Matematika IV – 13. přednáška
Bodové a intervalové odhady

Michal Bulant

Masarykova univerzita
Fakulta informatiky

12. 5. 2008

Obsah přednášky

Doporučené zdroje

- Martin Panák, Jan Slovák, **Drsná matematika**, *e*-text.
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Popisná statistika**, Masarykova univerzita, 3. vydání, 2002, 48 stran, ISBN 80-210-1831-3.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.

Základní statistiky

Nechť $X_1, \ldots, X_n$ je náhodný výběr. (**Náhodným výběrem rozsahu** n rozumíme n-tici **nezávislých** a **stejně rozdělených** náhodných veličin $X_1, \ldots, X_n \sim F_X(x)$).
Statistiku $M = \frac{1}{n} \sum_{i=1}^{n} X_i$ nazýváme **výběrový průměr**, statistiku $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$ **výběrový rozptyl** a statistiku $S = \sqrt{S^2}$ **výběrová směrodatná odchylka**. Analogicky se definují i výběrová kovariance, příp. výběrový korelační koeficient pro dvourozměrný náhodný výběr.

Věta

Nechť $X_1, \ldots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 . Pak platí:

- $E(M) = \mu$,
- $D(M) = \mathrm{var}(M) = \sigma^2/n$,
- $E(S^2) = \sigma^2$.

Náhodný výběr z normálního rozdělení

Uvažme nyní speciální případ, kdy je $X_1, \ldots, X_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$.

Věta

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

Poznámka

K odhadu μ, neznáme-li σ^2, slouží T, v opačném případě U.
K odhadu σ^2, neznáme-li μ, slouží K, v opačném případě následující (bezejmenná?) statistika, která je vlastně statistikou K, v níž místo odhadu M použijeme přímo μ.

Důkaz.

Položme $Z_i = (X_i - \mu)/\sigma$, což jsou zřejmě nezávislé náhodné veličiny s normovaným normálním rozdělením. Zřejmě je $X = a + \sigma E_n Z$, kde $a = (\mu, \ldots, \mu)$ je vektor samých μ, a proto má X podle věty z předchozí přednášky mnohorozměrné normální rozdělení. Je-li dále d vektor ze samých $1/n$, pak má náhodná veličina $M = d^T X$ (jednorozměrné) normální rozdělení se střední hodnotou $d^T a = \mu$ a rozptylem $d^T \sigma^2 E_n d = \sigma^2/n$.

Ostatní tvrzení se dokážou obdobně. □

Příklad

V roce 1951 bylo rozsáhlým statistickým průzkumem zjištěno, že střední hodnota výšky desetiletých chlapců je $136,1$ cm se směrodatnou odchylkou $\sigma = 6,4$ cm.
V roce 1961 byla zjištěna výška pouze u 15 náhodně vybraných chlapců:

130	140	136	141	139	133	149	151
139	136	138	142	127	139	147	

Otázkou je, zda se v porovnání s rokem 1951 změnila střední výška chlapců, pokud předpokládáme, že variabilita výšek se v různých generacích příliš nemění.

Řešení

Vzhledem k tomu, že základní soubor všech desetiletých chlapců je rozsáhlý, lze zmíněná data považovat za náhodný výběr[a]. Zjistíme, že výběrový průměr $M = 139,133$, $n = 15$ a s využitím statistiky U dostáváme, že s 95% pravděpodobností leží hodnota μ v intervalu

$$(M - 1,96\sigma/\sqrt{n}; M + 1,96\sigma/\sqrt{n}) = (135,9; 142,4).$$

Protože i střední hodnota výšek z roku 1951 leží v tomto intervalu, nemáme vážný důvod tvrdit, že se střední výška změnila. Pokud bychom ovšem připustili vyšší možnost omylu a stanovili interval se spolehlivostí pouze 90%, pak bychom na této hladině hypotézu, že střední výška se změnila, přijali – interval je nyní (136,41;141,85). Podobně, pokud nás zajímá pouze **dolní odhad** střední hodnoty výšek chlapců (a vůbec tedy nepřipouštíme možnost, že by se střední výška snížila), pak s 95% pravděpodobností je střední výška větší než 136,41, a tedy nyní opět přijímáme hypotézu, že se střední výška zvýšila.

