

Matematika III – 12. týden

Intervalové odhady, testování hypotéz, lineární modely

Jan Slovák

Masarykova univerzita
Fakulta informatiky

12.12. – 16.12. 2016

Obsah přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití

Plán přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice www.math.muni.cz/Matematika_drsne_svizne
- 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.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

Plán přednášky

- 1 Literatura
- 2 **Odhady**
- 3 Testování hypotéz
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití

Odhadování parametrů může být bodové nebo intervalové. V předchozím příkladu takovými byly výběrový průměr $\bar{x} = 139,133$ a interval spolehlivosti $(135,9,142,4)$.

Obecně postupujeme takto: Pro náhodný výběr rozsahu n X_1, \dots, X_n z rozdělení, které závisí na (vektorovém) parametru θ hledáme funkci náhodných veličin (říkáme též statistiku nebo výběrovou statistiku) $T(X_1, \dots, X_n)$, která bude mít v „rozumném smyslu“ blízko ke skutečné hodnotě θ .

Odhadování parametrů může být bodové nebo intervalové. V předchozím příkladu takovými byly výběrový průměr $\bar{x} = 139,133$ a interval spolehlivosti $(135,9,142,4)$.

Obecně postupujeme takto: Pro náhodný výběr rozsahu n X_1, \dots, X_n z rozdělení, které závisí na (vektorovém) parametru θ hledáme funkci náhodných veličin (říkáme též statistiku nebo výběrovou statistiku) $T(X_1, \dots, X_n)$, která bude mít v „rozumném smyslu“ blízko ke skutečné hodnotě θ .

Jakožto funkce náhodných veličin je T opět náhodnou veličinou (resp. náhodným vektorem). Konstanta (resp. konstantní vektor)

$$b = E T - \theta$$

se nazývá **vychýlení** odhadu T . **Nestranný** (nevychýlený) je takový odhad, kdy $b = 0$.

Nejlepší odhad

Máme-li k dispozici jistou třídu odhadů \mathcal{T} , říkáme že T je **nejlepším odhadem**, má-li mezi všemi nejmenší rozptyl.

Nejlepší odhad

Máme-li k dispozici jistou třídu odhadů \mathcal{T} , říkáme že T je **nejlepším odhadem**, má-li mezi všemi nejmenší rozptyl.
 $T = T_n$ je **konzistentním odhadem**, je-li pro každé $\epsilon > 0$

$$\lim_{n \rightarrow \infty} P(|T_n - \theta| < \epsilon) = 1.$$

Theorem

Je-li $\lim_{n \rightarrow \infty} E T_n = \theta$, $\lim_{n \rightarrow \infty} \text{var } T_n = 0$, pak je T_n konzistentním odhadem θ .

Plán přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz**
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití

Definition

Hypotézou rozumíme nějaké tvrzení o rozdělení určeném sdruženou distribuční funkcí $F_X(x)$ náhodného vektoru $X = (X_1, \dots, X_n)$. Rozhodujeme mezi tzv. **nulovou hypotézou** H_0 a **alternativní hypotézou** H_A , která bývá negací nulové hypotézy. Možnými rozhodnutími jsou **zamítnutí** nebo **nezamítnutí** nulové hypotézy.

Definition

Hypotézou rozumíme nějaké tvrzení o rozdělení určeném sdruženou distribuční funkcí $F_X(x)$ náhodného vektoru $X = (X_1, \dots, X_n)$. Rozhodujeme mezi tzv. **nulovou hypotézou** H_0 a **alternativní hypotézou** H_A , která bývá negací nulové hypotézy. Možnými rozhodnutími jsou **zamítnutí** nebo **nezamítnutí** nulové hypotézy.

Když nulovou hypotézu zamítneme, přestože ve skutečnosti platí, nastává **chyba prvního druhu**, když ji nezamítneme v situaci, kdy neplatí, hovoříme o **chybě druhého druhu**.

Statistické rozhodování se opírá o předem určený **kritický obor** W , tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítnat.

Statistické rozhodování se opírá o předem určený **kritický obor** W , tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítat.

Tvar kritického oboru oboru volíme tak, abychom platnou hypotézu zamítli s pravděpodobností nejvýše α . Tj. zadáváme předem ohraničení velikosti chyby prvního druhu tzv. hladinou testu α .

Zpravidla volíme $\alpha = 0,05$ nebo $\alpha = 0,01$.

