Mnohorozměrná statistická data

Ekonometrie

Jiří Neubauer

Katedra kvantitativních metod FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Statistický znak, statistický soubor

Jednotlivé objekty nebo subjekty, které jsou při statistickém zkoumání sledované, se nazývají **statistické jednotky**. Každá statistická jednotka musí být jednoznačně vymezena, aby nemohlo dojít k dvojímu nebo jinak zkreslenému výkladu zjištěných údajů. Statistické jednotky se vymezují z hlediska

- věcného,
- prostorového,
- časového.

Množina statistických jednotek stejného typu a shodného vymezení tvoří **statistický soubor**. V rámci statistického šetření budeme rozlišovat dva typy souborů:

- **základní soubor (populace)** – množina všech shodně vymezených statistických jednotek,
- **výběrový soubor (výběr, vzorek)** – podmnožina základního souboru, tj. vybraná část populace.

Statistický znak, statistický soubor

Vlastnosti, které u statistických jednotek budeme v rámci statistického šetření sledovat, nazýváme **statistické znaky** neboli **statistické proměnné**. Různé hodnoty, kterých může statistický znak nabývat, nazýváme **obměny** neboli **varianty**. Podle způsobu vyjadřování hodnot dělíme statistické znaky na **kvantitativní** – číselné a **kvalitativní** – slovní.
Podle typu vztahů mezi hodnotami a obměnami budeme rozlišovat statistické znaky

- **metrické**,
- **ordinální**,
- **nominální**.

Absolutní a relativní četnost – jednorozměrné bodové rozdělení četností

Mějme uspořádaný datový soubor o rozsahu n prvků.

- **Absolutní četnost** n_j představuje počet výskytů varianty x_j v souboru. Pro absolutní četnosti platí $\sum_{j=1}^{k} n_j = n$, kde k je počet variant.
- **Relativní četnost** p_j je dána vztahem

$$p_j = \frac{n_j}{n}$$

  a představuje podíl výskytů varianty x_j v souboru. Pro relativní četnosti platí $\sum_{j=1}^{k} p_j = 1$.
- **Absolutní kumulativní četnost** N_j je dána vztahem

$$N_j = n_1 + \cdots + n_j$$

  a udává součet četností všech pozorování, která nepřekračují hodnotu x_j.
- **Relativní kumulativní četnost** F_j je určena vztahem

$$F_j = \frac{N_j}{n} = p_1 + \cdots + p_j$$

  a udává podíl četností všech pozorování, která nepřekračují hodnotu x_j.

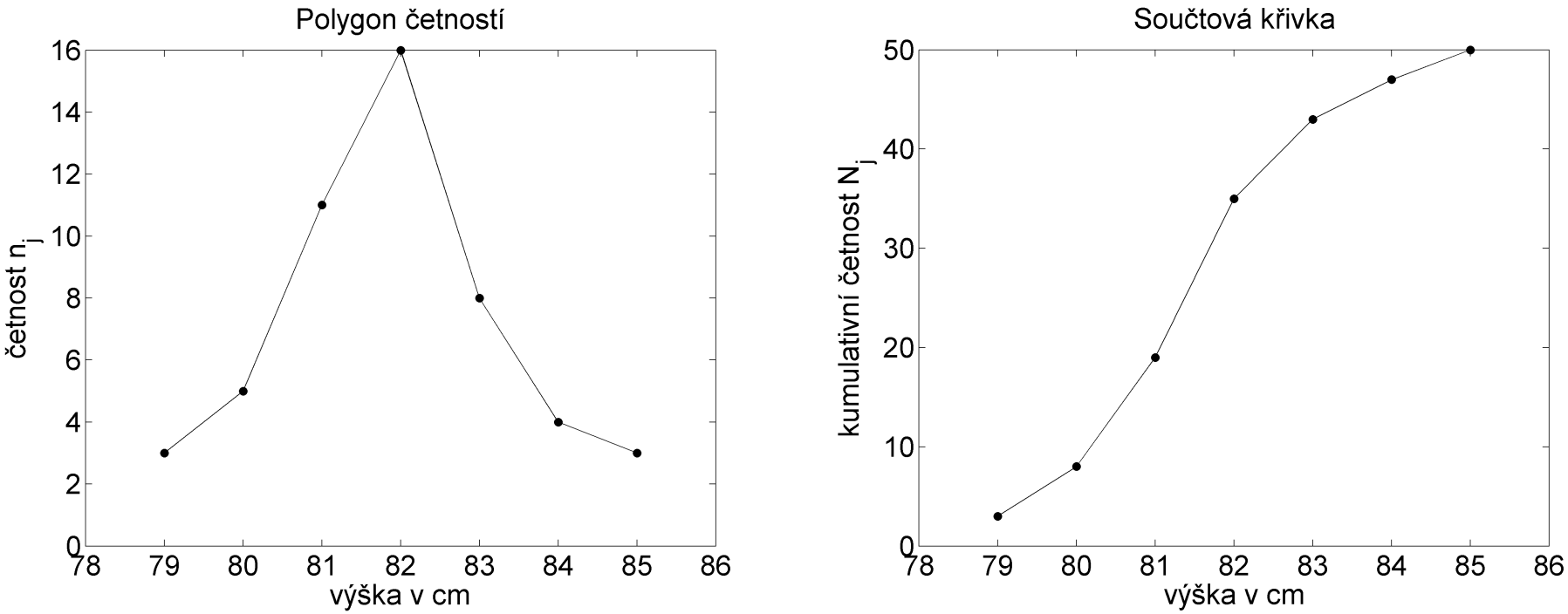
Absolutní a relativní četnost – jednorozměrné bodové rozdělení četností

