Popisná statistika

Lenka Křivánková

142474@mail.muni.cz

Přednáška Statistika 1 (BKMSTA1)

8. říjen 2016, Brno

Motivace

- Popisná statistika slouží zejména k prezentaci dat a výsledků.
- Číselné charakteristiky informují o úrovni, variabilitě a těsnosti závislosti znaků.
- V dalším budeme probírat analogické veličiny u náhodných výběrů.

Základní, výběrový a datový soubor

Základní a výběrový soubor

- **Základním souborem** rozumíme libovolnou neprázdnou množinu E. Její prvky značíme ϵ a nazýváme je objekty.
- Libovolnou neprázdnou podmnožinu $\{\epsilon_1, \ldots, \epsilon_n\}$ základního souboru E nazýváme **výběrový soubor** rozsahu n.
- Je-li $G \subseteq E$, 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

$$p(G) = \frac{N(G)}{n}.$$

Základní a výběrový soubor – příklad

Hodnocení finančního zdraví několika firem dvěma hodnotiteli.

I. hodnotitel	II. hodnotitel	I. hodnotitel	II. hodnotitel
2	2	4	2
1	3	4	4
4	3	2	2
1	1	4	3
1	2	2	3
4	4	4	4
3	3	1	1
3	4	4	3
1	1	4	4
1	1	1	3

Hodnocení I. hodnotitele budeme dále označovat X a hodnocení II. hodnotitele Y.

Datový soubor

Nechť je dán výběrový soubor $\{\epsilon_1, \ldots, \epsilon_n\} \subseteq E$. Hodnoty znaků X, Y, Z pro i-tý objekt označíme $x_i = X(\epsilon_i)$, $y_i = Y(\epsilon_i)$, $\ldots$, $z_i = Z(\epsilon_i)$, $i = 1, \ldots, n$.

Matice

$$\begin{bmatrix} x_1 & y_1 & \cdots & z_1 \\ x_2 & y_2 & \cdots & z_2 \\ \vdots & \vdots & & \vdots \\ x_n & y_n & \cdots & z_n \end{bmatrix}$$

typu $n \times p$ se nazývá **datový soubor**. Její řádky odpovídají jednotlivým objektům, sloupce znakům. Libovolný sloupec této matice nazýváme jednorozměrným datovým souborem.

Datový soubor

Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme **uspořádaný datový soubor**

$$\begin{bmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{bmatrix},$$

kde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Vektor

$$\begin{bmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{bmatrix},$$

kde $x_{[1]} < \cdots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá **vektor variant**.

Datový soubor – příklad

$$\begin{bmatrix} 2 \\ 1 \\ 4 \\ 1 \\ 1 \\ 4 \\ 3 \\ 3 \\ 1 \\ 1 \\ 4 \\ 4 \\ 2 \\ 4 \\ 2 \\ 4 \\ 1 \\ 4 \\ 4 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Bodové rozdělení četností

Bodové rozdělení četnosti

Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o **bodovém rozdělení četností**.

Bodové rozdělení četnosti

Existuje několik způsobů, jak graficky znázornit bodové rozdělení 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 četnost této varianty.

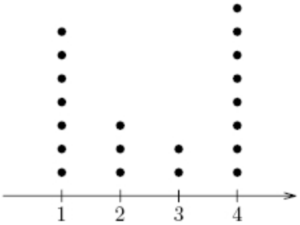
Bodové rozdělení četnosti

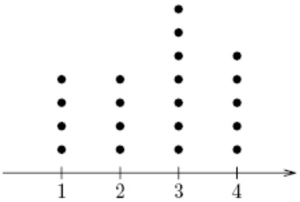
- **Sloupkový diagram**: je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní četnost této varianty.
- **Výsečový graf**: je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním četnostem variant znaku X.
- **Dvourozměrný tečkový diagram**: 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.

Bodové rozdělení četnosti – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem" sestrojte

jednorozměrné tečkové diagramy pro znak X a znak Y

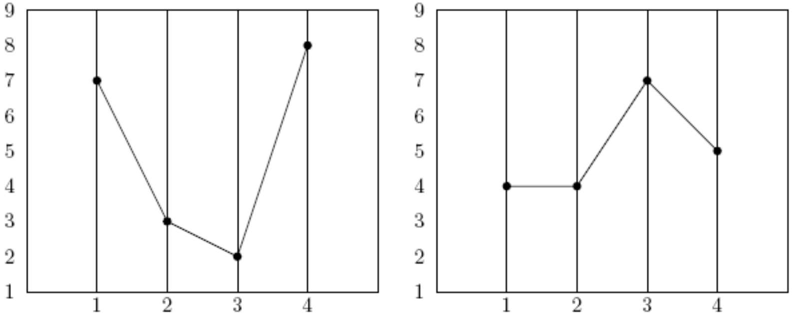




Bodové rozdělení četnosti – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem" sestrojte

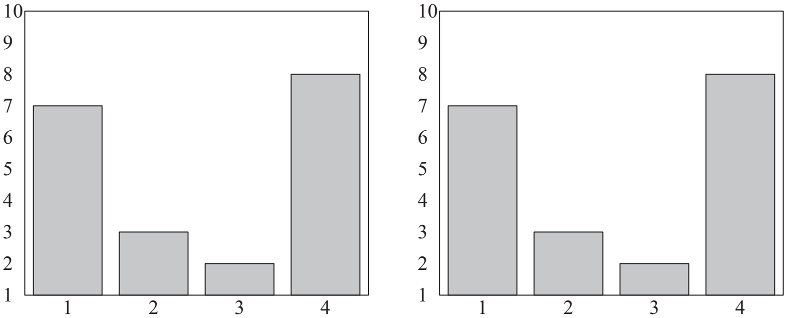
polygony četností pro znak X a znak Y



Bodové rozdělení četnosti – příklad

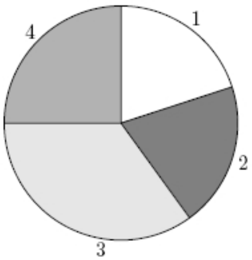
Pro datový soubor "hodnocení finančního zdraví několika firem" sestrojte

sloupkové diagramy pro znak X a znak Y



Pro datový soubor "hodnocení finančního zdraví několika firem" sestrojte

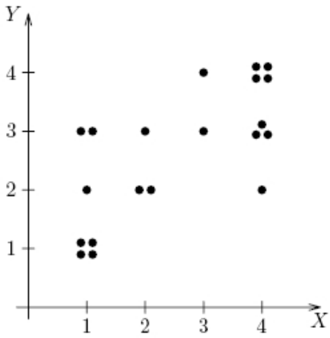
výsečové diagramy pro znak X a znak Y



Bodové rozdělení četnosti – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem" sestrojte

dvourozměrný tečkový diagram pro vektorový znak (X, Y)



