Mnohorozměrná statistická data

1 Statistický znak

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.

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**.

2 Jednorozměrné rozdělení

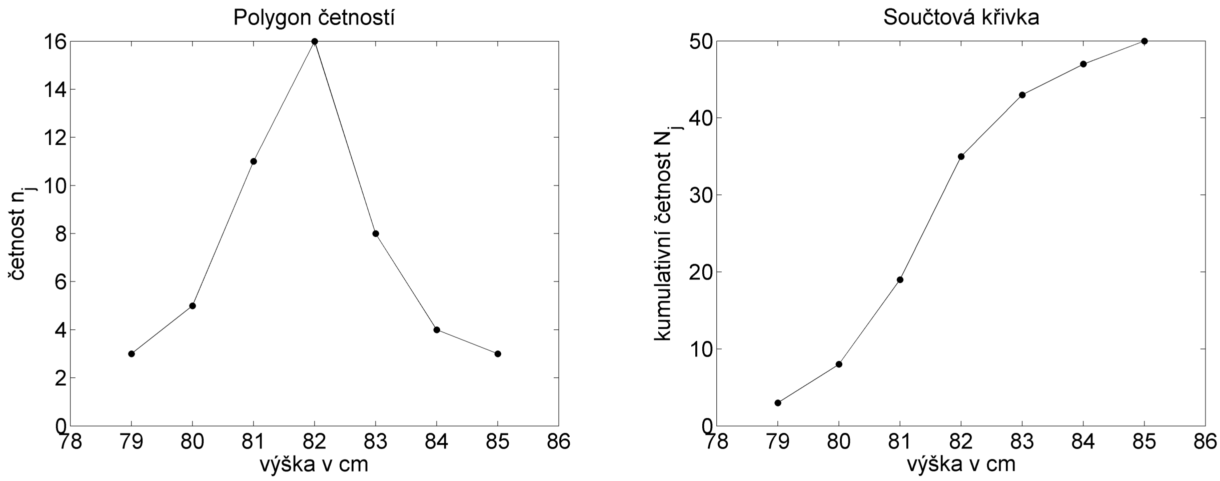
2.1 Jednorozměrné bodové rozdělení

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.

Operační program Vzdělávání pro konkurenceschopnost
Název projektu: Inovace magisterského studijního programu Fakulty ekonomiky a managementu
Registrační číslo projektu: CZ.1.07/2.2.00/28.0326

PROJEKT JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY.



Obr. 1: Polygon četností a součtová křivka výšky 15měsíčních dětí

- **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.

Příklad. 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):

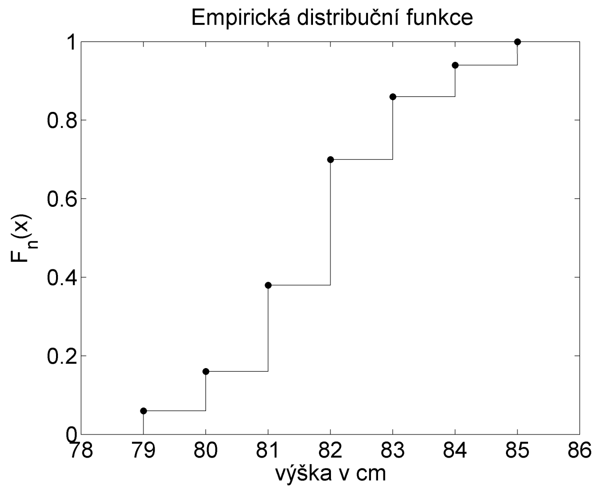
83 85 81 82 84 82 79 84 80 81 82 82 80 82 80 82 83 84 82 79
83 82 83 82 82 82 81 80 82 82 83 80 82 85 81 83 81 81 83 82
81 85 83 79 81 81 81 84 81 82

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	—	—

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....



Obr. 2: Empirická distribuční funkce

2.2 Jednorozměrné intervalové rozdělení

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 osu tedy rozdělíme na intervaly

