Matematika 6F — Definice

1. Centrální limitní věta

1.1. Skripta - Rogalewicz

Nechť $\{X_i\}_{i=1}^n$ je posloupnost vzájemně nezávislých náhodných veličin, které mají totéž rozdělení se střední hustotou μ a s konečným rozptylem σ^2. Potom

$$\lim_{n\to\infty} P\Big[\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le t\Big] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-u^2/2}\mathrm{d}u$$

1.2. Přednášky - Navara

Nechť $X_j, j \in \mathbb{N}$, jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou μ_X a směrodatnou odchylkou $\sigma_X \neq 0$. Pak normované náhodné veličiny

$$Y_N = \frac{\sqrt{N}}{\sigma_X}(\overline{X}_N - \mu_X)$$

konvergují k normalizovanému normálnímu rozdělení v následujícím smyslu:

$$\forall t \in \mathbb{R} : \lim_{N\to\infty} F_{Y_N}(t) = F_{N(0,1)}(t)$$

2. χ^2 rozdělení

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1}e^{-\frac{x}{2}} & \text{pro } x > 0 \\ 0 & \text{jinde} \end{cases}$$

Parametr: n - přirozené číslo (stupeň volnosti)

$$EX = n$$

$$\mu = 2n$$

χ^2 rozdělení se využívá v testu dobré shody.

Def. Nechť $np_1, np_2, \ldots, np_k$ a $o_1, o_2, \ldots, o_k$ jsou teoretické a napozorované četnosti k možných, navzájem se vylučujících výsledků pokusu, potom statistika

$$\sum_{i=1}^{k} \frac{(o_i - np_i)^2}{np_i}$$

má při $n \to \infty$ asymptoticky rozdělení χ^2 o $k-1$ stupni volnosti.

Hypotézu zamítáme, pokud platí:

$$\sum_{i=1}^{k} \frac{(o_i - np_i)^2}{np_i} > \chi^2_{1-\alpha}(k-1)$$

$\chi^2_\beta(\nu)$ je β-kvantil rozdělení χ^2 o $(k-1)$ stupních volnosti

Mnemotechnický tvar:

$$\sum \frac{(\text{pozorováno} - \text{teoreticky})^2}{\text{teoreticky}}$$

3. Znaménkový test

Mějme náhodný výběr $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ ze spojitého rozdělení s distribuční funkcí $F(x)$. Pro medián $x_{0,5}$ tohoto rozdělení platí

$$P[X < x_{0,5}] = 0,5 = P[X > x_{0,5}]$$

(tedy rozdělení musí být symetrické) Testujeme hypotézu $H_0 : x_{0,5} = x_0$ (x_0 je daná konstanta). Označíme $Z_i = X_i - x_0$ pro $i = 1, 2, \ldots, n$. Nechť Z je náhodná veličina, jejíž hodnota je rovna počtu kladných hodnot mezi $Z_1, Z_2, \ldots, Z_n$. Statistika Z nabývá hodnot $0, 1, 2, \ldots, n$ a za platnosti hypotézy má binomické rozdělení s parametry n a $p = \frac{1}{2}$. Uvažujme alternativní hypotézu $H_1 : x_{0,5} < x_0$ a hladinu významnosti $\alpha = 0,05$. Hypotézu H_0 zamítáme, pokud $Z \le c$, kde c dostaneme z nerovnice

$$\left(\frac{1}{2}\right)^n \sum_{j=0}^{c}\binom{n}{j} \le \alpha < \left(\frac{1}{2}\right)^n \sum_{j=0}^{c+1}\binom{n}{j}$$