Variační řada

- Bodové rozdělení četností lze znázornit nejenom graficky, ale též tabulkou zvanou **variační řada**, která obsahuje absolutní a relativní četnosti jednotlivých variant znaku v daném výběrovém souboru a též absolutní a relativní kumulativní četnosti.
- Pomocí relativních četností se zavádí četnostní funkce, pomocí relativních kumulativních četností empirická distribuční funkce (je pro ni typické, že má schodovitý průběh).

Variační řada

Nechť je dán jednorozměrný datový soubor, v němž znak X nabývá r variant. Pro $j = 1, \dots, r$ definujeme:

- **absolutní četnost** varianty $x_{[j]}$ ve výběrovém souboru
$$n_j = N(X = x_{[j]})$$
- **relativní četnost** varianty $x_{[j]}$ ve výběrovém souboru
$$p_j = \frac{n_j}{n}$$
- **absolutní kumulativní četnost** prvních j variant ve výběrovém souboru
$$N_j = N(X \leq x_{[j]}) = n_1 + \cdots + n_j$$
- **relativní kumulativní četnost** prvních j variant ve výběrovém souboru
$$F_j = \frac{N_j}{n} = p_1 + \cdots + p_j$$

Variační řada

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**.

Variační řada – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem" sestavte variační řadu pro znak X.

$x_{[j]}$	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	–	–

Četnostní a empirická distribuční funkce

Funkce

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

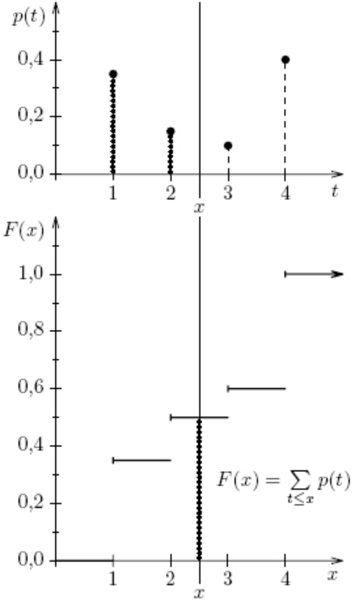
se nazývá **četnostní funkce**.

Funkce

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

se nazývá **empirická distribuční funkce**.

Četnostní a empirická distribuční funkce – příklad



Pro datový soubor "hodnocení finančního zdraví několika firem" nakreslete grafy četnostní funkce a empirické distribuční funkce znaku X.

Četnostní a empirická distribuční funkce – vlastnosti

- Četnostní funkce je
  - nezáporná ($\forall x \in R : p(x) \geq 0$) a
  - normovaná, tj.
$$\sum_{x=-\infty}^{\infty} p(x) = 1.$$
- Empirická distribuční funkce je
  - neklesající, tzn.
$$\forall x_1, x_2 \in R, x_1 < x_2 : F(x_1) \leq F(x_2),$$
  - zprava spojitá ($\forall x_0 \in R$ libovolné, ale pevně dané: $\lim_{x \to x_0} F(x) = F(x_0)$) a
  - normovaná ($\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$).

Dvourozměrný datový soubor

Nechť je dán dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde znak X má r variant a znak Y má s variant. Pak definujeme:

- **simultánní absolutní četnost** dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

$$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]}),$$

- **simultánní relativní četnost** dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

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

Dvourozměrný datový soubor

- **marginální absolutní četnost** varianty $x_{[j]}$

$$n_{j.} = N(X = x_{[j]}) = n_{j1} + \cdots + n_{js},$$

- **marginální relativní četnost** varianty $x_{[j]}$

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

- **marginální absolutní četnost** varianty $y_{[k]}$

$$n_{.k} = N(Y = y_{[k]}) = n_{1k} + \cdots + n_{rk},$$

- **marginální relativní četnost** varianty $y_{[k]}$

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

- **sloupcově podmíněná relativní četnost** varianty $x_{[j]}$ za předpokladu $y_{[k]}$

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

- **řádkově podmíněná relativní četnost** varianty $y_{[k]}$ za předpokladu $x_{[j]}$

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

Dvourozměrný datový soubor

Kteroukoliv ze simultánních četností či podmíněných relativních četností zapisujeme do kontingenční tabulky. **Kontingenční tabulka** simultánních absolutních četností má tvar

	y	$y_{[1]}$	$\cdots$	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	$\cdots$	n_{1s}	$n_{1.}$
$\vdots$		$\vdots$	$\cdots$	$\vdots$	$\vdots$
$x_{[r]}$		n_{r1}	$\cdots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	$\cdots$	$n_{.s}$	n

Simultánní četnostní funkce

Funkce

$$p(x,y) = \begin{cases} p_{jk} & \text{pro} \quad x = x_{[j]}, y = y_{[k]}, \quad j = 1,\ldots,r, \quad k = 1,\ldots,s \\ 0 & \text{jinak} \end{cases}$$

se nazývá **simultánní četnostní funkce**. Četnostní funkce pro znaky X a Y (tzv. **marginání četnostní funkce**) odlišíme indexem takto:

$$p_1(x) = \begin{cases} p_{j.} & \text{pro} \quad x = x_{[j]}, \quad j = 1,\ldots,r \\ 0 & \text{jinak} \end{cases}$$

$$p_2(y) = \begin{cases} p_{.k} & \text{pro} \quad y = y_{[k]}, \quad k = 1,\ldots,s \\ 0 & \text{jinak} \end{cases}$$

