Přednáška 08a

Od popisné statistiky k matematické statistice

Statistický soubor ... množina statistických jednotek

Měříme jeden nebo více znaků

Typy znaků
- nominální ($x_1 = x_2$?)
- ordinální ($x_1 < x_2$?)
- intervalové ... jsme navíc schopni posoudit rozdíl $x_1 - x_2$
- poměrové ... nerovnost, rozdíl, podíl

Nechť $x_1, x_2, \dots, x_n$ je soubor hodnot, které lze uspořádat, zjištěny měřením na n statistických jednotkách.

$$x_{(1)} \le x_{(2)} \le \dots \le x_{(n-1)} \le x_{(n)}$$

Různé hodnoty nechť jsou $a_1 < a_2 < \dots < a_m$, n_i je počet jednotek s naměřenou hodnotou a_i (důležitá četnost).

Průměry

① Aritmetický průměr

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\left(\sum_{j=1}^{m} n_j a_j\right)$$

Aritmetický průměr je invariantní vůči afinním transformacím

$$y_i = ax_i + b$$

Pak

$$\bar{y} = a\bar{x} + b$$

(2) Geometrický průměr pro kladná čísla

$$\bar{x}^G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

(3) Harmonický průměr pro kladná čísla

$$\bar{x}^H = \left\{ \frac{1}{n} \left(\sum_{i=1}^{n} \frac{1}{x_i} \right) \right\}^{-1}$$

Platí

$$\bar{x}^H \leq \bar{x}^G \leq \bar{x}.$$

Median, kvartil, percentil

Median souboru $x_1, x_2, \dots x_n$ je číslo $\tilde{x}$ takové, že 50% hodnot $x_1, x_2, \dots x_n$ je $\leq \tilde{x}$ a ostatní hodnoty jsou $> \tilde{x}$. Tím není $\tilde{x}$ určeno jednoznačně.

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

<u>n liché</u> $\qquad \tilde{x} = x_{\left(\frac{n+1}{2}\right)}$

<u>n sudé</u> $\qquad \tilde{x} = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$

Pro $0 < p < 1$ je p-tý kvantil, p-tý percentil číslo x_p takové, že p % hodnot je $\leq x_p$ a zbývající jsou $> x_p$. Není-li toto číslo určeno jednoznačně, bereme to jako aritmetický průměr dvou po sobě jdoucích

Hodnoty x_i 1 1 2 2 3 3 3 3

$$x_{0,5} = \tilde{x} = \frac{2+3}{2} = 2,5$$

$x_{0,25} = Q_1$ = dolní kvartil je $\frac{1+2}{2} = 1,5$

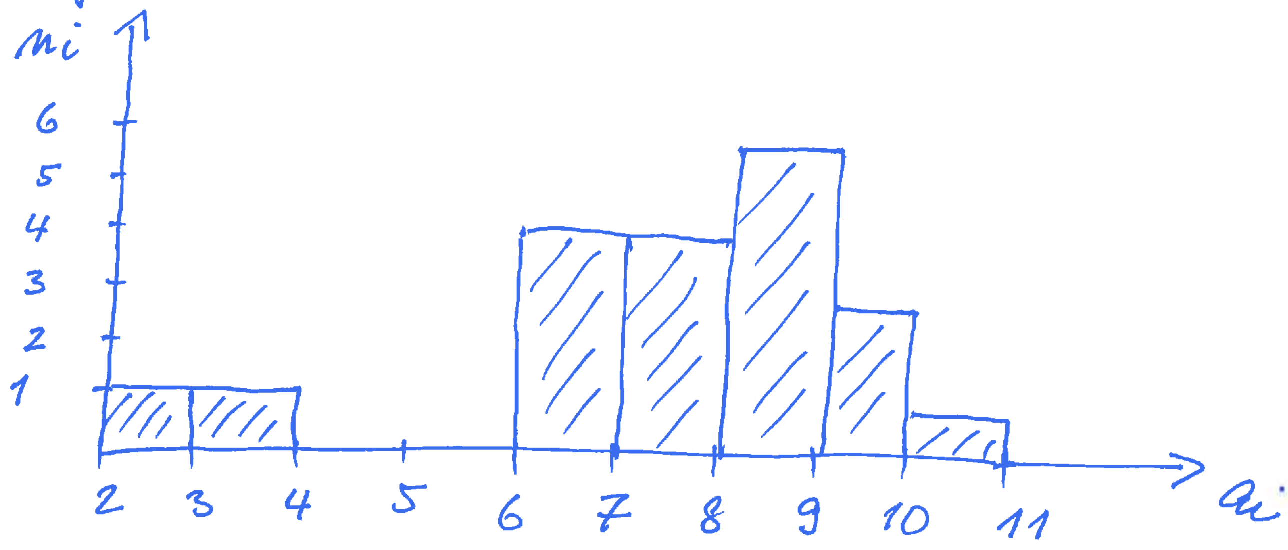
$x_{0,75} = Q_3$ = horní kvartil je $\frac{3+3}{2} = 3$

Modus je hodnota s největší četností $\hat{x} = 3$.

