Intervalové rozložení četností

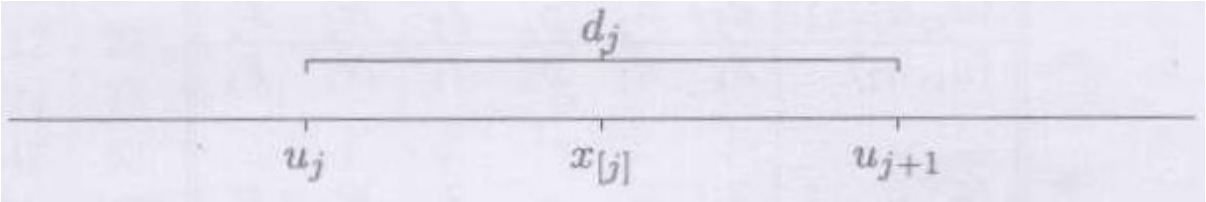
Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X je blízký rozsahu souboru, pak četnosti přiřazujeme nikoliv jednotlivým variantám, ale celým intervalům hodnot. Hovoříme pak o intervalovém rozložení četnosti.

Číselnou osu rozložíme na intervaly typu $(-\infty, u_1)$, $\langle u_1, u_2)$, ..., $\langle u_r, u_{r+}\rangle$, $\langle u_{r+1}, \infty)$ tak, aby okrajové intervaly neobsahovaly žádnou pozorovanou hodnotu znaku X. Užíváme označení:

$\langle u_j, u_{j+}\rangle$ – j-tý třídicí interval znaku X, j = 1, ..., r.

$d_j = u_{j+1} - u_j$ – délka j-tého třídicího intervalu znaku X

$x_{[j]} = \frac{u_j + u_{j+}}{2}$ – střed j-tého třídicího intervalu znaku X



Třídicí intervaly volíme nejčastěji stejně dlouhé. Jejich počet určíme např. pomocí Sturgersova pravidla: $r = 1 + 3{,}3 \log_{10} n$, kde n je rozsah souboru.

Sestavení tabulky rozložení četností

Hodnoty znaku X roztřídíme do r třídicích intervalů. Pro j = 1, ..., r definujeme:

$n_j = N(u_j < X \leq u_{j+1})$ – absolutní četnost j-tého třídicího intervalu ve výběrovém souboru

$p_j = \frac{n_j}{n}$ – relativní četnost j-tého třídicího intervalu ve výběrovém souboru

$f_j = \frac{p_j}{d_j}$ – četnostní hustota j-tého třídicího intervalu ve výběrovém souboru

$N_j = N(X \leq u_{j+1}) = n_1 + ... + n_j$ – absolutní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru

$F_j = \frac{N_j}{n} = p_1 + ... + p_j$ – relativní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru.

Tabulka typu

(u_j, u_{j+1})	d_j	n_j	p_j	f_j	N_j	F_j
(u_1, u_2)	d_1	n_1	p_1	f_1	N_1	F_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
(u_r, u_{r+1})	d_r	n_r	p_r	f_r	N_r	F_r
Součet		n	1			

se nazývá tabulka rozložení četností.

Příklad: Do laboratoře bylo dodáno 60 vzorků a byly zjištěny a hodnoty znaku X – mez plasticity (v kp/cm^2) a Y – mez pevnosti (v kp/cm^2). Datový soubor má tvar:

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

a) Pro znak X stanovte optimální počet třídicích intervalů dle Sturgersova pravidla.
b) Sestavte tabulku rozložení četností.

Řešení:

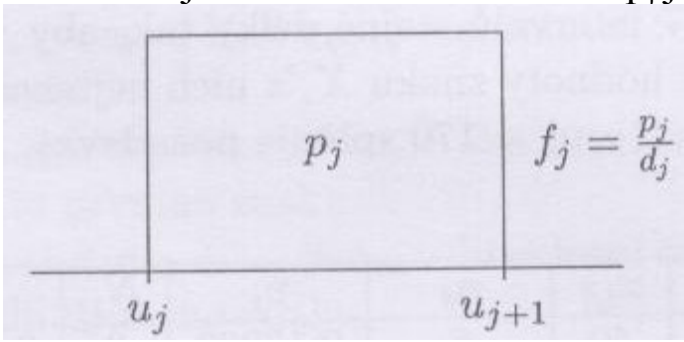
ad a) Rozsah souboru je 60. Podle Sturgersova pravidla je optimální počet třídicích intervalů r = 7. Budeme tedy volit 7 intervalů stejné délky tak, aby v nich byly obsaženy všechny pozorované hodnoty znaku X, z nichž nejmenší je 33, největší 160; volba $u_1 = 30, ..., u_8 = 170$ splňuje požadavky.

