Popisná statistika

Popisná statistika je disciplína, která popisuje a sumarizuje informace obsažené ve velkém množství dat pomocí tabulek, grafů, funkcionálních a číselných charakteristik. Činí tak pomocí základních matematických operací. Cílem popisné statistiky je zpřehlednit informace „ukryté´" v datových souborech.

Popisná statistika je velmi důležitá minimálně ze dvou důvodů:

- v praxi se často používá (všichni znají takové pojmy, jako je průměr, směrodatná ochylka, tabulka rozložení četností, výsečový graf apod.)
- motivuje pojmy, se kterými pak pracuje počet pravděpodobnosti (např. relativní četnost motivuje pravděpodobnost, hustota četnosti motivuje hustotu pravděpodobnosti, průměr motivuje střední hodnotu apod.)

Dobré pochopení pojmů popisné statistiky tedy velmi usnadní studium počtu pravděpodobnosti.

Základní, výběrový a datový soubor

Základním souborem rozumíme libovolnou neprázdnou množinu E.

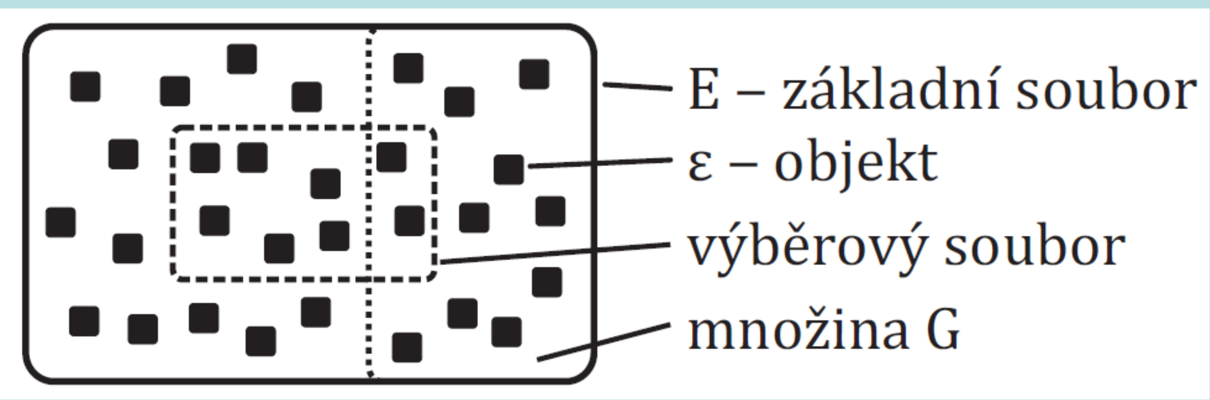
Prvky množiny E značíme ε a nazýváme je objekty.

Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, ..., \varepsilon_n\}$ základního souboru E nazýváme výběrový soubor rozsahu n.

Je-li množina $G \subset E$, pak symbolem N(G) rozumíme absolutní četnost množiny G ve výběrovém souboru, tj. počet těch objektů množiny G, které patří do výběrového souboru.

Relativní četnost množiny G ve výběrovém souboru zavedeme vztahem $p(G) = \frac{N(G)}{n}$.

Ilustrace



Příklad: Základním souborem E je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, ..., \varepsilon_{20}\}$. Z těchto 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

Řešení:

$$N(G_1) = 12, \quad N(G_2) = 15, \quad N(G_1 \cap G_2) = 11, \quad n = 20,$$

$$p(G_1) = \frac{12}{20} = 0{,}6,$$

$$p(G_2) = \frac{15}{20} = 0{,}75,$$

$$p(G_1 \cap G_2) = \frac{11}{20} = 0{,}55$$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

Vlastnosti relativní četnosti: Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

- $p(\emptyset) = 0$
- $p(G) \geq 0$ (nezápornost)
- $p(G) \leq 1$
- $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$
- $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$
- $p(G_1 \cup G_2) + 0 \leq p(G_1) + p(G_2)$ (subaditivita)
- $G_1 \cap G_2 = \emptyset \Rightarrow p(G_1 \cup G_2) = p(G_1) + p(G_2)$ (aditivita)
- $p(G_2 \setminus G_1) = p(G_2) - p(G_1 \cap G_2)$
- $G_1 \subset G_2 \Rightarrow p(G_2 \setminus G_1) = p(G_2) - p(G_1)$ (subtraktivita)
- $G_1 \subset G_2 \Rightarrow p(G_1) \leq p(G_2)$ (monotonie)
- $p(E) = 1$ (normovanost)
- $p(G) + p(\bar{G}) = 1$ (komplementarita)

Pojem podmíněné relativní četnosti: Pokud se v daném základním souboru zajímáme o dvě podmnožiny, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé podmnožiny.

