Odhady parametrů základního souboru

Úvodní poznámky

- **Základní soubor** můžeme popsat jeho parametry, např. střední hodnota μ, rozptyl σ^2 atd. Při praktických úlohách ovšem zpravidla nelze vyšetřit celou populaci, provádíme tzv. **náhodný výběr** z populace. Náhodný výběr popisujeme rovněž jeho parametry, např. výběrový průměr $\bar{x}$, výběrový rozptyl s^2 atd.

Úvodní poznámky

- Parametry populace jsou konstantní, parametry náhodného výběru jsou ovšem náhodné proměnné (provedením jiného náhodného výběru získáme jiné hodnoty parametrů) řídící se **výběrovým rozdělením**. Známe-li pro konkrétní výběrovou charakteristiku její výběrové rozdělení, jsem potom schopni odhadnout parametr celé populace.

Úvodní poznámky

- Rozlišujeme dva typy odhadů:
  1) **Bodový odhad** – parametr populace aproximujeme jedním číslem.
  2) **Intervalový odhad** – parametr populace aproximujeme intervalem, ve kterém jeho hodnota leží s určitou pravděpodobností.

Vlastnosti bodového odhadu

- Mějme náhodný výběr $x_1, x_2, \ldots, x_n$ pocházejícího z rozdělení pravděpodobnosti definovaného distribuční funkcí $F(x, \theta)$, kde θ je neznámý parametr. Bodový odhad parametru θ budeme značit $\hat{\theta}$.

Vlastnosti bodového odhadu

- Dobrý bodový odhad musí splňovat určité vlastnosti:
  1) **Nestrannost** (nevychýlenost, nezkreslenost).
  2) **Vydatnost** (eficience).
  3) **Konzistence**.
  4) **Dostatečnost**.

Vlastnosti bodového odhadu

- Odhad je nestranný, pokud se jeho střední hodnota rovná hledanému parametru, tedy:
$$E\hat{\theta} = \theta.$$
- Slabší formou nestrannosti je **asymptotická nestrannost**, odhad je asymptoticky nestranný, pokud:
$$\lim_{n \to \infty} E\hat{\theta} = \theta.$$
- Např. výběrový průměr je nestranným odhadem střední hodnoty.

Vlastnosti bodového odhadu

- Máme-li dva nestranné odhady, potom vybereme ten s menším rozptylem, tato vlastnost se nazývá vydatnost.
- Nestranný odhad, jehož rozptyl je nejmenší ze všech nestranných odhadů, se nazývá **nejlepší nestranný odhad**.
- Např. výběrový průměr je nejlepším nestranným odhadem střední hodnoty.

Vlastnosti bodového odhadu

- Odhad je konzistentní, pokud se s rostoucím rozsahem výběru zpřesňuje, musí tedy platit:

$$\lim_{n\to\infty} E\hat{\theta} = \theta \text{ a } \lim_{n\to\infty} D\hat{\theta} = 0.$$

- Např. výběrový průměr je konzistentním odhadem střední hodnoty.

Vlastnosti bodového odhadu

- Odhad je dostatečný, pokud obsahuje veškerou informaci o sledovaném parametru, kterou výběrový soubor poskytuje.

- Výběrový průměr je dostatečným odhadem střední hodnoty.

Konstrukce bodových odhadů

- Pro konstrukci bodových odhadů existuje více metod, my se budeme zabývat **metodou maximální věrohodnosti**.

Metoda maximální věrohodnosti

- Nechť náhodný výběr pochází z diskrétního rozdělení pravděpodobnosti definovaného pravděpodobnostní funkcí s neznámým parametrem θ. **Věrohodnostní funkce** je potom definována jako sdružená pravděpodobnostní funkce *n* nezávislých proměnných se stejným diskrétním rozdělením:

$$L(x_1,\ldots,x_n,\theta)=p(x_1,\theta)\cdot\ldots\cdot p(x_n,\theta)=\prod_{i=1}^{n}p(x_i,\theta).$$