ad b)

$\langle u_j, u_{j+} \rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 30,50\rangle$	20	40	8	$8/60 = 0,1\overline{3}$	8	$8/60 = 0,1\overline{3}$	$8/(60\cdot 20) = 0,00\overline{6}$
$\langle 50,70\rangle$	20	60	4	$4/60 = 0,0\overline{6}$	12	$12/60 = 0,2$	$4/(60\cdot 20) = 0,00\overline{3}$
$\langle 70,90\rangle$	20	80	13	$13/60 = 0,21\overline{6}$	25	$25/60 = 0,41\overline{6}$	$13/(60\cdot 20) = 0,0183$
$\langle 90,110\rangle$	20	100	15	$15/60 = 0,25$	40	$40/60 = 0,\overline{6}$	$15/(60\cdot 20) = 0,0125$
$\langle 110,130\rangle$	20	120	9	$9/60 = 0,15$	49	$49/60 = 0,81\overline{6}$	$9/(60\cdot 20) = 0,0075$
$\langle 130,150\rangle$	20	140	7	$7/60 = 0,11\overline{6}$	56	$56/60 = 0,9\overline{3}$	$7/(60\cdot 20) = 0,00583$
$\langle 150,170\rangle$	20	160	4	$4/60 = 0,0\overline{6}$	60	$60/60 =$	$4/(60\cdot 20) = 0,00\overline{3}$
Součty			60	1			

Histogram, hustota četnosti, intervalová empirická distribuční funkce

Intervalové rozložení četností graficky znázorňujeme pomocí histogramu. Je to graf skládající se z r obdélníků, sestrojených nad třídicími intervaly, přičemž obsah j-tého obdélníku je roven relativní četnosti p_j j-tého třídicího intervalu, j = 1, ..., r.



Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané hustota četnosti:

$$f(x) = \begin{cases} f_j \text{ pro } u_j < x \le u_{j+1},\ j = 1, \cdots, r \\ 0 \text{ jinak} \end{cases}$$

Pomocí hustoty četnosti zavedeme intervalovou empirickou distribuční funkci:

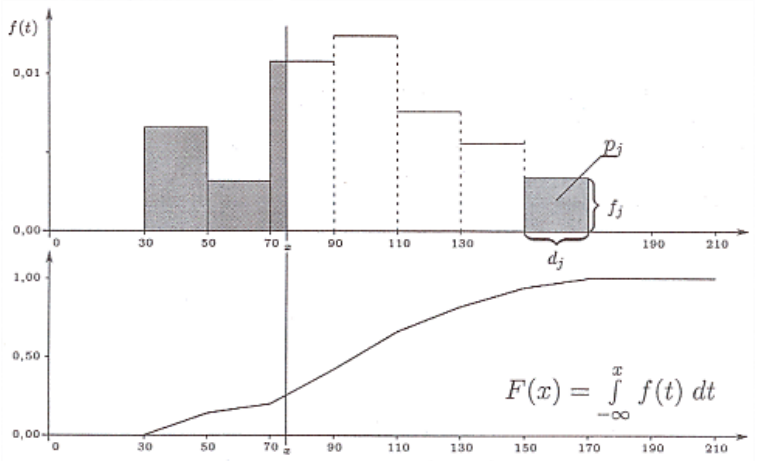
$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

Hustota četnosti je nezáporná ($\forall x \in \mathbb{R}: f(x) \ge 0$) a normovaná ($\int_{-\infty}^{\infty} f(x)\,dx = 1$). Intervalová empirická distribuční funkce je neklesající, spojitá a normovaná ($\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$).

Příklad: Pro mez plasticity oceli nakreslete histogram a pod histogram graf intervalové empirické distribuční funkce.

Řešení: Vyjdeme z tabulky rozložení četností.