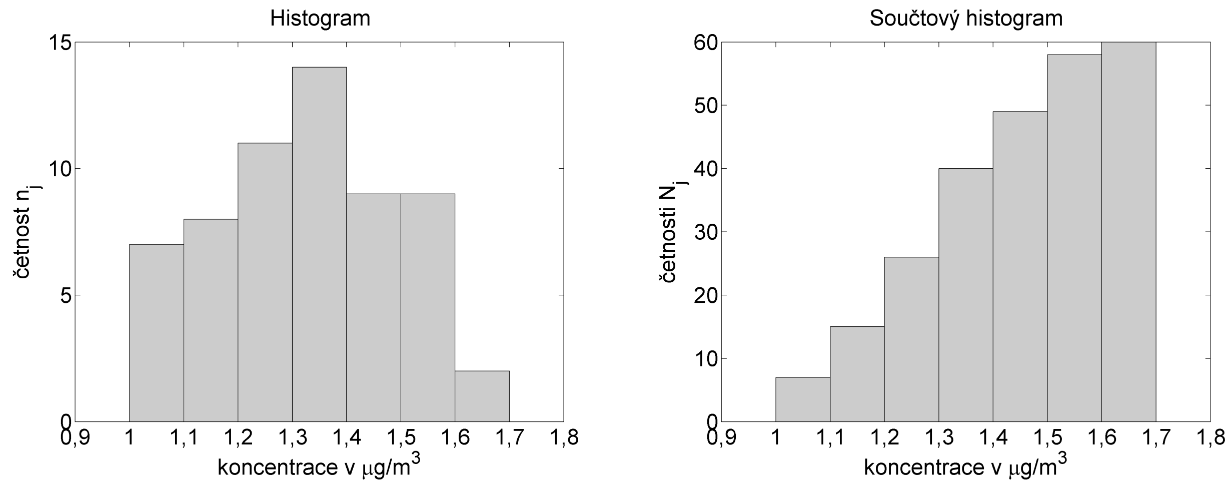
$$(-\infty, u_1\rangle, (u_1, u_2\rangle, \dots (u_r, u_{r+1}\rangle, (u_{r+1}, \infty)$$

a budeme zjišťovat četnosti v těchto intervalech. 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 $\mu\text{g/m}^3$):

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

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

Tab. 1: Tabulka intervalového rozdělení četností – množství prachových částic



Obr. 3: Histogram a součtový histogram koncentrace prachu

3 Dvourozměrné rozdělení četností

3.1 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
$$n_{j.} = N(X = x_j) = n_{j1} + \cdots + n_{js}.$$

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	

Tab. 2: Tabulka bodového 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

Tab. 3: Tabulka bodového rozdělení četností

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

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

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

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

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

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

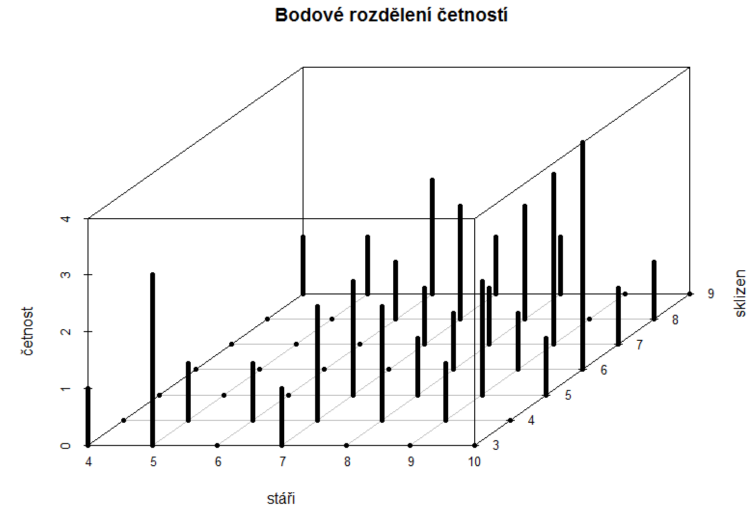
- **Sloupcově podmíněná relativní četnost** varianty x_j za předpokladu y_k je dána vztahem

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

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

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

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



Obr. 4: Grafické znázornění dvourozměrného bodového rozdělení četností

3.2 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 $\langle u_j, u_{j+1}\rangle, j = 1, \ldots, r$ a hodnoty znaku Y roztřídíme do s intervalů $\langle v_k, v_{k+1}\rangle, k = 1, \ldots, s$. Jednotlivé četnosti jsou potom vztaženy na četnosti hodnot v daných intervalech. Bylo provedeno 34 měření pH a množství hydrogenuhličitanu 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		

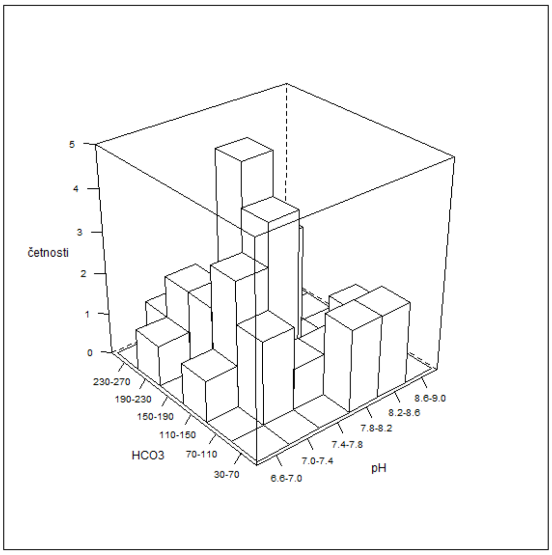
4 Číselné charakteristiky znaku

Číselné charakteristiky znaku

- charakteristiky polohy – průměry, kvantily, modus
- charakteristiky variability – rozptyl, sm. odchylka, výběrový rozptyl a sm. odchylka, kvantilové rozpětí …

