Drsná matematika IV – 10. přednáška
Matematická statistika

Jan Slovák

Masarykova univerzita
Fakulta informatiky

30. 4. 2012

Obsah přednášky

Kde je dobré číst?

- vlastní poznámky, texty současného přednášejícího, GOOGLE, atd.
- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- 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.

Zpravidla statistici nemají k dispozici znaky všech statistických jednotek v souboru a místo toho se musí spokojit je s daleko menším výběrem. Hovoříme o **populaci** místo základního velkého souboru a **výběru** (nebo výběrovém souboru). Všude dále předpokládáme že rozsah populace N je mnohem větší než rozsah n výběru z ní.

Naším úkolem je odhadovat charakteristiky jako jsou průměr μ hodnot znaku $\bar{x}$ nebo jejich rozptyl σ^2 pro celou populaci pomocí obdobných charakteristik pro náš daleko menší výběr, které budeme značit pomocí velkých písmen, např. $\bar{X}$, S^2.

Zde vstupuje do hry pravděpodobnost – budeme chtít znát pravděpodobnost přiblížení hodnot pro náš výběr těm pro celou populaci.

Zvolení výběru o rozsahu n považujeme za elemementární jev ω na vhodném pravděpodobnostním prostoru a hodnoty znaků považujeme za náhodný vektor

$$X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}),$$

kde indexy i_j odkazují na statistické jednotky v populaci, které byly do výběru zařazeny.

Výběry s vracením – bez vracení

- Výběr můžeme realizovat buď tak, že vybereme n–tici z N prvků, každý z elementárních jevů má tedy pravděpodobnost $1/\binom{N}{n}$, hovoříme o **výběru z konečné populace bez vracení**.
- Výběr také můžeme realizovat pomocí výběru s vracením (vybíráme měřené jednotky jednu po druhé a po změření znaku je opět vracíme do populace), hovoříme o **náhodném výběru** o rozsahu n.

Obě metodiky výběru se k sobě velice blíží, pokud je populace tak veliká, že ji můžeme vzhledem k rozsahu našeho výběru považovat za nekonečnou.
Říkáme, že číselná charakteristika $\bar{Y}$ výběru je nestranným odhadem náhodné veličiny y na celé populaci, jestliže je $E\bar{Y} = y$.

Theorem

V případě náhodného výběru jsou střední hodnota i rozptyl nestrannými odhady.

Pokud vybíráme z populace jejíž zkoumaný znak patří do předem známého rozdělení, máme dobrou znalost rozdělení náhodných veličin $\bar{X}$ apod.
Např. pro výběr z normálního rozdělení $N(\mu, \sigma^2)$ bude veličina $\bar{X}$ mít pro náhodný výběr rozsahu n rozdělení $N(\mu, \frac{\sigma^2}{n})$.

Výběr bez vracení z konečné populace

Výběr s popíšeme pomocí náhodných veličin W_i s hodnotou 1 je-li $i \in s$ a nula jinak.

Theorem

$$P(W_i = 1) = \frac{n}{N}, \quad \mathrm{E}\, W_i = \frac{n}{N},$$

$$\operatorname{var} W_i = \frac{n}{N}(1 - \frac{n}{N}), \quad \operatorname{cov}(W_i, W_j) = -\frac{n(N-n)}{N^2(N-1)}.$$

Všimněme si, že pochopitelně nejsou veličiny W_i po dvou nezávislé, když jejich součet je vždy právě n.

Výběrový průměr je $\bar{X} = \frac{1}{n}\sum_{i\in s} x_i = \frac{1}{n}\sum_{i=1}^{N} x_i W_i$.

Theorem

$$\mathrm{E}\,\bar{X} = \mu, \quad \mathrm{var}\,\bar{X} = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

Říkáme, že $\bar{X}$ je nestranným odhadem střední hodnoty znaku pro populaci.
U rozptylu o nestranný odhad nejde, koeficient je však dle našich předpokladů velice blízký jedné! Definujeme **výběrový rozptyl** jako $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$.

Theorem

$$\mathrm{E}\,S^2 = \frac{N}{N-1}\sigma^2.$$

Jako nestranný odhad σ^2 bychom tedy mohli vzít $\frac{N-1}{N}S^2$.