Nechť E je základní soubor, G_1, G_2 jeho podmnožiny, $\{\varepsilon_1, ..., \varepsilon_n\}$ výběrový soubor. Definujeme:

podmíněnou relativní četnost množiny G_1 ve výběrovém souboru za předpokladu G_2:

$$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)},$$

podmíněnou relativní četnost G_2 ve výběrovém souboru za předpokladu G_1:

$$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}.$$

Příklad: Pro údaje z příkladu o studentech vypočtěte podmíněnou relativní četnost úspěšných matematiků mezi úspěšnými angličtináři a podmíněnou relativní četnost úspěšných angličtinářů mezi úspěšnými matematiky.
(Připomínáme, že z 20 studentů 12 uspělo v matematice, 15 v angličtině a 11 v obou předmětech.)

Řešení:

$N(G_1) = 12,\ N(G_2) = 15,\ N(G_1 \cap G_2) = 11,\ n = 20$

$p(G_1/G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{11}{15} = 0{,}73$ (tzn., že 73% těch studentů, kteří byli úspěšní v angličtině, uspělo i v matematice)

$p(G_2/G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{11}{12} = 0{,}92$ (tzn., že 92% těch studentů, kteří byli úspěšní v matematice, uspělo i v angličtině)

Pojem četnostní nezávislosti dvou množin: O četnostní nezávislosti dvou množin v daném výběrovém souboru hovoříme tehdy, když informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ i z druhé množiny.

V příkladě se studenty by množiny úspěšných matematiků a úspěšných angličtinářů byly četnostně nezávislé, pokud podíl úspěšných matematiků mezi úspěšnými angličtináři by byl stejný jako podíl úspěšných matematiků mezi všemi zkoušenými studenty a stejně tak podíl úspěšných angličtinářů mezi úspěšnými matematiky by byl stejný jako podíl úspěšných angličtinářů mezi všemi zkoušenými studenty, tj.

$$\frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{N(G_1)}{n} \wedge \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{N(G_2)}{n}.$$

Po snadné úpravě dostaneme multiplikativní vztah

$$\frac{N(G_1 \cap G_2)}{n} = \frac{N(G_1)}{n} \cdot \frac{N(G_2)}{n}, \text{ tj. } p(G_1 \cap G_2) = p(G_1)p(G_2)$$

Řekneme tedy, že množiny G_1, G_2 jsou četnostně nezávislé v daném výběrovém souboru, jestliže $p(G_1 \cap G_2) = p(G_1)p(G_2)$.

(V praxi jen zřídka dojde k tomu, že uvedený vztah platí přesně. Většinou je jen naznačena určitá tendence četnostní nezávislosti.)

Příklad: Pro údaje z příkladu o studentech zjistěte, zda úspěchy v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé.

Řešení:

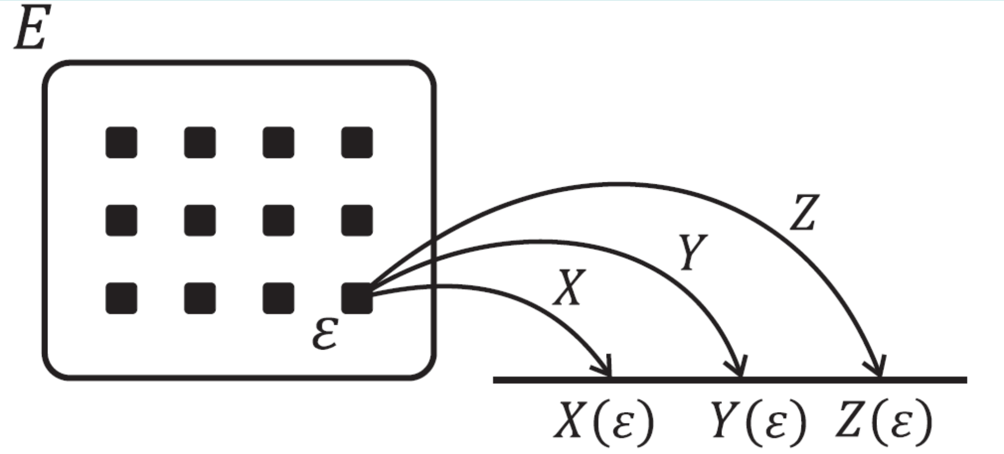
$p(G_1 \cap G_2) = 0{,}55$, $p(G_1)p(G_2) = 0{,}6 \times 0{,}75 = 0{,}45$,

tedy skutečná relativní četnost oboustranně úspěšných studentů je větší než by odpovídalo četnostní nezávislosti množin G_1, G_2 v daném výběrovém souboru. Znamená to, že úspěch v matematice se zpravidla sdružuje s úspěchem v angličtině a naopak.

