

Jednorozměrné bodové rozložení četností

Jestliže počet variant znaku X v jednorozměrném datovém souboru není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o **bodovém rozložení četností**.

Nechť je dán jednorozměrný datový soubor $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, v němž znak X nabývá r variant.

Pro $j = 1, \dots, r$ definujeme:

$n_j = N(X = x_{[j]})$ – **absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru**

$p_j = \frac{n_j}{n}$ – **relativní četnost varianty $x_{[j]}$ ve výběrovém souboru**

$N_j = N(X \leq x_{[j]}) = n_1 + \dots + n_j$ – **absolutní kumulativní četnost prvních j variant ve výběrovém souboru**

$F_j = \frac{N_j}{n} = p_1 + \dots + p_j$ – **relativní kumulativní četnost prvních j variant ve výběrovém souboru**

Tabulka typu

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
\vdots	\vdots	\vdots	\vdots	\vdots
$x_{[r]}$	n_r	p_r	N_r	F_r

se nazývá variační řada (nebo též tabulka rozložení četností).

Příklad: Máme jednorozměrný datový soubor, který obsahuje údaje o známkách z matematiky (znak X) u 20 studentů.

$$\begin{pmatrix} 2 \\ 1 \\ 4 \\ 4 \\ 1 \\ 1 \\ 4 \\ 3 \\ 3 \\ 1 \\ 1 \\ 4 \\ 4 \\ 4 \\ 2 \\ 4 \\ 4 \\ 2 \\ 4 \\ 1 \\ 4 \\ 4 \\ 1 \end{pmatrix}$$

Sestavte tabulku rozložení četností.

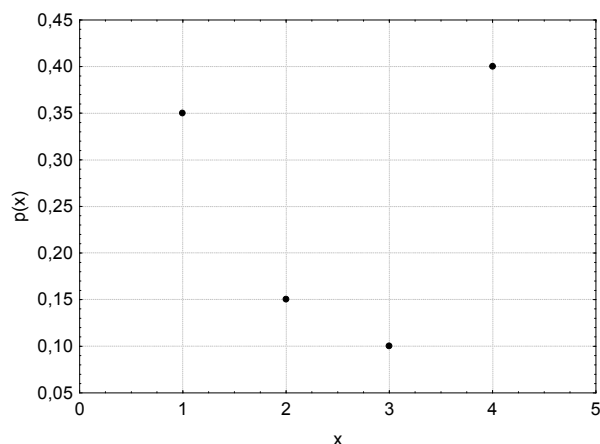
Řešení:

$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	0,35	7	0,35
2	3	0,15	10	0,50
3	2	0,10	12	0,60
4	8	0,40	20	1,00
-	20	1,00	-	-

Pomocí relativních četností zavedeme **četnostní funkci**.

Funkce $p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]}, j=1, \dots, r \\ 0 & \text{jinak} \end{cases}$ se nazývá četnostní funkce.

V našem případě má četnostní funkce tento graf:



Četnostní funkce je
nezáporná ($\forall x \in \mathbb{R}: p(x) \geq 0$)

a normovaná ($\sum_{x=-\infty}^{\infty} p(x) = 1$).

Pomocí kumulativních relativních četností zavedeme **empirickou distribuční funkci**.

Funkce $F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \leq x < x_{[j+1]}, j=1, \dots, r-1 \\ 1 & \text{pro } x \geq x_{[r]} \end{cases}$ se nazývá empirická distribuční funkce.