Podmíněné četnostní funkce

Funkce $p_{1|2}\left(x\,|y\right)$ zavedená vztahem $\forall x \in \mathbb{R}$:

$$p_{1|2}\left(x\,|y\right) = \begin{cases} \frac{p(x,y)}{p_2(y)} & \text{pro } p_2\left(y\right) > 0 \\ 0 & \text{jinak} \end{cases}$$

se nazývá **sloupcově podmíněná četnostní funkce.**

Funkce $p_{2|1}\left(y\,|x\right)$ zavedená vztahem $\forall y \in \mathbb{R}$:

$$p_{2|1}\left(y\,|x\right) = \begin{cases} \frac{p(x,y)}{p_1(x)} & \text{pro } p_1\left(x\right) > 0 \\ 0 & \text{jinak} \end{cases}$$

se nazývá **řádkově podmíněná četnostní funkce.**

Četnostní nezávislost

Znaky X, Y jsou v daném výběrovém souboru **četnostně nezávislé**, jestliže platí:

$$\forall j = 1, \ldots, r, \ \forall k = 1, \ldots, s : \quad p_{jk} = p_{j.} \cdot p_{.k}$$

neboli

$$\forall (x, y) \in \mathbb{R}^2 : \ p(x, y) = p_1(x) \cdot p_2(y).$$

Četnostní nezávislost – ekvivalentní definice

Znaky X, Y jsou v daném výběrovém souboru **četnostně nezávislé**, jestliže platí:

$$\forall y \in \mathbb{R},\ p_2(y) > 0: \quad p_{1|2}(x\,|y) = p_1(x)$$

resp.

$$\forall x \in \mathbb{R},\ p_1(x) > 0: \quad p_{2|1}(y\,|x) = p_2(y)\,.$$

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- sestavte kontingenční tabulku simultánních absolutních četností

x \ y (n_{jk})	1	2	3	4	$n_{j.}$
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	$n = 20$

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

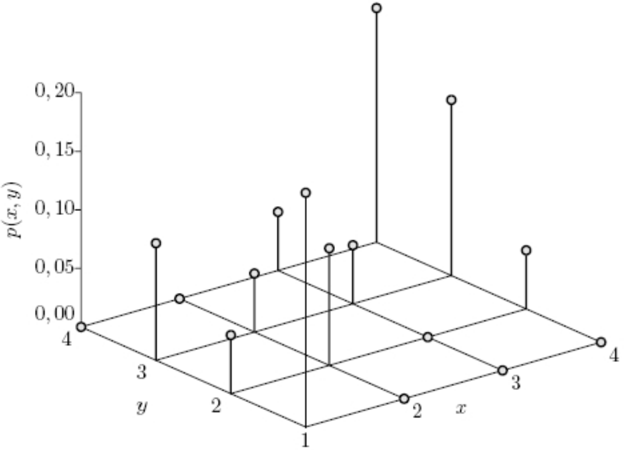
- sestavte kontingenční tabulku simultánních relativních četností

	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

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- nakreslete graf simultánní četnostní funkce $p(x, y)$



Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- sestavte kontingenční tabulku sloupcově podmíněných relativních četností

	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

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- sestavte kontingenční tabulku řádkově podmíněných relativních četností

	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

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- zjistěte, kolik procent firem, kterým první hodnotitel udělil jedničku, mělo od druhého hodnotitele dvojku

	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

Dvourozměrný datový soubor – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem"

- zjistěte, kolik procent firem, kterým druhý hodnotitel udělil jedničku, mělo od prvního hodnotitele dvojku

	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

Příklad 2

Na plicním oddělení jisté nemocnice bylo náhodně vybráno 20 pacientů a zjišťovalo se u nich pohlaví (znak X: 0 – muž, 1 – žena) a kuřáctví (znak Y: 0 – nekouří, 1 – kouří). Výsledky:

(0,0)	(1,0)	(1,1)	(1,0)	(0,1)	(0,1)	(1,0)	(0,1)	(1,0)	(0,0)
(1,0)	(0,1)	(0,1)	(1,0)	(1,0)	(1,1)	(0,0)	(0,0)	(1,0)	(1,1)

a) Sestrojte variační řady pro oba znaky

Variační řada pro znak X

	n_j	p_j	N_j	F_j
muž (0)	9	0,45	9	0,45
žena (1)	11	0,55	20	1,00

Variační řada pro znak Y

	n_j	p_j	N_j	F_j
nekouří (0)	12	0,6	12	0,6
kouří (1)	8	0,4	20	1,0

Příklad 2

b) Sestrojte kontingenční tabulku absolutních četností pro oba znaky

$X \setminus Y$	nekouří	kouří	$n_{i.}$
muž	4	5	9
žena	8	3	11
$n_{.j}$	12	8	20

Příklad 2

c) Zjistěte procento mužů, žen, kuřáků, nekuřáků.

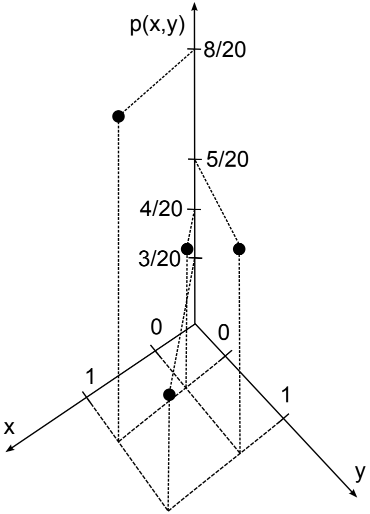
mužů je 45 %	kuřáků je 40 %
žen je 55 %	nekuřáků je 60 %

d) Kolik procent mužů kouří?
Mezi muži je $5/9 = 55{,}56$ % kuřáků. (z tabulky řádkově podmíněných četností)

e) Kolik procent kuřáků jsou muži?
Mezi kuřáky je $5/8 = 62{,}5$ % mužů. (z tabulky sloupcově podmíněných četností)

Příklad 2

f) Sestrojte graf dvourozměrného rozložení četností.



Intervalové rozdělení četností

Intervalové rozdělení četnosti

- V některých datových souborech je počet variant znaku příliš veliký a použití bodového rozdělení četností by vedlo k nepřehledným a roztříštěným výsledkům.
- Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X je blízký rozsahu souboru, pak přiřazujeme četnosti nikoliv jednotlivým variantám, ale celým intervalům hodnot. Hovoříme pak o **intervalovém rozdělení četností**.