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

Tab. 4: Tabulka intervalového rozdělení četností



Obr. 5: Grafické znázornění dvourozměrného intervalového rozdělení četností

- charakteristiky koncentrace – koeficient šikmosti a špičatosti
- charakteristiky těsnosti závislostí
- průměry:
  - aritmetický průměr $\overline{x} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$
  - harmonický průměr $\overline{x}_H = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}}$
  - 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
- variační rozpětí: $R = x_{\max} - x_{\min}$.

- kvantilová rozpětí: např. $R_Q = x_{0,75} - x_{0,25}$
- rozptyl (momentový): $s_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$
- směrodatná odchylka $s_n = \sqrt{s_n^2}$
- výběrový rozptyl $s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$
- výběrová směrodatná odchylka $s = \sqrt{s^2}$
- průměrná odchylka $d_{\overline{x}} = \frac{1}{n}\sum\limits_{i=1}^{n}|x_i - \overline{x}|$
- koeficient šikmosti: $a_3 = \dfrac{1}{n}\dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^3}{s_n^3}$
- koeficient špičatosti: $a_4 = \dfrac{1}{n}\dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^4}{s_n^4} - 3$

Mějme dvourozměrný datový soubor $\begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$, označme $\overline{x}$ a $\overline{y}$ průměry znaků a s_x, s_y směrodatné odchylky znaků X, Y. **Koeficient korelace (Pearsonův)** je definován vztahem

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \overline{x}}{s_x}\frac{y_i - \overline{y}}{s_y}.$$

Lze jej vyjádřit ve tvaru

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

kde

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

je kovariance znaků X a Y. V případě dvourozměrného souboru kvalitativních údajů, které jsou po složkách ordinálního typu, je možno zjistit stupeň závislosti těchto dvou znaků. K měření takovýchto závislostí se používá **Spearmanův korelační koeficient**. Hodnotám x_i, y_i přiřadíme pořadová čísla p_i, q_i (pořadí jednotlivých hodnot při uspořádaní podle velikosti). Spearmanův koeficient (koeficient pořadové korelace) je potom definován vztahem

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

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

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

Příklady k procvičení

---

1. U 30 žáků byly zjištěny hodnoty znaků X – známka z fyziky, Y – známka z chemie a Z – pohlaví (0. . . dívka, 1. . . chlapec).

Fyzika	Chemie	Pohlaví	Fyzika	Chemie	Pohlaví	fyzika	Chemie	Pohlaví
1	2	1	2	3	1	3	4	0
2	3	1	2	3	0	5	4	0
2	3	0	2	2	0	2	1	0
4	5	1	4	5	0	2	2	0
2	1	1	3	3	1	3	1	1
4	3	1	3	4	1	3	4	0
2	2	1	2	3	1	2	1	1
4	2	0	2	2	0	1	1	0
3	3	0	5	3	1	1	1	1
4	5	0	3	2	1	1	2	0

Sestavte tabulky rozdělení četností (jednorozměrné) pro známky z fyziky a chemie, zobrazte graficky. Spočítejte základní charakteristiky (průměr, výběrovou směrodatnou odchylku a rozptyl). Sestavte dvourozměrnou tabulku bodového rozdělení četností (známka z fyziky a chemie), vyjádřete graficky. To stejné proveďte zvlášť pro dívky a pro chlapce. [Datový soubor: znamky.txt]

2. U 40 pracovníků byla sledována závislost počtu chybných operací za směnu (Y) na délce zapracování v hodinách (X) s těmito výsledky

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
3	2	2	4	1	5	2	6	4	3	3	6	2	5	3	5
4	3	2	6	3	3	4	4	1	6	5	2	4	1	4	2
1	4	4	1	2	7	4	3	2	4	3	3	1	6	1	7
4	5	2	5	3	4	3	5	2	5	2	3	5	4	4	4
4	2	5	3	5	1	3	4	5	1	3	4	3	4	5	2

Sestavte dvourozměrnou tabulku bodového rozdělení četností, vyjádřete graficky. Sestavte tabulky rozdělení četností (jednorozměrné) pro počet chybných operací a délku zapracování, zobrazte graficky. Spočítejte základní charakteristiky (průměr, výběrovou směrodatnou odchylku a rozptyl).
[Datový soubor: chyby_zapracovani.txt]