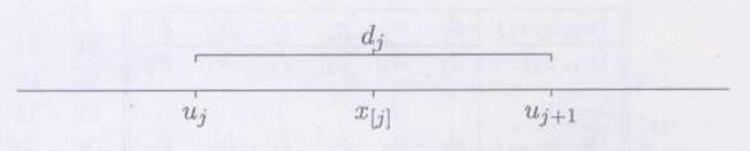
Intervalové rozložení četností - jednorozměrný případ

Je-li počet variant znaku X velký, přiřazujeme četnosti nikoli jednotlivým variantám, ale třídicím intervalům $\langle u_1, u_2 \rangle$, ..., $\langle u_r, u_{r+1} \rangle$ a hovoříme o intervalovém rozložení četností.

Ilustrace j-tého třídicícho intervalu:



Názvy četností jsou podobné jako u bodového rozložení četností, navíc zavádíme četnostní hustotu j-tého třídicího intervalu $f_j = \frac{p_j}{d_j}$, kde $d_j = u_{j+1} - u_j$.

Třídicí intervaly volíme nejčastěji stejně dlouhé. Stanovení jejich počtu je dosti subjektivní záležitost.

Často se používá Sturgesovo pravidlo: $r = 1 + 3{,}3 \log_{10} n$ (n je rozsah souboru) nebo se doporučuje volit r blízké $\sqrt{n}$.

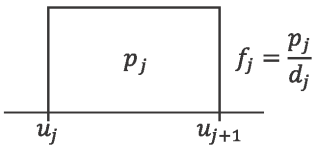
Tabulka rozložení četností:

$\langle u_j, u_{j+1} \rangle$	$x_{[j]}$	d_j	n_j	p_j	N_j	F_j	f_j
$\langle u_1, u_2 \rangle$	$x_{[1]}$	d_1	n_1	p_1	N_1	F_1	f_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\langle u_r, u_{r+1} \rangle$	$x_{[r]}$	d_r	n_r	p_r	N_r	F_r	f_r

Hustota četnosti: $f(x) = \begin{cases} f_j \text{ pro } u_j < x \le u_{j+1},\ j = 1, \cdots, r \\ 0 \text{ jinak} \end{cases}$, intervalová empirická distribuční funkce: $F(x) = \int_{-\infty}^{x} f(t)dt$.

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.

Ilustrace konstrukce histogramu:



Příklad: U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	$\langle 35,65\rangle$	$\langle 65,95\rangle$	$\langle 95,125\rangle$	$\langle 125,155\rangle$	$\langle 155,185\rangle$	$\langle 185,215\rangle$
Počet dom.	7	16	27	14	4	2

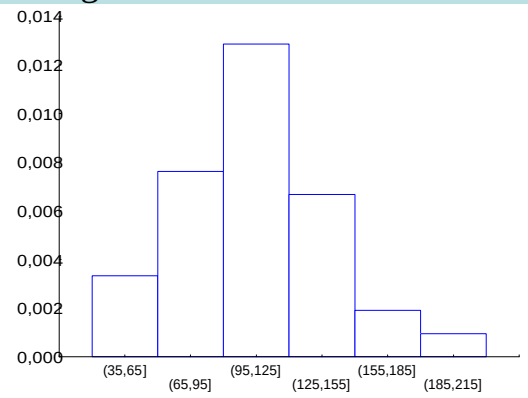
Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

Řešení:

