Popisná statistika

základní soubor X **výběrový soubor**

Naměřili jsme n hodnot
$$x_1, x_2, \ldots, x_n,$$
počet prvků souboru je tzv. **rozsah** souboru. Pro lepší zpracování data uspořádáme:
$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$
a dostaneme **uspořádaný soubor hodnot**

Míry polohy

Průměr (resp. výběrový, aritmetický průmer)
$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

p-kvantil (výběrový p-kvantil)
$$\tilde{x}_p = \begin{cases} x_{([np]+1)} & np \neq [np] \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}) & np = [np] \end{cases},$$
kde $[a]$ značí celou část z a a $0 < p < 1$.

Míry variability

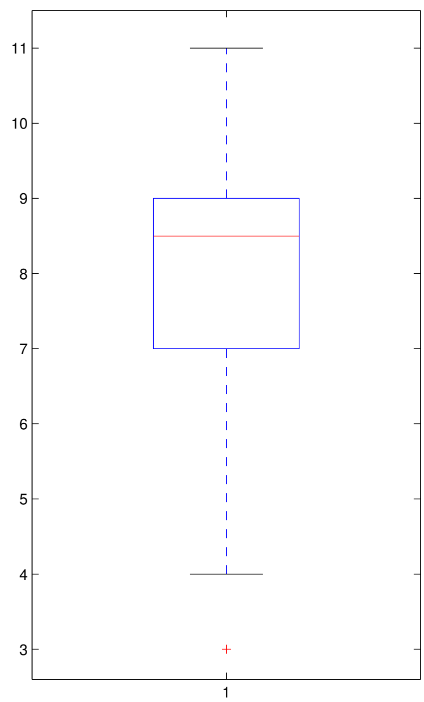
Rozptyl (výběrový rozptyl)
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right)$$

Kvartilové rozpětí
$$R_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$$

Krabicový diagram (box plot, box and whisker plot, vousatá krabička)

”Krabička” je ohraničena hodnotami kvartilů a je zobrazen medián. Vousky znázorňují hodnoty, které nejsou od jednotlivých kvartilů vzdálené o více jak 1,5 násobek R_Q. Jednotlivě jsou vyznačena pozorování, která jsou ve větší vzdálenosti.

1. Byly naměřeny hodnoty nějakého jevu:
10; 7; 7; 8; 8; 9; 10; 9; 4; 9; 10; 9; 11; 9; 7; 8; 3; 9; 8; 7
Určete průměr, medián, kvartily, rozptyl, mezikvartilové rozpětí a hodnoty znázorněte pomocí krabicového diagramu.



Náhodný výběr

Náhodným výběrem (rozsahu n) nazýváme posloupnost n stochasticky nezávislých náhodných veličin $X_1, X_2 \ldots, X_n$, které mají stejné rozložení, tedy $X_i \sim F(x_i),\ i = 1, 2, \ldots, n$.

Pozn.: Prakticky se s náhodným výběrem setkáváme při nezávislém vícenásobném opakování téhož pokusu.

Statistika: Náhodná veličina, která vznikne transformací náhodného výběru, se nazývá statistika.

Významné statistiky:

- *Výběrový průměr*

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- *Výběrový rozptyl*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - nM^2 \right)$$

- *Výběrová směrodatná odchylka*

$$S = \sqrt{S^2}$$

2. Nechť $X_1, X_2 \ldots, X_n$ je náhodý výběr z rozložení, které má střední hodnotu μ a rozptyl σ^2. Vypočítejte střední hodnotu a rozptyl výběrového průměru M.

3. Předpokládejme, že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Určete pravděpodobnost, že

a) náhodně vybraný student bude mít výsledek nad 80 bodů

b) průměr výsledků náhodného výběru 10 studentů bude větší než 80 bodů.

Bodové a intervalové odhady

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z rozložení daného distribuční funkcí $F(x_i)$.

Nestranný odhad: Statistika $T = g(X_1, X_2, \ldots, X_n)$ (kde g je borelovská funkce) je nestranný odhad parametru θ, právě když platí $E(T) = \theta$.

- Jsou-li T_1, T_2 dva nestranné odhady parametru θ, pak řekneme, že T_1 je lepší nestranný odhad než T_2, právě když platí $D(T_1) < D(T_2)$.
- Řekneme, že T^* je **nejlepší nestranný odhad** parametru θ, pokud je nestranným odhadem a pokud platí $D(T^*) \leq D(T)$, kde T je jakýkoli nestranný odhad parametru θ.