$\langle u_j, u_{j+} \rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 0,50 \rangle$	20	40	8	$8/60 = 0{,}1\overline{3}$	8	$8/60 = 0{,}1\overline{3}$	$8/(60\cdot 20) = 0{,}00\overline{6}$
$\langle 0,70 \rangle$	20	60	4	$4/60 = 0{,}0\overline{6}$	12	$12/60 = 0{,}2$	$4/(60\cdot 20) = 0{,}00\overline{3}$
$\langle 70,90 \rangle$	20	80	13	$13/60 = 0{,}21\overline{6}$	25	$25/60 = 0{,}41\overline{6}$	$13/(60\cdot 20) = 0{,}0183$
$\langle 90,110 \rangle$	20	100	15	$15/60 = 0{,}25$	40	$40/60 = 0{,}\overline{6}$	$15/(60\cdot 20) = 0{,}0125$
$\langle 110,130 \rangle$	20	120	9	$9/60 = 0{,}15$	49	$49/60 = 0{,}81\overline{6}$	$9/(60\cdot 20) = 0{,}0075$
$\langle 130,150 \rangle$	20	140	7	$7/60 = 0{,}11\overline{6}$	56	$56/60 = 0{,}9\overline{3}$	$7/(60\cdot 20) = 0{,}00583$
$\langle 150,170 \rangle$	20	160	4	$4/60 = 0{,}0\overline{6}$	60	$60/60 =$	$4/(60\cdot 20) = 0{,}00\overline{3}$
Součty			60	1			

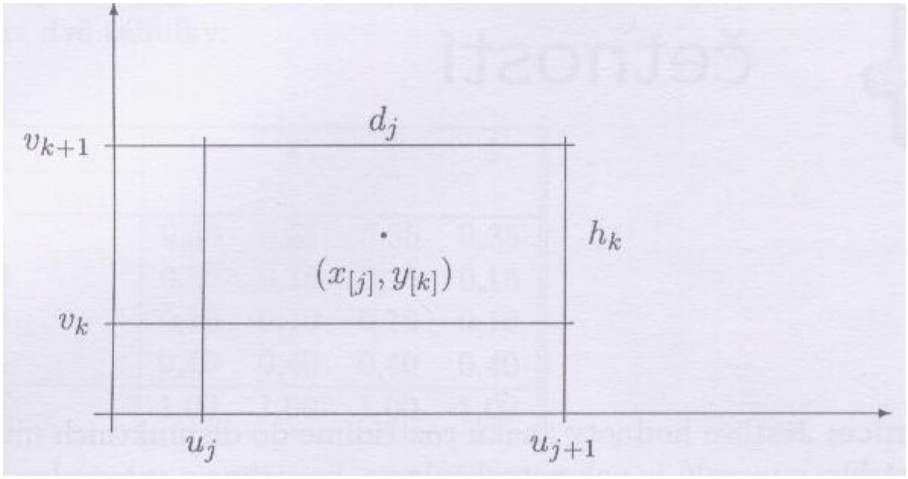


Dvourozměrné intervalové rozložení četností

Dále se budeme věnovat dvourozměrnému intervalovému rozložení četností, tj. budeme pracovat s dvourozměrným datovým souborem. Zavedeme podobné pojmy jako u dvourozměrného bodového rozložení četností

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$, kde hodnoty znaku X roztřídíme do r třídicích intervalů $\langle u_j, u_{j+} \rangle$, j = 1, ..., r s délkami d_1, ..., d_r a hodnoty znaku Y roztřídíme do s třídicích intervalů $\langle v_k, v_{k+} \rangle$, k = 1, ..., s s délkami h_1, ..., h_s.

Obdélník $\langle u_j, u_{j+} \rangle \times \langle v_k, v_{k+} \rangle$ se nazývá (j,k)-tý dvourozměrný třídicí interval.



Simultánní a marginální četnosti

$n_{jk} = N(u_j < X \leq u_{j+1} \wedge v_k < Y \leq v_{k+1})$ – simultánní absolutní četnost (j, k)-tého třídicího intervalu.

$p_{jk} = \frac{n_{jk}}{n}$ – simultánní relativní četnost (j, k)-tého třídicího intervalu.