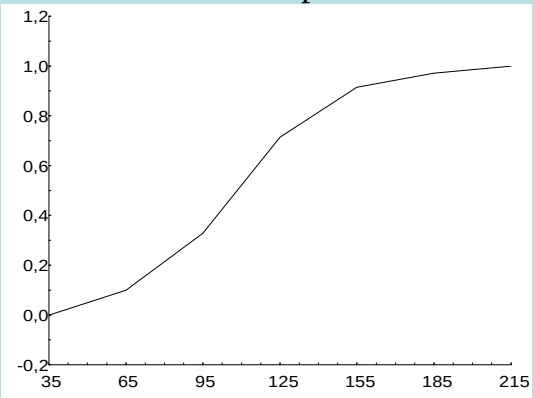
Tabulka rozložení četností

$(u_j,u_{j+1}]$	$x_{[j]}$	$d_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 35,65\rangle$	50	30	7	7/70	7	7/70	7/2100
$\langle 65,95\rangle$	80	30	16	16/70	23	23/70	16/2100
$\langle 95,125\rangle$	110	30	27	27/70	50	50/70	27/2100
$\langle 125,155\rangle$	140	30	14	14/70	64	64/70	14/2100
$\langle 155,185\rangle$	170	30	4	4/70	68	68/70	4/2100
$\langle 185,215\rangle$	200	30	2	2/70	70	1	2/2100

Histogram



Graf intervalové empirické distribuční funkce



Intervalové rozložení četností - dvourozměrný případ

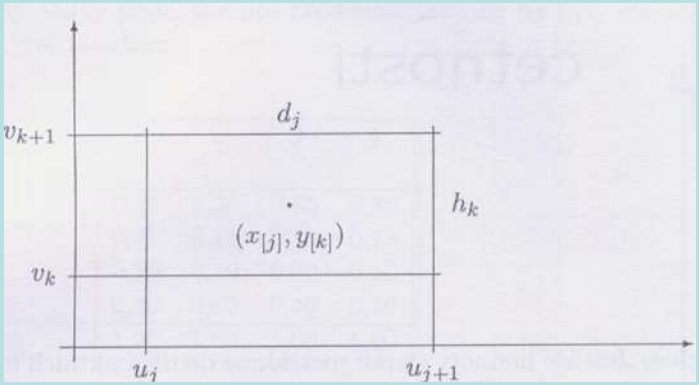
Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$.

Hodnoty znaku X roztřídíme do r třídicích intervalů $\langle u_j, u_{j+1} \rangle$, j = 1, ..., r s délkami $d_1, ..., d_r$,

hodnoty znaku Y roztřídíme do s třídicích intervalů $\langle v_k, v_{k+1} \rangle$, k = 1, ..., s s délkami $h_1, ..., h_s$.

Získáme tak r x s dvourozměrných třídicích intervalů.

Ilustrace dvourozměrného třídicícho intervalu



Pak definujeme:

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

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

$n_{j.} = n_{j1} + ... + n_{js}$ – 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.

$n_{.k} = n_{1k} + ... + n_{rk}$ – 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:

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

Příklad: U 60 náhodně vybraných manželských párů byl zjišťován průměrný čistý měsíční příjem (v Kč). Příjem manžela považujeme za znak X, příjem manželky za znak Y. Pro oba znaky X, Y najděte podle Sturgesova pravidla optimální počet třídicích intervalů a sestavte kontingenční tabulku simultánních absolutních četností.

příjem manžela	příjem manželky	příjem manžela	příjem manželky	příjem manžela	příjem manželky
16210	13710	31760	30250	24420	14640
30310	27960	38620	21980	15460	12800
33900	24930	27030	25410	37600	24200
40580	36720	43670	37540	42190	28650
19070	12940	45270	30580	15960	14500
29800	25810	39210	25470	18650	20210
26000	24590	14470	10550	26020	30150
37500	34810	23630	14820	23570	18840
21950	18860	15840	16340	20630	12760
19020	21530	25720	18700	31450	26840
17460	19870	17290	11560	19950	17960
13840	14320	18900	12080	16840	20900
29200	21200	47920	35620	16790	15740
14400	17300	29740	31420	26930	23980
15340	11930	13930	15790	46090	27960
23400	13220	25920	12870	22020	17400
18780	12760	21770	15980	31230	13580
33290	27140	17670	14320	20320	18490
31890	36970	19880	14800	19960	20500
18990	15470	14880	12680	36550	24360

Řešení:

Rozsah datového souboru je 60, tedy podle Sturgesova 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 13840, největší 47270; volba $u_1 = 13000, ..., u_8 = 48000$ splňuje požadavky. Délka třídicích intervalů: $d_j = 5000$.