Empirická distribuční funkce je

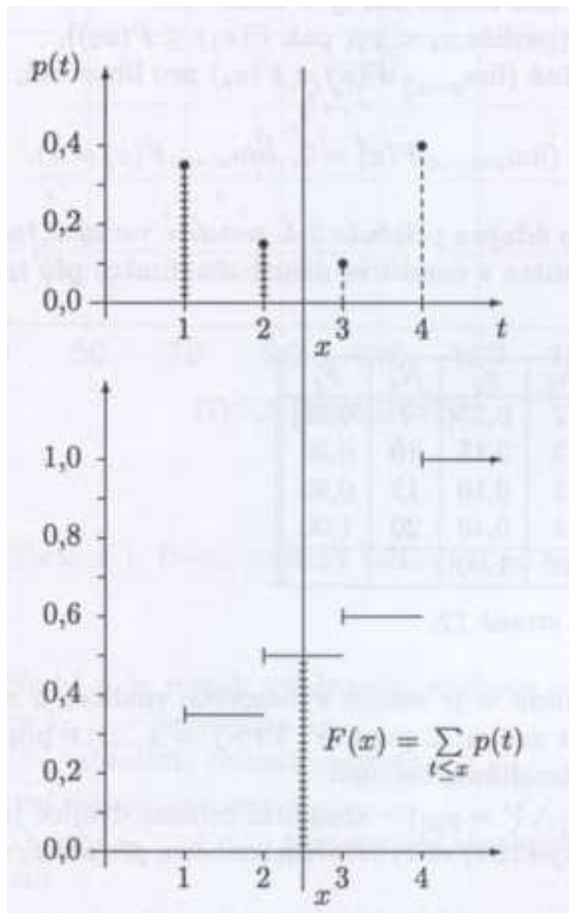
neklesající ($\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2: F(x_1) \leq F(x_2)$),

zprava spojitá ($\forall x_0 \in \mathbb{R}$ libovolné, ale pevně dané: $\lim_{x \rightarrow x_0+} F(x) = F(x_0)$)

a normovaná ($\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$).

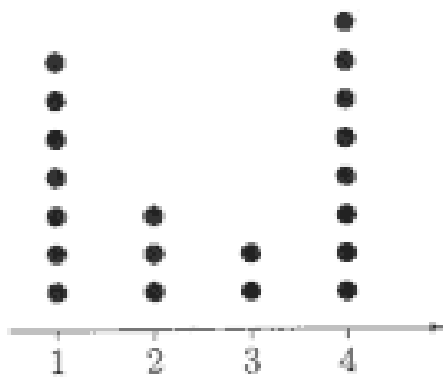
Platí $\forall x \in \mathbb{R}: F(x) = \sum_{t \leq x} p(t)$.

Vztah mezi četnostní funkcí a empirickou distribuční funkcí zachycuje obrázek:

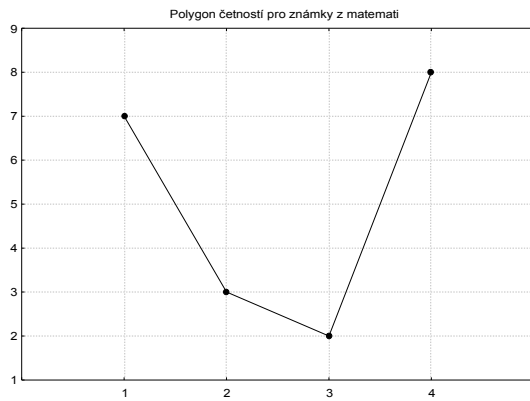


Existuje několik způsobů, jak graficky znázornit jednorozměrné bodové rozložení četností.

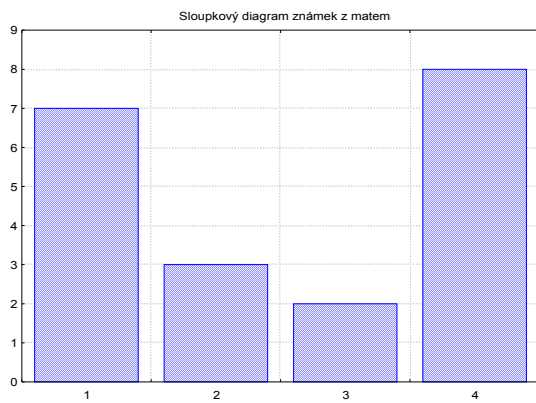
Tečkový diagram: na číselné ose vyznačíme jednotlivé varianty znaku X a nad každou variantu nakreslíme tolik teček, jaká je její absolutní četnost.



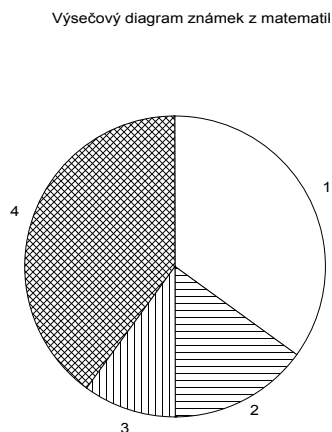
Polygon četnosti: je lomená čára spojující body, jejichž x-ová souřadnice je varianta znaku X a y-ová souřadnice je absolutní či relativní četnost této varianty.