$n_{j.} = n_{j1} + \ldots + n_{js}$ – marginální absolutní četnost j-tého třídicího intervalu pro znak X.

$p_{j.} = \frac{n_{j.}}{n}$ – marginální relativní četnost j-tého třídicího intervalu pro znak X.

$n_{.k} = n_{1k} + \ldots + n_{rk}$ – marginální absolutní četnost k-tého třídicího intervalu pro znak Y.

$p_{.k} = \frac{n_{.k}}{n}$ – marginální relativní četnost k-tého třídicího intervalu pro znak Y.

$f_{jk} = \frac{p_{jk}}{d_j h_k}$ – simultánní četnostní hustota v (j, k)-tém třídicím intervalu.

$f_{j.} = \frac{p_{j.}}{d_j}$ – marginální četnostní hustota v j-tém třídicím intervalu pro znak X.

$f_{.k} = \frac{p_{.k}}{h_k}$ – marginální četnostní hustota v k-tém třídicím intervalu pro znak Y.

Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky.

Uveďme kontingenční tabulku simultánních absolutních četností:

	(v_k, v_{k+1})	(v_1, v_2)	$\ldots$	(v_s, v_{s+1})	
(u_j, u_{j+1})	n_{jk}				$n_{j.}$
(u_1, u_2)		n_{11}	$\ldots$	n_{1s}	$n_{1.}$
$\vdots$					$\vdots$
(u_r, u_{r+1})		n_{r1}	$\ldots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	$\ldots$	$n_{.s}$	n

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

a) stanovte dle Sturgersova pravidla optimální počet třídicích intervalů pro znak Y

b) sestavte kontingenční tabulku simultánních absolutních četností.

Řešení:

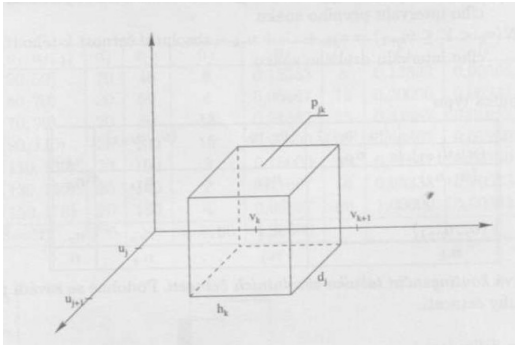
ad a) Rozsah datového souboru je 60. Podle Sturgersova pravidla je tedy optimální počet třídicích intervalů 7. Nejmenší hodnota je 52 a největší 189. Volíme $v_1 = 50$, $v_2 = 70$, ..., $v_8 = 190$.

ad b)

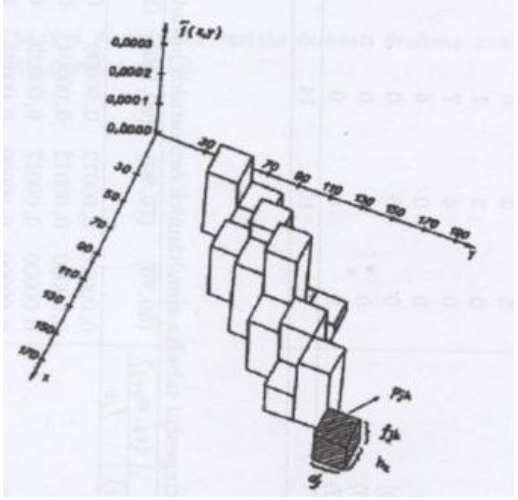
(u_j, u_{j+1}) \ (v_k, v_{k+1})	$(50, 70)$	$(70, 90)$	$(90, 110)$	$(110, 130)$	$(130, 150)$	$(150, 170)$	$(170, 190)$	$n_{j\cdot}$
$(30, 50)$	5	3	0	0	0	0	0	8
$(50, 70)$	0	3	1	0	0	0	0	4
$(70, 90)$	0	4	7	1	1	0	0	13
$(90, 110)$	0	0	6	8	1	0	0	15
$(110, 130)$	0	0	0	4	5	0	0	9
$(130, 150)$	0	0	0	0	2	5	0	7
$(150, 170)$	0	0	0	0	0	1	3	4
$n_{\cdot k}$	5	10	14	13	9	6	3	$n = 60$

Stereogram

Dvourozměrné intervalové rozložení četností graficky znázorňujeme pomocí stereogramu. Je to graf skládající se z r x s kvádrů, sestrojených nad dvourozměrnými třídicími intervaly, přičemž objem (j, k)-tého kvádru je roven relativní četnosti p_{jk} (j, k)-tého třídicího intervalu,
j = 1, ..., r, k = 1, …, s. Výška kvádru tedy vyjadřuje simultánní četnostní hustotu.