Stanovení třídících intervalů

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

- j-tý třídicí interval znaku X, $j = 1, \dots, r$:

$$(u_j, u_{j+1}],$$

- délka j-tého třídicího intervalu znaku X:

$$d_j = u_{j+1} - u_j,$$

- střed j-tého třídicího intervalu znaku X:

$$x_{[j]} = \frac{1}{2}(u_j + u_{j+1}).$$

Stanovení třídících intervalů

Třídicí intervaly volíme nejčastěji stejně dlouhé. Jejich počet určíme např. pomocí **Sturgesova pravidla**:

$$r = 1 + 3,3 \log v,$$

kde v je **rozsah souboru**.

Charakteristiky intervalových dat

Nechť je dán jednorozměrný datový soubor rozsahu n. Hodnoty znaku X roztřídíme do r třídicích intervalů. Pro $j = 1, \ldots, r$ definujeme:

- **absolutní četnost** j-tého třídicího intervalu ve výběrovém souboru
$$n_j = N(u_j < X \leq u_{j+1}),$$
- **relativní četnost** j-tého třídicího intervalu ve výběrovém souboru
$$p_j = \frac{n_j}{n},$$
- **četnostní hustota** j-tého třídicího intervalu ve výběrovém souboru
$$f_j = \frac{p_j}{d_j},$$

Charakteristiky intervalových dat

- **absolutní kumulativní četnost** prvních j třídicích intervalů ve výběrovém souboru

$$N_j = N(X \leq u_{j+1}) = n_1 + \cdots + n_j,$$

- **relativní kumulativní četnost** prvních j třídicích intervalů ve výběrovém souboru

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

Charakteristiky intervalových dat

Tabulka typu

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

se nazývá **tabulka rozdělení četností**.

Histogram

Intervalové rozdělení 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, $j = 1, \dots, r$.

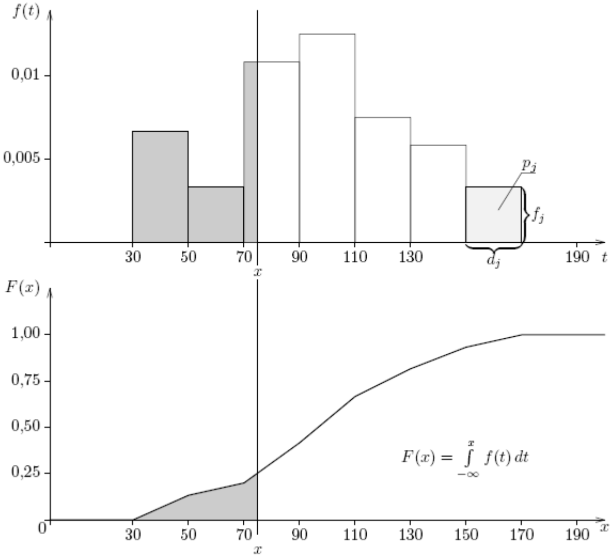
Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané **hustota četnosti**

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

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

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

Vztah histogramu a empirické distribuční funkce



Dvourozměrný soubor intervalových dat

Nechť je dán dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde hodnoty znaku X roztřídíme do r třídicích intervalů $(u_j, u_{j+1}]$, $j = 1, \ldots, r$ s délkami $d_1, \ldots, d_r$ a hodnoty znaku Y roztřídíme do s třídicích intervalů $(v_k, v_{k+1}]$, $k = 1, \ldots, s$ s délkami $h_1, \ldots, h_s$.
Pak definujeme:

- **simultánní absolutní četnost** (j, k)-tého třídicího intervalu:

$$n_{jk} = N(u_j < X \leq u_{j+1} \wedge v_k < Y \leq v_{k+1}),$$

- **simultánní relativní četnost** (j, k)-tého třídicího intervalu:

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

Dvourozměrný soubor intervalových dat

- **marginální absolutní četnost** j-tého třídicího intervalu pro znak X:
$$n_{j.} = n_{j1} + \cdots + n_{js},$$
- **marginální relativní četnost** j-tého třídicího intervalu pro znak X:
$$p_{j.} = \frac{n_{j.}}{n},$$
- **marginální absolutní četnost** k-tého třídicího intervalu pro znak Y:
$$n_{.k} = n_{1k} + \cdots + n_{rk},$$
- **marginální relativní četnost** k-tého třídicího intervalu pro znak Y:
$$p_{.k} = \frac{n_{.k}}{n},$$

- **simultánní četnostní hustota** v (j,k)-tém třídicím intervalu:

$$f_{jk} = \frac{p_{jk}}{d_j h_k},$$

- **marginální četnostní hustota** v j-tém třídicím intervalu pro znak X:

$$f_{j.} = \frac{p_{j.}}{d_j},$$

- **marginální četnostní hustota** v k-tém třídicím intervalu pro znak Y:

$$f_{.k} = \frac{p_{.k}}{h_k}.$$

Dvourozměrný datový soubor – kontingenční tabulka

Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky. Uveďme kontingenční tabulku simultánních absolutních četností:

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

Simultánní hustota četnosti

Funkce

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

se nazývá simultánní **hustota četnosti**.

