Odhady parametrů

Na statistický soubor $(x_1, \ldots, x_n)$, který dostaneme statistickým šetřením, se můžeme dívat jako na výběrový soubor získaný realizací náhodného výběru z náhodné veličiny X.

Obdobně: Na dvourozměrný statistický soubor $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$, který dostaneme statistickým šetřením, se můžeme dívat jako na výběrový soubor získaný realizací náhodného výběru z náhodného vektoru (X, Y).

Nyní se nabízí otázka: Nelze pomocí parametru statistického souboru $(x_1, \ldots, x_n)$, resp. $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$ přesně určit parametry náhodné veličiny X, resp. náhodného vektoru (X, Y)?

1. Poznámka Parametry, které nás nejvíce zajímají jsou číselné charakteristiky náhodné veličiny (náhodného vektoru), např. střední hodnota $E(X)$, rozptyl $D(X)$, koeficient korelace $\varrho\,(X,Y)$ apod.

Odpověď na výše položenou otázku je *ne*, parametry náhodné veličiny X či náhodného vektoru (X, Y) nelze přesně určit, ale lze je *odhadnout*.

2. Příklad Jestliže znám hmotnosti 50 hřídelí stejného typu vyrobených za stejných podmínek nemohu přesně určit střední hodnotu náhodné veličiny X, která reprezentuje hmotnost všech hřídelí vyrobených za stejných podmínek. Ale mohu střední hodnotu této veličiny odhadnout.

Bodové odhady

V technické praxi se nejčastěji používají bodové odhady parametrů náhodné veličiny.

3. Pojmy **Odhadem** T parametru ϑ je statistika $T(X_1, \ldots, X_n)$, která na celém parametrickém prostoru nabývá hodnot blízkých parametru ϑ. Používáme zejména tyto odhady:

1. Odhad T parametru ϑ je **nestranný** (**nevychýlený**), jestliže jeho střední hodnota $E(T) = \vartheta$. Pokud je $E(T) \neq \vartheta$, jde o **stranný** (**vychýlený**) odhad.
2. Je-li rozptyl nestranného odhadu T nejmenší z rozptylů všech nestranných odhadů téhož parametru ϑ, je T **nejlepší nestranný odhad**.
3. Odhad T je **konzistentní**, jestliže $\lim\limits_{n\to\infty} P\left(|T - \vartheta| < \varepsilon\right) = 1$ pro libovolné reálné číslo $\varepsilon > 0$.

4. Bodové odhady základních číselných charakteristik náhodné veličiny X na základě výběrových charakteristik:

a) $\bar{X}$ je nestranný konzistentní odhad střední hodnoty $E(X)$;

b) $\frac{n}{n-1}S^2$ je nestranný konzistentní odhad rozptylu $D(X)$.

Odhady a) a b) jsou pro normální rozdělení X také nejlepší.

5. Bodové odhady základních číselných charakteristik náhodné veličiny X resp. náhodného vektoru (X,Y) na základě empirických charakteristik statistického souboru:

a) $\bar{x}$ je bodovým odhadem střední hodnoty$E\,(X)$;

b) $\frac{n}{n-1}s^2$ je bodovým odhadem rozptylu $D\,(X)$;

c) $\sqrt{\frac{n}{n-1}}s$ je bodovým odhadem směrodatné odchylky$\sigma\,(X)$;

d) r je bodovým odhadem korelačního koeficientu $\varrho\,(X, Y)$;

kde $\bar{x}$, s^2, s resp. r jsou empirické charakteristiky získané ze statistického souboru $(x_1, \ldots, x_n)$ resp. $((x_1, y_1), \ldots, (x_n, y_n))$.

Intervalové odhady

6. Poznámka V dalším textu symbol = mezi parametrem náhodné veličiny ϑ a odhadem T neznačí ekvivalenci, ale odhad. Tedy např. $E(X) = \bar{x}$, čteme jako, střední hodnota $E(X)$ je odhadnuta aritmetickým průměrem $\bar{x}$.