Výpočetní síla dnes umožňuje úkol obrátit a pro daná data se ptát, na jaké **nejmenší** hladině bychom ještě hypotézu zamítli. Hovoříme o **dosažené hladině testu** nebo také **p -hodnotě** (v angličtině P -value nebo Sig. level).

Statistické rozhodování se opírá o předem určený **kritický obor** W , tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítat.

Tvar kritického oboru oboru volíme tak, abychom platnou hypotézu zamítli s pravděpodobností nejvýše α . Tj. zadáváme předem ohraničení velikosti chyby prvního druhu tzv. hladinou testu α .

Zpravidla volíme $\alpha = 0,05$ nebo $\alpha = 0,01$.

Výpočetní síla dnes umožňuje úkol obrátit a pro daná data se ptát, na jaké **nejmenší** hladině bychom ještě hypotézu zamítli. Hovoříme o **dosažené hladině testu** nebo také **p -hodnotě** (v angličtině P -value nebo Sig. level).

Mezi všemi kritickými obory na dané hladině testu ale pochopitelně přitom chceme vybrat ten, který bude minimalizovat chybu druhého druhu.

Předpokládejme, že náhodný vektor X má hustotu rozdělení $f(x, \theta)$ závislou na (vektorovém) parametru. Za nulové hypotézy je to rozdělení s hustotou $f(x, \theta_0)$, za alternativní s hustotou $f(x, \theta_1)$.

Theorem (Neymanovo-Pearsonovo lemma)

Nechť k danému $\alpha \in (0, 1)$ existuje $c > 0$ takové, že pro množinu $W_c = \{x : f(x, \theta_1) \geq cf(x, \theta_0)\}$ platí $\int_{W_c} f(x, \theta_0) dx = \alpha$. Pak pro každou měřitelnou množinu W takovou, že je $\int_W f(x, \theta_0) dx = \alpha$, platí

$$\int_{W_c} f(x, \theta_1) dx \geq \int_W f(x, \theta_1) dx$$

V případě intervalových odhadů můžeme problém přeformulovat jako hypotézy H_0 – „střední hodnota je μ_0 “ a H_A – „střední hodnota je μ_1 “. Kritický obor je pak dán požadavkem

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right| \geq z(\alpha/2)$$

a nezávisí na konkrétní hodnotě μ_1 .

Example

Úkol v našem předchozím příkladu o výšce desetiletých chlapců lze formulovat tak, že nulovou hypotézou je nezměněná výška populace, zatímco alternativní je, že se výška změnila (tj. náš kritický obor je symetrický). Hladinu testu pak spočteme na 6,66%, takže je přirozené, že jsme nulovou hypotézu na úrovni 5% nezamítli.

V případě intervalových odhadů můžeme problém přeformulovat jako hypotézy H_0 – „střední hodnota je μ_0 “ a H_A – „střední hodnota je μ_1 “. Kritický obor je pak dán požadavkem

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right| \geq z(\alpha/2)$$

a nezávisí na konkrétní hodnotě μ_1 .

Example

Úkol v našem předchozím příkladu o výšce desetiletých chlapců lze formulovat tak, že nulovou hypotézou je nezměněná výška populace, zatímco alternativní je, že se výška změnila (tj. náš kritický obor je symetrický). Hladinu testu pak spočteme na 6,66%, takže je přirozené, že jsme nulovou hypotézu na úrovni 5% nezamítli.

Když interpretujeme zadání tak, že buď se výška nezměnila, nebo vzrostla, bude náš kritický obor nesymetrický a dojdeme k hladině testu 3,33%. Nulovou hypotézu proto na hladině 5% zamítneme.

Plán přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz
- 4 Lineární modely**
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití

Uvažujme náhodný vektor $Y = (Y_1, \dots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodnotí k , β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $N(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.

Uvažujme náhodný vektor $Y = (Y_1, \dots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodnotí k , β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $N(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.

V praktických problémech zpravidla známe veličiny x_{ij} a snažíme se odhadnout nebo predikovat hodnotu Y .

Chceme přitom mít jasno o pravděpodobnostních charakteristikách těchto odhadů.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i -tého studenta v j -tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i -tého studenta v j -tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).

K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ pro vektor β tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i -tého studenta v j -tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).

K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ pro vektor β tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.

To je ale jednoduchá úloha lineární algebry a víme, že jde o nalezení kolmého průmětu vektoru Y do podprostoru $\langle X \rangle \subset \mathbb{R}^n$ generovaném sloupci matice X .