4. Wilcoxonův test

4.1. Skripta - Rogalewicz

Mějme náhodný výběr $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ ze spojitého rozdělení symetrického podle mediánu $x_0, 5$. Testujeme opět hypotézu $H_0 : x_{0,5} = x_0 = \text{konst}$. Uvažujme nyní absolutní hodnoty rozdílů $|Z_i| = |X_i - x_{0,5}|$. Označme R_i^* pořadí $|Z_i|$ pro $i = 1, 2, \ldots, n$ Položme

$$T = \sum_{i=1}^{n} a_i R_i^*,$$

kde $a_i = 1$, jestliže $Z_i = X_i - \alpha_{0,5} > 0$ a $a_i = 0$, jestliže $Z_i \le 0$. Tedy T je součet pořadí pro kladné rozdíly. Hodnoty T_P splňující za platnosti hypotézy H_0 vztahy

$$P[T \le T_P] \le P, \qquad P[T \le T_P + 1] > P$$

jsou tabelovány.

4.2. Skripta - Němičková

Uvažujme náhodný výběr $X_1, X_2, \ldots, X_n$ ze spojitého rozdělení symetrického kolem mediánu $x_{0,5}$. Testujeme hypotézu $H_0 : X_{0,5} = x_0$ (x_0 je dané číslo).
Vypočteme rozdíly $Y_i = X_i - x_0, i = 1, 2, \ldots, n$ a jejich absolutní hodnoty $|Y_i|$ seřadíme podle velikosti. Označíme R_i pořadí $|Y_i|$ a

$$T^+ = \sum_{Y_i \ge 0} R_i, \qquad T^- = \sum_{Y_i < 0} R_i$$

Přitom $T^+ + T^- = \frac{n(n+1)}{2}$, což lze využít pro kontrolu. Je-li číslo $T = min(T^+, T^-)$ menší nebo rovno tabelované kritické hodnotě, hypotéza se zamítá.

5. Metoda momentů

5.1. Skripta - Rogalewicz

Používá si při praktické konstrukci odhadů. Je výpočetně jednoduchá a rychlá, dává však jen velmi přibližný odhad, který se hodí například pro předběžné posouzení, vyslovení hypotéz nebo jako první přiblížení pro iterační metody.
Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z populace s rozdělením $F(x)$ s parametry $\theta_1, \theta_2, \ldots, \theta_k$. Nechť Y je (nějaká) náhodná veličina s rozdělením $F(x)$.

p-tý obecný moment: $m_p = EY^p$

$$\text{p-tý výběrový moment} \quad M_p = \hat{m}_p = \frac{1}{n}\sum_{i=1}^{n} X_i^p$$

Metoda momentů spočívá v tomt, že porovnáváme k prvnmích obecných momentů s hodnotami jejich výběrových protějšků. Tím dostaneme k rovnic v proměnných $\theta_1, \theta_2, \ldots, \theta_k$ a jejich řešení můžeme považovat za bodové odhady $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ parametrů $\theta_1, \theta_2, \ldots, \theta_k$.

5.2. Přednášky - Navara

Pro $k = 1, 2, \ldots$ je k-tý obecný moment funkcí Θ, $\mu_{k,X}(\Theta) = \mu_{k,X}(\theta_1, \ldots, \theta_i)$ (lze stanovit dle pravděpodobnostního modelu). Lze též odhadnout pomocí výběrového k-tého obecného momentu $m_{k,X}$.
Metoda momentů doporučuje odhad $(\hat{\theta}_1, \ldots, \hat{\theta}_i)$ takový, že

$$\mu_{k,X}(\hat{\theta}_1, \ldots, \hat{\theta}_i) = m_{k,X} = \frac{1}{N}\sum_{j=1}^{N} x_j^k, \quad k = 1, 2, \ldots$$

K jednoznačnému určení i proměnných obvykle potřebujeme (prvních) i rovnic pro $k = 1, 2, \ldots, i$.