Hustoty četnosti pro znaky X a Y (tzv. **marginální hustoty četnosti**) odlišíme indexem takto:

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

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

Podmíněné hustoty četnosti

Funkce $f_{1|2}\left(x\,|y\right)$ zavedená vztahem $\forall x \in \mathbb{R}$:

$$f_{1|2}\left(x\,|y\right) = \begin{cases} \frac{f(x,y)}{f_2(y)} & \text{pro } f_2\left(y\right) > 0 \\ 0 & \text{jinak} \end{cases}$$

se nazývá **sloupcově podmíněná hustota četnosti.**

Funkce $f_{2|1}\left(y\,|x\right)$ zavedená vztahem $\forall y \in \mathbb{R}$:

$$f_{2|1}\left(y\,|x\right) = \begin{cases} \frac{f(x,y)}{f_1(x)} & \text{pro } f_1\left(x\right) > 0 \\ 0 & \text{jinak} \end{cases}$$

se nazývá **řádkově podmíněná hustota četnosti.**

Četnostní nezávislost

Řekneme, že znaky X, Y jsou v daném výběrovém souboru **četnostně nezávislé** při intervalovém rozložení četností, jestliže

$$\forall j = 1, \ldots, r, \ \forall k = 1, \ldots, s : \quad f_{jk} = f_{j.} \cdot f_{.k}$$

neboli

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

Četnostní nezávislost – ekvivalentní definice

Znaky X, Y jsou v daném výběrovém souboru **četnostně nezávislé** při intervalovém rozložení četností, jestliže platí:

$$\forall y \in \mathbb{R},\ f_2(y) > 0 : \quad f_{1|2}(x\,|y) = f_1(x)$$

resp.

$$\forall x \in \mathbb{R},\ f_1(x) > 0 : \quad f_{2|1}(y\,|x) = f_2(y)\,.$$

Dvourozměrný datový soubor – příklad

U 50 náhodně vybraných srovnatelných firem byly zjišťovány náklady na reklamu v tisících Kč (znak X) a hrubý zisk opět v tisících Kč (znak Y).

$$
\begin{bmatrix} 58 & 178 \\ 68 & 173 \\ 56 & 170 \\ 60 & 170 \\ 61 & 173 \\ 71 & 181 \\ 85 & 184 \\ 80 & 170 \\ 52 & 172 \\ 72 & 182 \end{bmatrix}
\begin{bmatrix} 65 & 170 \\ 57 & 169 \\ 65 & 169 \\ 60 & 170 \\ 54 & 162 \\ 52 & 169 \\ 83 & 182 \\ 60 & 168 \\ 68 & 173 \\ 63 & 171 \end{bmatrix}
\begin{bmatrix} 72 & 177 \\ 90 & 192 \\ 57 & 176 \\ 51 & 168 \\ 81 & 190 \\ 73 & 177 \\ 75 & 179 \\ 71 & 180 \\ 66 & 178 \\ 67 & 182 \end{bmatrix}
\begin{bmatrix} 72 & 191 \\ 57 & 174 \\ 57 & 160 \\ 56 & 170 \\ 56 & 172 \\ 52 & 165 \\ 72 & 185 \\ 75 & 170 \\ 52 & 163 \\ 63 & 184 \end{bmatrix}
\begin{bmatrix} 63 & 172 \\ 58 & 163 \\ 64 & 174 \\ 52 & 168 \\ 55 & 164 \\ 67 & 173 \\ 60 & 170 \\ 55 & 160 \\ 62 & 172 \\ 70 & 171 \end{bmatrix}
$$

Dvourozměrný datový soubor – příklad

Pro znak X stanovte optimální počet třídicích intervalů podle Sturgesova pravidla, sestavte tabulku rozdělení četnosti, nakreslete histogram a graf intervalové empirické distribuční funkce.

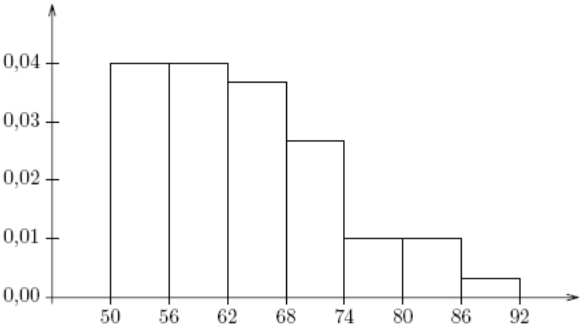
Optimální počet třídicích intervalů je 7. Tabulka rozdělení četností:

$(u_j, u_{j+1}\rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$(50, 56\rangle$	6	53	12	0,24000	12	0,24000	0,04000
$(56, 62\rangle$	6	59	12	0,24000	24	0,48000	0,04000
$(62, 68\rangle$	6	65	11	0,22000	35	0,70000	0,03667
$(68, 74\rangle$	6	71	8	0,16000	43	0,86000	0,02666
$(74, 80\rangle$	6	77	3	0,06000	46	0,92000	0,01000
$(80, 86\rangle$	6	83	3	0,06000	49	0,98000	0,01000
$(86, 92\rangle$	6	89	1	0,02000	50	1,00000	0,00333

Dvourozměrný datový soubor – příklad

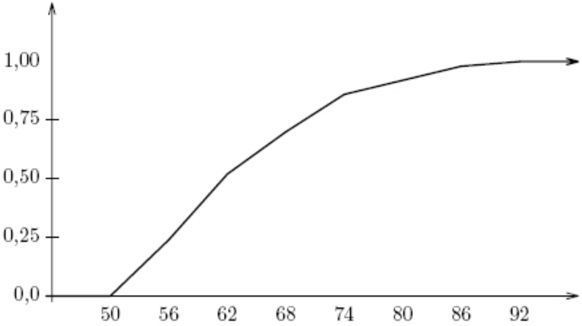
Pro znak X stanovte optimální počet třídicích intervalů podle Sturgesova pravidla, sestavte tabulku rozdělení četnosti, nakreslete histogram a graf intervalové empirické distribuční funkce.

Histogram:



Pro znak X stanovte optimální počet třídicích intervalů podle Sturgesova pravidla, sestavte tabulku rozdělení četnosti, nakreslete histogram a graf intervalové empirické distribuční funkce.

Graf intervalové empirické distribuční funkce:



Dvourozměrný datový soubor – příklad

Pro vektorový znak (X, Y) sestavte kontingenční tabulku absolutních četností a nakreslete dvourozměrný tečkový diagram.

