12} = \frac{1}{n} \sum_{i=1}^n x_i y_i - m_1 m_2 = \frac{1}{60} (154 \cdot 178 + 133 \cdot 164 + \dots + 85 \cdot 91) - 95,5 \cdot 114,4 = 985,76$$

Pearsonův koeficient korelace

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:

$$r_{12} = \frac{1}{n} \sum_{i=1}^n \frac{x_i - m_1}{s_1} \frac{y_i - m_2}{s_2}. \text{ Je to průměr součinů standardizovaných hodnot. Počítá se podle vzorce } r_{12} = \frac{s_{12}}{s_1 s_2}.$$

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočítejte 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:

$$r_{12} = \frac{s_{12}}{s_1 s_2} = \frac{985,76}{32,4 \cdot 32,5} = 0,936$$

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, \dots, x_n a y_1, \dots, y_n existuje úplná lineární závislost, tj. existují konstanty a , b tak, že $y_i = a + b x_i$, $i = 1, \dots, n$, 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. $r_{12} = r_{21}$.



Počtení pravidla pro číselné charakteristiky

Počtení 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 $m_2 = a + bm_1$, 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 $m_3 = m_1 + m_2$, 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

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) $s_1^2 + s_2^2 = 120$, $s_1^2 \times s_2^2 = 1000$, $s_{1+2}^2 = 100 = s_1^2 + s_2^2 + 2s_{12} \Rightarrow s_{12} = \frac{1}{2}(s_{1+2}^2 - s_1^2 - s_2^2) = \frac{1}{2}(100 - 120) = -10$

$$r_{12} = \frac{s_{12}}{s_1 \times s_2} = \frac{-10}{\sqrt{1000}} = -0,3162.$$



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

Příklad

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) Vypočítejte průměry a směrodatné odchylky znaků X a Y.
 b) Vypočítejte a interprejte 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

$$\text{ad a) } m_1 = \frac{1}{27} (1 \cdot 9 + 2 \cdot 11 + 3 \cdot 7) = \frac{52}{27} = 1,926, \quad m_2 = \frac{1}{27} (1 \cdot 11 + 2 \cdot 7 + 3 \cdot 9) = \frac{52}{27} = 1,926$$

$$s_1^2 = \frac{1}{27} (1^2 \cdot 9 + 2^2 \cdot 11 + 3^2 \cdot 7) - \left(\frac{52}{27} \right)^2 = \frac{116}{27} - \frac{2704}{729} = \frac{428}{729}, \quad s_1 = 0,766$$

$$s_2^2 = \frac{1}{27} (1^2 \cdot 11 + 2^2 \cdot 7 + 3^2 \cdot 9) - \left(\frac{52}{27} \right)^2 = \frac{120}{27} - \frac{2704}{729} = \frac{536}{729}, \quad s_2 = 0,857$$

ad b)

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

$$= \frac{102}{27} - \frac{2704}{729} = \frac{2754 - 2704}{729} = \frac{50}{729} = 0,0685871$$

$$r_{12} = \frac{\frac{50}{729}}{\sqrt{\frac{428}{729} \cdot \frac{536}{729}}} = 0,10439.$$

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



Koeficient variace, geometrický průměr

Pro poměrové znaky používáme jako charakteristiku variability **koeficient variace** $\frac{s}{m}$. 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** $\sqrt[n]{x_1 \cdot \dots \cdot x_n}$.

Příklad: Pro datový soubor obsahující údaje o mezi plasticity (znak X) a mezi pevnosti oceli (znak Y) vypočtete 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:

$$cv_1 = \frac{s_1}{m_1} = \frac{32,4}{95,5} = 0,339, cv_2 = \frac{s_2}{m_2} = \frac{32,5}{114,4} = 0,284$$



Výpočty zavedením pomocné proměnné

- pomocná proměnná $\Rightarrow v_i = \frac{x_i - a}{h}$
- konstanty:
 - $a \rightarrow$ střed třídy s nejvyšší četností
 - $h \rightarrow$ šířka třídy



Výpočty zavedením pomocné proměnné

$$\bar{v} = \frac{\bar{x} - a}{h} \implies \bar{x} = \bar{v}h + a$$

$$s_v^2 = \frac{s_x^2}{h^2} \implies s_x^2 = h^2 s_v^2$$



Příklad

x_i	n_i
<30 – 40)	10
<40 – 50)	31
<50 – 60)	27
<60 – 70)	19
<70 – 80)	13
Celkem	100

Vypočítejte:

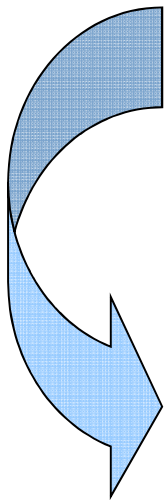
- aritmetický průměr, rozptyl, směrodatnou odchylku a variační koeficient zavedením pomocné proměnné