Nyní vhodně stanovíme třídicí intervaly pro znak Y, tj. pro příjem manželky. Bude jich 7, stejně jako pro znak X.

Minimální hodnota je 10 550 Kč, maximální 37 550 Kč. Vhodná volba třídicích intervalů bude např. $v_1 = 10000, ..., 38000$.

Délka třídicích intervalů: $h_k = 4000$.

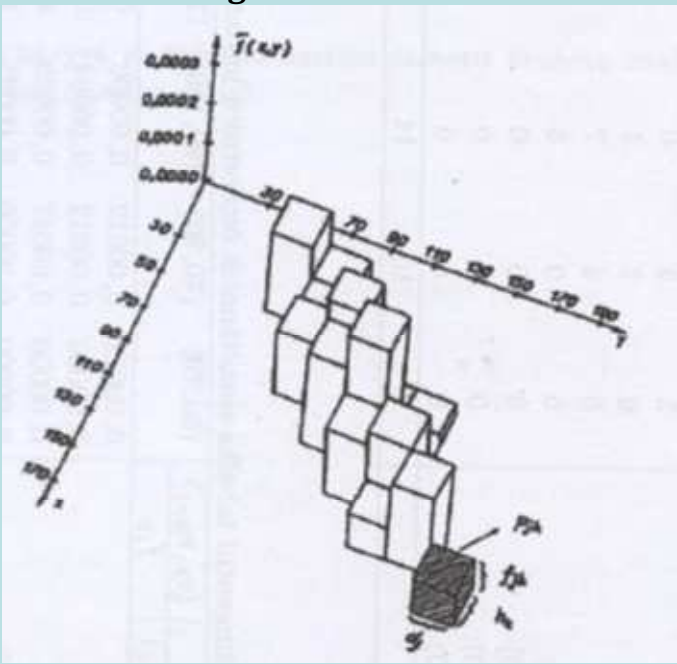
Kontingenční tabulka simultánních absolutních četností (symboly RX, RY označují středy třídicích intervalů):

RX	RY 12000	RY 16000	RY 20000	RY 24000	RY 28000	RY 32000	RY 36000	Řádk. součty
15500	6	7	2	0	0	0	0	15
20500	4	5	5	0	0	0	0	14
25500	2	2	2	3	0	1	0	10
30500	1	0	1	1	2	2	1	8
35500	0	0	0	3	1	0	1	5
40500	0	0	1	1	1	0	1	4
45500	0	0	0	0	1	1	2	4
Celková četn.	13	14	11	8	5	4	5	60

Z této kontingenční tabulky můžeme např. zjistit, že v našem výběrovém souboru je 6 manželských párů, kde muž má průměrný měsíční příjem mezi 13 000 Kč až 18 000 Kč a současně žena má průměrný měsíční příjem mezi 10 000 Kč až 14 000 Kč. Rovněž je patrno, že nenulové četnosti se vyskytují především kolem hlavní diagonály této kontingenční tabulky, tedy nízké (vysoké) příjmy manželů mají tendenci se vyskytovat společně s nízkými (vysokými) příjmy manželek.

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.

Příklad stereogramu:



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:

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

Pomocí simultánních a marginálních četnostních hustot 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 $\forall (x, y) \in R^2$: $f(x, y) = f_1(x)\, f_2(y)$.