V rámci antropometrického průzkumu bylo podle metodiky lékařské komory provedeno měření tělesné výšky u 15měsíčních dětí. U 50 vybraných chlapců byly naměřeny tyto hodnoty (v cm):

83 85 81 82 84 82 79 84 80 81 82 82 80 82 80 82 83 84 82 79
83 82 83 82 82 82 81 80 82 82 83 80 82 85 81 83 81 81 83 82
81 85 83 79 81 81 81 84 81 82

Hodnota znaku x_j	*Absolutní četnost* n_j	*Relativní četnost* p_j	*Abs. kum. četnost* N_j	*Rel. kum. četnost* F_j
79	3	0,06	3	0,06
80	5	0,10	8	0,16
81	11	0,22	19	0,38
82	16	0,32	35	0,70
83	8	0,16	43	0,86
84	4	0,08	47	0,94
85	3	0,06	50	1,00
$\sum$	50	1,00	—	—

Absolutní a relativní četnost – jednorozměrné bodové rozdělení četností



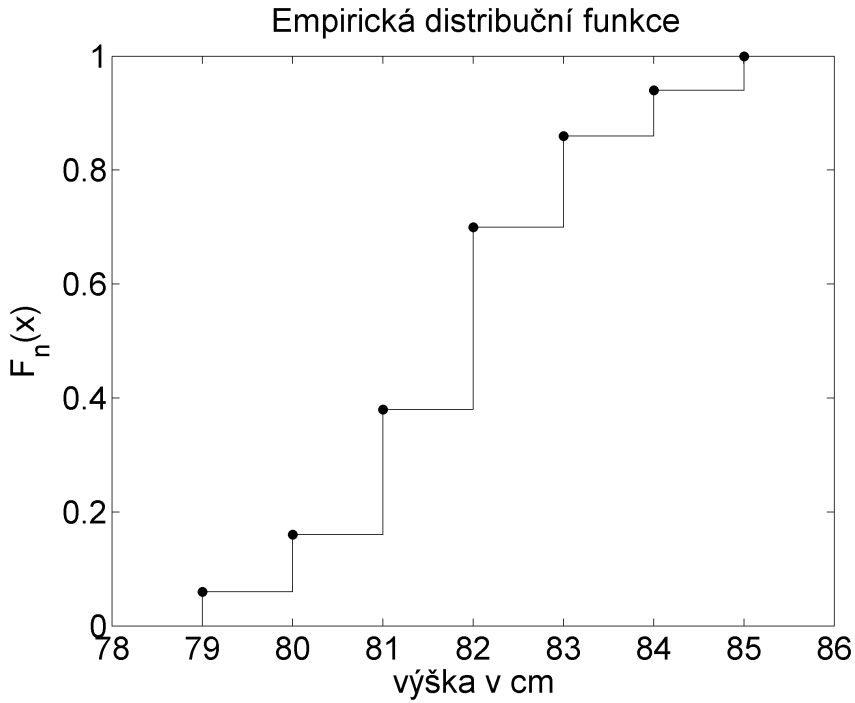
Obrázek: Polygon četností a součtová křivka výšky 15měsíčních dětí

Absolutní a relativní četnost – jednorozměrné bodové rozdělení četností

Rozdělení četností je také možné znázornit pomocí **empirické distribuční funkce**, kterou můžeme definovat vztahem

$$F_n(x) = \frac{N(x_i \leq x)}{n},$$

kde výraz v čitateli značí počet prvků výběru, jejichž hodnota je menší nebo rovna x. Je to neklesající funkce s hodnotami mezi 0 a 1. . . .