Pojem skalárního a vektorového znaku: Vlastnosti objektů vyjadřujeme číselně pomocí znaků.

Nechť E je základní soubor. Funkce X: E → R, Y: E → R, ..., Z: E → R, které každému objektu přiřazují číslo, se nazývají (skalární) znaky. Uspořádaná p-tice (X, Y, ..., Z) se nazývá vektorový znak.

Ilustrace



Označení: Nechť je dán výběrový soubor $\{\varepsilon_1, ..., \varepsilon_n\} \subset$ E. Hodnoty znaků X, Y, ..., Z pro i-tý objekt označíme

$x_i = X(\varepsilon_i)$, $y_i = Y(\varepsilon_i)$, ..., $z_i = Z(\varepsilon_i)$, i = 1, ..., n.

Pojem datového souboru:

Matice $\begin{pmatrix} x_1 & y_1 & \cdots & z_1 \\ x_2 & y_2 & \cdots & z_2 \\ \ldots & \ldots & \ldots & \ldots \\ x_n & y_n & \cdots & z_n \end{pmatrix}$ typu n × p se nazývá datový soubor. Její řádky odpovídají jednotlivým objektům, sloupce znakům.

Libovolný sloupec této matice nazýváme jednorozměrným datovým souborem.

Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme uspořádaný datový soubor $\begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{pmatrix}$, kde $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Vektor $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$, kde $x_{[1]} < \ldots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá vektor variant.

Příklad: Pro studenty z výběrového souboru uvedeného výše byly zjišťovány hodnoty znaků X – známka z matematiky v prvním zkušebním termínu, Y – známka z angličtiny v prvním zkušebním termínu, Z – pohlaví studenta (0 ... žena, 1 ... muž). Byl získán datový soubor

$$\begin{pmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{pmatrix}$$

Utvořte jednorozměrný uspořádaný i neuspořádaný datový soubor pro známky z matematiky a vektor variant pro známky z matematiky.

Řešení:

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

Pojem jevu:

Nechť $\{\varepsilon_1, ..., \varepsilon_n\}$ je výběrový soubor, X, Y, ..., Z jsou znaky, B, B_1, B_p jsou číselné množiny.

Zápis $\{X \in B\}$ znamená jev „znak X nabyl hodnoty z množiny B“ .

Zápis $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ znamená jev „znak X nabyl hodnoty z množiny B_1 a současně znak Y nabyl hodnoty z množiny B_2 atd. až znak Z nabyl hodnoty z množiny B_p“.

Symbol $N(X \in B)$ značí absolutní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. počet těch objektů ve výběrovém souboru, pro něž $x_i \in B$.

Symbol $p(X \in B)$ znamená relativní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj. $p(X \in B) = \frac{N(X \in B)}{n}$.

Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ resp.

$p(X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge ... \wedge Z \in B_p\}$ ve výběrovém souboru.

Příklad: Pro datový soubor s údaji o známkách najděte relativní četnost

a) matematických jedničkářů

b) úspěšných matematiků

c) oboustranně neúspěšných studentů.

Datový soubor má tvar:

$$\begin{pmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{pmatrix}$$

Řešení:

ad a) $p(X = 1) = \frac{7}{20} = 0{,}35;$

ad b) $p(X \leq 3) = \frac{12}{20} = 0{,}60;$

ad c) $p(X = 4 \wedge Y = 4) = \frac{4}{20} = 0{,}20.$

Zjistili jsme, že jedničku z matematiky mělo 35% studentů, zkoušku z matematiky úspěšně složilo 60% studentů a oboustranně neúspěšných bylo 20% studentů.

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, ..., r definujeme:

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

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

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

$F_j = \dfrac{N_j}{n} = p_1 + ... + p_j$ – 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
:	:	:	:	:
$x_{[r]}$	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.

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

Sestavte tabulku rozložení četností.

Řešení:

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

Četnostní funkce, empirická distribuční funkce

Pomocí relativních četností zavedeme četnostní funkci.

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

Četnostní funkce je

nezáporná ($\forall x \in R$: $p(x) \geq 0$)

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

Pomocí kumulativních relativních četností zavedeme empirickou distribuční funkci.

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

Empirická distribuční funkce je

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

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

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

Příklad: Pro známky z matematiky nakreslete graf četnostní funkce a empirické distribuční funkce.

