Přednáška 11a + b

Markovova nerovnost

Pro libovolnou nezápornou náhodnou veličinu X platí

$$P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon}.$$

Důkaz: Pro spojitou náh. veličinu máme:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x f(x)\,dx = \int_{0}^{\varepsilon} x f(x)\,dx + \int_{\varepsilon}^{\infty} x f(x)\,dx$$

$$\geq \int_{\varepsilon}^{\infty} \varepsilon f(x)\,dx = \varepsilon \int_{\varepsilon}^{\infty} f(x)\,dx = \varepsilon\, P(X \geq \varepsilon). \quad \blacksquare$$

Lze použít pro každou nezápornou náhodnou veličinu, známe-li její střední hodnotu.
Známe-li kromě střední hodnoty i rozptyl lze použít Čebyševovu nerovnost (uvedena dříve).

Příklad Ve fotbalové lize padne v průměru λ branek v zápase. Odhadněte pravděpodobnost, že padne aspoň 3λ branek.

Řešení pomocí Markovovy nerovnosti:

$$P(X \geq 3\lambda) \leq \frac{E(X)}{3\lambda} = \frac{\lambda}{3\lambda} = \frac{1}{3}.$$

Lepší odhad: Víme, že náhodná veličina X = počet hanek v zápase má Poissonovo rozdělení s pravděpodobnostní funkcí

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Toto rozdělení má střední hodnotu

$$E(X) = \lambda$$

a rozptyl

$$\text{var}(X) = \lambda.$$

Řešení úlohy pomocí Čebyševovy nerovnosti je pak

$$P(X \geq 3\lambda) = P(|X - \lambda| \geq 2\lambda) =$$

$$= P(|X - EX| \geq 2\lambda) \leq \frac{\text{var}(X)}{(2\lambda)^2} = \frac{\lambda}{4\lambda^2} = \frac{1}{4\lambda}.$$

■

Matematická statistika

- zkoumáme výběr z nějakého souboru (populace)
- chceme vědět, do jaké míry jsou zjištěné výsledky platné pro celý soubor (populaci)
- případně ze zjištěných dat se snažíme upřesnit předný teoretický model pro chování celého souboru a z něho pak odhadnout pravděpodobnost budoucího jevu

2 přístupy

- frekvenční (klasická) statistika
- bayesovská statistika

Frekvenční statistika

- vychází z mat. abstrakce, že teoretická pravděpodobnosti jsou dány četností výskytu jevů v dlouhých řadách dat, které můžeme dobře aproximovat nekonečnými modely a využívá pro odhady spolehlivosti centrální limitní věty.
- má své limity, pokud nejsou data spolehlivá nebo je experiment nevhodný
- nelze použít pro odhad výsledků jednorázových

děje.

Náhodný výběr

Základní soubor (populace) má N jednotek. Pro každou jednotku můžeme určit číselný znak $x_1, x_2, \dots, x_N$. Ze základního souboru vybereme $n < N$ jednotek jednu po druhé a každou jednotku pak do souboru vrátíme. Tedy každá jednotka má pravděpodobnost výběru $\frac{1}{N}$. Hovoříme o náhodném výběru.

V matematické idealizaci pracujeme s n-ticí nezávislých náhodných veličin $X_1, X_2, \dots, X_n$, které mají všechny stejné (tj. stejnou distribuční funkci $F_X(x)$). Tuto n-tici náhodných veličin nazýváme NÁHODNÝ VÝBĚR ROZSAHU n.

V matematické statistice často pracujeme s transformacemi náhodného výběru

$$Y = f(X_1, X_2, \dots, X_n),$$

takovým náhodným veličinám (případně vektorům) říkáme statistiky.

Zavedeme několik důležitých statistik a ukážeme jejich souvislost s číselnými charakteristikami.

Definice Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr. Statistiku

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i$$

nazýváme výběrovým průměrem.

Statistiku

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$$

výběrovým rozptylem a statistiku $S = \sqrt{S^2}$ výběrovou směrodatnou odchylkou.

Statistiky jsou opět náhodné veličiny, proto se lze ptát na jejich číselné charakteristiky - střední hodnota, rozptyl, momenty, rozdělení.

Věta Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2. Pak platí

(1) $EM = \mu$

(2) $\operatorname{var} M = \frac{\sigma^2}{n}$

(3) $ES^2 = \sigma^2$

Důkaz: (1) a (2) plyne z vlastností střední hodnoty a rozptylu.