Absolutní a relativní četnost – jednorozměrné intervalové rozdělení četností

Pokud datový soubor, který máme zpracovat, má větší rozsah (zpravidla $n > 30$) a data reprezentují spojitý znak nebo diskrétní znak s velkým počtem variant (obměn), je vhodné nejprve data uspořádat podle velikosti a zjistit nejmenší a největší hodnotu $x_{\min}$ a $x_{\max}$ sledovaného znaku. Odtud lze určit **variační rozpětí** $R = x_{\max} - x_{\min}$ udávající šířku intervalu, ve kterém se data nacházejí.
Pro určení optimálního počtu (k) intervalů existuje několik pravidel, např.:

- Sturgesovo pravidlo $k \approx 1 + 3{,}32 \log n$,
- Yuleovo pravidlo $k \approx 2{,}5\sqrt[4]{n}$,
- jiná pravidla $k \approx \sqrt{n}$, příp. $k \approx 5 \log n$.

Odtud zvolíme podle uvážení vhodné k a orientačně stanovíme šířku intervalů ze vztahu $h = \frac{R}{k}$. Dále stanovíme počátek prvního intervalu (ozn. a) a šířku intervalů zvolíme tak, aby nejmenší a největší hodnota padly do prvního a posledního intervalu.
Číselnou usu tedy rozdělíme na intervaly

$$(-\infty, u_1\rangle, (u_1, u_2\rangle, \ldots (u_r, u_{r+1}\rangle, (u_{r+1}, \infty)$$

a budeme zjišťovat četnosti v těchto intervalech.

Absolutní a relativní četnost – jednorozměrné bodové rozdělení četností

Při kontrole dodržování hygienických norem v kuchyni se prováděl odběr vzduchu a pomocí filtru Pallflex se měřilo množství prachových částic. Ze 60 vzorků vzduchu jsme dostali následující výsledky (v $\mu\mathrm{g}/\mathrm{m}^3$):

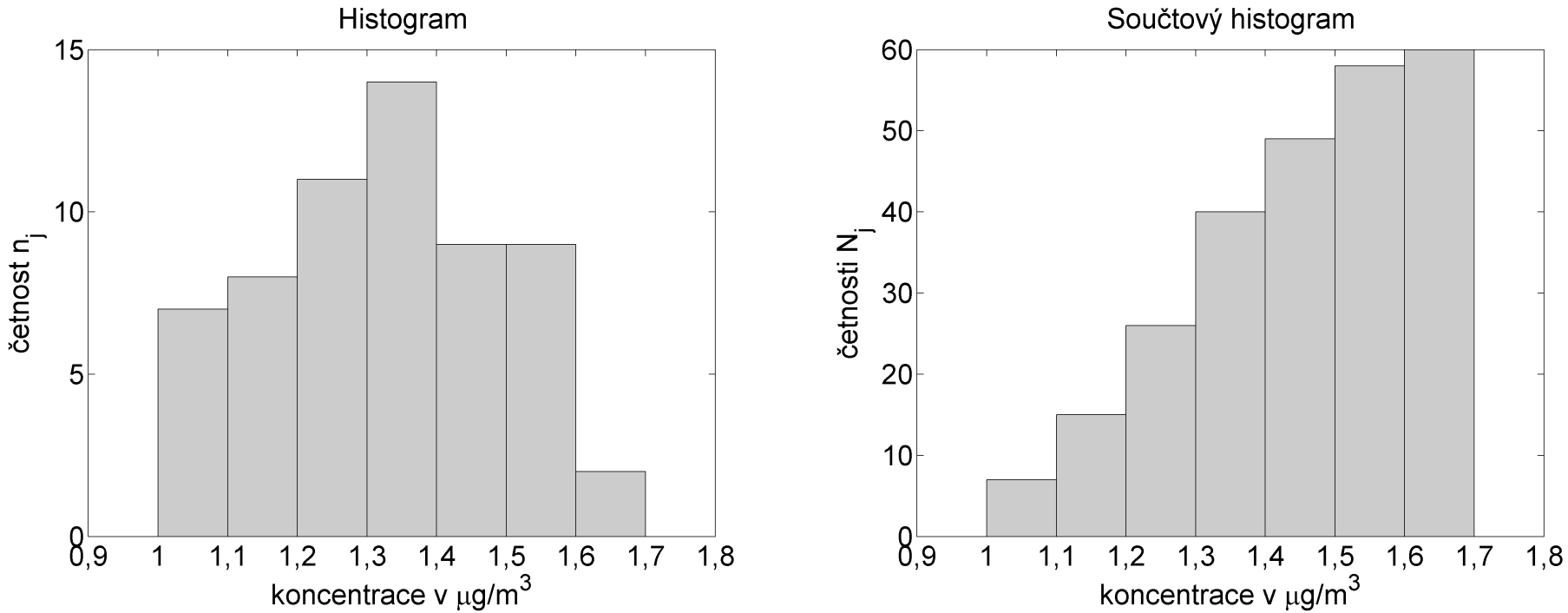
1,23	1,10	1,54	1,34	1,06	1,09	1,41	1,48	1,52	1,37	1,37	1,63
1,51	1,53	1,31	1,23	1,31	1,27	1,17	1,27	1,34	1,27	1,09	1,01
1,41	1,22	1,27	1,37	1,14	1,22	1,43	1,40	1,41	1,51	1,51	1,47
1,14	1,34	1,16	1,51	1,58	1,33	1,31	1,04	1,58	1,12	1,19	1,17
1,47	1,24	1,45	1,29	1,17	1,63	1,39	1,02	1,38	1,39	1,43	1,28