Dva nezávislé výběry z normálního rozdělení

Věta

Nechť je $X_{11}, \ldots, X_{m1}$ *náhodný výběr rozsahu* m *z rozdělení* $N(\mu, \sigma_1^2)$ *a* $X_{12}, \ldots, X_{n2}$ *je na něm nezávislý náhodný výběr rozsahu* n *z rozdělení* $N(\mu, \sigma_2^2)$*, přičemž* $m, n \geq 2$*. Označme* M_1, M_2 *jejich výběrové průměry a* S_1^2, S_2^2 *výběrové rozptyly. Dále nechť je*

$$S_*^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

vážený průměr výběrových rozptylů. *Pak platí:*

- $M_1 - M_2$ *a* S_*^2 *jsou stochasticky nezávislé,*
- $M_1 - M_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$,
- *je-li* $\sigma_1^2 = \sigma_2^2 = \sigma^2$*, pak*
  $K = (m+n-2)S_*^2/\sigma^2 \sim \chi^2(m+n-2)$,
- $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(m-1, n-1)$.

Užití statistik dvou nezávislých výběrů

- Statistika U, vzniklá normováním $M_1 - M_2$, se používá pro odhad rozdílu $\mu_1 - \mu_2$, známe-li rozptyly σ_1^2, σ_2^2.
- Je-li $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak statistika T (vzniklá z U nahrazením teoretického společného rozptylu σ^2 váženým průměrem výběrových rozptylů S_*^2) slouží pro odhad rozdílu $\mu_1 - \mu_2$, neznáme-li rozptyl σ^2.
- Statistika $K = (m + n - 2)S_*^2/\sigma^2$ slouží k odhadu společného rozptylu σ^2.
- Statistika $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$ slouží k odhadu podílu rozptylů σ_1^2/σ_2^2.

Příklad

Mějme dva nezávislé náhodné výběry; první rozsahu 10 z rozdělení $N(2; 1,5)$ a druhý rozsahu 5 z rozdělení $N(3, 4)$. Určete pravděpodobnost, že výběrový průměr prvního výběru bude menší než výběrový průměr druhého výběru.

Řešení

$$P(M_1 < M_2) = P(M_1 - M_2 < 0) =$$

$$= P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} < \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}\right) =$$

$$P\left(U < \frac{-2+3}{\sqrt{\frac{1,5}{10} + \frac{4}{5}}}\right) = P(U < 1,05) =$$

$$= \Phi(1,05) = 0,853.$$

Náhodný výběr je n-tice nezávislých náhodných veličin se stejným rozdělením, které záleží na jednom nebo více parametrech. Obvykle přitom hodnotu těchto parametrů neznáme, ale můžeme tuto hodnotu nebo hodnotu nějaké jeho funkce (tzv. parametrické funkce) z náhodného výběru odhadnout.

Definice

Mějme náhodný výběr $X_1, \ldots, X_n$, které závisí na (obecně vektorovém) parametru θ. **Bodovým odhadem parametru** θ rozumíme statistiku $T((X_1, \ldots, X_n)$, která je v nějakém smyslu blízko parametru θ. Rozdíl (příp. vektorový) $E(T) - \theta$ nazveme **vychýlení**, je-li $E(T) = \theta$, pak odhad T nazveme **nestranným**. **Intervalovým odhadem parametru** θ rozumíme (obecně vícerozměrný) 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 interval spolehlivosti $1 - \alpha$ pro parametr θ .

Definice

Jsou-li T_1, T_2 nestranné odhady parametru θ, říkáme, že odhad T_1 je **lepší** než odhad T_2, pokud $D(T_1) < D(T_2)$, příp.
var $T_1 <$ var T_2 (tj. matice var $T_2 -$ var T_1 je pozitivně definitivní).
O posloupnosti T_n odhadů θ říkáme, že je **asymptoticky nestranná** , pokud $\lim_{n\to\infty} E(T_n) = \theta$.
O posloupnosti T_n odhadů θ říkáme, že je **konzistentní**, pokud $\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1$.

Věta

Je-li posloupnost T_n odhadů parametru θ asymptoticky nestranná a platí-li $\lim_{n\to\infty} D(T_n) = 0$, *pak T_n je konzistentním odhadem θ.*

Důkaz.