Řešení:

Variační řada

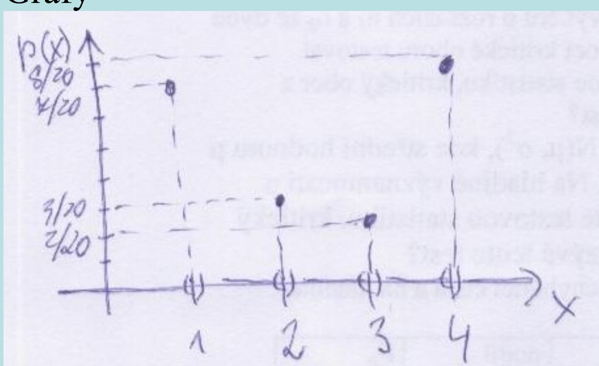
$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	7/20=0,35	7	7/20=0,35
2	3	3/20=0,15	10	10/20=0,50
3	2	2/20=0,10	12	12/20=0,60
4	8	8/20=0,40	20	20/20=1,00
∑	20	1,00	-	-

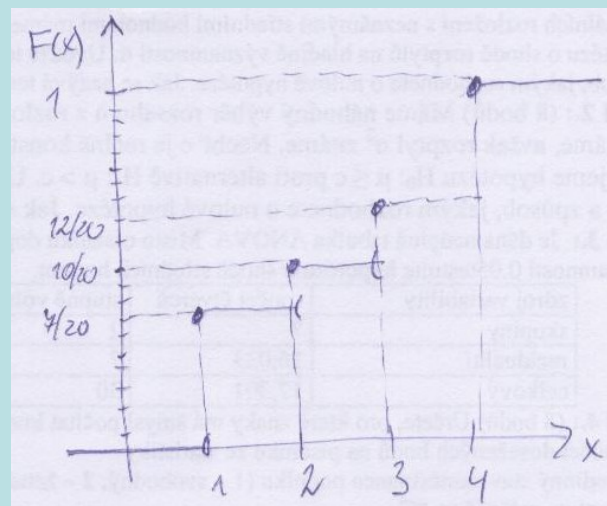
Vzorce

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

$$F(x)=\begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \le x < x_{[j+1]},\ j=1,\dots,r-1 \\ 1 & \text{pro } x \ge x_{[r]} \end{cases}$$

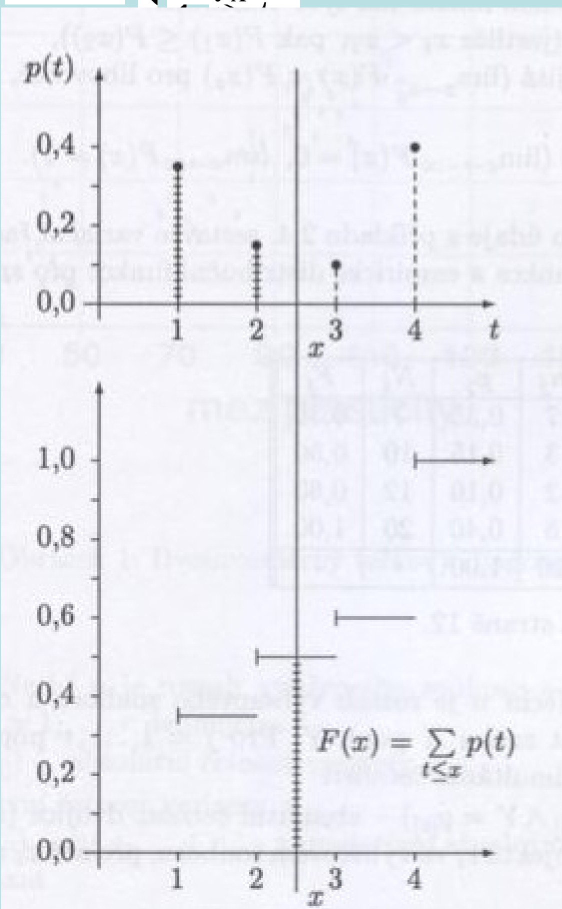
Grafy





Vztah mezi četnostní funkcí a empirickou distribuční funkcí

$\forall x \in R: F(x) = \sum_{t \le x} p(t)$



Grafické znázornění bodového 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.

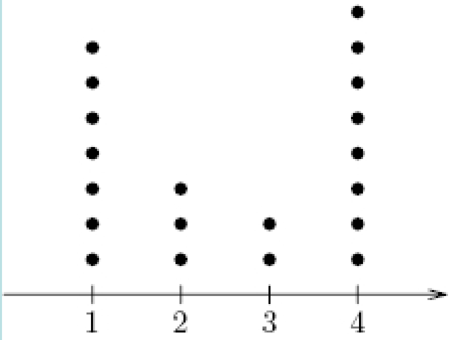
Sloupkový diagram: je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní či relativní četnost této varianty.