Absolutní a relativní četnost – jednorozměrné intervalového rozdělení četností

Interval	*Střed intervalu* x_j^*	*Absolutní četnost* n_j	*Relativní četnost* p_j	*Abs. kum. četnost* N_j	*Rel. kum. četnost* F_j
$(1{,}00; 1{,}10\rangle$	1,05	7	0,177	7	0,117
$(1{,}10; 1{,}20\rangle$	1,15	8	0,133	15	0,250
$(1{,}20; 1{,}30\rangle$	1,25	11	0,183	26	0,433
$(1{,}30; 1{,}40\rangle$	1,35	14	0,233	40	0,667
$(1{,}40; 1{,}50\rangle$	1,45	9	0,150	49	0,817
$(1{,}50; 1{,}60\rangle$	1,55	9	0,150	58	0,967
$(1{,}60; 1{,}70\rangle$	1,65	2	0,033	60	1,000
$\sum$	—	60	1	—	—

Tabulka: Tabulka intervalového rozdělení četností – množství prachových částic

Absolutní a relativní četnost – jednorozměrné intervalového rozdělení četností



Obrázek: Histogram a součtový histogram koncentrace prachu

Dvourozměrné bodové rozdělení četností

Mějme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$, kde znak X má r variant a znak Y má s variant.

- **Simultánní absolutní četnost** n_{jk} představuje počet výskytů dvojice (x_j, y_k) v souboru, tedy $n_{jk} = N(X = x_j \wedge Y = y_k)$.
- **Simultánní relativní četnost** dvojice (x_j, y_k) je dána vztahem

$$p_{jk} = \frac{n_{jk}}{n}.$$

- **Marginální absolutní četnost** varianty x_j je definována jako

$$n_{j.} = N(X = x_j) = n_{j1} + \cdots + n_{js}.$$

- **Marginální relativní četnost** varianty x_j je definována jako

$$p_{j.} = \frac{n_{j.}}{n} = p_{j1} + \cdots + p_{js}.$$

Dvourozměrné bodové rozdělení četností

- **Marginální absolutní četnost** varianty y_j je definována jako

$$n_{.k} = N(Y = y_k) = n_{1k} + \cdots + n_{rk}.$$

- **Marginální relativní četnost** varianty y_k je definována jako

$$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \cdots + p_{rk}.$$

- **Sloupcově podmíněná relativní četnost** varianty x_j za předpokladu y_k je dána vztahem

$$p_{j(k)} = \frac{n_{jk}}{n_{.k}}.$$

- **Sloupcově podmíněná relativní četnost** varianty y_k za předpokladu x_j je dána vztahem

$$p_{(j)k} = \frac{n_{jk}}{n_{j.}}.$$

Dvourozměrné bodové rozdělení četností

Příklad: U 42 zákrsku jabloní bylo zaznamenáno stáří stromu v letech (znak X) a roční sklizeň (znak Y).