Metoda maximální věrohodnosti

- Pochází-li výběr ze spojitého rozdělení s neznámým parametrem θ, je věrohodnostní funkce definována jako sdružená hustota pravděpodobnosti *n* nezávislých proměnných se stejným spojitým rozdělením:

$$L(x_1,\ldots,x_n,\theta)=f(x_1,\theta)\cdot\ldots\cdot f(x_n,\theta)=\prod_{i=1}^{n} f(x_i,\theta).$$

Metoda maximální věrohodnosti

- Jelikož je věrohodnostní funkce neznámého parametru θ, je nyní úkolem najít $\hat{\theta}$ tak, aby se maximalizovala hodnota věrohodnostní funkce, tedy:

$$L\left(x_1,\ldots,x_n,\hat{\theta}\right)=\max_{\theta}\left\{L\left(x_1,\ldots,x_n,\theta\right)\right\}.$$

- Při praktických výpočtech se místo věrohodnostní funkce pracuje s jejím přirozeným algoritmem.

Metoda maximální věrohodnosti

- Podmínku optimality můžeme tedy vyjádřit ve tvaru:

$$\frac{\partial \ln L(x_1,\ldots,x_n,\theta)}{\partial \theta}=0.$$

- Řešením této rovnice (v případě více neznámých parametrů řešením soustavy rovnic) získáme odhady neznámého parametru, resp. parametrů.

Metoda maximální věrohodnosti

- **Př. 1:** Je dán výběr $x_1, x_2, \dots, x_n$ pocházející z Poissonova rozdělení. Metodou maximální věrohodnosti odhadněte parametr λ.

- Poissonovo rozdělení je definováno pravděpodobnostní funkcí ve tvaru:

$$p(X = x_i) = \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda} \text{ pro } x_i = 0,1,2,\dots \text{ a } \lambda > 0.$$

Metoda maximální věrohodnosti

- Pro věrohodnostní funkci můžeme psát:

$$L(x_1,\dots,x_n,\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}\cdot e^{-\lambda}.$$

- Věrohodnostní funkci zlogaritmujeme a dále upravujeme:

$$\ln L(x_1,\dots,x_n,\lambda)=\ln\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}\cdot e^{-\lambda}=\sum_{i=1}^{n}\ln\left(\frac{\lambda^{x_i}}{x_i!}\cdot e^{-\lambda}\right)=$$

$$=\sum_{i=1}^{n}\left(\ln\lambda^{x_i}+\ln e^{-\lambda}-\ln x_i!\right)=\sum_{i=1}^{n}\left(x_i\cdot\ln\lambda-\lambda\cdot\ln e-\ln x_i!\right)=$$

Metoda maximální věrohodnosti

$$= \sum_{i=1}^{n} x_i \cdot \ln \lambda - \sum_{i=1}^{n} \lambda - \sum_{i=1}^{n} \ln x_i! = \ln \lambda \cdot \sum_{i=1}^{n} x_i - n \cdot \lambda - \sum_{i=1}^{n} \ln x_i!\,.$$

- Upravený logaritmus věrohodnostní funkce derivujeme podle neznámého parametru λ a vyhledáme extrém:

$$\frac{\partial L(x_1, \ldots, x_n, \lambda)}{\partial \lambda} = \frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i - n = 0$$

Metoda maximální věrohodnosti

- Získanou rovnici upravujeme:

$$\frac{1}{\lambda}\cdot\sum_{i=1}^{n}x_i-n=0,$$

$$\sum_{i=1}^{n}x_i=n\cdot\lambda\Rightarrow\hat{\lambda}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i=\bar{x}.$$

- Na závěr ověříme, zda nalezený extrém je maximem:

$$\frac{\partial^2 L(x_1,\ldots,x_n,\lambda)}{\partial\lambda^2}=-\frac{1}{\lambda^2}\cdot\sum_{i=1}^{n}x_i<0,$$ jedná se tedy o maximum.

Metoda maximální věrohodnosti

- **Př. 2:** Je dán výběr x_1, x_2,…, x_n pocházející z exponenciálního rozdělení. Metodou maximální věrohodnosti odhadněte parametr μ.

- Exponenciální rozdělení je definováno hustotou pravděpodobnosti ve tvaru:

$f\left(x_i\right) = \mu \cdot e^{-\mu \cdot x_i}$ pro $x_i > 0$ a $\mu > 0$.