Buď $\epsilon > 0$ libovolné. Z Čebyševovy nerovnosti máme $P(|T_n - E(T_n)| < \epsilon/2) \geq 1 - D(T_n)/(\epsilon/2)^2$. Zároveň pro dostatečně velké n máme $|E(T_n) - \theta| < \epsilon/2$. Proto

$$P(|T_n - \theta| < \epsilon) \geq P(|T_n - E(T_n)| < \epsilon/2, |E(T_n) - \theta| < \epsilon/2) = \\ = P(|T_n - E(T_n)| < \epsilon/2),$$

která konverguje k 1, což znamená, že T_n konverguje podle pravděpodobnosti k θ . □

Příklad

Uvažujme opakovaná nezávislá měření určité konstanty μ, popsaná náhodným výběrem $X_1, \ldots, X_n$ z rozdělení se střední hodnotou $E(X_i) = \mu$ a rozptylem $D(X_i) = \sigma^2$. Dokažte, že statistiky $M = \frac{1}{n}\sum X_i$ a $L = \frac{1}{2}(X_1 + X_n)$ jsou nestrannými odhady μ a rozhodněte, který z odhadů je lepší.

Řešení

$$E(M) = \frac{1}{n}\sum E(X_i) = \frac{1}{n} n\mu = \mu$$

$$E(L) = \frac{X_1 + X_n}{2} = \frac{1}{2}E(X_1 + X_n) = \frac{1}{2}(\mu + \mu) = \mu$$

$$D(M) = \frac{1}{n^2}\sum D(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$D(L) = D(1/2(X_1 + X_n)) = 1/4 D(X_1 + X_n) =$$

$$= \frac{1}{4}(D(X_1) + D(X_n)) = \frac{\sigma^2}{2}.$$

Poznámka

Odpověď na otázku z minulé přednášky je, že výběrová směrodatná odchylka S **není** nestranným odhadem směrodatné odchylky σ. Kdyby totiž $E(S) = \sigma$, pak by $D(S) = E(S^2) - E(S)^2 = \sigma^2 - \sigma^2 = 0$, což by znamenalo, že S je konstanta, a to je spor, protože rozptyl S je nenulový.

Poznámka

Jak jsme viděli dříve, není statistika $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$ náhodného výběru z normálního rozdělení nestranným odhadem rozptylu σ^2 – je totiž $E(s_n^2) = E(\frac{n-1}{n}S^2) = \frac{n-1}{n}\sigma^2$. Zřejmě je ale $\lim_{n\to\infty} E(s_n^2) = \sigma^2$ a protože

$$D(S^2) = \frac{2\sigma^2}{n-1},$$

je i $\lim_{n\to\infty} D(s_n^2) = \lim_{n\to\infty} D((n-1)S^2/n) = 0$, a je tedy posloupnost s_n^2 konzistentním odhadem rozptylu σ^2.

Intervaly spolehlivosti (*confidence intervals*)

Připomeňme, že pro náhodný výběr $X_1, \ldots, X_n$ závislý na parametru θ jsme definovali intervalový odhad parametru θ pomocí statistik T_L, T_U výběru tak, že $P(T_L \leq \theta \leq T_U) = 1 - \alpha$. Jde o tzv. **oboustranný interval spolehlivosti pro** θ . Podobně definujeme **levostranný interval spolehlivosti** (T_L, ∞) pomocí $P(T_L < \theta) = 1 - \alpha$, analogicky **pravostranný interval spolehlivosti**. Číslo α se nazývá riziko (obvykle se používá $\alpha = 0,05$), číslo $1 - \alpha$ spolehlivost.

Algoritmus konstrukce intervalu spolehlivosti

1. Zvolíme statistiku V, která je nestranným bodovým odhadem parametru θ.
2. Najdeme tzv. *pivotovou statistiku*, která je transformací V se známým rozdělením, nezávisící na neznámé hodnotě θ (např. M, K, T, F).
3. Najdeme příslušné kvantily rozdělení statistiky W tak, že

$$P(w_{\alpha/2} \leq W \leq w_{1-\alpha/2}) = 1 - \alpha.$$

4. Nerovnost $w_{\alpha/2} \leq W \leq w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost $T_L \leq \theta \leq T_U$.
5. Z daného výběru zjistíme konkrétní číselné realizace statistik T_L, T_U a dostaneme tak intervalový odhad požadované spolehlivosti $1 - \alpha$.