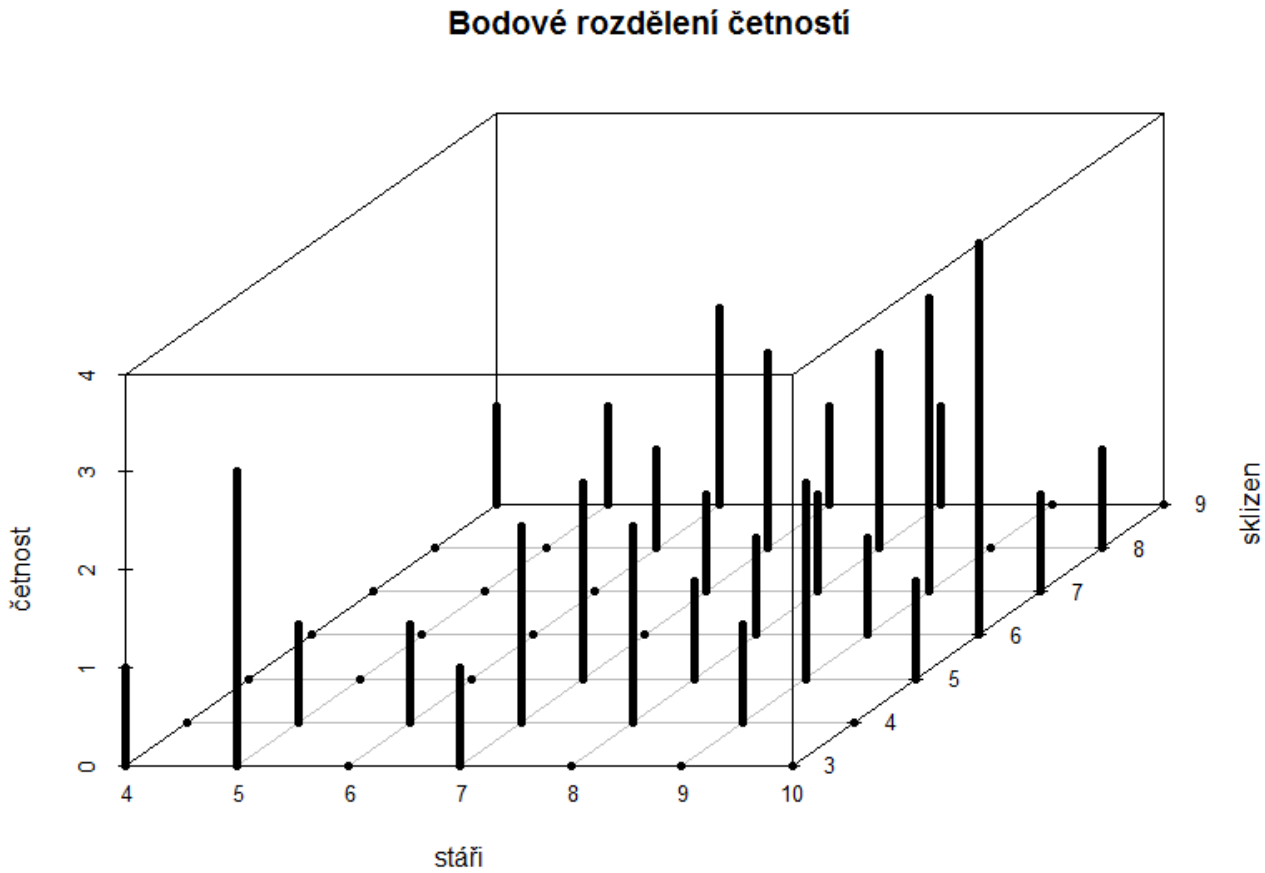
x_j	y_i						
3	4	7	5	5	5		
4	9	5	7	6	8	7	8
5	9	8	9	10	7	7	
6	10	8	10	10	10	9	
7	9	7	8	9	10	9	
8	8	7	7	8	6	10	
9	5	4	6	7	6	8	

Dvourozměrné bodové rozdělení četností

stáří/sklizeň	4	5	6	7	8	9	10	$n_{j.}$
3	1	3	0	1	0	0	0	5
4	0	1	1	2	2	1	0	7
5	0	0	0	2	1	2	1	6
6	0	0	0	0	1	1	4	6
7	0	0	0	1	1	3	1	6
8	0	0	1	2	2	0	1	6
9	1	1	2	1	1	0	0	6
$n_{.k}$	2	5	4	9	8	7	7	42

Tabulka: Tabulka bodového rozdělení četností

Dvourozměrné bodové rozdělení četností



Obrázek: Grafické znázornění dvourozměrného bodového rozdělení četností

Dvourozměrné intervalové rozdělení četností

Mějme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$, kde hodnoty znaku X roztřídíme do r třídících intervalů $(u_j, u_{j+1}\rangle, j = 1, \ldots, r$ a hodnoty znaku Y roztřídíme do s intervalů $(v_k, v_{k+1}\rangle, k = 1, \ldots, s$. Jednotlivé četnosti jsou potom vztaženy na četnosti hodnot v daných intervalech.