Můžeme snadno ověřit, že v našem příkladu nejsou příjmy manželů a manželek četnostně nezávislé v daném výběrovém souboru při intervalovém rozložení četností:

RX	RY 12000	RY 16000	RY 20000	RY 24000	RY 28000	RY 32000	RY 36000	Řádk. součty
15500	6	7	2	0	0	0	0	15
20500	4	5	5	0	0	0	0	14
25500	2	2	2	3	0	1	0	10
30500	1	0	1	1	2	2	1	8
35500	0	0	0	3	1	0	1	5
40500	0	0	1	1	1	0	1	4
45500	0	0	0	0	1	1	2	4
Celková četn.	13	14	11	8	5	4	5	60

$$f_{11} = \frac{n_{11}}{n \cdot d_1 \cdot h_1} = \frac{6}{60 \cdot 5000 \cdot 4000} = 0{,}00025, f_{1.} = \frac{n_{1.}}{n \cdot d_1} = \frac{15}{60 \cdot 5000} = 0{,}00005, f_{.1} = \frac{n_{.1}}{n \cdot h_1} = \frac{13}{60 \cdot 4000} = 0{,}000542,$$

$$f_{1.} \cdot f_{.1} = 0{,}0000000027 \neq 0{,}00025 = f_{11}$$

Číselné charakteristiky znaků

Doposud jsme se zabývali funkcionálními charakteristikami znaků, jako jsou:

simultánní četnostní funkce $p(x,y)$,

marginální četnostní funkce $p_1(x)$, $p_2(y)$,

empirická distribuční funkce $F(x)$,

simultánní hustota 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 rys 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 rys poměrových znaků: Poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.
Mimo uvedenou klasifikaci stojí alternativní znaky, které nabývají jen dvou hodnot, např. 0,1, což znamená absenci a prezenci nějakého jevu. Například 0 bude znamenat neúspěch, l úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

Číselné charakteristiky nominálních znaků

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

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

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

Příklad na stanovení modu a výpočet mutability:

20 náhodně vybraných osob mělo odpovědět na otázku, který z pěti výrobků (označíme je A, B, C, D, E) preferují.

Výsledky máme v tabulce:

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

Stanovte modus a vypočtěte mutabilitu.

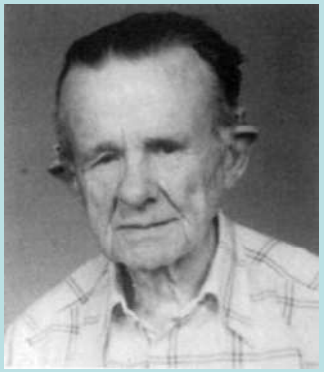
Řešení:

Modus = D

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

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

Charakteristika těsnosti závislosti dvou nominálních znaků: Cramérův koeficient kontingence.



Carl Harald Cramér (1893 – 1985): Švédský matematik

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

	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.

Upozornění: Pro tabulky typu 2 x 2 se Cramérův koeficient značí Φ a vzorec pro jeho výpočet se zjednoduší takto:

$\Phi = \sqrt{\frac{K}{n}}$.

Příklad na výpočet koeficientu Φ:

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 koeficient Φ.

Řešení:

Nejprve vypočteme teoretické četnosti:

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

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

Nyní dosadíme do vzorce pro výpočet statistiky K:

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

Nakonec vypočteme koeficient Φ:

$$\Phi=\sqrt{\frac{371{,}456}{686}}=0{,}7358$$

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

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se třemi proměnnými X (vlastnictví auta, 1 – ano, 2 – ne), Y (ochota používat MHD, 1 – ano, 2 – ne), četnost a o čtyřech případech:

	1 x	2 y	3 cetnost
1	1	1	56
2	1	2	312
3	2	1	283
4	2	2	35

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fí (tabulky 2 x 2) & Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

	Statist. : x(2) x y(2) (Tabulka1)		
Statist.	Chí-kvadr.	sv	p
Pearsonův chí-kv.	371,4530	df=1	p=0,0000
M-V chí-kvadr.	416,5616	df=1	p=0,0000
Fí pro tabulky 2 x 2	-,735851		
Tetrachorická korelace	-,917021		
Kontingenční koeficient	,5926815		

Vidíme, že koeficient Φ nabývá hodnoty 0,7358.