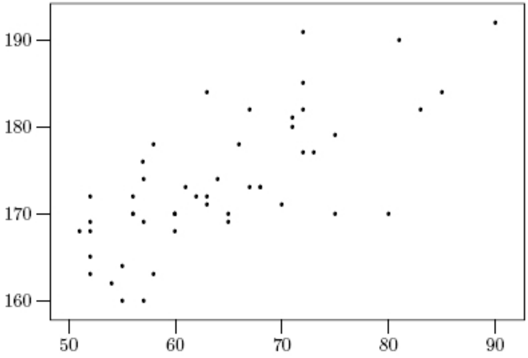
Optimální počet třídicích intervalů pro znak Y je 7. Kontingenční tabulka absolutních četností

$(u_j, u_{j+1}\rangle$ \ $(v_k, v_{k+1}\rangle$, n_{jk}	$(159, 164\rangle$	$(164, 169\rangle$	$(169, 174\rangle$	$(174, 179\rangle$	$(179, 184\rangle$	$(184, 189\rangle$	$(189, 194\rangle$	$n_{j.}$
$(50, 56\rangle$	4	4	4	0	0	0	0	12
$(56, 62\rangle$	2	2	6	2	0	0	0	12
$(62, 68\rangle$	0	1	7	1	2	0	0	11
$(68, 74\rangle$	0	0	1	2	3	1	1	8
$(74, 80\rangle$	0	0	2	1	0	0	0	3
$(80, 86\rangle$	0	0	0	0	2	0	1	3
$(86, 92\rangle$	0	0	0	0	0	0	1	1
$n_{.k}$	6	7	20	6	7	1	3	50

Dvourozměrný datový soubor – příklad

Pro vektorový znak (X, Y) sestavte kontingenční tabulku absolutních četností a nakreslete dvourozměrný tečkový diagram.

Dvourozměrný tečkový diagram



Číselné charakteristiky znaků

Typy znaků

Podle stupně kvantifikace znaky třídíme takto:

(n) **Nominální znaky** připouštějí obsahovou interpretaci jedině relace rovnosti $x_1 = x_2$ (popřípadě $x_1 \neq x_2$), tj. hodnoty znaku představují jen číselné kódy kvalitativních pojmenování.

Např. městské tramvaje jsou očíslovány, ale např. č. 4 a 12 říkají jen to, že jde o různé tratě: nic jiného se z nich o vztahu obou tratí nedá vyčíst.

Typy znaků

(o) **Ordinální znaky** připouštějí obsahovou interpretaci kromě relace rovnosti i v případě relace uspořádání $x_1 < x_2$ (popřípadě $x_1 > x_2$), tj. jejich uspořádání vyjadřuje větší nebo menší intenzitu zkoumané vlastnosti.

Např. školní klasifikace vyjadřuje menší nebo větší znalosti zkoušených (jedničkář je lepší než dvojkař), ale intervaly mezi známkami nemají obsahové interpretace (netvrdíme, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem. Podobný charakter mají různá bodování ve sportovních, uměleckých a jiných soutěžích.

Typy znaků

(i) **Intervalové znaky** připouštějí obsahovou interpretaci kromě relace rovnosti a uspořádání též u operace rozdílu $x_1 - x_2$ (popřípadě součtu $x_1 + x_2$), tj. stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v extenzitě zkoumané vlastnosti.

Např. teplota měřená ve stupních Celsia představuje intervalový znak. Naměříme-li ve čtyřech dnech polední teploty 0, 2, 4, 6, znamená to, že každým dnem stoupla teplota o 2 stupně Celsia. Bylo by však chybou interpretovat tyto údaje tvrzením, že ze druhého na třetí den vzrostla teplota dvakrát, kdežto ze třetího na čtvrtý pouze jedenapůlkrát.

Typy znaků

(p) **Poměrové znaky** umožňují obsahovou interpretaci kromě relace rovnosti a uspořádání a operace rozdílu ještě u operace podílu x_1/x_2 (popřípadě součinu $x_1 \cdot x_2$), tj. stejný poměr mezi jednou dvojicí hodnot a druhou dvojicí hodnot znamená i stejný podíl v extenzitě zkoumané vlastnosti.

Např. má-li jedna osoba hmotnost 150 kg a druhá 75 kg, má smysl prohlásit, že první je dvakrát hmotnější než druhá.

Typy znaků

Zvláštní postavení mají:

(a) **Alternativní znaky**, které nabývají jen dvou hodnot, např. 0, 1, což znamená absenci a prezenci nějakého jevu.

Například 0 bude znamenat neúspěch, 1 úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

Charakteristiky polohy

- Pro nominální znaky používáme jako charakteristiku polohy **modus**. U bodového rozdělení četností je to nejčetnější varianta znaku, u intervalového střed nejčetnějšího třídicího intervalu.
- Pro ordinální znaky používáme jako charakteristiku polohy α-**kvantil**. Jeli $\alpha \in (0, 1)$, 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 $1 - \alpha$ všech dat. Pro výpočet α-kvantilu slouží algoritmus:
  - $n\alpha$ je celé číslo c: $x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}$
  - $n\alpha$ je necelé číslo: zaokrouhlíme nahoru na nejbližší celé číslo c a $x_\alpha = x_{(c)}$.

  Pro speciálně zvolená α užíváme názvů: $x_{0.50}$ – medián, $x_{0.25}$ – dolní kvartil, $x_{0.75}$ – horní kvartil, $x_{0.1}, \ldots, x_{0.9}$ – decily, $x_{0.01}, \ldots, x_{0.99}$ – percentily.

Charakteristiky polohy

- Pro intervalové a poměrové znaky slouží jako charakteristika polohy aritmetický průměr

$$m_x = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Lze ho interpretovat jako těžiště jednorozměrného tečkového digramu.

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

I. hodnotitel	I. hodnotitel
2	4
1	4
4	2
1	4
1	2
4	4
3	1
3	4
1	4
1	1

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

α	$n\alpha$	c		x_α
0.25	$20 \cdot 0.25 = 5$	5	$\frac{(1+1)}{2}$	1

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

α	$n\alpha$	c		x_α
0.50	$20 \cdot 0.5 = 10$	10	$\frac{(2+3)}{2}$	2.5

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

α	$n\alpha$	c		x_α
0.75	$20 \cdot 0.75 = 15$	15	$\frac{(4+4)}{2}$	4

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

α	$n\alpha$	c		x_α
0.36	$20 \cdot 0.36 = 7.2$	8	2	2

Charakteristiky polohy – příklad

Pro datový soubor "hodnocení finančního zdraví několika firem I. hodnotitelem" vypočtěte medián a oba kvartily.