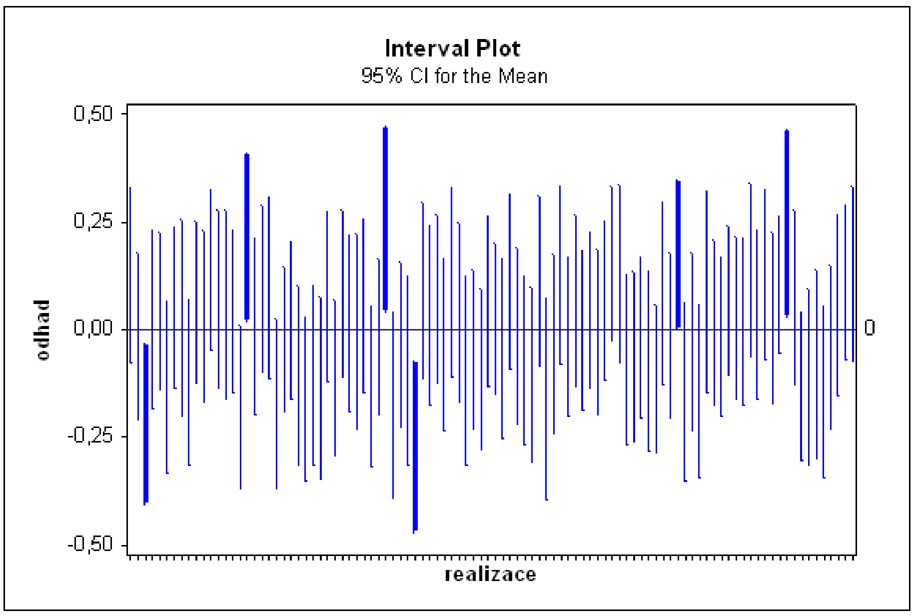
Nevýhodou je, že spolehlivost bodových odhadů (pravděpodobnost, že určíme parametr náhodné veličiny přesně) je nulová. Proto zavádíme intervalové odhady.

7. Pojmy **Interval spolehlivosti** (**konfidenční interval**) pro parametr ϑ **se spolehlivostí** $1-\alpha$, kde $\alpha \in \langle 0; 1\rangle$, je dvojice takových statistik $(T_1; T_2)$, že

$$P(T_1 \leq \vartheta \leq T_2) = 1 - \alpha$$

pro libovolnou hodnotu parametru ϑ. **Intervalový odhad** parametru ϑ **se spolehlivostí** $1-\alpha$ je interval $\langle t_1; t_2\rangle$ a píšeme $\vartheta \in \langle t_1; t_2\rangle$, kde t_1, t_2 jsou hodnoty statistik T_1, T_2 na daném statistickém souboru $(x_1, \ldots, x_n)$.

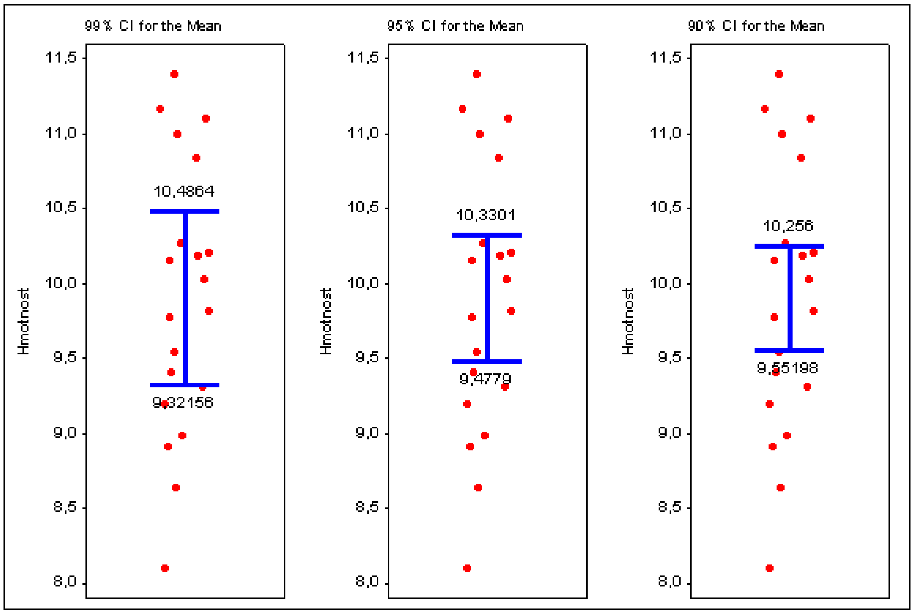
Spolehlivost $1-\alpha$ volíme blízkou jedné, podle konvence obvykle $0,95$ nebo $0,99$, a uvádíme ji také v %. Spolehlivost $1-\alpha$ znamená, že při mnoha opakovaných výběrech s konstantním rozsahem n z daného základního souboru zhruba $(1-\alpha)100$ % všech intervalových odhadů obsahuje skutečnou hodnotu parametru ϑ a naopak $\alpha 100$ % jich tuto hodnotu neobsahuje. Situaci ilustruje počítačově simulovaný příklad na Obrázku 1, kde $\vartheta = 0$ a tučně jsou vyznačeny případy odpovídající riziku chybného odhadu α, tj. intervalové odhady, které nezachytily hodnotu parametru ϑ.