Příklad

a_i	3	4	7	8	9	10	11
n_i	1	1	4	4	6	3	1

Histogram

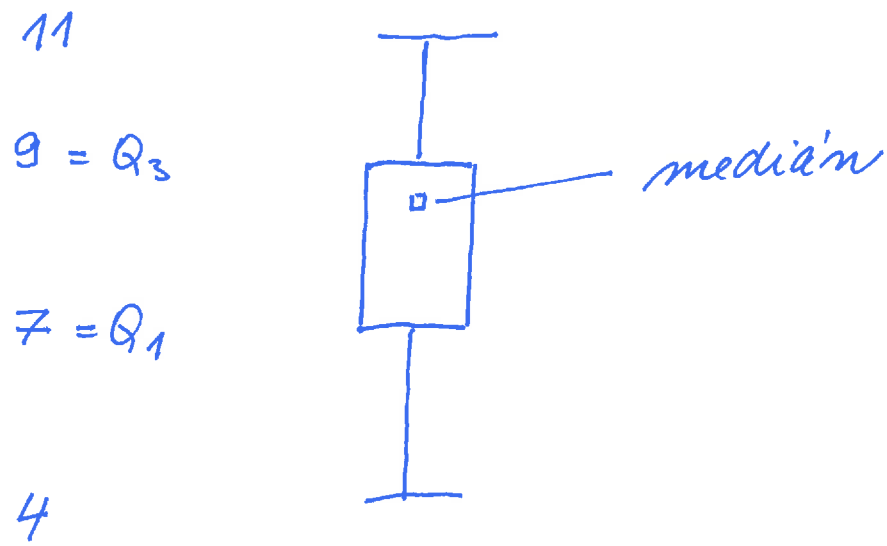


Krabicový graf

$n = 20$, $\bar{x} = 8{,}1$, $\tilde{x} = \dfrac{x_{(10)} + x_{(11)}}{2} = \dfrac{8+9}{2} = 8{,}5$

$$x_{0,25} = Q_1 = \frac{x_{(5)} + x_{(6)}}{2} = 7$$

$$x_{0,75} = Q_3 = \frac{x_{(15)} + x_{(16)}}{2} = 9$$



Rozptyl a směrodatná odchylka

Rozptyl $\quad s_x^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$s_x \geq 0$ je směrodatná odchylka

Průměrná odchylka

$$D_x = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

Věta (a) Funkce $S(t) = \frac{1}{n} \sum_{i=1}^{n} (x_i - t)^2$ nabývá svého minima pro $t = \bar{x}$.

(b) Funkce $D(t) = \frac{1}{n} \sum_{i=1}^{n} |x_i - t|$ nabývá svého minima pro $t = \tilde{x}$.

Důkaz (a)

$$\sum_{i=1}^{n} (x_i - t)^2 = \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - t)^2 =$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - t)^2 + 2 \sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - t)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - t)^2 + 2 \underbrace{\left(\sum_{i=1}^{n} x_i - n\bar{x} \right)}_{= 0} (\bar{x} - t)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - t)^2$$

Odtud je vidět, že výraz nabývá svého minima pro $t = \bar{x}$.

(b) Zkoumáme postupně součty

$$|x_{(1)} - t| + |x_{(n)} - t|$$

$$|x_{(2)} - t| + |x_{(n-1)} - t|$$

atd.

Platí

$$|x_{(1)} - t| + |x_{(n)} - t| = |x_{(1)} - \tilde{x}| + |x_{(n)} - \tilde{x}| = x_{(n)} - x_{(1)} \quad \text{pro } t \in [x_{(1)}, x_{(n)}]$$

$$> |x_{(1)} - \tilde{x}| + |x_{(n)} - \tilde{x}| \quad \text{pro } t \notin [x_{(1)}, x_{(n)}]$$

$$|x_{(2)} - t| + |x_{(n-1)} - t| = |x_{(2)} - \tilde{x}| + |x_{(n-1)} - \tilde{x}| = x_{(n-1)} - x_{(2)}$$

$$\text{pro } t \in [x_{(2)}, x_{(n-1)}]$$

a

$$|x_{(2)} - t| + |x_{(n-1)} - t| > |x_{(2)} - \tilde{x}| + |x_{(n-1)} - \tilde{x}|$$

$$\text{pro } t \notin [x_{(2)}, x_{(n-1)}] .$$

atd. Odtud dostaneme, že $D(t)$ nabývá svého minima pro $t = \tilde{x}$.