Číselné charakteristiky ordinálních znaků

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

Pro výpočet α-kvantilu slouží algoritmus:

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

Pro speciálně zvolená α užíváme názvů:

$x_{0,50}$ – medián (median),

$x_{0,25}$ – dolní kvartil (lower quartile),

$x_{0,75}$ – horní kvartil (upper quartile),

$x_{0,1}$, ..., $x_{0,9}$ – decily (deciles),

$x_{0,01}$, ..., $x_{0,99}$ – percentily (percentiles).

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

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

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

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

Řešení:

Pro snadnější výpočet tabulku doplníme ještě o absolutní kumulativní četnosti:

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

Rozsah souboru n = 50

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

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

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 žáků jsou aspoň tři čtvrtiny takových, kteří mají z matematiky dvojku či horší známku).

Výpočet pomocí systému STATISTICA:

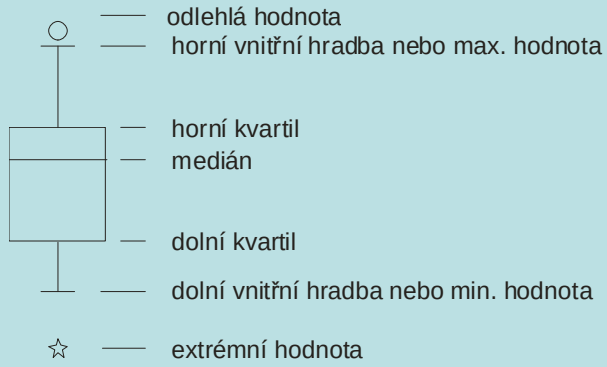
Vytvoříme nový datový soubor o pěti případech a dvou proměnných nazvaných X a četnost a vepíšeme zjištěné hodnoty. Statistika – Základní statistiky/tabulky – Popisné statistiky – Proměnné X – OK – klikneme na ikonu závaží – Proměnná vah četnost – OK – Stav Zapnuto – OK – Detailní výsledky – zaškrtneme Medián, Dolní a horní kvartily, Kvartil. rozpětí – Výpočet.

Proměnná	Medián	Dolní kvartil	Horní kvartil	Kvartilové rozpětí
X	3	2	3	1

Grafické znázornění ordinálních dat pomocí krabicového diagramu (box plot)

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

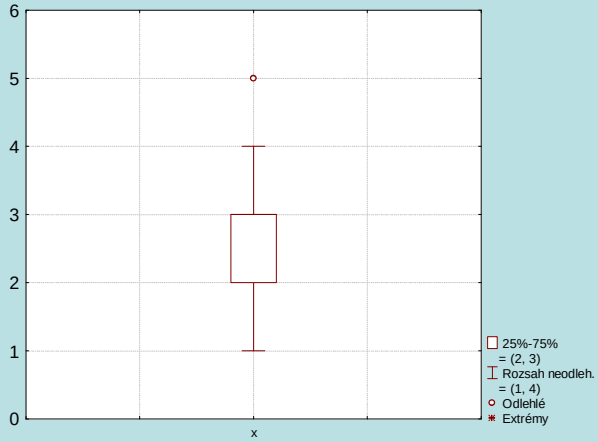
Způsob konstrukce



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

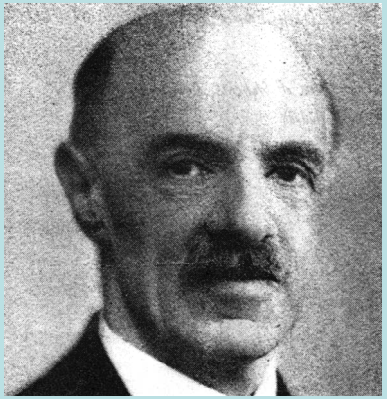
Příklad na konstrukci krabicového diagramu
Pro datový soubor známek z matematiky 50 žáků 7. ročníku ZŠ sestrojte krabicový diagram

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



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

Charakteristika těsnosti závislosti dvou ordinálních znaků: Spearmanův koeficient pořadové korelace (Spearman Rank Correlation Coefficient)



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

Nejprve je nutné vysvětlit pojem pořadí čísla v posloupnosti čísel.