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

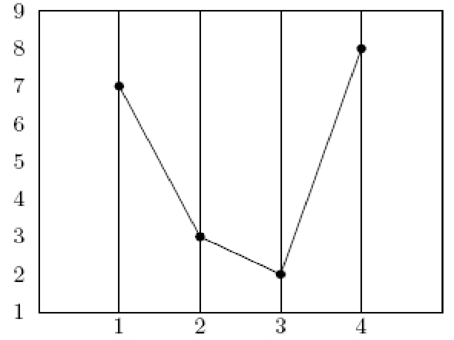
Příklad: Pro jednorozměrný datový soubor známek z matematiky sestrojte tečkový diagram, polygon četností, sloupkový diagram a výsečový graf.

Řešení:

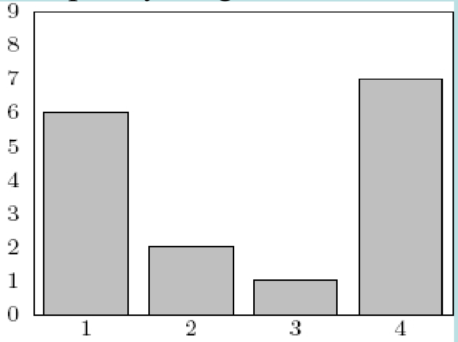
Tečkový diagram



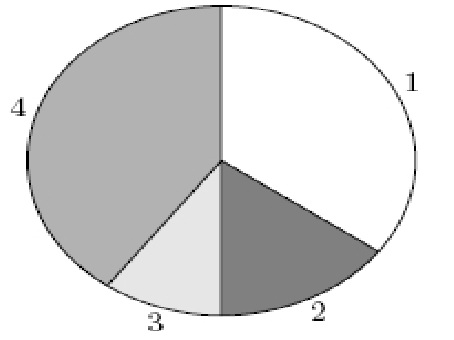
Polygon četností



Sloupkový diagram



Výsečový graf



Dvourozměrné bodové rozložení četností

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

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

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

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

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

Simultánní četností zapisujeme do kontingenční tabulky.

Kontingenční tabulka simultánních absolutních četností má tvar:

x \ y (n_{jk})	$y_{[1]}$	...	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	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:

X	2	1	4	1	1	4	3	3	1	1	4	4	2	4	2	4	1	4	4	1
Y	2	3	3	1	2	4	3	4	1	1	2	4	2	3	3	4	1	3	4	3
Z	0	1	1	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	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	$n_{j.}$
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	$n = 20$

Kontingenční tabulka simultánních relativních četností

	y	1	2	3	4	$p_{j.}$
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
$p_{.k}$		0,20	0,20	0,35	0,25	1,00

Simultánní a marginální četnostní funkce

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:

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

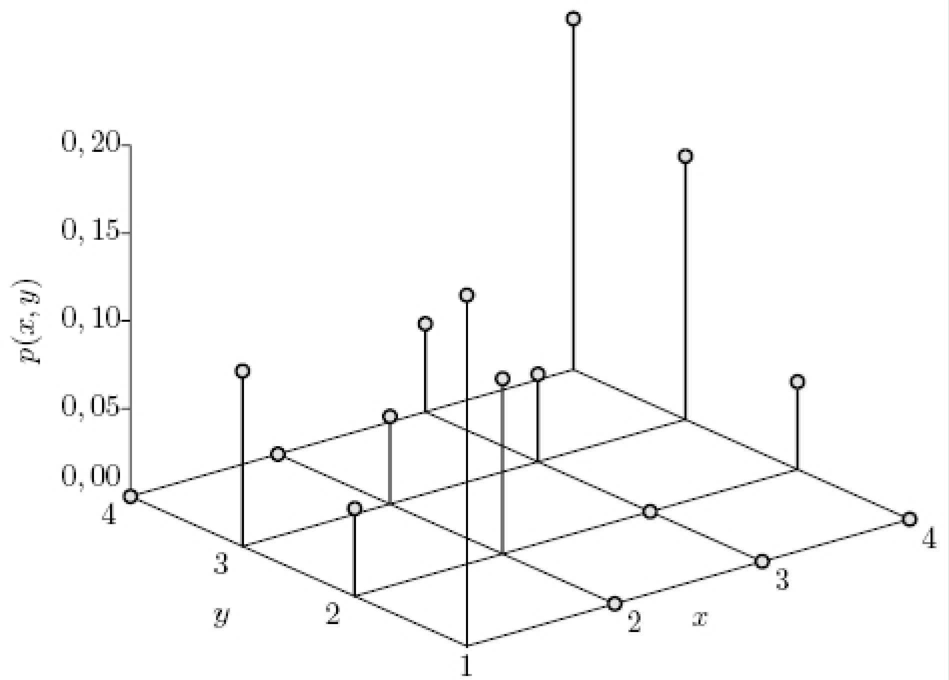
Mezi simultánní četnostní funkcí a marginálními četnostními funkcemi platí vztahy:

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

Příklad: Sestrojte graf simultánní četnostní funkce pro známky z matematiky a angličtiny.