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^k x_{ij}\beta_j \right)^2.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^k x_{ij}\beta_j \right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^k x_{ij}\beta_j \right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Víme, že $\hat{Y} = Xb$ a že, díky našemu předpokladu o maximální hodnotě X , je matice $X^T X$ invertibilní. Můžeme proto rovnou spočítat $b = (X^T X)^{-1} X^T \hat{Y}$.

Theorem

Nechť A je libovolná matice typu m/n nad reálnými nebo komplexními skaláry. Pak existují čtvercové unitární matice U a V dimenzí m a n , a reálná diagonální matice s nezápornými prvky D dimenze r , $r \leq \min\{m, n\}$, takové, že

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

a r je hodnost matice AA^ . Přitom je S určena jednoznačně až na pořadí prvků a prvky diagonální matice D jsou druhé odmocniny vlastních čísel d_i matice AA^* . Pokud je A reálná matice, pak i matice U a V jsou ortogonální.*

Definition

Nechť A je reálná matice typu m/n a necht'

$$A = USV^*, \quad S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

je její singulární rozklad (zejména D je invertibilní). Matici

$$A^\dagger := VS'U^*, \quad S' = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

nazýváme **pseudoinverzní matice** k matici A .

Jak ukazuje následující věta, je pseudoinverze důležité zobecnění pojmu inverzní matice, včetně přímočarých aplikací.

Theorem

Nechť A je reálná nebo komplexní matice typu m/n . Pak pro její pseudoinverzní matici platí:

- 1 *Je-li A invertibilní (zejména tedy čtvercová), pak*

$$A^\dagger = A^{-1}.$$

- 2 *Pro pseudoinverzi A^\dagger platí, že $A^\dagger A$ i AA^\dagger jsou hermiteovské (v reálném případě symetrické) a*

$$AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger.$$

- 3 *Pseudoinverzní matice A^\dagger je čtyřmi vlastnosti z předchozího bodu určena jednoznačně. Pokud tedy nějaká matice B typu $n \times m$ splňuje, že BA i AB jsou hermiteovské, $ABA = A$ a $BAB = B$, pak $B = A^\dagger$.*

Theorem (Pokračování)

- 1 *Je-li A matice systému lineárních rovnic $Ax = b$, s pravou stranou $b \in \mathbb{K}^m$, pak vektor $y = A^\dagger b \in \mathbb{K}^n$ minimalizuje velikost $\|Ax - b\|$ pro všechny vektory $x \in \mathbb{K}^n$.*
- 2 *System lineárních rovnic $Ax = b$ s $b \in \mathbb{K}^m$ je řešitelný, právě když platí $AA^\dagger b = b$. V tomto případě jsou všechna řešení dána výrazem*

$$x = A^\dagger b + (E - A^\dagger A)u,$$

kde $u \in \mathbb{K}^n$ je libovolné.

Z bodu (4) předchozí věty plyne, že matice AA^\dagger je maticí kolmé projekce z vektorového prostoru \mathbb{R}^n , kde n je počet řádků matice A na podprostor generovaný sloupci matice A (tato interpretace má samozřejmě smysl pouze pro matice mající více řádků než sloupců). Dále pro matice A , jejichž sloupce tvoří nezávislé vektory, má smysl výraz $(A^T A)^{-1} A^T$ a není těžké ověřit, že tato matice splňuje všechny vlastnosti z (1) a (2) z předchozí věty, jedná se tedy o pseudoinverzi k matici A .

Z předchozí věty vyplývají také následující vlastnosti pseudoinverze:

- Pro všechny matice A platí $(A^\dagger)^\dagger = A$,
- pokud má matice A , typu $m \times n$, plnou řádkovou hodnost m , pak $A^\dagger = A^*(AA^T)^{-1}$,
- pokud má matice A , typu $m \times n$, plnou sloupcovou hodnost n , pak $A^\dagger = (A^T A)^{-1}A^*$.

Plán přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$**
- 6 Hlavní věta a použití

Jestliže má náhodný vektor $Z = (Z_1, \dots, Z_n)$ nezávislé komponenty $Z_i \sim N(0, 1)$, je jeho varianční matice jednotkovou maticí, tj. $\text{var } Z = \mathbb{I}_n$.

Uvažme vektor $U = a + BZ$, kde a je libovolný konstantní vektor v \mathbb{R}^m a B je konstantní matice typu (m, n) .