V našem příkladě s mezí plasticity a mezí pevnosti oceli bude mít stereogram tvar:



Simultánní a marginální hustota četnosti

Pomocí simultánních četnostních hustot zavedeme simultánní hustotu četnosti:

Funkce $f(x,y) = \begin{cases} f_{jk} \text{ pro } u_j < x \le u_{j+1},\ v_k < y \le v_{k+1},\ j=1,\cdots,r,\ k=1,\cdots,s \\ 0 \text{ jinak} \end{cases}$ se nazývá simultánní hustota četnosti. Jejím grafem je schodovitá plocha shora omezující stereogram.

Hustoty četnosti pro znaky X a Y odlišíme indexem takto:

$$f_1(x) = \begin{cases} f_{j.} \text{ pro } u_j < x \le u_{j+1},\ j=1,\cdots,r \\ 0 \text{ jinak} \end{cases},$$

$$f_2(y) = \begin{cases} f_{.k} \text{ pro } v_k < y \le v_{k+1},\ k=1,\cdots,s \\ 0 \text{ jinak} \end{cases}.$$

Mezi simultánní hustotou četnosti a marginálními hustotami četnosti platí vztahy:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy, \quad f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx.$$

Četnostní nezávislost znaků v daném výběrovém souboru při intervalovém rozložení četností

Pomocí simultánních a marginálních četnostních zavedeme pojem četnostní nezávislosti znaků v daném výběrovém souboru při intervalovém rozložení četností: Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé při intervalovém rozložení četností, jestliže pro všechna j = 1, ..., r a všechna k = 1, ..., s platí multiplikativní vztah: $f_{jk} = f_{j.}\, f_{.k}$ neboli pro $\forall\ x, y) \in\ ^2$: $f(x, y) = f_1(x)\, f_2(y)$.

V našem příkladě nejsou mez pevnosti a mez plasticity četnostně nezávislé, protože už pro j = 1, k = 1 je multiplikativní vztah porušen:

$\langle u_j, u_{j+1}\rangle$ \ $\langle v_k, v_{k+1}\rangle$	n_{jk}	(50, 70⟩	(70, 90⟩	(90, 110⟩	(110, 130⟩	(130, 150⟩	(150, 170⟩	(170, 190⟩	$n_{j.}$
(30, 50⟩		5	3	0	0	0	0	0	8
(50, 70⟩		0	3	1	0	0	0	0	4
(70, 90⟩		0	4	7	1	1	0	0	13
(90, 110⟩		0	0	6	8	1	0	0	15
(110, 130⟩		0	0	0	4	5	0	0	9
(130, 150⟩		0	0	0	0	2	5	0	7
(150, 170⟩		0	0	0	0	0	1	3	4
$n_{.k}$		5	10	14	13	9	6	3	$n = 60$

$f_{11} = \frac{5}{60 \cdot 20 \cdot 20} =$),000208, $f_{1.} = \frac{8}{60 \cdot 20} =$),006667, $f_{.1} = \frac{5}{60 \cdot 20} =$),004167, tudíž

0,000208 ≠ 0,006667.0,004167 = 0,000028

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.

Charakteristika variability: mutabilita $M = \frac{n^2 - \sum_{j=} \quad ^2}{n(n-\quad)}$, 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

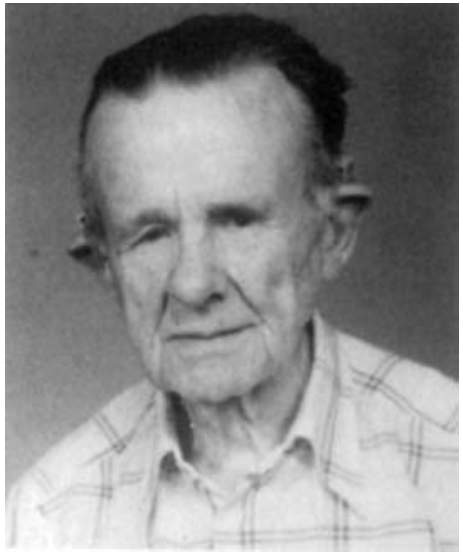
Stanovte modus a vypočtěte mutabilitu.