Dvourozměrné intervalové rozdělení četností

Bylo provedeno 34 měření pH a množství hydrogenuhličitanového aniontu ve studniční vodě

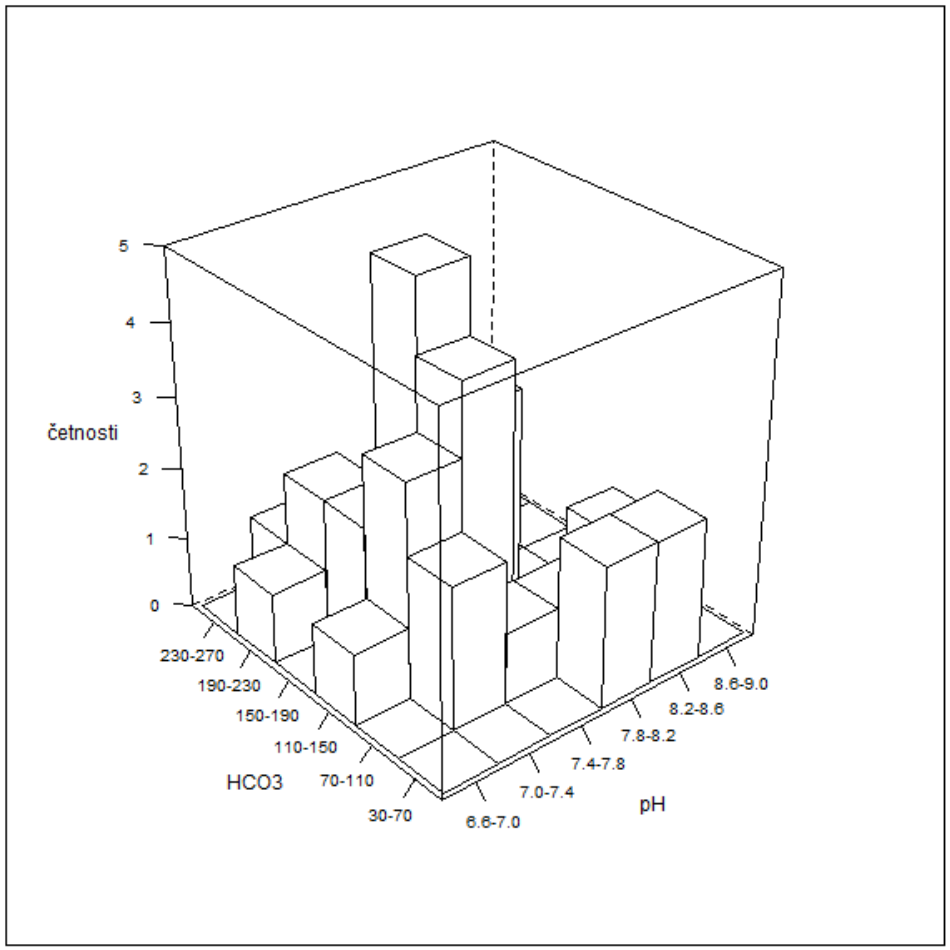
pH	HCO_3^-	pH	HCO_3^-	pH	HCO_3^-	pH	HCO_3^-
7,6	157	7,5	190	8,2	202	7,4	125
7,1	174	8,1	215	7,9	155	7,3	76
8,2	175	7,0	199	7,6	157	8,5	48
7,5	188	7,3	262	8,8	147	7,8	147
7,4	171	7,8	105	7,2	133	6,7	117
7,8	143	7,3	121	7,9	53	7,1	182
7,3	217	8,0	81	8,1	56	7,3	87
8,0	190	8,5	82	7,7	113		
7,1	142	7,1	210	8,4	35		

Dvourozměrné intervalové rozdělení četností

pH/HCO_3^-	30–70	70–110	110–150	150–190	190–230	230–270	$n_{j.}$
6,6–7,0	0	0	1	0	1	0	2
7,0–7,4	0	2	3	2	2	1	10
7,4–7,8	0	1	4	5	0	0	10
7,8–8,2	2	1	0	3	2	0	8
8,2–8,6	2	1	0	0	0	0	3
8,6–9,0	0	0	1	0	0	0	1
$n_{.k}$	4	5	9	10	5	1	34

Tabulka: Tabulka intervalového rozdělení četností

Dvourozměrné intervalové rozdělení četností



Obrázek: Grafické znázornění dvourozměrného intervalového rozdělení četností

Číselné charakteristiky

Číselné charakteristiky znaku

- charakteristiky polohy – průměry, kvantily, modus
- charakteristiky variability – rozptyl, sm. odchylka, výběrový rozptyl a sm. odchylka, kvantilové rozpětí . . .
- charakteristiky koncentrace – koeficient šikmosti a špičatosti
- charakteristiky těsnosti závislostí