Nechť $x_1, \ldots, x_n$ je posloupnost reálných čísel.

a) Jsou-li čísla navzájem různá, pak pořadím R_i čísla x_i rozumíme počet těch čísel $x_1, \ldots, x_n$, která jsou menší nebo rovna číslu x_i.

b) Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

Příklad na stanovení pořadí

a) Jsou dána čísla 9, 4, 5, 7, 3, 1.

b) Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.

Stanovte pořadí těchto čísel.

Řešení

ad a)

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

ad b)

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

Zavedení 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,

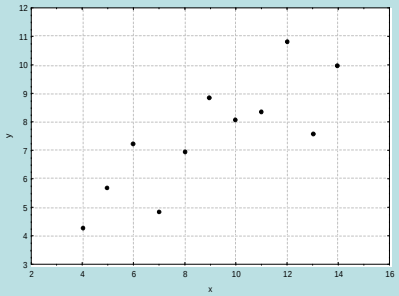
mezi 0,1 až 0,3 … slabá pořadová závislost,

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

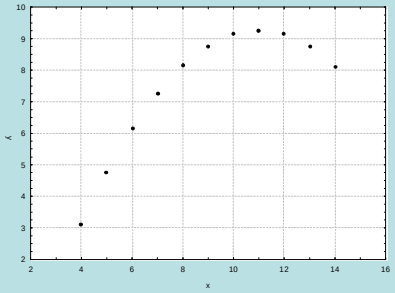
mezi 0,7 až 1 … silná pořadová závislost.

Ilustrace významu Spearmanova koeficientu pořadové korelace

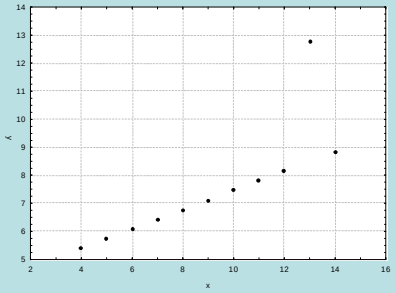
$r_S = 0{,}82$



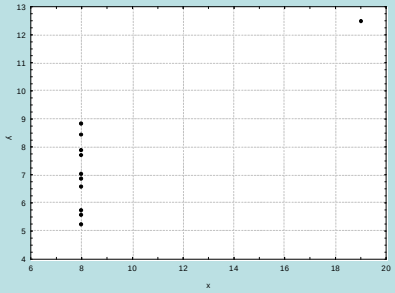
$r_S = 0{,}69$



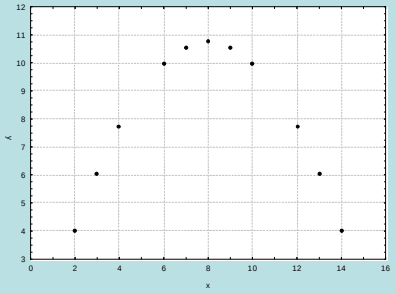
$r_S = 0{,}99$



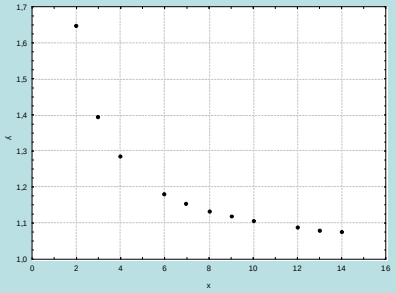
$r_S = 0{,}5$



$r_S = 0$



$r_S = -1$



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

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

Znamená to, že mezi znaky X a Y existuje slabá přímá pořadová závislost.
Výpočet pomocí systému STATISTICA:
Vytvoříme nový datový soubor o 5 případech a dvou proměnných X, Y.
Statistiky – Neparametrická statistika – Korelace – OK – Proměnné X, Y – OK – Spearman R.

Proměnná	X	Y
X	1,000000	0,300000
Y	0,300000	1,000000