Nominální znaky

n_i počet jednotek s nominálním znakem a_i

a_1	a_2	a_3	...	a_k
n_1	n_2	n_3		n_k

$$\sum_{i=1}^{k} n_i = n$$

Míra variability znaků vyjadřuje entropie

$$H_X = - \sum_{i=1}^{k} \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right)$$

$k = 1 \qquad H_X = 0$

$k = 2 \qquad n_1 = n_2$

$$H_X = - \sum \frac{1}{2} \ln \frac{1}{2} = \ln 2 > 0 .$$

Entropie nabývá svého maxima při stejné četnosti znaků:

Hledáme maximum funkce

$$H(p_1, \dots p_k) = - \sum_{i=1}^{k} p_i \ln p_i$$

Za podmínky $\sum_{i=1}^{k} p_i = 1$ (a samozřejmě $p_i > 0$). Pomocí Lagrangeova multiplikátoru λ

$$\frac{\partial H}{\partial p_i} = -\ln p_i - 1 = \lambda$$

$$p_i = e^{\lambda-1}$$

$$k e^{\lambda-1} = 1$$

$$\lambda = 1 - \ln k$$

$$p_i = \frac{1}{k} .$$

V bodě $\left[\frac{1}{k}, \frac{1}{k}, \dots \frac{1}{k} \right]$ nabývá H svého lokálního maxima. Funkci $f(y) = -y \ln y$ lze spojitě rozšířit na $y=0$ neboť $\lim_{y \to 0_+} (-y \ln y) = 0$.
H nabývá na kompaktní množině

$$\left\{ \sum p_i = 1, \quad p_i \geq 0 \right\}$$

svého maxima, a to musí být v bodě $\left[\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k} \right]$.

Pravděpodobnost

Házíme kostkou - možné výsledky jsou 1, 2, 3, 4, 5, 6.
Můžeme mít situace, kdy možných výsledků je nekonečně mnoho :
Opakované hody mincí - rub a líc u jednoho hodu.
$w_k \in \mathbb{N} \cup \{\infty\}$ líc padne poprvé v k-tém hodu

Pravděpodobnost $P(w_k) = \frac{1}{2^k}$. $P(w_\infty) = 0$

$$\sum_{k=1}^{\infty} P(w_k) = 1.$$

Jevové pole – σ-algebra

Pracujeme se základní množinou Ω. Vybrané podmnožiny reprezentují jevy.

<u>Jevové pole</u> na Ω je systém podmnožin $\mathcal{A}$ množiny Ω takový, že

- $\Omega \in \mathcal{A}$
- $A, B \in \mathcal{A} \Rightarrow A \setminus B \in \mathcal{A}$
- $A_i \in \mathcal{A}$, I nejvýše spočetná $\Rightarrow \bigcup_{i \in I} A_i \in \mathcal{A}$

$\mathcal{A}$ nazýváme také σ-algebrou.

<u>Důsledky</u>:

- $\emptyset \in \mathcal{A}$ (nemožný jev, Ω je jistý jev)
- $A \in \mathcal{A}$, $A^c = \Omega \setminus A \in \mathcal{A}$ (doplněk)
- $A, B \in \mathcal{A} \Rightarrow A \cap B = A \setminus (\Omega \setminus B) \in \mathcal{A}$

Pravděpodobnostní prostor

je jevové pole $\mathcal{A}$ společně s funkcí $P: \mathcal{A} \to [0,1]$ takovou, že

(1) $P(\Omega) = 1$,

(2) A_i po dvou disjunktní, I nejvýše spočetná

$$P\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} P(A_i) .$$

Důsledky:

- $P(A^c) = 1 - P(A)$, $P(\emptyset) = 0.$
- $$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) - \sum_{1\le i<j\le k} P(A_i \cap A_j)$$
$$+ \sum_{1\le i<j<l\le k} P(A_i \cap A_j \cap A_l) - \dots$$

Věta

(1) Nechť $A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots \in \mathcal{A}$

Pak $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim\limits_{n\to\infty} P(A_n)$

(2) Nechť $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots \in \mathcal{A}$

Pak $P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim\limits_{n\to\infty} P(A_n).$