Řešení:
Modus = D

Mutabilita: $M = \frac{n^2 - \sum_{j=} \quad ^2}{n(n-\quad)} = \frac{20^2 - (\quad^2 + \quad^2 + \quad^2 + \quad^2 + \quad^2)}{20 \cdot 19} = \quad,821$

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:

	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: $V = \sqrt{\frac{K}{n(m-\ )}}$, 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:

$$\frac{n_{1.}n_{.1}}{n} = \frac{368 \cdot 339}{686} = 81{,}8542, \quad \frac{n_{1.}n_{.2}}{n} = \frac{368 \cdot 347}{686} = 86{,}1458,$$

$$\frac{n_{2.}n_{.1}}{n} = \frac{318 \cdot 339}{686} = 57{,}1458, \quad \frac{n_{2.}n_{.2}}{n} = \frac{318 \cdot 347}{686} = 60{,}8542$$

Nyní dosadíme do vzorce pro výpočet statistiky K:

$$K = \frac{(56 - 81{,}8542)^2}{181{,}8542} + \frac{(312 - 86{,}1458)^2}{186{,}1485} + \frac{(283 - 57{,}1458)^2}{157{,}1458} + \frac{(35 - 60{,}8542)^2}{160{,}8542} = 71{,}456$$

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 $\alpha \in \langle 0;1 \rangle$, 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:

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

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	$\frac{x_{(5)}+x_{(6)}}{2}=\frac{1+2}{2}=$
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	$x_{(38)} = 3$

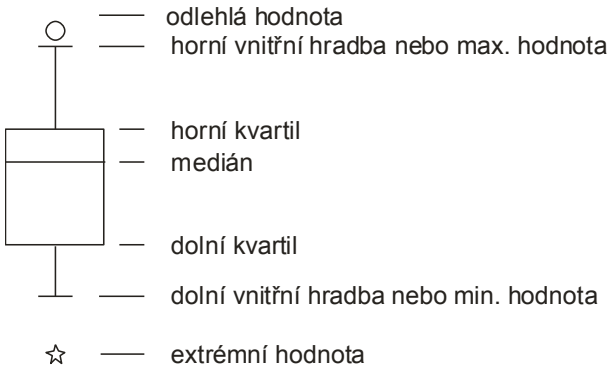
Kvartilová odchylka: q = 3 – 2 = 1.

Interpretace např. dolního kvartilu: V souboru 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

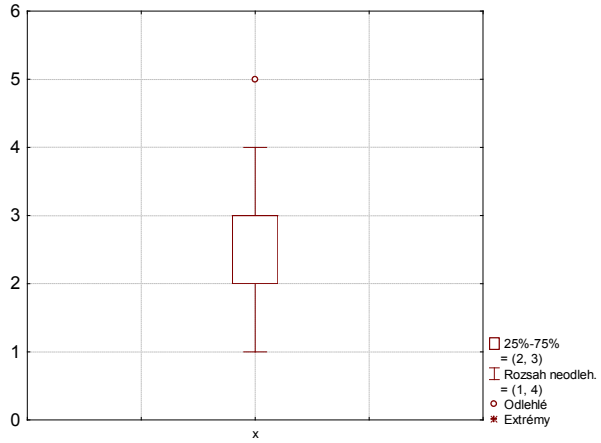
dolní vnitřní hradba: $x_{0,25} - 1,5q = 2 - 1,5.1 = 0,5$,

horní vnitřní hradba: $x_{0,75} + 1,5q = 3 + 1,5.1 = 4,5$,

dolní vnější hradba: $x_{0,25} - 3q = 2 - 3.1 = -1$,

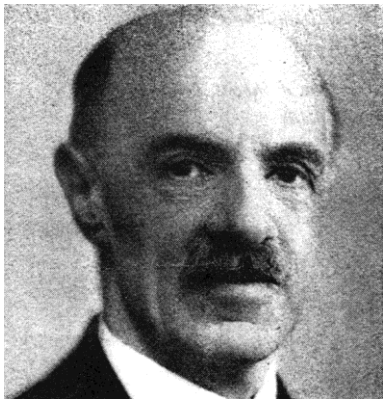
horní vnější hradba: $x_{0,75} + 3q = 3 + 3.1 = 6$.