Víme $E U = a$ a $\text{var } U = V = BB^T$ (protože varianční matice Z je identická). Je tedy tato varianční matice vždy pozitivně semidefinitní.

Říkáme, že náhodný vektor U má **mnohoměrné normální rozdělení** $N_m(a, V)$.

Pro libovolné mnohoměrné normální rozdělení $N_m(a, V)$ znovu vezmeme afinní transformaci

$$W = c + DU$$

s vektorem konstant $c \in \mathbb{R}^k$ a libovolnou konstantní maticí typu (k, m) . Přímým výpočtem

$$W = c + D(a + BZ) = (c + Da) + (DB)Z,$$

což je samořejmě náhodný vektor $W \sim N_k(c + Da, DB^T B D^T)$. Chová se tedy kovarianční matice mnohoměrného normálního rozdělení při afinních transformacích jako kvadratická forma.

Dokázali jsme, že jakákoliv lineární kombinace složek náhodného vektoru s mnohoměrným normálním rozdělením je náhodná veličina s normálním rozdělením. Stejně je každý vektor vzniklý výběrem jen některých komponent vektoru U opět náhodným vektorem s mnohoměrným normálním rozdělením.

Plán přednášky

- 1 Literatura
- 2 Odhady
- 3 Testování hypotéz
- 4 Lineární modely
- 5 Vícerozměrné $N_m(\mu, V)$
- 6 Hlavní věta a použití**

Theorem

V lineárním modelu $Y = X\beta + \sigma Z$ platí pro vhodné matice P a R :

(1) Pro odhad \hat{Y} platí

$$\hat{Y} = X\beta + \sigma PP^T Z, \quad \hat{Y} \sim N(X\beta, \sigma^2 PP^T).$$

(2) Reziduální součet čtverců RSS a normovaný čtverec velikosti rezidua mají rozdělení:

$$Y - \hat{Y} \sim N(0, \sigma^2 RR^T), \quad \|Y - \hat{Y}\|^2 / \sigma^2 \sim \chi_{n-k}^2.$$

(3) Náhodná veličina $b = \beta + \sigma(P^T X)^{-1} P^T Z$ má rozdělení

$$b \sim N(\beta, \sigma^2(X^T X)^{-1}).$$

(4) Pro reziduální rozptyl platí $(n - k)S^2 / \sigma^2 \sim \chi_{n-k}^2$.

(5) Střední hodnota reziduálního rozptylu je $E S^2 = \sigma^2$.

(6) Veličiny b a S^2 jsou nezávislé.

Úplně nejjednodušší je to v případě jediného výběru, kdy testujeme, zda jediný parametr β je roven dané hodnotě β_0 .

Volíme matici X s jediným sloupcem plným jedniček. Výraz

$$Y = X\beta + \sigma Z$$

komponenty v Y jsou nezávislé veličiny $Y_i \sim N(\beta, \sigma^2)$, jde o náhodný výběr rozsahu n z normálního rozdělení.

Obecná věta dává odhad

$$b = (X^T X)^{-1} X^T Y = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}$$

$$S^2 = \frac{1}{n-1} \|Y - X\bar{Y}\|^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

což jsou právě výběrový průměr a rozptyl, se kterými jsme již počítali.

Zajímá nás přitom statistika

$$T = \frac{\bar{Y} - \beta_0}{S} \sqrt{n}$$

Testování hypotézy $\beta = \beta_0$ se nazývá **jednovýběrový t-test**. Na hladině α hypotézu zamítáme, když je $|T| \geq t_{n-1}(\alpha)$.

párový t-test

Je vhodný na případy, kdy testujeme dvojice náhodných vektorů $W_1 = (W_{i1})$ a $W_2 = (W_{i2})$, o rozdílech jejichž komponent $Y_i = W_{i1} - W_{i2}$ víme, že mají rozdělení $N(\beta, \sigma^2)$. Potřebujeme navíc, aby byly veličiny Y_i nezávislé (což neříká, že musí být nezávislé jednotlivé dvojice W_{i1} a W_{i2} !). Můžeme si představit třeba hodnocení dvou různých vyučujících týmž studentem. Testujeme hypotézu, že pro všechna i je $E W_{i1} = E W_{i2}$. Používáme statistiku

$$T = \frac{\bar{W}_1 - \bar{W}_2}{S} \sqrt{n}.$$