Intervalový odhad: Nechť $\alpha \in (0; 1)$ je libovolné číslo a $D = g_1(X_1, X_2, \ldots, X_n)$, $H = g_2(X_1, X_2, \ldots, X_n)$ jsou statistiky. Interval (D, H) se nazývá $100(1-\alpha)\%$ **interval spolehlivosti** pro parametr θ, právě když platí:

$$P(D < \theta < H) \geq 1 - \alpha$$

Statistika H se nazývá **horní odhad** parametru θ na hladině významnosti α, právě když platí:

$$P(\theta < H) \geq 1 - \alpha$$

Intervalové odhady pro paramtery μ a σ^2 jednoho normálního rozložení

1. Odhad parametru μ

   - pokud σ^2 známe

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$U = \frac{M-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

   $D = M - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, \quad H = M + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$

   - pokud σ^2 neznáme

$$T = \frac{M-\mu}{S/\sqrt{n}} \sim t(n-1)$$

   $D = M - \frac{S}{\sqrt{n}}t_{1-\alpha/2}(n-1), \quad H = M + \frac{S}{\sqrt{n}}t_{1-\alpha/2}(n-1)$

2. Odhad paramteru σ^2

   - pokud μ známe

$$W = \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} U_i^2 \sim \chi^2(n)$$

   $D = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \quad H = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}$

   - pokud μ neznáme

$$K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

   $D = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \quad H = \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}$

Intervaly spolehlivosti pro parametry dvou normálních rozložení

1. Interval spolehlivost $c_1\mu_1 + c_2\mu_2$

   - pokud σ_1, σ_2 známe

$$V = c_1M_1 + c_2M_2 = \frac{c_1}{n_1}\sum_{i=1}^{n} X_{1i} + \frac{c_2}{n_2}\sum_{i=1}^{n} X_{2i} \sim N\left(c_1\mu_1 + c_2\mu_2, \frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}\right)$$

$$U = \frac{(c_1M_1 + c_2M_2) - (c_1\mu_1 + c_2\mu_2)}{\sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}}} \sim N(0,1)$$

   $D = c_1M_1 + c_2M_2 - \sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2}$

   $H = c_1M_1 + c_2M_2 + \sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2}$

- pokud σ_1, σ_2 neznáme, ale víme, že jsou si rovny

$$T = \frac{(c_1 M_1 + c_2 M_2) - (c_1 \mu_1 + c_2 \mu_2)}{S_* \sqrt{c_1^2/n_1 + c_2^2/n_2}} \sim t(n_1 + n_2 - 2),$$

kde $S_*^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$
$D = c_1 M_1 + c_2 M_2 - t_{1-\alpha/2}(n_1 + n_2 - 2) S_* \sqrt{c_1^2/n_1 + c_2^2/n_2}$
$H = c_1 M_1 + c_2 M_2 + t_{1-\alpha/2}(n_1 + n_2 - 2) S_* \sqrt{c_1^2/n_1 + c_2^2/n_2}$

2. Interval spolehlivosti pro $\frac{\sigma_1^2}{\sigma_2^2}$

$$W = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

$D = \frac{S_1^2/S_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)}, \quad H = \frac{S_1^2/S_2^2}{F_{\alpha/2}(n_1-1, n_2-1)}$

4. Rychlost letadla byla určována v pěti zkouškách a z jejich výsledků byl vypočten odhad $m = 870,3\, m \cdot s^{-1}$*. Najděte 95% interval spolehlivosti pro* μ*, je-li známo, že rozptýlení rychlosti se řídí normálním rozložením se směrodatnou odchylkou* $\sigma = 2,1\, m \cdot s^{-1}$.

5. Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky v gramech jsou následující:
(62,52)',(54,56)',(55,49)',(60,50)',(53,51)',(58,50)'
Sestrojte 95% interval spolehlivosti pro $\mu = \mu_1 - \mu_2$.

6. Nechť $X_1, \dots X_n$ *je náhodný výběr z rozložení* $N(\mu; 0,04)$*. Zvolme hladinu významnosti* $\alpha = 0,05$*. Jaký musí být nejmenší počet měření, aby šířka intervalu spolehlivosti pro neznámou střední hodnotu* μ *nepřesáhla číslo 0,16?*

7. Při zjišťování přesnosti nově zaváděné metody pro stanovení obsahu manganu v oceli bylo rozhodnuto provést čtyři nezávislá měření u oceli se známým obsahem manganu, který je roven 0,30 %. Stanovte dolní odhad pro σ *na hladině významnosti* $\alpha = 0,05$*, když výsledky měření byly:*
0,31 %, 0,30 %, 0,29 %, 0,32 %. Údaje o obsahu manganu v oceli považujeme za realizace náhodného výběru rozsahu 4 z $N(\mu, \sigma^2)$

7. V tabulce jsou uvedeny výsledky analýz niklu získané dvěma analytickými metodami. Stanovte horní odhad pro podíl směrodatných odchylek obou metod při riziku $\alpha = 0,05$*, jestliže tyto výsledky považujeme za realizace nezávislých náhodných výběrů rozsahu 4 z* $N(\mu_1, \sigma_1^2)$ *a* $N(\mu_2, \sigma_2^2)$.
Metoda I: 3,26; 3,26; 3,27; 3,27
Metoda II: 3,23; 3,27; 3,29; 3,29