Použitelnost metody momentů

- Řešení neexistuje ⇒ zkusme ubrat rovnice
- Je nekonečně mnoho řešení ⇒ zkusme přidat další rovnice
- Je více než jedno řešení (např. soustavy kvadratických rovnic).
- Je jediné řešení, ale je obtížné je nalézt.
- Soustava je špatně podmíněná (typicky pro velký počet parametrů)
- Našli jsme jediné řešení, které však *nesplňuje předpoklady* (např parametry nemohou být libovolná čísla) ⇒ NELZE! Vždy zkontrolovat řešení
- všem rovnicím je přikládána stejná důležitost, což bývá nežádoucí (typicky pro velký počet parametrů)
- Nelze použít pro nenumerická data (pokud je nelze smysluplně očíslovat)

Výhoda: Lze použít pro diskrétní, spojité i smíšené rozdělení beze změn

6. Metoda maximální věrohodnosti

6.1. Skripta - Rogalewicz

Nechť $f(\mathbf{t}, \Theta)$ je sdružená hustota (resp $P(\mathbf{t}, \Theta)$ sdružená pravděpodobnostní funkce) náhodného výběru $\mathbb{X}$. Nechť $\mathbf{x}$ jsou pevné experimentální hodnoty náhodného výběru $\mathbb{X}$. Funkci $L(\Theta) = f(\mathbf{t}, \Theta)$ (*resp.* $L(\Theta) = P(\mathbf{t}, \Theta)$), tj. tuto sdruženou hustotu (resp. pravděpodobnostní funkci) uvažovanou jako funkci proměnných Θ při pevném $\mathbf{x}$, nazýváme **funkcí věrohodnosti**.

$$f(\mathbf{t}, \Theta) = \prod_{i=1}^{n} f(t_i, \Theta)$$

Vektor statistik $\hat{\Theta} = (\hat{\theta}_1(\mathbb{X}), \hat{\theta}_2(\mathbb{X}), \ldots, \hat{\theta}_k(\mathbb{X}))$ nazýváme **maximálně věrohodným odhadem** vektoru parametrů Θ, jestliže $L(\hat{\Theta}) \geq L(\Theta)$ pro každý možný vektor parametrů Θ.

6.2. Přednášky - Navara

Pro diskrétní rozdělení Pravděpodobnost realizace,

$$p_{\vec{X}}(\vec{x}|\Theta) = P[X_1 = x_1 \wedge \cdots \wedge X_N = x_N|\Theta] = \prod_{j=1}^{N} P[X_j = x_j|\Theta] = \prod_{j=1}^{N}$$

je funkce $l : \mathbb{R}^i \to \langle 0, 1\rangle$ parametrů $\Theta = (\theta_1, \ldots, \theta_i)$ zvaná věrohodnost. Maximalizujeme buď ji nebo její logaritmus.

$$L(\Theta) = \ln l(\Theta) = \sum_{j=1}^{N} \ln p_X(x_j|\Theta)$$

(Nutno vyloučit případ $p_X(x_j|\Theta) = 0$, který však nevede na maximum)

Pro spojité rozdělení Každá realizace má nulovou púravděpodobnost, proto místo ní použijeme hustotu pravděpodobnosti, což ale vede na zcela *jiný pojem*

$$f_{\vec{X}}(\vec{x}|\Theta) = \prod_{j=1}^{N} f_X(x_j|\Theta) = l(\Theta)$$

Nicméně i tato funkce $l : \mathbb{R}^i \to \langle 0, \infty)$ se nazývá *věrohodnost*

$$L(\Theta) = \ln l(\Theta) \sum_{j=1}^{N} \ln f_X(x_j|\Theta)$$

(Nutno vyloučit případ, kdy $f_X(x_j|\Theta) = 0$, který však nevede na maximum)

Pro smíšené rozdělení není věrohodnost definována

Použitelnost metody maximální věrohodnosti Možné problémy

- Je více než jedno řešení
- Je jedíné řešení, ale je obtížné je nalézt
- Soustava je špatně podmíněná (typické pro velký počet parametrů)
- Nesmí se použít pro smíšené rozdělení