Sloupkový diagram: je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní či relativní četnost této varianty.



Výsečový graf: je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním četnostem variant znaku X.



Dvourozměrné bodové rozložení četností

Nechť je dán dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$, kde znak X má r variant a znak Y má

s variant. Pak definujeme:

$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]})$ – **simultánní absolutní četnost dvojice $(x_{[j]}, y_{[k]})$** ve výběrovém souboru

$p_{jk} = \frac{n_{jk}}{n}$ – **simultánní relativní četnost dvojice $(x_{[j]}, y_{[k]})$** ve výběrovém souboru

$n_{.j} = N(X = x_{[j]}) = n_{j1} + \dots + n_{js}$ – **marginální absolutní četnost varianty $x_{[j]}$**

$p_{.j} = \frac{n_{.j}}{n} = p_{j1} + \dots + p_{js}$ – **marginální relativní četnost varianty $x_{[j]}$**

$n_{.k} = N(Y = y_{[k]}) = n_{1k} + \dots + n_{rk}$ – **marginální absolutní četnost varianty $y_{[k]}$**

$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \dots + p_{rk}$ – **marginální relativní četnost varianty $y_{[k]}$**

Simultánní četností zapisujeme do kontingenční tabulky.

Kontingenční tabulka simultánních absolutních četností má tvar:

	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

Příklad: Máme datový soubor, který obsahuje údaje o známkách z matematiky (znak X), z angličtiny (znak Y) a pohlaví studenta (znak Z, 0 – žena, 1 – muž) u 20 studentů:

(2	2	0)
	1	3	1	
	4	3	1	
	1	1	0	
	1	2	1	
	4	4	1	
	3	3	1	
	3	4	0	
	1	1	0	
	1	1	0	
	4	2	1	
	4	4	0	
	2	2	0	
	4	3	1	
	2	3	1	
	4	4	0	
	1	1	0	
	4	3	1	
	4	4	1	
	1	3	0)

Vytvořte kontingenční tabulku simultánních absolutních a relativních četností pro známky z matematiky a angličtiny.

Řešení:

Kontingenční tabulka simultánních absolutních četností

	y	1	2	3	4	$n_{j\cdot}$
x	n_{jk}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
$n_{\cdot k}$		4	4	7	5	$n = 20$

Kontingenční tabulka simultánních relativních četností

	y	1	2	3	4	$p_{j\cdot}$
x	p_{jk}					
1		0,20	0,05	0,10	0,00	0,35
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
$p_{\cdot k}$		0,20	0,20	0,35	0,25	1,00

Pomocí simultánních relativních četností zavedeme **simultánní četnostní funkci**:

Funkce $p(x, y) = \begin{cases} p_{jk} & \text{pro } x = x_{[j]}, y = y_{[k]}, j = 1, \dots, r, k = 1, \dots, s \\ 0 & \text{jinak} \end{cases}$ se nazývá simultánní četnostní funkce.

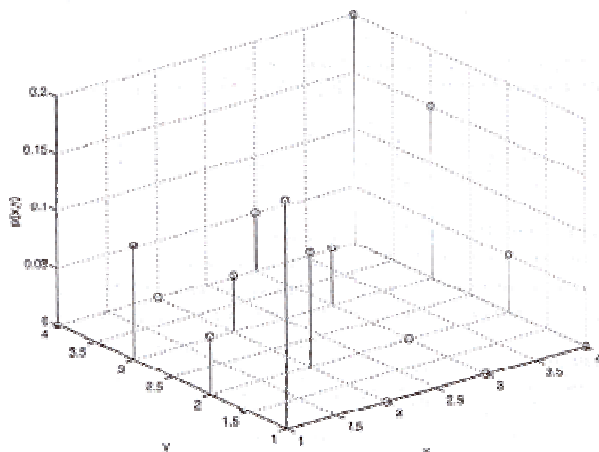
Pomocí marginálních relativních četností zavedeme **marginální četnostní funkce pro znaky X a Y**. Odlišíme je indexem takto:

$$p_1(x) = \begin{cases} p_{j\cdot} & \text{pro } x = x_{[j]}, j = 1, \dots, r \\ 0 & \text{jinak} \end{cases}, p_2(y) = \begin{cases} p_{\cdot k} & \text{pro } y = y_{[k]}, k = 1, \dots, s \\ 0 & \text{jinak} \end{cases}.$$