Hodnoty	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
Pořadí	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

α	$n\alpha$	c		x_α
0.25	$20 \cdot 0.25 = 5$	5	$\frac{(1+1)}{2}$	1
0.50	$20 \cdot 0.5 = 10$	10	$\frac{(2+3)}{2}$	2.5
0.75	$20 \cdot 0.75 = 15$	15	$\frac{(4+4)}{2}$	4
0.36	$20 \cdot 0.36 = 7.2$	8	2	2

Jako charakteristika variability může sloužit **kvartilová odchylka**

$$IQR = x_{0.75} - x_{0.25}.$$

Nejpoužívanější charakteristikou variability je však **rozptyl**

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2$$

či **směrodatná odchylka** $s_x = \sqrt{s_x^2}$.

Charakteristiky variability

- Pomocí průměru a směrodatné odchylky zavedeme **standardizovanou hodnotu** $\frac{x_i - m_x}{s_x}$ (vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru).
- Rozptyl vychází v kvadrátech jednotek, v nichž byl měřen znak X, proto raději používáme směrodatnou odchylku s.
- Pro poměrové znaky používáme jako charakteristiku variability **koeficient variace** $\frac{s_x}{m_x}$. Je to bezrozměrné číslo, které se často vyjadřuje v procentech. Umožňuje porovnat variabilitu několika znaků.
- Jsou-li všechny hodnoty poměrového znaku kladné, pak jako charakteristiku polohy lze užít geometrický průměr $\sqrt[n]{x_1 \cdot \dots \cdot x_n}$.

Dvourozměrný datový soubor – charakteristiky

Pro dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde znaky X a Y jsou intervalového či poměrového typu, používáme jako charakteristiku společné variability znaků X a Y kolem jejich průměrů **kovarianci**

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_x)(y_i - m_y).$$

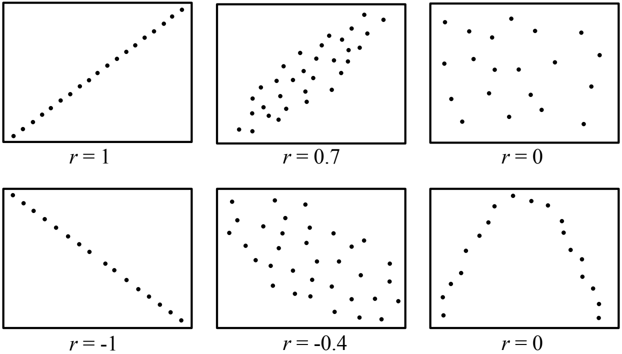
Jsou-li směrodatné odchylky s_x, s_y nenulové, pak definujeme **koeficient korelace** znaků X, Y vzorcem

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

Pro koeficient korelace platí $-1 \leq r_{xy} \leq 1$ a rovnosti je dosaženo právě když mezi hodnotami $x_1, \ldots, x_n$ a $y_1, \ldots, y_n$ existuje úplná lineární závislost, tj. existují konstanty a, b tak, že $y_i = a + bx_i$, $i = 1, \ldots, n$, přičemž znaménko $+$ platí pro $b > 0$, znaménko $-$ pro $b < 0$.

Dvourozměrný datový soubor – charakteristiky

Představu o významu hodnot koeficientu korelace podávají následující dvourozměrné tečkové diagramy.



Vážené číselné charakteristiky

- Vážený aritmetický průměr

$$m = \frac{1}{n} \sum_{j=1}^{r} n_j x_{[j]}$$

- Vážený rozptyl

$$s^2 = \frac{1}{n} \sum_{j=1}^{r} n_j (x_{[j]} - m)^2$$

- Vážená kovariance

$$s_{12} = \frac{1}{n} \sum_{j=1}^{r} \sum_{k=1}^{r} n_{jk} (x_{[j]} - m_1)(y_{[k]} - m_2)$$

Vážené číselné charakteristiky - použití

Mějme data zadaná následujícím způsobem:

Výše dotace (v milionech)	1	2	5
Počet	4	3	1

- Hodnot je celkem 8, nikoliv 3 (častá chyba).
- Pokud máme spočítat průměr, můžeme to provést obvyklým způsobem:

$$m = \frac{1+1+1+1+2+2+2+5}{8},$$

- anebo úsporněji podle vzorce pro vážený průměr:

$$m = \frac{4 \cdot 1 + 3 \cdot 2 + 1 \cdot 5}{8}.$$

Regresní přímka

Regresní přímka

- Cílem regresní analýzy je vystižení závislosti hodnot znaku Y na hodnotách znaku X. Při tom je nutné vyřešit dva problémy:
  - jaký typ funkce použít k vystižení dané závislosti a
  - jak stanovit konkrétní parametry zvoleného typu funkce?
- Typ funkce určíme buď logickým rozborem zkoumané závislosti nebo se ho snažíme odhadnout pomocí dvourozměrného tečkového diagramu.

Zde se omezíme na lineární závislost $y = \beta_0 + \beta_1 x$. Odhady b_0 a b_1 neznámých parametrů β_0, β_1 získáme na základě dvourozměrného datového souboru metodou nejmenších čtverců. Požadujeme, aby průměr součtu čtverců odchylek skutečných a odhadnutých hodnot byl minimální, tj. aby výraz

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

nabýval svého minima vzhledem k β_0 a β_1. Tento výraz je minimální, jsou-li jeho první derivace podle β_0 a β_1 nulové. Stačí tyto derivace spočítat, položit je rovny 0 a řešit systém dvou rovnic o dvou neznámých, tzv. systém normálních rovnic.

Regresní přímka

Nechť je dán dvourozměrný datový soubor a přímka $y = \beta_0 + \beta_1 x$. Výraz

$$C = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

se nazývá rozptyl hodnot znaku Y kolem přímky $y = \beta_0 + \beta_1 x$. Přímka $y = b_0 + b_1 x$, jejíž parametry minimalizují rozptyl

$$y = \beta_0 + \beta_1 x$$

v celém dvourozměrném prostoru, se nazývá **regresní přímka** znaku Y na znak X.