Řešení: Vyjdeme z kontingenční tabulky simultánních relativních četností.

	y	1	2	3	4	$p_{j.}$
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
$p_{.k}$		0,20	0,20	0,35	0,25	1,00



Četnostní nezávislost znaků v daném výběrovém souboru

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé, právě když

pro všechna j = 1, ..., r a všechna k = 1, ..., s platí multiplikativní vztah: $p_{jk} = p_{j.}\, p_{.k}$

neboli pro $\forall (x, y) \in R^2$: $p(x, y) = p_1(x)\, p_2(y)$.

Příklad: Ověřte, zda v našem datovém souboru jsou známky z matematiky a angličtiny četnostně nezávislé.

Řešení: Vyjdeme z kontingenční tabulky simultánních relativních četností:

	y	1	2	3	4	$p_{j.}$
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
$p_{.k}$		0,20	0,20	0,35	0,25	1,00

Známky z matematiky a angličtiny nejsou četnostně nezávislé, protože už pro j = 1, k = 1 je multiplikativní vztah porušen:

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

Řádkově a sloupcově podmíněné relativní četnosti

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

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

Podmíněné relativní četnosti zapisujeme do kontingenční tabulky. Často je vyjadřujeme v procentech.

Příklad: Pro datový soubor známek z matematiky a angličtiny sestavte kontingenční tabulku sloupcově a poté řádkově podmíněných relativních četností.

Řešení:

Nejprve vypočítáme sloupcově podmíněné relativní četnosti. Použijeme kontingenční tabulku 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
$\sum$		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	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	$n = 20$

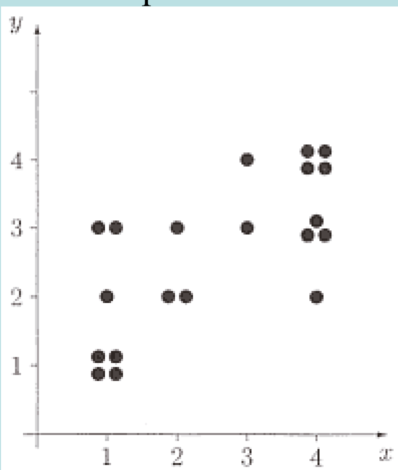
	y	1	2	3	4	$\sum$
x	$p_{(j)k}$					
1		0,57	0,14	0,29	0,00	1,00
2		0,00	0,67	0,33	0,00	1,00
3		0,00	0,00	0,50	0,50	1,00
4		0,00	0,12	0,38	0,50	1,00

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.

Dvourozměrný tečkový diagram

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.

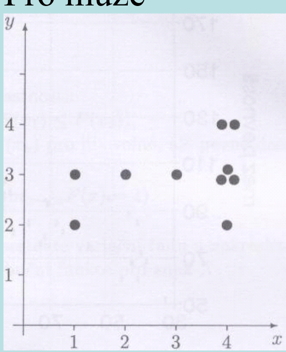
V našem příkladě se studenty dostaneme tento diagram:



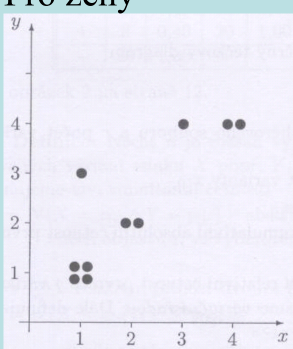
Dvourozměrný tečkový diagram svědčí o nepříliš výrazné tendenci k podobné klasifikaci v obou předmětech.

Zcela odlišný vzhled však mají diagramy pro muže a pro ženy:

Pro muže



Pro ženy



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 četnosti.

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

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

$d_j = u_{j+1} - u_j$ – délka j-tého třídicího intervalu znaku X, $x_{[j]} = \frac{u_j + u_{j+1}}{2}$ – střed j-tého třídicího intervalu znaku X

Třídicí 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 třídicích intervalů. Pro j = 1, ..., r definujeme:

$n_j = N(u_j < X \le u_{j+1})$ – 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

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

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

Tabulka typu

$(u_j, u_{j+1}\rangle$	d_j	n_j	p_j	f_j	N_j	F_j
$(u_1, u_2\rangle$	d_1	n_1	p_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(u_r, u_{r+1}\rangle$	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 Sturgesova pravidla.
b) Sestavte tabulku rozložení četností.