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

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,

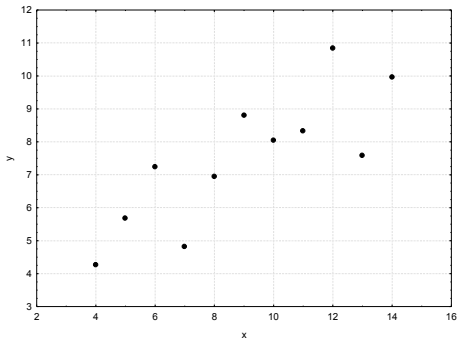
mezi 0,1 až 0,3 … slabá pořadová závislost,

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

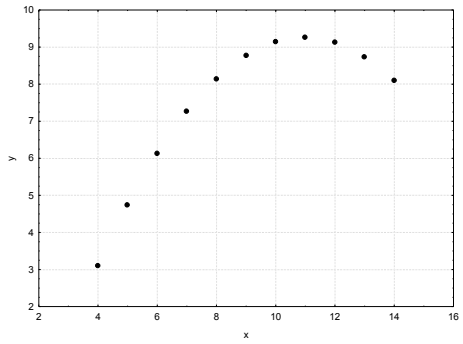
mezi 0,7 až 1 … silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace:

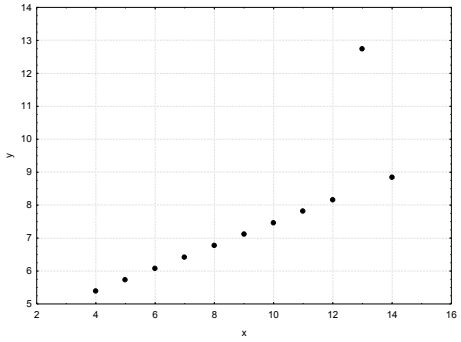
$r_S = 0{,}82$



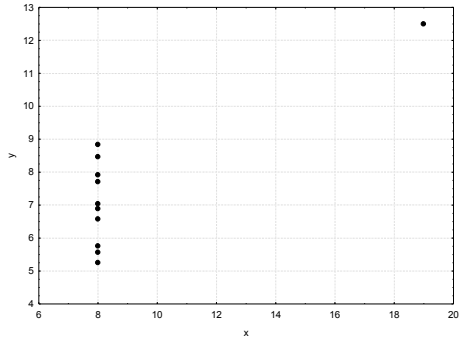
$r_S = 0{,}69$



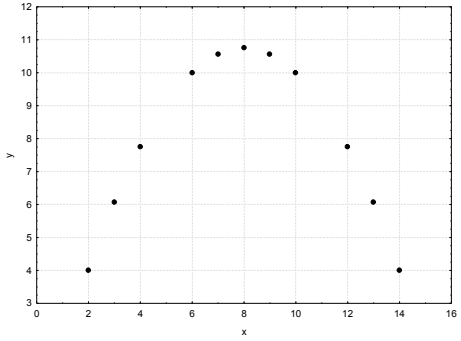
$r_S = 0{,}99$



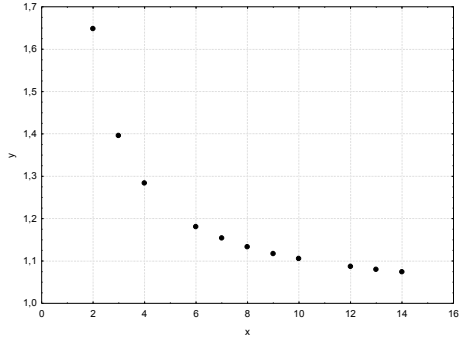
$r_S = 0{,}5$



$r_S = 0$



$r_S = -1$



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

$$r_S = -\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2 = -\frac{6}{5\cdot 24}(1+4+0+9+0) = -\frac{6\cdot 14}{5\cdot 24} = -0,7$$

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