Metoda maximální věrohodnosti

- Pro věrohodnostní funkci můžeme psát:

$$L(x_1,\dots,x_n,\mu)=\prod_{i=1}^{n}\mu\cdot e^{-\mu\cdot x_i}\,.$$

- Věrohodnostní funkci zlogaritmujeme a dále upravujeme:

$$\ln L(x_1,\dots,x_n,\mu)=\ln\prod_{i=1}^{n}\mu\cdot e^{-\mu\cdot x_i}=\sum_{i=1}^{n}\ln\left(\mu\cdot e^{-\mu\cdot x_i}\right)=$$

$$=\sum_{i=1}^{n}\left(\ln\mu+\ln e^{-\mu\cdot x_i}\right)=\sum_{i=1}^{n}\left(\ln\mu-\mu\cdot x_i\cdot\ln e\right)=\sum_{i=1}^{n}\ln\mu-\sum_{i=1}^{n}\mu\cdot x_i=$$

$$=n\cdot\ln\mu-\mu\cdot\sum_{i=1}^{n}x_i\,.$$

Metoda maximální věrohodnosti

- Upravený logaritmus věrohodnostní funkce derivujeme podle neznámého parametru μ a vyhledáme extrém:

$$\frac{\partial nL(x_1,\ldots,x_n,\mu)}{\partial\mu}=\frac{n}{\mu}-\sum_{i=1}^{n}x_i=0.$$

Metoda maximální věrohodnosti

- Získanou rovnici upravujeme:

$$\frac{n}{\mu}-\sum_{i=1}^{n}x_i=0,$$

$$\frac{n}{\mu}=\sum_{i=1}^{n}x_i \Rightarrow \hat{\mu}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\bar{x}}.$$

- Ověříme, zda nalezený extrém je maximem: $\frac{\partial^2 L(x_1,\ldots,x_n,\mu)}{\partial\mu^2}=-\frac{n}{\mu^2}<0,$ jedná se tedy o maximum.

Intervalové odhady

- Hledaný parametr aproximujeme intervalem (nazývá se **interval spolehlivosti**, resp. **konfidenční interval**), ve kterém jeho hodnota leží s určitou pravděpodobností – **spolehlivost odhadu**.
- Spolehlivost odhadu označujeme $1 - \alpha$, kde α se nazývá **hladina významnosti**, zpravidla se volí $\alpha = 0{,}05$ nebo $0{,}01$.
- S rostoucí spolehlivostí odhadu roste i šířka intervalu spolehlivosti.

Intervalové odhady

- Označme dolní mez konfidenčního intervalu T_d a horní mez T_h.
- Rozlišujeme 3 druhy intervalů spolehlivosti:
  1) Levostranný interval spolehlivosti:
     $$P(\theta > T_d) = 1 - \alpha,$$
  2) Pravostranný interval spolehlivosti:
     $$P(\theta < T_h) = 1 - \alpha,$$
  3) Oboustranný interval spolehlivosti:
     $$P(\theta < T_d) = P(\theta > T_h) = \frac{\alpha}{2} \Rightarrow P(T_d < \theta < T_h) = 1 - \alpha.$$

Intervalový odhad střední hodnoty

- Předpokladem je, že náhodný výběr pochází z normálního rozdělení.
- Mohou vzniknout dva případy:
  1) Známe směrodatnou odchylku σ normálního rozdělení, ze kterého pochází náhodný výběr.
  2) Neznáme směrodatnou odchylku σ normálního rozdělení, ze kterého pochází náhodný výběr.

Intervalový odhad střední hodnoty

- ad 1) Mějme náhodný výběr o rozsahu *n* a s průměrem $\bar{x}$ pocházející z normálního rozdělení se známým rozptylem σ^2.
- Víme, že pro náhodnou proměnnou Z platí:

$$Z = \frac{\bar{x} - \mu}{\sigma} \cdot \sqrt{n} \rightarrow N(0{,}1),$$

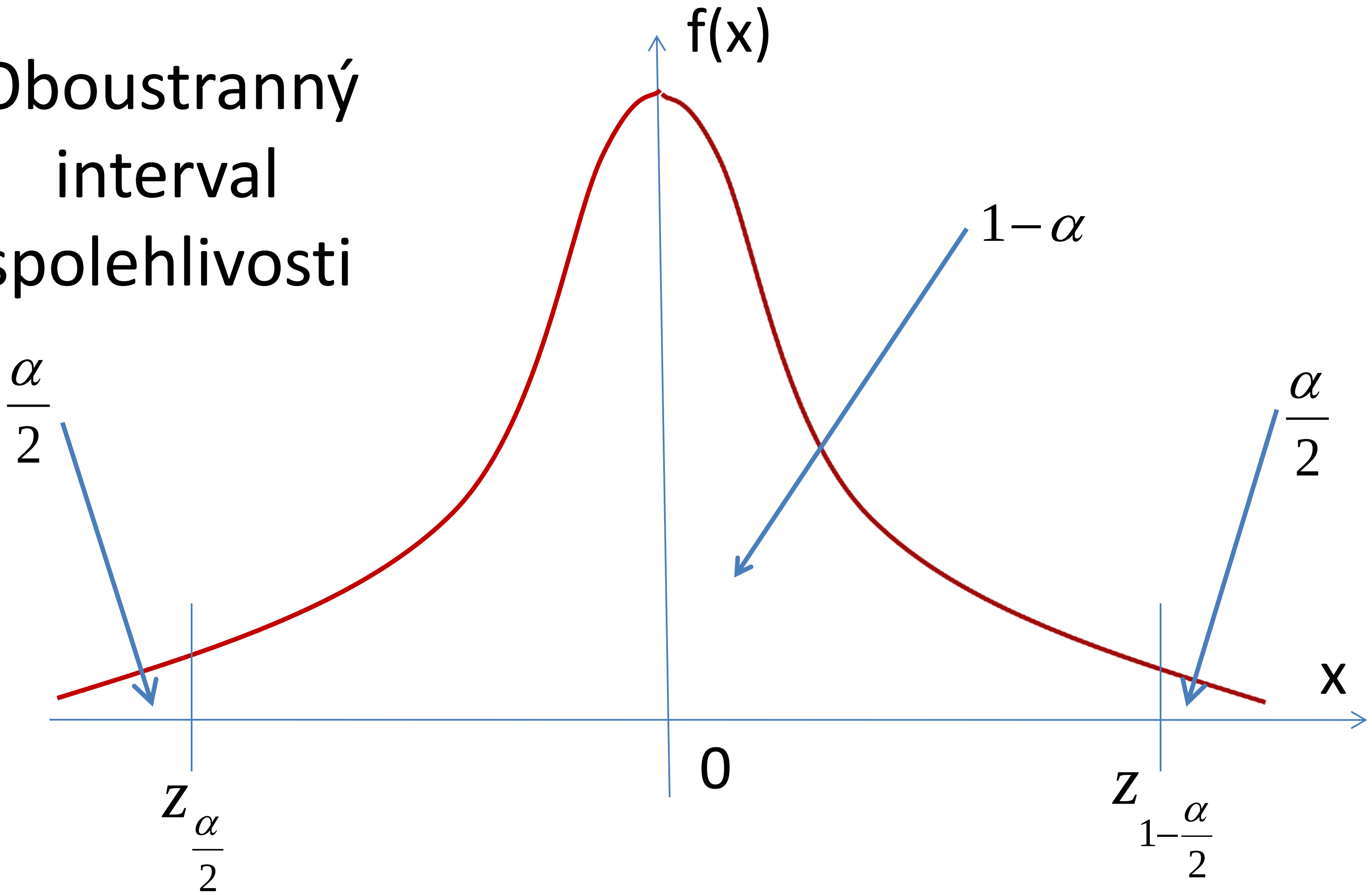
takto definovaná náhodná proměnná se tedy řídí normovaným rozdělením pravděpodobnosti.

Intervalový odhad střední hodnoty

- Odvoďme si nyní oboustranný interval spolehlivosti pro střední hodnotu. Označme $100 \cdot \frac{\alpha}{2}$ %-ní kvantil normovaného rozdělení jako $z_{\frac{\alpha}{2}}$ a $100 \cdot \left(1-\frac{\alpha}{2}\right)$ %-ní kvantil jako $z_{1-\frac{\alpha}{2}}$.