Řešení:

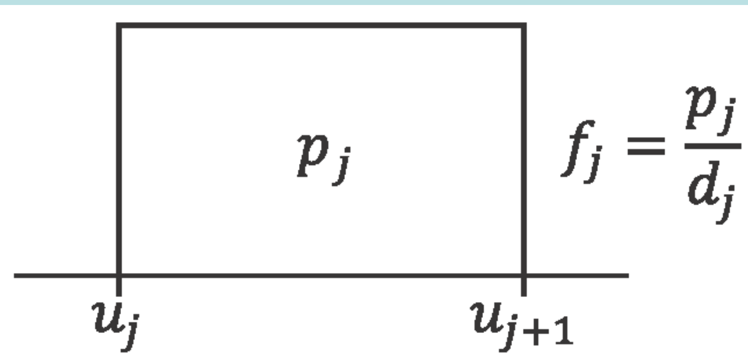
ad a) Rozsah souboru je 60. Podle Sturgesova pravidla je optimální počet třídicích intervalů $r = 7$. Budeme tedy volit 7 intervalů stejné délky tak, aby v nich byly obsaženy všechny pozorované hodnoty znaku X, z nichž nejmenší je 33, největší 160; volba $u_1 = 30$, ..., $u_8 = 170$ splňuje požadavky.

ad b)

$\langle u_j, u_{j+1})$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 30, 50)$	20	40	8	8/60 = [illegible]	8	8/60 = [illegible]	8/(60.20) = [illegible]
$\langle 50, 70)$	20	60	4	4/60 = [illegible]	12	12/60 = [illegible]	4/(60.20) = [illegible]
$\langle 70, 90)$	20	80	13	13/60 = [illegible]	25	25/60 = [illegible]	13/(60.20) = [illegible]
$\langle 90, 110)$	20	100	15	15/60 = [illegible]	40	40/60 = [illegible]	15/(60.20) = [illegible]
$\langle 110, 130)$	20	120	9	9/60 = [illegible]	49	49/60 = [illegible]	9/(60.20) = [illegible]
$\langle 130, 150)$	20	140	7	7/60 = [illegible]	56	56/60 = [illegible]	7/(60.20) = [illegible]
$\langle 150, 170)$	20	160	4	4/60 = [illegible]	60	60/60 = [illegible]	4/(60.20) = [illegible]
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:

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

Pomocí hustoty četnosti zavedeme intervalovou empirickou distribuční funkci:

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

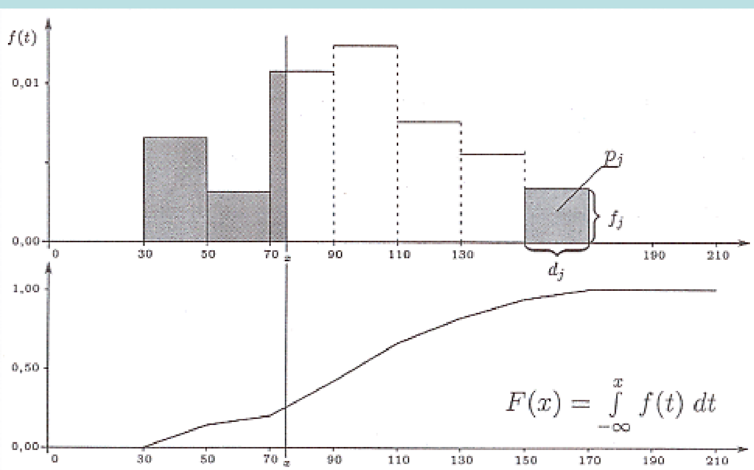
Hustota četnosti je nezáporná ($\forall x \in \mathbb{R}: f(x) \ge 0$) a normovaná ($\int_{-\infty}^{\infty} f(x)\,dx = 1$).

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.

$\langle u_j, u_{j+1})$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$\langle 30, 50)$	20	40	8	8/60 [illegible]	8	8/60 [illegible]	8/(60.20) [illegible]
$\langle 50, 70)$	20	60	4	4/60 [illegible]	12	12/60 [illegible]	4/(60.20) [illegible]
$\langle 70, 90)$	20	80	13	13/60 [illegible]	25	25/60 [illegible]	13/(60.20) [illegible]
$\langle 90, 110)$	20	100	15	15/60 [illegible]	40	40/60 [illegible]	15/(60.20) [illegible]
$\langle 110, 130)$	20	120	9	9/60 [illegible]	49	49/60 [illegible]	9/(60.20) [illegible]
$\langle 130, 150)$	20	140	7	7/60 [illegible]	56	56/60 [illegible]	7/(60.20) [illegible]
$\langle 150, 170)$	20	160	4	4/60 [illegible]	60	60/60 [illegible]	4/(60.20) [illegible]
Součty			60	1			