Důkaz (1) $A = \bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \dots$

Proto

$$P(A) = P(A_1) + \sum_{i=1}^{\infty} P(A_{i+1} \setminus A_i) = P(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n} P(A_{i+1} \setminus A_i)$$

$$= P(A_1) + \lim_{n\to\infty} \{P(A_2) - P(A_1) + P(A_3) - P(A_2) + \dots + P(A_{n+1}) - P(A_n)\}$$

$$= P(A_1) + \lim_{n\to\infty}\big(P(A_{n+1}) - P(A_1)\big) = \lim_{n\to\infty} P(A_{n+1}).$$

(2) $$P\Big(\bigcap_{i=1}^{\infty} A_i\Big) = 1 - P\Big(\bigcup_{i=1}^{\infty}(\Omega \setminus A_i)\Big) =$$

$$= 1 - \lim_{n\to\infty} P(\Omega \setminus A_n) = 1 - 1 + \lim_{n\to\infty} P(A_n)$$

$$= \lim_{n\to\infty} P(A_n)$$

V druhé rovnosti jsme použili (1).

Nezávislost jevů A a B

Jevy A a B jsou stochasticky nezávislé, jestliže

$$P(A \cap B) = P(A) \cdot P(B)$$

Základní představa o pravděpodobnostním prostoru: Ω konečná množina, $\mathcal{A}$ systém všech jejích podmnožin a

$$P(A) = \frac{|A|}{|\Omega|}.$$

Podmíněná pravděpodobnost

H je jev s nenulovou pravděpodobností.
Podmíněná pravděpodobnost jevu A za podmínky H je pravděpodobnost

$$P(A|H) = \frac{P(A \cap H)}{P(H)}$$

Lemma Nechť B je disj. sjednocením jevů $B_1 \cup B_2 \cup \ldots \cup B_n$. Pak

$$P(A|B) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i|B).$$

Speciálně pro $\Omega = B_1 \cup B_2 \cup \ldots \cup B_n$ disj.

$$P(A) = \sum P(A|B_i)\, P(B_i).$$

Důkaz:

$$P(A|B) = \frac{P(A \cap (B_1 \cup \ldots \cup B_n))}{P(B_1 \cup \ldots \cup B_n)} = \sum_{i=1}^{n} \frac{P(A \cap B_i)}{P(B)}$$

$$= \sum_{i=1}^{n} \frac{P(A \cap B_i)}{P(B_i)} \, \frac{P(B_i)}{P(B)} = \sum_{i=1}^{n} P(A|B_i)\, P(B_i|B).$$

Bayesova věta

(A) Bayesův vzorec pro inverzní pravděpodobnost

$$p(A|B) = \frac{p(A)\,p(B|A)}{P(B)}$$

(B) 1. Bayesův vzorec

$$P(A|B) = \frac{P(A)\,p(B|A)}{p(A)\,p(B|A) + p(A^c)\,p(B|A^c)}$$

(C) obecnější verze: Je-li $\Omega = A_1 \cup A_2 \cup \dots \cup A_n$ disjunktně, pak

$$P(A_i|B) = \frac{p(A_i)\,p(B|A_i)}{\sum_{j=1}^{n} p(A_j)\,p(B|A_j)}$$

Důkaz: (A)

$$\frac{p(A)\,p(B|A)}{p(B)} = \frac{p(A)\,\frac{p(B\cap A)}{p(A)}}{P(B)} = \frac{P(A\cap B)}{P(B)}$$

$$= P(A|B)$$

(B) a (C) za $P(B)$ dosadíme z lemmatu.

Příklad: v předmětu X je úspěšnost 40%, v předmětu Y je 80%. Každý předmět v daný den dělal stejný počet studentů.

Jaká je pravděpodobnost, že student, který uspěl, dělal předmět X?

A student dělal předmět X

B student u zkoušky uspěl

A^c student dělal předmět Y

Víme, že $P(B|A) = 0{,}4$, $P(B|A^c) = 0{,}8$.

Dále $P(A) = P(A^c) = 0{,}5$. Chceme spočítat $P(A|B)$. Dosazením do vzorce

$$P(A|B) = \frac{p(A)\cdot p(B|A)}{p(A)\,p(B|A) + p(A^c)\,p(B|A^c)}$$

$$= \frac{0{,}5\cdot 0{,}4}{0{,}5\cdot 0{,}4 + 0{,}5\cdot 0{,}8} = \frac{1}{3}.$$

■

Nezávislost jevů A a B pro $P(B) \neq 0$ je ekvivalentní s rovností

$$P(A) = P(A|B).$$