Charakteristiky polohy

- průměry:
  - aritmetický průměr $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - harmonický průměr $\overline{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$
  - geometrický průměr $\overline{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$
- kvantily: x_p je hodnota znaku, pro kterou platí, že $100p\,\%$ jednotek uspořádaného souboru má hodnotu menší nebo rovnu x_p a $100(1-p)\,\%$ jednotek má hodnotu větší nebo rovnu x_p.
- modus: $\hat{x}$ je hodnota znaku s největší četností

Charakteristiky variability

- variační rozpětí: $R = x_{\max} - x_{\min}$.
- kvantilová rozpětí: např. $R_Q = x_{0,75} - x_{0,25}$
- rozptyl (momentový): $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- směrodatná odchylka $s_n = \sqrt{s_n^2}$
- výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- výběrová směrodatná odchylka $s = \sqrt{s^2}$
- průměrná odchylka $d_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$

Charakteristiky koncentrace

- koeficient šikmosti: $a_3 = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{s_n^3}$
- koeficient špičatosti: $a_4 = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{s_n^4} - 3$

Charakteristiky těsnosti závislosti

Kovariance hodnot znaků X a Y je definována vztahem

$$s_{n,xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

kde n je rozsah souboru, $\overline{x}$ značí aritmetický průměr znaku X a $\overline{y}$ značí aritmetický průměr znaku Y.

Výběrová kovariance hodnot znaků X a Y je definována vztahem

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Charakteristiky těsnosti závislosti

Korelační koeficient (Pearsonův) je definován vztahem

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_{n,x}} \cdot \frac{y_i - \overline{y}}{s_{n,y}},$$

kde $s_{n,x}$ a $s_{n,y}$ jsou směrodatné odchylky hodnot znaků X a Y.

Korelační koeficient lze vyjádřit jako podíl kovariance a součinu obou směrodatných odchylek ve tvaru

$$r_{xy} = \frac{s_{n,xy}}{s_{n,x} \cdot s_{n,y}}, \quad \text{resp.} \quad r_{xy} = \frac{s_{xy}}{s_x \cdot s_y},$$

kde s_x a s_y jsou výběrové směrodatné odchylky hodnot znaků X a Y.
Pro korelační koeficient tedy platí

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}},$$

což je vzorec vhodný zejména pro praktický výpočet (výpočetní tvar).

Charakteristiky těsnosti závislosti

Vlastnosti Pearsonova korelačního koeficientu:

1. koeficient korelace je symetrický, tedy $r_{xy} = r_{yx}$,
2. pro koeficient korelace platí $-1 \leq r_{xy} \leq 1$,
3. koeficient korelace je roven 1, resp. -1, právě když mezi hodnotami $x_1, x_2, \ldots, x_n$ a $y_1, y_2, \ldots, y_n$ existuje úplná lineární závislost, tedy existují reálné konstanty a, b tak, že $y_i = a + bx_i$, $i = 1, 2, \ldots, n$, přičemž hodnota korelačního koeficientu 1 odpovídá $b > 0$ a hodnota -1 pak $b < 0$.

Hodnoty korelačního koeficientu blízké nule značí, že mezi znaky neexistuje lineární vazba, může však existovat vazba jiného typu.

Příklad

Určete kovarianci, výběrovou kovarianci a korelační koeficient (Pearsonův) pro hodnotou pH a množství hydrogenuhličitanového aniontu HCO_3^- ve studniční vodě.

- kovariance $s_{n,xy} = -9{,}21964$
- výběrová kovariance $s_{xy} = -9{,}49902$
- korelační koeficient $r_{xy} = -0{,}33951$

Charakteristiky těsnosti závislosti

Těsnost závislosti mezi znaky lze vyjádřit prostřednictvím *Spearmanova korelačního koeficientu*, který popisuje vazbu mezi dvěma spojitými znaky pomocí pořadí. Je vhodný v případě, kdy vztah mezi dvěma proměnnými může být popsán pomocí monotónní funkce.

Hodnotám x_i, y_i přiřadíme čísla p_i, q_i udávající pořadí jednotlivých hodnot při vzestupném uspořádaní hodnot znaků podle velikosti. **Spearmanův koeficient pořadové korelace** je potom definován vztahem

$$r_{xy}^s = 1 - \frac{6\sum_{i=1}^{n}(p_i - q_i)^2}{n(n^2-1)}.$$

Spearmanův korelační koeficient je roven Pearsonovu korelačnímu koeficientu spočítanému pro pořadí p_i, q_i. I tento korelační koeficient nabývá hodnot od -1 do 1, hodnota blízká 1 značí silnou přímou závislost, hodnota blízká -1 potom silnou nepřímou závislost.