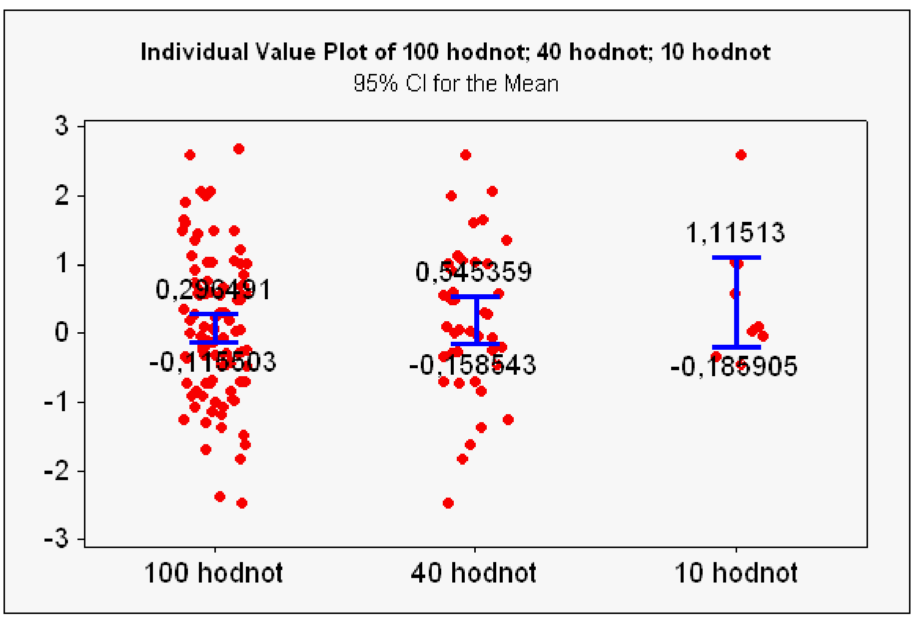
Obrázek 1: 6 intervalových odhadu ze 100 provedených intervalových odhadů se spolehlivostí 0,95 neobsahují odhadovanou hodnotu 0

S rostoucí spolehlivostí roste i rozpětí intervalového odhadu, takže pokud chceme intervalový odhad zúžit je možné snížit spolehlivost tohoto odhadu (viz Obrázek 2).

Tento přístup ovšem nedoporučujeme, protože chceme udržet relativně vysokou spolehlivost. Podstatně lepší je zvětšit rozsah souboru n, ovšem s ohledem na „kletbu statistiky“, neboť velikost intervalového odhadu se zmenší víceméně úměrně $\sqrt{n}$ (viz Obrázek 3).



Obrázek 2: Intervalové odhady střední hodnoty se spolehlivostí 0,99, 0,95, a 0,9 pro stejný statistický soubor



Obrázek 3: Intervalové odhady střední hodnoty pro náhodné výběry různého rozsahu

Odhady parametrů normálního rozdělení

8. Předpokládáme, že pozorovaná náhodná veličina X, resp. náhodný vektor (X, Y), má normální rozdělení pravděpodobnosti s parametry μ, σ^2, resp. ϱ.

Bodové odhady jsou

$$\mu = \bar{x},\ \sigma^2 = \frac{n}{n-1}s^2,\ \ \sigma = \sqrt{\frac{n}{n-1}}s,\ \ \varrho = r \quad .$$

Intervalový odhad střední hodnoty μ se spolehlivostí $1-\alpha$, při neznámém rozptylu σ^2 je

$$\left\langle \bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n-1}};\ \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n-1}} \right\rangle,$$

kde $t_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$- kvantil Studentova rozdělení $S(k)$ s $k = n-1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T2** Statistických tabulek, které jsou k dispozici na Matematice Online.

9. Intervalový odhad rozptylu σ^2 se spolehlivostí $1-\alpha$ je

$$\left\langle \frac{ns^2}{\chi^2_{1-\alpha/2}};\ \frac{ns^2}{\chi^2_{\alpha/2}} \right\rangle,$$

kde χ^2_P je P- kvantil Pearsonova rozdělení $\chi^2(k)$ s $k = n-1$ stupni volnosti. Kvantily tohoto rozdělení jsou uvedeny v tabulce **T3**. Z uvedeného intervalového odhadu získáme po odmocnění jeho mezí **intervalový odhad směrodatné odchylky σ se spolehlivostí $1-\alpha$**.