(3) Platí $\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - M)^2 + n(M-\mu)^2$

Na levé straně se po umocnění vyskytují výrazy typu

L: X_i^2, $-2X_i\mu$, $n\mu^2$

Na pravé straně se vyskytují výrazy

P: X_i^2, $-2MX_i$, $2nM^2$, $-2nM\mu$, $n\mu^2$

První a poslední výrazy se okrádají. Dále

$$\sum_{i=1}^{n} (-2X_i\mu) = -2n\,\frac{X_1 + \dots + X_n}{n}\,\mu$$

a $2nM^2 - 2M\left(\sum X_i\right) = 2nM^2 - 2nM\,\frac{\sum X_i}{n} =$

$= 2nM^2 - 2nM^2 = 0$

Důkaz

$$E(S^2) = E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2\right\} =$$

$$= E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{n-1}(M-\mu)^2\right\}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i-\mu)^2 - \frac{n}{n-1}E(M-\mu)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \operatorname{var} X_i - \frac{n}{n-1}\operatorname{var} M =$$

$$= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 = \sigma^2. \qquad \blacksquare$$

Dokázali jsme, že výběrový průměr M splňuje

$$E(M) = \mu,$$

jeho střední hodnota je tedy rovna odhadovanému parametru μ. Říkáme, že statistika M je <u>nestranným odhadem</u> parametru μ.

Obdobně statistika S^2 je nestranným odhadem parametru σ^2.

<u>Poznámka</u> „Přirozenější" definovaná statistika $\frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2$ není nestranným odhadem σ^2, její střední hodnota je $\frac{n-1}{n}\sigma^2$.

Náhodný výběr z normálního rozdělení

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr a nechť každá z náhodných veličin má normální rozdělení $N(\mu, \sigma^2)$.

Časem dokážeme a doplníme následující

Věta:

(1) M a S^2 jsou nezávislé náhodné veličiny.

(2) M má normální rozdělení se střední hodnotou μ a rozptylem σ^2/n, tj.

$$M \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(3) Náhodná veličina $U = \dfrac{\sqrt{n}(M-\mu)}{\sigma}$

má normální rozdělení se střední hodnotou 0 a rozptylem 1.

$$U \sim N(0,1).$$

V následujícím příkladu si ukážeme, jak lze větu použít k odhadu μ, známe-li σ^2.

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. Naměřené hodnoty byly:

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 rozptyl výšek se v uvažovaných generacích příliš nemění.

Řešení Data lze považovat za náhodný výběr. Zjistíme výběrový průměr $M = 139,133$, $n = 15$. $X_1, X_2, \dots X_{15}$ je náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$. Střední hodnotu μ pro rok 1961 neznáme, σ považujeme za dejme tomu jako v roce 1951, tj. $\sigma = 6,4$. Zajímá nás interval $(M - \varepsilon, M + \varepsilon)$ takový, že

$$P(M-\varepsilon < \omega < M+\varepsilon) = 0,95$$

Nerovnosti $M-\varepsilon < \omega < M+\varepsilon$ jsou ekvivalentní s nerovností

$$|M-\omega| < \varepsilon$$

a ta je ekvivalentní s nerovností

$$|U| = \left|\frac{\sqrt{n}(M-\omega)}{\sigma}\right| < \frac{\sqrt{n}}{\sigma}\varepsilon$$

Náhodná veličina U má stand. normální rozdělení s distribuční funkcí Φ

Proto chceme najít ε tak, aby

$$P\left(-\frac{\sqrt{n}}{\sigma}\varepsilon < U < \frac{\sqrt{n}}{\sigma}\varepsilon\right) = 0,95$$

Pomocí distribuční funkce Φ to je

$$\Phi\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) - \Phi\left(-\frac{\sqrt{n}}{\sigma}\varepsilon\right) = 0,95$$

$$2\Phi\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) - 1 = 0,95$$

$$\Phi\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) = \frac{1,95}{2} = 0,975$$

$$\frac{\sqrt{n}}{\sigma}\varepsilon = \Phi^{-1}(0,975) = 1,96$$

$$\varepsilon = \frac{1,96\cdot\sigma}{\sqrt{n}} = 3,23$$

Interval $(M - 3{,}23,\ M + 3{,}23)$ je interval s okraji zadanými náhodnými veličinami. V praxi nahradíme M použitým průměrem $139{,}13$. Pak dostaneme interval

$$(139{,}13 - 3{,}23,\ 139{,}13 + 3{,}23) = (135{,}90;\ 142{,}36)$$

Střední hodnota výšek z roku 1951 leží v tomto intervalu, nemáme tedy se spolehlivostí 95% zjištěno, že se průměrná výška změnila.

Pokud si vezmeme spolehlivost pouze 90%, pak je interval

$$(136{,}41\ ;\ 141{,}85).$$

V něm už střední hodnota 136,1 z roku 1951 neleží. Tedy střední hodnota výšky se změnila se spolehlivostí 90%.