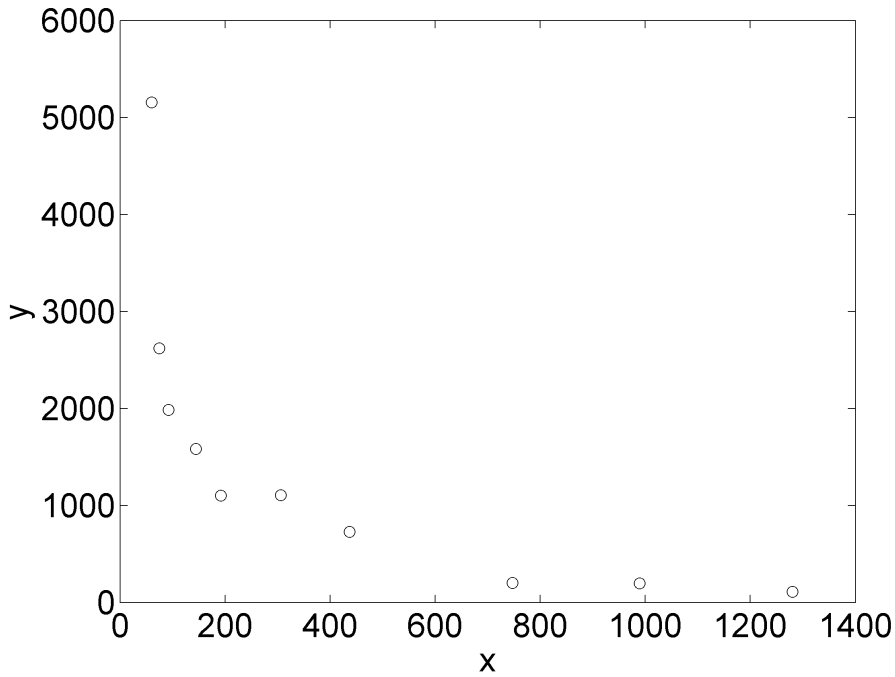
Příklad

V podniku se sledovala závislost vlastních nákladů připadajících na jednotku produkce (znak Y) na objemu produkce v 1000 ks (znak X). Zjištěné údaje jsou uvedeny v tabulce. Určete Pearsonův a Spearmanův korelační koeficient.

x_i	60	75	92	144	192	306	437	747	989	1280
y_i	5157	2620	1986	1582	1100	1105	729	200	196	110

Dostáváme hodnoty $r_{xy} = -0{,}695$, $r^s_{xy} = -0{,}988$.

Příklad



Obrázek: Závislost vlastních nákladů připadajících na jednotku produkce (znak Y) na objemu produkce v 1000 ks (znak X)

Charakteristiky těsnosti závislosti

Pokud se v měřených datech vyskytuje mnoho shodných hodnot, doporučuje se použít korigovaný Spearmanův korelační koeficient definovaný vzorcem

$$r^{S}_{xy\,/\,kor} = 1 - \frac{6\sum_{i=1}^{n}(p_i - q_i)^2}{n\left(n^2-1\right) - n_x - n_y}\ ,$$

kde $n_x = \frac{1}{2}\sum_{j=1}^{r}\left(n_{x,j}^3 - n_{x,j}\right)$ a $n_y = \frac{1}{2}\sum_{k=1}^{s}\left(n_{y,k}^3 - n_{y,k}\right)$. Symbol $n_{x,j}$ značí počty stejně velkých x-ových hodnot a $n_{y,k}$ značí počty stejně velkých y-ových hodnot.

Charakteristiky těsnosti závislosti

Mějme dvourozměrný datový soubor. Řekneme, že dvojice (x_i, y_i) a (x_j, y_j) jsou ve shodě (concordant), pokud platí, že $x_i > x_j$ a zároveň $y_i > y_j$ nebo $x_i < x_j$ a zároveň $y_i < y_j$. Řekneme, že nejsou ve shodě (discordant), pokud $x_i < x_j$ a zároveň $y_i > y_j$ nebo $x_i > x_j$ a zároveň $y_i < y_j$. V případě, že $x_i = x_j$ nebo $y_i = y_j$ nemluvíme ani o shodě, ani o neshodě. Označme počet dvoji ve shodě n_c a počet dvojic, které ve shodě nejsou n_d. **Kendallův korelační koeficient** je definován vztahem

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}.$$