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*
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
Pokud $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, pak $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- *Výběrový rozptyl*
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$
- *Výběrová směrodatná odchylka*
$$S = \sqrt{S^2}$$
- *Výběrová kovariance*
$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$$

Bodové odhady parametrů

Parametrický prostor, parametrická funkce: Je dán měřitelný prostor $(\Omega, \mathcal{A})$, náh. veličina X a množina pravděpodobností $\{P_\theta; \theta \in \Theta\}$. Množina Θ se nazývá *parametrický prostor* a její prvky *parametry*. Jakékoli zobrazení $h : \Theta \to \mathbb{R}^q$, kde $q \in \mathbb{N}$, se nazývá *parametrická funkce*.

- Statistika $T = g(X_1, \ldots, X_n)$ se nazývá **nestranný odhad** parametrické funkce $h(\theta)$, právě když pro každé $\theta \in \Theta$ platí $E_\theta(T) = h(\theta)$.
- T_1, T_2 jsou dva různé nestranné odhady param. fce $h(\theta)$. Řekneme, že T_1 je *lepší* nestranný odhad než T_2, právě když pro každé $\theta \in \Theta$ platí $D_\theta(T_1) < D_\theta(T_2)$.
- Posloupnost $T_1, \ldots, T_n, \ldots$ statistik tvoří posloupnost **asymptoticky nestranných odhadů** parametrické funkce $h(\theta)$, právě když pro každé $\theta \in \Theta$ platí:
$$\lim_{n\to\infty} E_\theta(T_n) = h(\theta)$$
- Posloupnost $T_1, \ldots, T_n, \ldots$ statistik tvoří posloupnost **konzistentních odhadů** parametrické funkce $h(\theta)$, právě když pro každé $\theta \in \Theta$ a $\varepsilon > 0$ platí:
$$\lim_{n\to\infty} P_\theta(|T_n - h(\theta)| < \varepsilon) = 1$$

Intervalové odhady parametrů

Intervalový odhad: Nechť $\alpha \in (0;1)$ je libovolné číslo a $D = g_1(X_1, X_2, \dots, X_n)$, $H = g_2(X_1, X_2, \dots, X_n)$ jsou statistiky. Interval (D, H) se nazývá $100 \cdot (1-\alpha)\%$ **interval spolehlivosti (konfidenční, toleranční)** 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:

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

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

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

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

1. Odhad parametru μ
   - pokud σ^2 známe
     $$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
     $D = \overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \quad H = \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$
   - pokud σ^2 neznáme
     $$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$
     $D = \overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1), \quad H = \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1)$
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
     $$K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$
     $D = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \quad H = \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}$

1. Odvoďte vztahy pro horní a dolní odhady parametrů μ a σ^2.
2. Rychlost letadla byla určována v pěti zkouškách a z jejich výsledků byl určen odhad $\overline{x} = 870,3\,\mathrm{m} \cdot \mathrm{s}^{-1}$. Určete 95% interval spolehlivosti pro μ, je-li známo, že rozptýlení rychlosti se řídí normálním rozdělením se směrodatnou odchylkou $\sigma = 2,1\,\mathrm{m} \cdot \mathrm{s}^{-1}$.

3. 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 s rizikem 0,05, když výsledeky měření byly: 0,31 %, 0,30 %, 0.29 %, 0,32 %. Údaje o obsahu manganu v oceli považujeme za realizace náhodného výběru rozsahu 4 z $N(\mu, \sigma^2)$.

4. Z populace stejně starých selat téhož plemene bylo vylosováno 6 selat a po dobu půl roku jim byla podávána táž výkrmná dieta. Byly zaznamenány průměrné denní přírůstky v dg. Z dřívějších pokusů je známo, že v populaci mívají takové přírůstky normální rozložení, avšak střední hodnota i rozptyl se měnívají. Přírůstky v dg: 62, 54, 55, 60, 53, 58. Při riziku $\alpha = 0.05$ odvoďte:

   (a) dolní odhad neznámé střední hodnoty μ při neznámé směrodatné odchylce σ;

   (b) intervalový odhad směrodatné odchylky σ.

5. Nechť $X_1, \ldots X_n$ je náhodný výběr z rozložení $N(\mu; 0,04)$. Zvolme riziko $\alpha = 0,05$. Jaký musí být nejmenší počet měření, aby šířka intervalu spolehlivost pro neznámou střední odnotu μ nepřesáhla číslo 0,16?

6. Hloubka moře se měří přístrojem, jehož systematická chyba je nulová a náhodné chyby měření mají normální rozložení se směrodatnou odchylkou $\sigma = 1\,\mathrm{m}$. Kolik měření je nutno provést, aby se hloubka stanovila s chybou nejvýše $\pm 0,25\,\mathrm{m}$ při riziku 0,05?

Intervaly spolehlivosti pro parametry dvou normálních rozložení

(a) Interval spolehlivost $c_1\mu_1 + c_2\mu_2$

- pokud σ_1, σ_2 známe

  $$V = c_1\overline{X}_1 + c_2\overline{X}_2 = \frac{c_1}{n_1}\sum_{i=1}^{n} X_{1i} + \frac{c_2}{n_2}\sum_{i=1}^{n} X_{2i} \sim N\left(c_1\mu_1 + c_2\mu_2, \frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}\right)$$

  $$U = \frac{(c_1\overline{X}_1 + c_2\overline{X}_2) - (c_1\mu_1 + c_2\mu_2)}{\sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}}} \sim N(0,1)$$

  $D = c_1\overline{X}_1 + c_2\overline{X}_2 - \sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2}$

  $H = c_1\overline{X}_1 + c_2\overline{X}_2 + \sqrt{\frac{c_1^2\sigma_1^2}{n_1} + \frac{c_2^2\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2}$

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

  $$T = \frac{(c_1\overline{X}_1 + c_2\overline{X}_2) - (c_1\mu_1 + c_2\mu_2)}{S_*\sqrt{c_1^2/n_1 + c_2^2/n_2}} \sim t(n_1 + n_2 - 2),$$

  kde $S_*^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

  $D = c_1\overline{X}_1 + c_2\overline{X}_2 - t_{1-\alpha/2}(n_1 + n_2 - 2)S_*\sqrt{c_1^2/n_1 + c_2^2/n_2}$

  $H = c_1\overline{X}_1 + c_2\overline{X}_2 + t_{1-\alpha/2}(n_1 + n_2 - 2)S_*\sqrt{c_1^2/n_1 + c_2^2/n_2}$

(b) Interval spolehlivosti pro $\frac{\sigma_1^2}{\sigma_2^2}$

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

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

7. Byla provedena čtyři nezávislá stanovení obsahu manganu u dvou vzorků oceli s různými obsahy manganu a byly získány výsledky:
   1. vzorek: 0,31 %, 0,30 %, 0,29 %, 0,32 %
   2. vzorek: 0,59 %, 0,57 %, 0,58 %, 0,57 %
   Stanovte 95% interval spolehlivosti pro rozdíl středních hodnot obsahu manganu $\mu_1 - \mu_2$. Údaje o obsahu manganu představují realizace náhodných výběrů rozsahu 4 z $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$ s neznámými, avšak shodnými rozptyly.

8. 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 ralizace nezávislých náhodných výběrů rozsahu 4 z $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.
   Metoda 1: 3.26, 3.26, 3.27, 3.27
   Metoda 2: 3.23, 3.27, 3.29, 3.29

9. 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 $\mu = \mu_1 - \mu_2$.