Regresní přímka

- **Regresní odhad** i-té hodnoty znaku Y značíme

$$\hat{y}_i = b_0 + b_1 x_i, \quad i = 1, \ldots, n.$$

- Kvadrát koeficientu korelace znaků X, Y se nazývá **index determinace** a značí se ID^2.
  - Index determinace udává, jakou část variability hodnot znaku Y vystihuje regresní přímka.
  - Nabývá hodnot z intervalu $\langle 0, 1 \rangle$.
  - Čím je bližší 1, tím lépe vystihuje regresní přímka závislost Y na X.

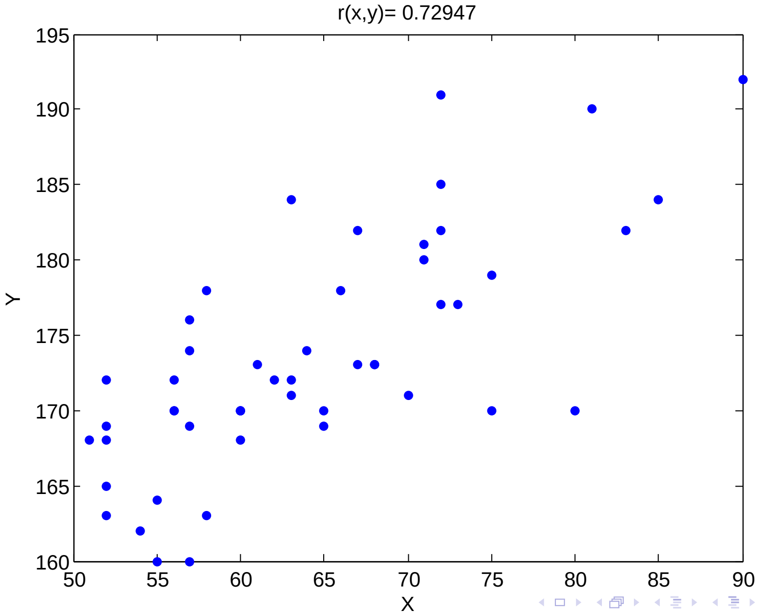
Regresní přímka

Nechť $y = b_0 + b_1 x$ je regresní přímka znaku Y na znak X. Pak použitím metody nejmenších čtverců dostaneme

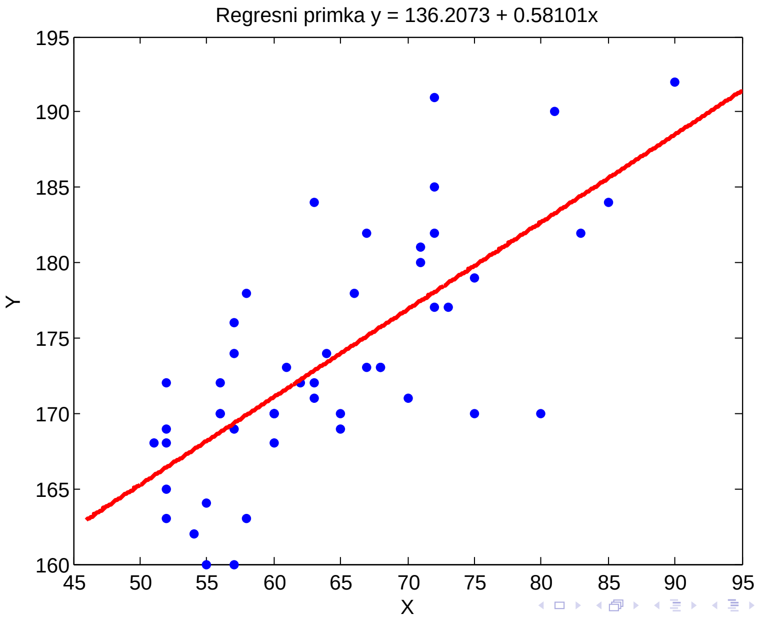
$$b_1 = \frac{s_{xy}}{s_x^2}, \quad b_0 = m_y - \frac{s_{xy}}{s_x^2} m_x.$$

- Parametr b_0 udává velikost posunutí regresní přímky na svislé ose (tj. udává, jaký je regresní odhad hodnoty znaku Y, nabývá-li znak X hodnoty 0).
- Směrnice b_1 udává, o kolik jednotek se změní hodnota znaku Y, změní-li se hodnota znaku X o jednotku.
- Jestliže je $b_1 > 0$, dochází s růstem X k růstu Y a hovoříme o přímé závislosti hodnot znaku Y na hodnotách znaku X.
- Je-li $b_1 < 0$, dochází s růstem X k poklesu Y a hovoříme o nepřímé závislosti hodnot znaku Y na hodnotách znaku X.

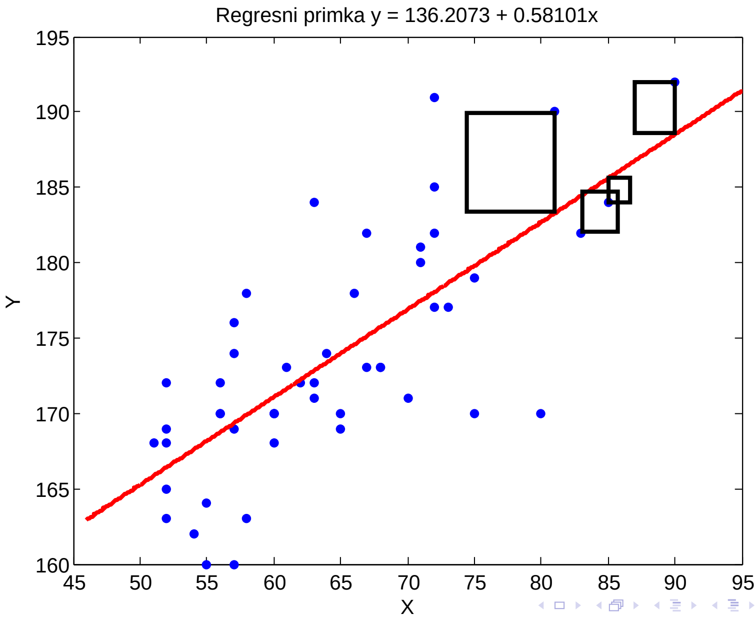
Regresní přímka – příklad



Regresní přímka – příklad



Regresní přímka – příklad



DĚKUJI ZA POZORNOST