Výhody

- Hledání optima je o něco snažší než řešení soustavy rovnic
- Různým datům je dán společný (srovnatelný) význam
- Lze použít i na nenumerická data

7. Chyby 1. a 2. druhu

7.1. Skripta - Rogalewicz

Chyby mohou nastat i při testech významnosti, kdy testujeme nulovou hypotézu H_0 proti alternativní H_1. Skutečnou situaci, zda platí H_0 nebo H_1 neznáme a rozhodujeme se na základě nepřesného testu. Chyby rozlišujeme - viz tabulka:

na základě testu se	ve skutečnosti platí	
experimentátor rozhodne pro	H_0	H
H_0	v pořádku	chyba 2. d
H_1	chyba 1. druhu (α)	v poř

Obě chyby nelze současně snížit na minimum. Lze snížit pouze jednu na úkor druhé. Chyba 1. druhu je závažnější, statistik volí maximální velikost chyby 1. druhu α. Toto číslo se nazývá **hladina významnosti**. Obvykle se volí $\alpha = 0,05$ nebo $\alpha = 0,01$.

7.2. Skripta - Němičková

Při rozhodování o H_0 se můžeme dopustit dvou typů chyb:

- Můžeme zamítnout hypotézu, která platí. Pak mluvíme o chybě 1. druhu a označíme její pravděpodobnost α
- Mužeme přijmout hypotézu, která neplatí. Pak mluvíme o chybě 2. druhu a označíme její pravděpodobnost β

Obě chyby současně minimalizovat nelze. Chyba 1. druhu je závažnější, proto stanovíme její maximální velikost a mezi testy hledáme ten, pro který je minimální chyba 2. druhu při dané hodnotě chyby 1. druhu. Pravděpodobnost chyby 1. druhu α nazýváme hladinou významnosti příslušného testu.

8. Nestranný odhad a nejlepší nestranný odhad

Odhad je **nestranný**, jestliže střední hodnota jeho výběrového rozdělení je rovna hledanému parametru. Tedy $\hat{\theta}$ je nestranným odhadem θ, pokud $E\hat{\theta} = \theta$. Není-li odhad nestranný, nazveme jej **vychýleným**. **Nejlepším nestranným odhadem** nazýváme nestranný odhad s nejmenším rozptylem. Odhad $\hat{\theta}$ se nazývá **konzistentní**, jestliže

1. $E\hat{\theta} \to \theta$ pro $n \to \infty$
2. $var\hat{\theta} \to 0$ pro $n \to \infty$

Nechť $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ je náhodný výběr z rozdělení $N(\mu, \sigma^2)$. Potom

1. $\bar{X}$ je nejlepším nestranným odhadem μ
2. S^2 je nejlepším nestranným odhadem σ^2

9. F-rozdělení

F-rozdělení se používá při testu rovnosti rozptylů normálního rozdělení. Jestliže náhodné veličiny X_1 a X_2 mají obě rozdělení χ^2 a υ_1, resp. υ_2 stupni volnosti a jsou nezávislé, pak řekneme, že statistika $F = \frac{X_1/\upsilon_1}{X_2/\upsilon_2}$ má F-rozdělení s υ_1 a υ_2 stupni volnosti. Toto rozdělení závisí na parametrech υ_1 a υ_2. Statistika

$$F = \frac{S_1^2}{S_2^2} = \frac{\frac{(m-1)S_1^2}{\sigma^2} \cdot \frac{1}{m-1}}{\frac{(n-1)S_2^2}{\sigma^2} \cdot \frac{1}{n-1}}$$

se řídí F-rozdělením o $(m-1)$ a $(n-1)$ stupních volnosti.