Mezi simultánní četnostní funkcí a marginálními četnostními funkcemi platí vztahy:

$$p_1(x) = \sum_{y=-\infty}^{\infty} p(x, y), p_2(y) = \sum_{x=-\infty}^{\infty} p(x, y).$$

Graf simultánní četnostní funkce pro známky z matematiky a angličtiny:



Pomocí simultánních a marginálních relativních četností zavedeme pojem **četnostní nezávislosti znaků v daném výběrovém souboru**:

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé, právě když pro všechna $j = 1, \dots, r$ a všechna $k = 1, \dots, s$ platí multiplikativní vztah: $p_{jk} = p_j \cdot p_k$ neboli pro $\forall (x, y) \in R^2$: $p(x, y) = p_1(x) p_2(y)$.

V našem příkladě nejsou známky z matematiky a angličtiny četnostně nezávislé, protože už pro $j = 1, k = 1$ je multiplikativní vztah porušen:

$$p_{11} = 0,20, p_{1.} = 0,35, p_{.1} = 0,20, \text{ tudíž } 0,20 \neq 0,35 \cdot 0,20$$

Nyní zavedeme řádkově a sloupcově podmíněné relativní četnosti.

$$p_{j(k)} = \frac{n_{jk}}{n_{.k}} - \text{sloupcově podmíněná relativní četnost varianty } x_{[j]} \text{ za předpokladu } y_{[k]}$$

$$p_{(j)k} = \frac{n_{jk}}{n_{j.}} - \text{řádkově podmíněná relativní četnost varianty } y_{[k]} \text{ za předpokladu } x_{[j]}$$

V našem příkladě vypočítáme nejprve sloupcově podmíněné relativní četnosti. Vyjdeme z kontingenční tabulky simultánních absolutních četností.

	y	1	2	3	4	$n_{j.}$
x	n_{jk}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
$n_{.k}$		4	4	7	5	$n = 20$

	y	1	2	3	4
x	$p_{j(k)}$				
1		1,00	0,25	0,29	0,00
2		0,00	0,50	0,14	0,00
3		0,00	0,00	0,14	0,20
4		0,00	0,25	0,43	0,80
Σ		1,00	1,00	1,00	1,00

Interpretujeme např. třetí sloupec: z těch studentů, kteří měli trojku z angličtiny, mělo $2/7 = 29\%$ jedničku z matematiky, $1/7 = 14\%$ dvojku z matematiky, $1/7 = 14\%$ trojku z matematiky a $3/7 = 43\%$ čtyřku z matematiky.

Nyní vypočítáme řádkově podmíněné relativní četnosti. Opět použijeme kontingenční tabulku simultánních absolutních četností.

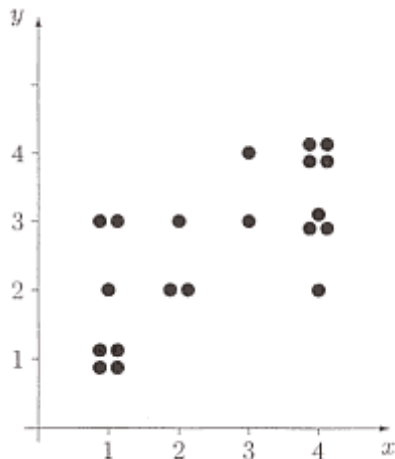
		y				n _{j.}
		1	2	3	4	
x	n _{j.k}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
n _{.k}		4	4	7	5	n = 20

		y				Σ
		1	2	3	4	
x	P _{(j)k}					
1		0,57	0,14	0,29	0,00	1,00
2		0,00	0,67	0,33	0,00	1,00
3		0,00	0,00	0,50	0,50	1,00
4		0,00	0,12	0,38	0,50	1,00

Interpretujeme např. první řádek: z těch studentů, kteří měli jedničku z matematiky, mělo $4/7 = 57\%$ jedničku z angličtiny, $1/7 = 14\%$ dvojku z angličtiny a $2/7 = 29\%$ trojku z angličtiny.

