Cvičení 11: Nerovnosti, Popisná statistika a normální rozdělení

Teorie:

Datový soubor tvoří naměřené hodnoty: $x_1, x_2, \ldots, x_n$. Po seřazení označujeme

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

Aritmetický průměr: $\bar{x} = m = \frac{1}{n}\sum_{i=1}^{n} x_i$.
Modus $\hat{x}$ je nejčastější hodnota znaku. Průměrná odchylka: $o = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$.
Rozptyl $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, směrodatná odchylka $s = \sqrt{s^2}$.
Modifikovaný rozptyl $\frac{n}{n-1}s^2$.
Centrované hodnoty $x_i - \bar{x}$, , standardizované hodnoty $\frac{x_i - \bar{x}}{s}$.
p-**kvantil** $(0 < p < 1)$ je

$$\tilde{x}_p = \begin{cases} x_{([np]+1} & \text{pro } np \notin \mathbb{Z} \\ \frac{1}{2}(x_{(np)} + x_{(np)+1}) & \text{pro } np \in \mathbb{Z}, \end{cases}$$

kde $[a]$ značí celou část čísla a. Speciálně $\tilde{x}_{0,5}$ je medián, $\tilde{x}_{0,25}$, resp. $\tilde{x}_{0,75}$, dolní, resp. horní kvartil, $\tilde{x}_{0,1}, \tilde{x}_{0,2}, \ldots, \tilde{x}_{0,9}$ decily apod.

Mezikvartilové rozpětí je definováno jako $Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$.

Pro **dvourozměrný datový soubor** $[x_i, y_i]$, kde $1 \leq i \leq n$ definujeme *kovarianci prvního a druhého znaku* jako

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

a koeficient korelace mezi prvním a druhým znakem jako $r_{12} = \frac{s_{12}}{s_x s_y}$.

Krabicový diagram (box plot): dolní a horní strana základního obdélníka (krabice) odpovídá dolnímu a hornímu kvartilu, vodorovná čára uvnitř krabice mediánu (výška krabice je tedy mezikvartilové rozpětí). Dolní svislá úsečka (dolní fous) odpovídá hodnotám, které leží „pod krabicí“ ve vzdálenosti nejvýše 1,5 násobku její výšky, obdobně horní fous. Mimo fousy se znázorňují ostatní body (tzv. odlehlá pozorování). Křížek uvnitř krabice znázorňuje aritmetický průměr.

Příklad 1. Byly naměřeny následující hodnoty nějakého znaku

$$10; 7; 7; 8; 8; 9; 10; 9; 4; 9; 10; 9; 11; 9; 7; 8; 3; 9; 8; 7.$$

Určete aritmetický průměr, medián, kvartily, rozptyl a příslušný krabicový diagram.

Příklad 2. Mějme nezápornou náhodnou veličinu X se střední hodnotou μ.

1. Bez dalších informací o rozdělení X odhadněte $P(X > 3\mu)$.

2. Víte-li, že $X \sim \mathrm{Ex}(\frac{1}{\mu})$, vypočtěte $P(X > 3\mu)$.

Příklad 3. Určete pravděpodobnost, že při 1200 hodech kostkou padne šestka alespoň 150 krát a nejvýše 250 krát

1. pomocí Čebyševovy nerovnosti,

2. pomocí de Moivre-Laplaceovy věty.

Příklad 4. Průměrná rychlost větru je na určitém místě 20 km/hod.

- Bez ohledu na rozdělení rychlosti větru jako náhodné veličiny určete pravděpodobnost, že při jednom pozorování rychlost větru nepřesáhne 60 km/h.

- Určete interval, v němž se bude rychlost větru nacházet s pravděpodobností alespoň 0,9, víte-li navíc, že směrodatná odchylka $\sigma = 1$ km/hod.

Příklad 5. Na FI je 10% studentů s prospěchem do 1,2. Jak velkou skupinu je třeba vybrat, aby s pravděpodobností 0,95 v ní bylo 8-12% studentů s prospěchem do 1,2? Úlohu řešte zvlášť pomocí Čebyševovy a zvlášť pomocí Moivre-Laplaceovy věty.

Náhodným výběrem rozsahu n rozumíme n-tici **stochasticky nezávislých** a náhodných veličin $X_1, \ldots, X_n$, které mají totéž rozdělení. S náhodným výběrem se obvykle setkáváme při opakovaném provádění téhož pokusu.

Statistika je náhodná veličina vzniklá transformací náhodného výběru.

- Výběrový průměr $M = \frac{1}{n}\sum_{i=1}^{n} X_i$, a jsou-li navíc $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, pak $M \sim N(\mu, \sigma^2/n)$.

- Výběrový rozptyl $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 = \frac{1}{n-1}\big(\sum_{i=1}^{n} X_i^2 - nM\big), S = \sqrt{S^2}$.

Intervalovým odhadem parametru θ rozumíme interval (T_L, T_U), kde $T_L(X_1, \ldots, X_n)$ a $T_U(X_1, \ldots, X_n)$ jsou statistiky výběru $(X_1, \ldots, X_n)$. Platí-li

$$P(T_L \leq \theta \leq T_U) = 1 - \alpha,$$

říkáme, že (T_L, T_U) je $100 \cdot (1 - \alpha)\%$ interval spolehlivosti pro parametr θ. **Horním odhadem** parametru θ na hladině významnosti $1 - \alpha$ je statistika U, pro níž

$$P(\theta < U) \geq 1 - \alpha,$$

dolním odhadem θ na hladině významnosti $1 - \alpha$ je pak statistika L, pro níž

$$P(L < \theta) \geq 1 - \alpha.$$

Případ, kdy je $X_1, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$:

- M a S^2 jsou nezávislé náhodné veličiny.
- $M \sim N(\mu, \sigma^2/n)$, a tedy $U = (M-\mu)/(\sigma/\sqrt{n}) \sim N(0,1)$.
- $K = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$.
- $\sum (X_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.
- $T = (M-\mu)/(S/\sqrt{n}) \sim t(n-1)$.

Příklad 6. Pravděpodobnost, že zasazený strom se ujme, je 0,8. Jaká je pravděpodobnost, že z 500 zasazených stromů se jich ujme aspoň 380?

Výsledek. 0,987.

Příklad 7. Ke každému jogurtu běžné značky je náhodně (rovnoměrně) přibalen obrázek některého z 26 hokejových mistrů světa. Kolik jogurtů si fanynka Věrka musí koupit, aby s pravděpodobností 0,95 získala alespoň 5 kartiček Jaromíra Jágra?

Příklad 8. Při 600 hodech kostkou padla jednička pouze 45 krát. Rozhodněte, jestli je možné tvrdit, že jde o ideální kostku na hladině $\alpha = 0{,}01$. Vše zdůvodněte a svůj závěr explicitně formulujte.

Příklad 9. Předpokládejme, že výška desetiletých chlapců má normální rozdělení $N(\mu, \sigma^2)$. S neznámou střední hodnotou μ a rozptylem $\sigma^2 = 39{,}112$. Změřením výšky 15 chlapců jsme určili výběrový průměr $M = 139{,}13$. Určete

a) 99% oboustranný interval spolehlivosti pro parametr μ,

b) dolní odhad μ na hladině významnosti 95%.

Výsledek. a) $(134{,}97; 143{,}29)$; b) 136,474.