Intervalový odhad střední hodnoty

Oboustranný interval spolehlivosti



Intervalový odhad střední hodnoty

- Na základě obrázku můžeme psát:

$$P\left(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha .$$

- Dosadíme a postupně upravujeme:

$$P\left(z_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\sigma} \cdot \sqrt{n} < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha ,$$

$$P\left(\frac{\sigma}{\sqrt{n}} \cdot z_{\frac{\alpha}{2}} < \bar{x}-\mu < \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right) = 1-\alpha ,$$

Intervalový odhad střední hodnoty

$$P\left(\frac{\sigma}{\sqrt{n}}\cdot z_{\frac{\alpha}{2}}-\bar{x}<-\mu<\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}-\bar{x}\right)=1-\alpha\,,$$

$$P\left(\bar{x}-\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}<\mu<\bar{x}-\frac{\sigma}{\sqrt{n}}\cdot z_{\frac{\alpha}{2}}\right)=1-\alpha\,,$$

$$P\left(\bar{x}-\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}<\mu<\bar{x}+\frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right)=1-\alpha\,.$$

---

pozn. Jelikož je normované rozdělení souměrné, platí mezi kvantily vztah:

$$z_{\frac{\alpha}{2}}=-z_{1-\frac{\alpha}{2}}\,.$$

Intervalový odhad střední hodnoty

- Vidíme tedy, že dolní mez konfidenčního intervalu stanovíme dle vztahu:

$$T_d = \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

a horní mez podle vztahu:

$$T_h = \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}.$$

Intervalový odhad střední hodnoty

- Příslušnou hodnotu kvantilu normovaného rozdělení získáme buď z tabulek nebo s využitím funkce Excelu NORMSINV:

$$z_{1-\frac{\alpha}{2}} = NORMSINV\left(1-\frac{\alpha}{2}\right).$$

Intervalový odhad střední hodnoty

- V případě levostranného intervalu spolehlivosti získáme dolní hranici podle vztahu:

  $$T_d = \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$$

  v případě pravostranného intervalu získáme horní hranici podle vzorce:

  $$T_h = \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}.$$

Intervalový odhad střední hodnoty

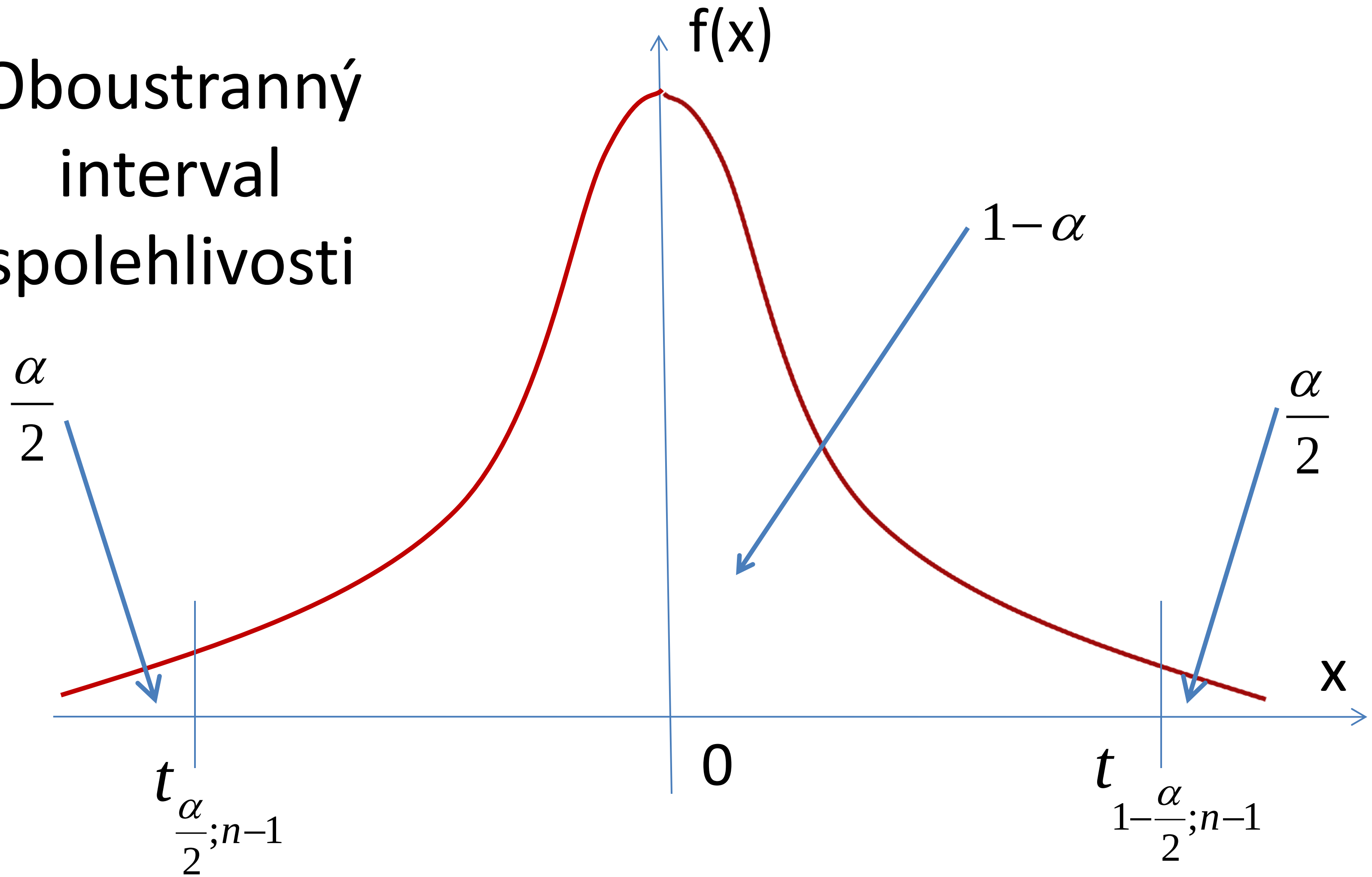
- ad 2) Mějme náhodný výběr o rozsahu n a s průměrem $\bar{x}$ pocházející z normálního rozdělení s neznámým rozptylem.
- Víme, že platí:

  $$T_{n-1} = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n} \rightarrow t_{n-1},$$

  testová statistika se tedy řídí Studentovým rozdělením pravděpodobnosti s $n - 1$ stupni volnosti.

Intervalový odhad střední hodnoty

Oboustranný interval spolehlivosti



Intervalový odhad střední hodnoty

- Na základě obrázku můžeme psát:

$$P\left(t_{\frac{\alpha}{2};n-1} < T_{n-1} < t_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha.$$

- Dosazením a analogickými úpravami (Studentovo rozdělení je rovněž souměrné) získáme konečný vztah:

$$P\left(\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};n-1} < \mu < \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};n-1}\right) = 1-\alpha.$$

Intervalový odhad střední hodnoty

- Vidíme tedy, že dolní mez konfidenčního intervalu stanovíme dle vztahu:

$$T_d = \bar{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};n-1}$$

a horní mez podle vztahu:

$$T_h = \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2};n-1}.$$

Intervalový odhad střední hodnoty

- Příslušnou hodnotu kvantilu Studentova rozdělení získáme buď z tabulek nebo s využitím funkce Excelu TINV:

$$t_{1-\frac{\alpha}{2};n-1} = TINV(\alpha; n-1).$$

Intervalový odhad střední hodnoty

- V případě levostranného intervalu spolehlivosti získáme dolní hranici podle vztahu:

  $$T_d = \bar{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha;n-1}$$

  v případě pravostranného intervalu získáme horní hranici podle vzorce:

  $$T_h = \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha;n-1}.$$

- Příslušný kvantil pomocí Excelu získáme:

  $$t_{1-\alpha;n-1} = TINV(2\alpha; n-1).$$

Intervalový odhad střední hodnoty

- V případě, když neznáme směrodatnou odchylku normálního rozdělení, ze kterého pochází náhodný výběr, ale máme výběr velkého rozsahu (tj. $n \geq 30$), můžeme Studentovo rozdělení aproximovat normovaným rozdělením, pro výpočet konfidenčního intervalu můžeme použít první vztahy ($\sigma = s$).