Dvourozměrné rozložení četností lze znázornit pomocí **dvourozměrného tečkového diagramu**. Na vodorovnou osu vyneseme varianty znaku X, na svislou varianty znaku Y a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dané dvojice.

V našem příkladě se studenty dostaneme tento diagram:



Dvourozměrný tečkový diagram svědčí o nepříliš výrazné tendenci k podobné klasifikaci v obou předmětech.

Intervalové rozložení četností

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

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

(u_j, u_{j+1}) – **j -tý třídící interval znaku X** , $j = 1, \dots, r$.

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

$x_{[j]} = \frac{u_j + u_{j+1}}{2}$ – **střed j -tého třídícího intervalu znaku X**

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

Hodnoty znaku X roztřídíme do r třídících intervalů. Pro $j = 1, \dots, r$ definujeme:

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

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

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

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

$F_j = \frac{N_j}{n} = p_1 + \dots + p_j$ – **relativní kumulativní četnost prvních j třídících 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
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
(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	157	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
113	125	107	118	136	155
85	97	98	140	82	81
41	73	97	115	136	153
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	130	125	149	85	91

- a) Pro znak X stanovte optimální počet třídících intervalů dle Sturgesova pravidla.
b) Sestavte tabulku rozložení četností.

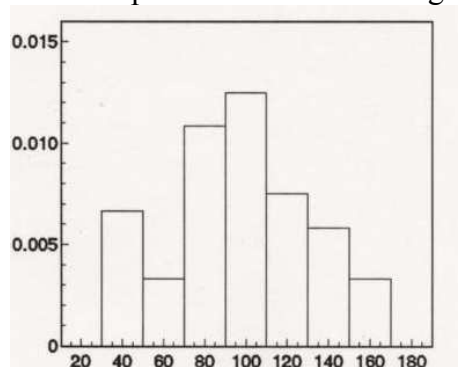
Řešení:

ad a) Rozsah souboru je 60. Podle Sturgesova pravidla je optimální počet třídících 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 $u_1 = 30, \dots, u_8 = 170$ splňuje požadavky.

ad b)

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

V našem příkladě bude mít histogram pro mez plasticity oceli tvar:



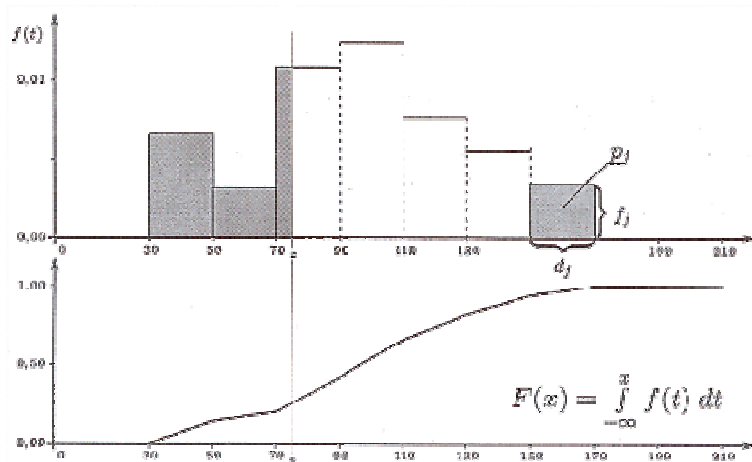
Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané **hustota četnosti**:

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

Pomocí hustoty četnosti zavedeme **intervalovou empirickou distribuční funkci**:

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

Vztah mezi hustotou četnosti a intervalovou empirickou distribuční funkcí je znázorněn na obrázku:



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

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

třídících intervalů (u_j, u_{j+1}) , $j = 1, \dots, r$ s délkami d_1, \dots, d_r a hodnoty znaku Y roztřídíme do s třídících intervalů (v_k, v_{k+1}) , $k = 1, \dots, s$ s délkami h_1, \dots, h_s . 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 třídícího intervalu**.

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

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

$p_j = \frac{n_j}{n}$ – marginální relativní četnost j-tého třídícího intervalu pro znak X.