Příklad

$$a = 45$$

$$h = 10$$



x_i	n_i	v_i
35	10	-1
45	31	0
55	27	1
65	19	2
75	13	3
Součet	100	x



Příklad

x_i	n_i	v_i	$v_i n_i$
35	10	-1	-10
45	31	0	0
55	27	1	27
65	19	2	38
75	13	3	39
Součet	100	x	94

$$\begin{aligned}\bar{v} = 0,94 &\Rightarrow \bar{x} = \bar{v}h + a = \\ &= 0,94 \cdot 10 + 45 = 54,4\end{aligned}$$



Příklad

x_i	n_i	v_i	$v_i^2 n_i$
35	10	-1	10
45	31	0	0
55	27	1	27
65	19	2	76
75	13	3	117
Součet	100	x	230

$$s_v^2 = \frac{1}{n} \sum_{i=1}^n (v_i^2 n_i) - \bar{v}^2 = 2,3 - 0,8836 = 1,4164$$

$$\Rightarrow s_x^2 = h^2 \cdot s_v^2 = 10^2 \cdot 1,4164 = 141,64$$



Příklad

$$s_x = \sqrt{s_x^2} = \sqrt{141,64} = 11,9$$

nebo $s_x = h \cdot s_v^2 = 10 \cdot \sqrt{1,4164} = 11,9$

$$cv_x = \frac{s_x}{\bar{x}} = \frac{11,9}{55,4} = 0,2188$$

nebo $cv_x = \frac{h \cdot s_v}{\bar{v} \cdot h + a} = \frac{10 \cdot 1,19}{0,94 \cdot 10 + 45} = 0,2188$



Společný rozptyl

$$s^2 = \bar{s}^2 + s_{\bar{x}}^2$$

\bar{s}^2 vnitroskupinová variabilita ($s_{\#}^2$)¹

$s_{\bar{x}}^2$ meziskupinová variabilita (s_*^2)

¹⁾ Značení ze skript „Popisná statistika“



Společný rozptyl

- **vnitroskupinová variabilita**

$$\bar{s}^2 = \frac{1}{n} \sum_{i=1}^k s_i^2 \cdot n_i$$

- **meziskupinová variabilita**

$$s_{\bar{x}}^2 = \frac{1}{n} \sum_{i=1}^k (\bar{x}_i - \bar{\bar{x}})^2 \cdot n_i$$



Příklad

D1:	104	108	79	155
D2:	93	65	76	111

Vypočítejte:

- **dílčí průměry,**
- **společný průměr,**
- **dílčí rozptyly,**
- **společný rozptyl.**



Příklad

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i^1 =$$

$$= \frac{1}{4} \cdot (104 + 108 + 79 + 155) = \underline{\underline{111,5}}$$

$$\bar{x}_2 = \frac{1}{4} \cdot (93 + 65 + 76 + 111) = \underline{\underline{86,25}}$$

$$\bar{x} = \frac{1}{8} \cdot (111,5 \cdot 4 + 86,25 \cdot 4) = \underline{\underline{98,875}}$$



Příklad

$$s_{x1}^2 = \frac{1}{n} \sum_{i=1}^{n_1} (x_i^1 - \bar{x}_1)^2 =$$
$$= \frac{7,5^2 + 3,5^2 + 32,5^2 + 43,5^2}{4} = \underline{\underline{754,25}}$$

$$s_{x2}^2 = \frac{6,75^2 + 21,25^2 + 10,25^2 + 24,75^2}{4} = \underline{\underline{303,69}}$$



Příklad

$$\begin{aligned}\bar{s}^2 &= \frac{1}{n} \sum_{i=1}^k s_i^2 \cdot n_i = \frac{1}{8} (s_{x_1}^2 \cdot 4 + s_{x_2}^2 \cdot 4) = \\ &= \frac{1}{8} \cdot (754,25 \cdot 4 + 303,69 \cdot 4) = \underline{\underline{528,97}}\end{aligned}$$

$$\begin{aligned}s_{\bar{x}}^2 &= \frac{1}{n} \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 \cdot n_i = \\ &= \frac{(111,5 - 98,875)^2 \cdot 4 + (86,25 - 98,875)^2 \cdot 4}{8} = \underline{\underline{159,39}}\end{aligned}$$



Příklad

$$\begin{aligned} s^2 &= \bar{s}^2 + s_{\bar{x}}^2 = \\ &= 528,97 + 159,39 = \underline{\underline{688,36}} \end{aligned}$$

Pro kontrolu ještě spočteme rozptyl přímo:

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{8} \cdot 83717 - 98,875^2 \\ &= 10464,63 - 9776,27 = \underline{\underline{688,36}} \end{aligned}$$