Intervaly spolehlivosti pro parametry normálního rozdělení

μ (známe σ^2)	$(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2})$
μ (neznáme σ^2)	$(M - \frac{S}{\sqrt{n}} t_{1-\alpha/2}, M + \frac{S}{\sqrt{n}} t_{1-\alpha/2})$
σ^2 (neznáme μ)	$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \right)$
σ^2 (známe μ)	$\left(\frac{\sum (X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum (X_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \right)$

Příklad

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení $N(\mu; 0, 1)$. Jaký musí být minimální rozsah výběru, aby velikost 95% intervalu spolehlivosti pro μ nepřesáhla číslo 0,03?

Řešení

Podle předchozí tabulky dostáváme (pro $\alpha = 0,05$)

$$0,03 \geq M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} - (M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}) =$$
$$= 2\frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}.$$

Proto

$$n \geq \frac{4\sigma^2 u_{1-\alpha/2}^2}{0,03^2} = 170,7$$

a rozsah výběru tedy musí splňovat $n \geq 171$.

Intervaly spolehlivosti pro parametry 2 normálních rozdělení

$\mu_1 - \mu_2$ (známe σ_1^2, σ_2^2)	$M_1 - M_2 \pm \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\, u_{1-\alpha/2}$
$\mu_1 - \mu_2$ (neznámé $\sigma_1^2 = \sigma_2^2$)	$M_1 - M_2 \pm S_* \sqrt{\frac{1}{m} + \frac{1}{n}}\, t_{1-\alpha/2}$
společný rozptyl σ^2	$\left(\frac{(m+n-2)S_*^2}{\chi^2_{1-\alpha/2}(m+n-2)}, \frac{(m+n-2)S_*^2}{\chi^2_{\alpha/2}(m+n-2)} \right)$
podíl rozptylů σ_1^2/σ_2^2	$\left(\frac{S_1^2/S_2^2}{F_{1-\alpha/2}(m-1,n-1)}, \frac{S_1^2/S_2^2}{F_{\alpha/2}(m-1,n-1)} \right)$

Poznámka

Pokud a priori nevíme, jestli jsou rozptyly shodné, můžeme to ověřit tak, že nejprve sestrojíme interval spolehlivosti pro σ_1^2/σ_2^2. Obsahuje-li 1, lze (s pravděpodobností $1 - \alpha$) považovat rozptyly za shodné a tento rozptyl odhadovat pomocí statistiky K jak je uvedeno v tabulce.

Inteval spolehlivosti pro výběr z dvourozměrného rozdělení

Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je výběr z rozdělení

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right).$$

Označíme $\mu = \mu_1 - \mu_2$ a zavedeme *rozdílový výběr* $Z_i = X_i - Y_i$. Pak statistika $T = \frac{M-\mu}{S/\sqrt{n}}$ výběru Z má t-rozdělení s $n-1$ stupni volnosti, proto jsou hranice intervalu spolehlivosti $1-\alpha$ pro μ rovny

$$M \pm \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1).$$

Příklad

U šesti nových automobilů bylo testováno, nakolik se sjíždějí pneumatiky na předních kolech. Byly naměřeny tyto hodnoty (v mm):

číslo auta	1	2	3	4	5	6
sjetí pravé pneu	1,8	1,0	2,2	0,9	1,5	1,6
sjetí levé pneu	1,5	1,1	2,0	1,1	1,4	1,4

Předpokládejte, že jde o realizaci náhodného výběru z dvourozměrného normálního rozdělení a rozhodněte, jestli nedochází k výraznějšímu nesymetrickému sjíždění pneumatik (tj. sestrojte 95% interval spolehlivosti pro $\mu = \mu_1 - \mu_2$).

Řešení

Postupně vypočteme: $Z = (0,3; -0,1; 0,2; -0,2; 0,1; 0,2)$, $M = 0,0833$, $S = 0,1941$. Pak jsou krajními body hledaného 95% intervalu spolehlivosti
$M \pm \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0,0833 \pm 0,1941 \cdot 2,5706/\sqrt{6}$, tj. $(-0,12; 0,29)$.

Poznamenejme, že snadno odvodíme i míru rizika, se kterou bychom mohli tvrdit, že je $\mu_1 > \mu_2$, tj. že pravé pneumatiky se sjíždějí více než levé. Je to takové číslo α, aby příslušný interval spolehlivosti neobsahoval číslo 0 – v našem případě je $\alpha = 0,34$, což je riziko příliš vysoké.