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

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

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

$f_j = \frac{p_j}{d_j}$ – marginální četnostní hustota v j-tém třídícím intervalu pro znak X.

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

Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky.

Uvedme kontingenční tabulku simultánních absolutních četností:

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

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

- stanovte dle Sturgesova pravidla optimální počet třídících intervalů pro znak Y
- sestavte kontingenční tabulku simultánních absolutních četností.

Řešení:

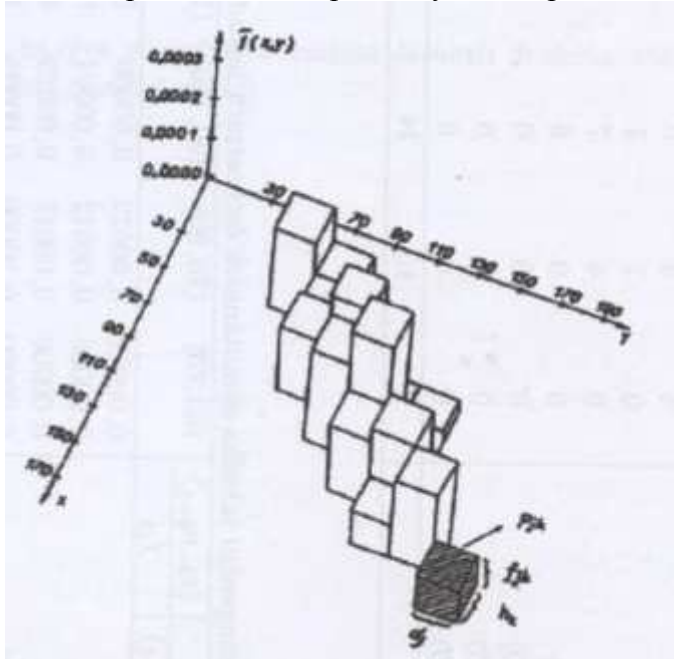
ad a) Rozsah datového souboru je 60. Podle Sturgesova pravidla je tedy optimální počet třídících intervalů $s = 7$. Nejmenší hodnota je 52 a největší 189. Volíme $v_1 = 50, v_2 = 70, \dots, v_8 = 190$.

ad b)

	(v_k, v_{k+1})	(50, 70)	(70, 90)	(90, 110)	(110, 130)	(130, 150)	(150, 170)	(170, 190)	
(u_j, u_{j+1})	n_{jk}								$n_{j.}$
(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	$n = 60$

Dvourozměrné intervalové rozložení četností graficky znázorňujeme pomocí **stereogramu**. Je to graf skládající se z $r \times s$ kvádrů, sestavených nad dvourozměrnými třídícími intervaly, přičemž objem (j, k)-tého kvádrů je roven relativní četnosti p_{jk} (j, k)-tého třídícího intervalu, $j = 1, \dots, r, k = 1, \dots, s$. Výška kvádrů tedy vyjadřuje simultánní četnostní hustotu.

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



Pomocí simultánních četnostních hustot zavedeme **simultánní hustotu četnosti**:

Funkce $f(x, y) = \begin{cases} f_{jk} \text{ pro } u_j < x \leq u_{j+1}, v_k < y \leq v_{k+1}, j=1, \dots, r, k=1, \dots, s \\ 0 \text{ jinak} \end{cases}$ se nazývá simul-

tánní hustota četnosti. Jejím grafem je schodovitá plocha shora omezující stereogram.

Hustoty četnosti pro znaky X a Y odlišíme indexem takto:

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

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

Mezi simultánní hustotou četnosti a marginálními hustotami četnosti platí vztahy:

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

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 $j = 1, \dots, r$ a všechna $k = 1, \dots, s$ platí multiplikační vztah: $f_{jk} = f_{.j} \cdot f_{.k}$ neboli pro $\forall (x, y) \in \mathbb{R}^2$: $f(x, y) = f_1(x) f_2(y)$.

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

$$f_{11} = \frac{5}{60 \cdot 20 \cdot 20} = 0,000208, \quad f_{.1} = \frac{8}{60 \cdot 20} = 0,006667, \quad f_{.1} = \frac{5}{60 \cdot 20} = 0,004167, \text{ tudíž}$$

$$0,000208 \neq 0,006667 \cdot 0,004167 = 0,000028$$