10. Příklad Měřením délky 10 válečků byl získán statistický soubor s empirickými charakteristikami $\bar{x} = 5,37$mm, $s^2 = 0,0019$mm^2 a $s = 0,044$mm. Určete bodové odhady střední hodnoty, rozptylu a směrodatné odchylky. Za předpokladu, že naměřená délka X má normální rozdělení pravděpodobnosti, určete intervalové odhady těchto číselných charakteristik se spolehlivostí 0,95.

Řešení Bodové odhady jsou:
střední délka válečku $\mu = 5,37$mm,
rozptyl délky válečku $\sigma^2 = \frac{10}{9}0,0019 = 0,00211$mm^2,
směrodatná odchylka délky válečku $\sigma = \sqrt{0,00211} \approx 0,046$mm.

Intervalový odhad střední délky válečku μ se spolehlivostí 0,95 je, neboť $t_{0,975}$= 2,262 pro 9 stupňů volnosti z tabulky **T2**,

$$\mu \in \left\langle 5,37 - 2,262\frac{\sqrt{0,0019}}{\sqrt{10-1}}; 5,37 + 2,262\frac{\sqrt{0,0019}}{\sqrt{10-1}} \right\rangle \approx \langle 5,337; 5,403\rangle \text{mm}.$$

Intervalový odhad rozptylu délky válečku σ^2 se spolehlivostí 0,95 je, neboť $\chi^2_{0,025} = 2,700$ a $\chi^2_{0,975} = 19,023$ pro 9 stupňů volnosti z tabulky **T3**,

$$\sigma^2 \in \left\langle \frac{10.0,0019}{19,023}; \frac{10.0,0019}{2,700} \right\rangle \approx \langle 0,00100; 0,00704\rangle \text{mm}^2,$$

takže intervalový odhad směrodatné odchylky délky válečku σ je

$$\sigma \in \left\langle \sqrt{0,00100}; \sqrt{0,00704} \right\rangle \approx \langle 0,0316; 0,0839\rangle \text{mm}.$$

Intervalový odhad koeficientu korelace ϱ se spolehlivostí $1-\alpha$ pro $n \geq 10$ je

$$\langle \operatorname{tgh} z_1; \operatorname{tgh} z_2 \rangle\,,$$

kde

$$z_1 = w - \frac{u_{1-\alpha/2}}{\sqrt{n-3}},\quad z_2 = w + \frac{u_{1-\alpha/2}}{\sqrt{n-3}},\quad w = \frac{1}{2}\left(\ln\frac{1+r}{1-r} + \frac{r}{n-1}\right),\quad \operatorname{tgh} z = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z}-1}{e^{2z}+1}$$

a $u_{1-\alpha/2}$ je $\left(1-\frac{\alpha}{2}\right)$- kvantil normovaného normálního rozdělení N(0;1), jehož hodnoty lze získat z tabulky **T1** s hodnotami distribuční funkce $\Phi(u)$. Pro $1-\alpha = 0,95$ je $u_{0,975} = 1,960$ a pro $1-\alpha = 0,99$ je $u_{0,995} = 2,576$. Uvedený odhad je pouze přibližný, avšak jeho přesnost je v praktických úlohách postačující (přesný odhad není znám).

11. Příklad Sledováním nákladů a ceny stejného výrobku u 10 výrobců byl získán dvourozměrný statistický soubor s koeficientem korelace $r = 0,82482$. Určete bodový odhad a intervalový odhad se spolehlivosti 0,99 koeficientu korelace ϱ základního souboru.

Řešení

Bodový odhad koeficientu korelace nákladů a ceny je $\varrho = 0,82482$.
Po dosazení je

$$w = \frac{1}{2}\left(\ln\frac{1+0,82482}{1-0,82482} + \frac{0,82482}{10-1}\right) \approx 1,21753.$$

Z tabulky **T1** je $u_{0,995} = 2,576$, takže

$$z_1 = 1,21753 - \frac{2,576}{\sqrt{10-3}} \approx 0,24397, \qquad z_2 = 1,21753 + \frac{2,576}{\sqrt{10-3}} \approx 2,19110$$

a intervalový odhad koeficientu korelace nákladů a ceny ϱ se spolehlivostí 0,99 je

$$\varrho \in \langle \operatorname{tgh} 0,24397; \operatorname{tgh} 2,19110\rangle \approx \langle 0,239242; 0,975313\rangle.$$