Velmi častou úlohou je pro spočtenou hodnotu $\bar{X}$ výběrového průměru určit interval, ve kterém se skutečná hodnota průměru veličiny pro celou populaci nachází s předem danou (vysokou) pravděpodobností.
Např. pro náhodnou veličinu X s normálním rozdělením máme její normovanou veličinu $Z = \frac{X - \mathrm{E}\,X}{\sqrt{\mathrm{var}\,X}}$ s výběrovým průměrem $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ a chceme najít takovýto interval pro pravděpodobnost $1 - \alpha$, $\alpha \in (0, 1)$.
Potřebujeme tedy znát hodnotu $z(\alpha)$ takovou, že $P(Z > z(\alpha)) = \alpha$.
Je-li $F(x)$ spojitá rostoucí distribuční funkce naší veličiny, pak zjevně $z(\alpha) = F^{-1}(1 - \alpha)$. Pro normální rozdělení splňuje distribuční funkce Φ tento požadavek. Takto definovaným hodnotám $z(\alpha)$ se říká **kritické hodnoty**.
Protože je hustota pro normální rozdělení symetrická kolem jeho střední hodnoty, dostáváme $1 - \alpha = P(|Z| < z(\alpha/2))$.

$$1-\alpha = P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right| < z(\alpha/2)\right)$$
$$= P\left(\bar{X} - \frac{\sigma}{\sqrt{n}}z(\alpha/2) < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}z(\alpha/2)\right)$$

což je interval s náhodnými konci, který s námi určenou pravděpodobností pokrývá neznámý parametr μ. V kontextu takových úloh hovoříme o **intervalu spolehlivosti** s **koeficientem spolehlivosti** $1-\alpha$.

Pro normální rozdělení je velice populární kritická hodnota $z(0,025) = 1,96$, která odpovídá naší úloze se zvolenou pravděpodobností 95%.

Kritické hodnoty jsou dány pomocí tzv. **kvantilové funkce**

$$F^{-1}(u) = \inf\{x \in \mathbb{R}; F(x) \geq u\}, \quad 0 < u < 1.$$

Kvantilová funkce skutečně dává přímo příslušné kvantily, např. $F^{-1}(0,5)$ je medián, atd.

Example

Před deseti lety byl uskutečněn rozsáhlý výzkum výšky desetiletých chlapců a zjistilo se, že střední výška byla $\mu_0 = 136,1$cm se směrodatnou odchylkou $\sigma = 6,4$cm. Nyní byly na náhodném výběru 15 desetiletých chlapců zjištěny následující výšky: 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147. Je známo, že variabilita výšek v populaci se mění velice pomalu, zatímco výšky se mohou měnit rychle. **Otázka: došlo ke změně střední výšky populace desetiletých chlapců?**

Ze zadání předpokládáme, že výběr 15 hodnot je z normálního rozdělení se známým rozptylem σ^2 a otázku si upřesníme tak, že hledáme v jakém intervalu je nyní střední hodnota výšky populace se spolehlivostí 95% : $\bar{x} = 139,133$ a tedy interval spolehlivosti je $(139,133 - (6,4/\sqrt{15})1,96, 139,133 + (6,4/\sqrt{15})1,96) = (135,9, 142,4)$.

Protože tento interval pokrývá i populační průměr před deseti lety, nemůžeme na této hladině spolehlivosti tvrdit, že se populační výška změnila.

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 $X_1, \ldots, X_n$ z rozdělení, které závisí na (vektorovém) parametru θ hledáme funkci náhodných veličin (říkáme též statistiku nebo výběrovou statistiku) $T(X_1, \ldots, X_n)$, 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)

$$b = \mathsf{E}\, T - \theta$$

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.
$T = T_n$ je **konzistentním odhadem**, je-li pro každé $\epsilon > 0$

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

Theorem

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

Obdobně pro intervalové odhady (Zvára-Štěpán, str. 164).

Definition

Hypotézou rozumíme nějaké tvrzení o rozdělení určeném sdruženou distribuční funkcí $F_X(x)$ náhodného vektoru $X = (X_1, \ldots, X_n)$. 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 zamítat.
Tvar kritického oboru oboru volíme tak, aby do něj náhodný vektor X padal za platnosti alternativní hypotézy co nejčastěji (tj. s co největší pravděpodobonností), kdežto při platnosti nulové hypotézy jen zřídka (tj. s malou pravděpodobností). Velikost W pak volíme tak, abychom platnou nulovou hypotézu zamítali s pravděpodobností nejvýše α. Této pravděpodobnosti α se říká **hladina testu**.
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 apod.).
Jinými slovy: dosažená hladina testu je právě pravděpodobnost, že za platnosti nulové hypotézy dostaneme právě náš vektor X nebo vektor ještě více odporující testované hypotéze.

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.

Naopak, pokud věrně interpretujeme zadání tak, že víme předem, 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.

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

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