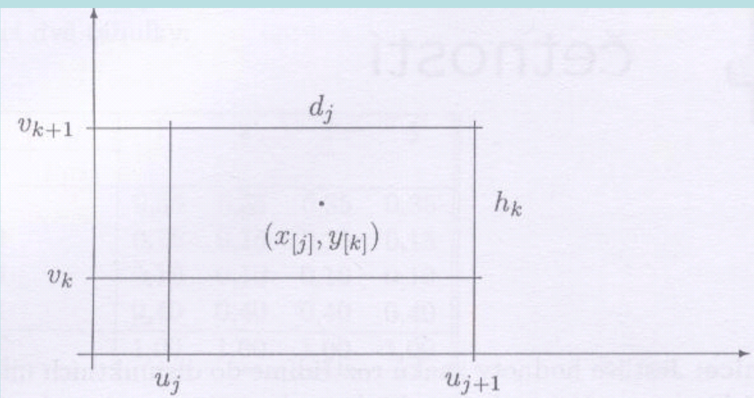


Dvourozměrné intervalové rozložení četností

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

Obdélník $\langle u_j, u_{j+1}) \times \langle v_k, v_{k+1})$ se nazývá (j,k)-tý dvourozměrný třídicí interval.



Simultánní a marginální četnosti

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

$p_{jk} = \dfrac{n_{jk}}{n}$ – 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.

$p_{j.} = \dfrac{n_{j.}}{n}$ – 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.

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

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

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

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

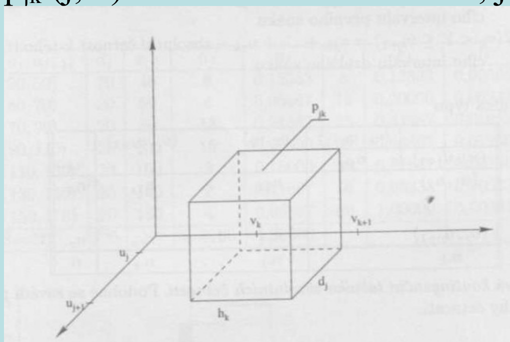
Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky.

Uveďme kontingenční tabulku simultánních absolutních četností:

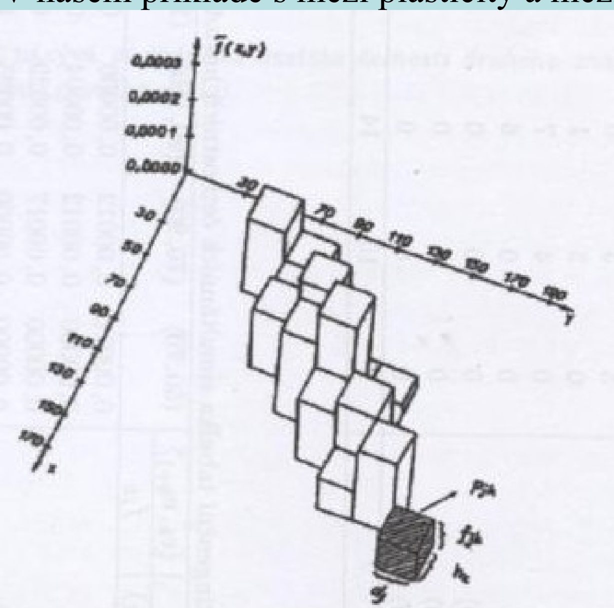
	$\langle v_k, v_{k+1}\rangle$	$\langle v_1, v_2\rangle$	$\ldots$	$\langle v_s, v_{s+1}\rangle$	
$\langle u_j, u_{j+1}\rangle$	n_{jk}				$n_{j.}$
$\langle u_1, u_2\rangle$		n_{11}	$\ldots$	n_{1s}	$n_{1.}$
$\vdots$					$\vdots$
$\langle u_r, u_{r+1}\rangle$		n_{r1}	$\ldots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$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

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

se nazývá simultánní hustota četnosti. Jejím grafem je schodovitá plocha shora omezující stereogram.

Hustoty četnosti pro znaky X a Y odlišíme indexem takto:

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

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

Mezi simultánní hustotou četnosti a marginálními hustotami četnosti platí vztahy:

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

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 $\forall (x, y) \in \mathbb{R}^2$: $f(x, y) = f_1(x)\, f_2(y)$.

Příklad: Zjistěte, zda mez pevnosti a mez plasticity jsou četnostně nezávislé.

Řešení: Vyjdeme z kontingenční tabulky simultánních absolutních četností.

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

Vidíme, že už pro j = 1, k = 1 je multiplikativní vztah porušen:
$f_{11} = \frac{5}{60 \cdot 20 \cdot 20} = 0{,}0002\ldots$, $f_{1.} = \frac{8}{60 \cdot 20} = 0{,}0066\ldots$, $f_{.1} = \frac{5}{60 \cdot 20} = 0{,}0041\ldots$, tudíž
$0{,}000208 \neq 0{,}006667 \cdot 0{,}004167 = 0{,}000028$ a mez pevnosti a mez plasticity nejsou četnostně nezávislé.