Zvolíme testovací hypotézu $H_0 : \sigma_1^2 = \sigma_2^2$ proti alternativě $H_1 : \sigma_1^2 > \sigma_2^2$ nebo $H_1' : \sigma_1^2 \neq \sigma_2^2$
H_0 zamítneme, pokud platí buď:

- alternativa H_1: hodnota statistiky F je větší než $F_{1-\alpha}(m-1, n-1)$
- alternativa H_1': hodnota statistiky F je větší než $F_{1-\alpha/2}(m-1, n-1)$ nebo menší než $F_{\alpha/2}(m-1, n-1)$

Při hledání kvantilu můžeme využít vztah:

$$F_\beta(\upsilon_1, \upsilon_2) = \frac{1}{F_{1-\beta}(\upsilon_2, \upsilon_1)}$$

10. t-rozdělení

t-rozdělení se používá při testu rovnosti střední hodnoty normálního rozdělení.
Předpokládejme, že $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ je náhoný výběr z rozdělení $N(\mu, \sigma^2)$ a my testujeme hypotézu: $H_0 : \mu = \mu_0$ pro pevně danou hodnotu μ_0. Alternativní hypotéza v případě oboustranného testu: $H_1 : \mu \neq \mu_0$. V případě jednostranného testu: $H_1' : \mu > \mu_0$ nebo $H_1'' : \mu < \mu_0$. Testovaná statistika je: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
Nulovou hypotézu zamítneme na hladině významnosti α v těchto případech:

- alternativa H_1, pokud $|Z| > u_{1-\alpha/2}$
- alternativa H_1' (horní jednostranný test), pokud $Z > u_{1-\alpha}$
- alternativa H_1'' (dolní jednostranný test), pokud $Z < u_\alpha = -u_{1-\alpha}$

Pokud neznáme skutečnou hodnotu parametru σ^2, nahradíme ji jejím odhadem S^2. Potom však musíme použít statistiku $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$, tedy kvantily Studentova rozdělení o $(n-1)$ stupních volnosti.

Při porovnávání dvou normálních rozdělení se musíme ujistit, že jejich rozptyly jsou SHODNÉ (tedy provést test pomocí F-rozdělení). Testujeme hypotézu $H_0 : \mu_X - \mu_Y = 0$. Testovaná statistika je $Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}$, pokud je rozptyl $\sigma^2 = \sigma_X^2 = \sigma_Y^2$ známý. Pokud známý není, použijeme odhady S_X a S_Y. Pak použijeme odhad založený na obou rozptylech:

$$S = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{(m-1) + (n-1)}$$

A testovaná statistika bude $T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$, kterou porovnáváme s kvantilem Studentova rozdělení o $(m+n-2)$ stupních volnosti. (se stejnými hranicemi pro jedno- a oboustranné odhady jako v předchozím případě)

11. Charakteristická funkce

Definice. Funkce $\Psi(t) = Ee^{itX}$ se nazývá **charakteristickou funkcí** náhodné veličiny X.
V diskrétním případě dostáváme:

$$\Psi(t) = \sum_j e^{itx_j} \cdot P[X = x_j]$$

a pro spojité náhodné veličiny s hustotou $f(x)$:

$$\Psi(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx$$

Charakteristická funkce je tedy obraz hustoty při Fourierově transformaci.

Distribuční funkce je jednoznačně určena její charakteristickou funkcí.

Vlastnosti charakteristické funkce Nechť X je náhodná veličina a Ψ její charakteristická funkce. Potom:

(i) $\Psi(t)$ existuje pro každé $t \in R$

(ii) $\Psi(0) = 1$

(iii) $|\Psi(t)| \leq 1$

Nechť X je náhodná veličina a nechť existuje její k-tý obecný moment $m_k = EX^k$. Potom $m_k = \frac{1}{i^k}\Psi^{(k)}(0)$.

12. Čebyševova nerovnost

Nechť náhodná veličina X má střední hodnotu EX a rozptyl $var\, X$. Potom pro každé $\varepsilon > 0$ platí

$$P[|X - EX| \geq \varepsilon] \leq \frac{var\, X}{\varepsilon^2}$$