Pokud nás zajímá pouze dolní odhad, tj. interval

$$(M - \varepsilon,\ \infty)$$

takový, že $P(M - \varepsilon < \mu) = 0{,}95$,

dostaneme interval

$$(136{,}41\ ,\ \infty)$$

Tedy s pravděpodobností 95% je střední výška v roce 1961 větší než 136,41 cm a tedy i větší než v roce 1951. ■

Teoretické shrnutí předchozího příkladu

Interval spolehlivosti pro střední hodnotu μ

Nechť $X_1, X_2, \dots X_n$ jsou náhodné veličiny s rozdělením $N(\mu, \sigma^2)$. Nechť $\alpha \in (0,1)$

~~Oboustranný~~ Oboustranný interval pro střední hodnotu μ spolehlivosti $1-\alpha$ je

$$(M-\varepsilon,\ M+\varepsilon)$$

s vlastností

$$P(M-\varepsilon < \mu < M+\varepsilon) = 1-\alpha$$

jednostranný interval spolehlivosti je levostranný

$$(M-\varepsilon,\ \infty), \text{ kde } P(M-\varepsilon < \mu) = 1-\alpha$$

a pravostranný $(-\infty, M+\varepsilon)$, kde $P(\mu < M+\varepsilon) = 1-\alpha$.

Stejně jako v příkladu spočteme, že

$$\Phi\left(\frac{\sqrt{n}}{\sigma}\varepsilon\right) = 1-\frac{\alpha}{2}$$

$$\frac{\sqrt{n}}{\sigma}\varepsilon = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

Tedy oboustranný interval spolehlivosti $1-\alpha$ je

$$\left(M - \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1-\frac{\alpha}{2}\right),\ M + \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \right)$$

Levostranný interval spolehlivosti je

$$\left(M - \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\alpha),\ \infty \right),$$

pravostranný pak

$$\left(-\infty,\ M + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1-\alpha) \right).$$

■

Vrátíme se k větě před příkladem

Věta: Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr s rozložením $N(\mu, \sigma^2)$. Pak M má normální rozložení $N(\mu, \frac{\sigma^2}{n})$ a

$$U = \frac{\sqrt{n}\,(M-\mu)}{\sigma} \sim N(0,1).$$

Důkaz pro libovolné n – dále již známe. Víme, že normální rozložení $Z \sim N(0,1)$ má momentovou funkci $E(e^{tZ}) = e^{\frac{t^2}{2}}$.

Normální rozložení $X \sim N(\mu, \sigma^2)$ má momentovou funkci

$$M_X(t) = M_{\mu+\sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} \cdot e^{\frac{(\sigma t)^2}{2}}$$

Nechť $X_1 \sim N(\mu_1, \sigma_1)$ a $X_2 \sim N(\mu_2, \sigma_2)$.

Potom

$$M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t) =$$

$$= e^{\mu_1 t} \cdot e^{\frac{\sigma_1^2 t^2}{2}} \cdot e^{\mu_2 t} \cdot e^{\frac{\sigma_2^2 t^2}{2}} =$$

$$= e^{(\mu_1+\mu_2)t} e^{\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}} = M_Y(t)$$

kde $Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. Tedy součet nezávislých náhodných veličin s normálním rozdělením je opět náhodná veličina s normálním rozdělením.

Proto jestliže $X_i \sim N(\mu, \sigma^2)$, pak

$$X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$$

a $$M = \frac{1}{n}(X_1 + X_2 + \dots + X_n) \sim N(\mu, \frac{\sigma^2}{n}).$$ ∎

Budeme se zabývat dalšími rozděleními, které jsou z normálního rozložení odvozeny nebo jsou jeho zobecněním.

Γ rozdělení $\Gamma(a,b)$

má hustotu

$$f(x) = \begin{cases} 0, & x \le 0 \\ c\,x^{a-1} e^{-bx}, & x > 0, \end{cases}$$

kde $c = \frac{b^a}{\Gamma(a)}$, $a \ge 0$, $b \ge 0$,

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt .$$

Poznámka: Funkce Γ je zobecněním faktoriálu, neboť $\Gamma(n) = (n-1)!$ pro $n \in \mathbb{N}$.
Dále platí $\Gamma(1/2) = \sqrt{\pi}$

$$\Gamma(a+1) = a \cdot \Gamma(a).$$

Věta Momentová vytvořující funkce rozdělení $\Gamma(a,b)$ je

$$M(t) = \left(\frac{b}{b-t}\right)^a,$$

střední hodnota je a/b a rozptyl a/b^2.

Důkaz za DÚ

Rozdělení χ^2 – chí kvadrát $\chi^2(1)$

Nechť Z je náhodná veličina s rozdělením $N(0,1)$. Spočítáme hustotu f pro náhodnou veličinu Z^2:
Zřejmě $f(x) = 0$ pro $x \le 0$. Pro kladná x máme

$$F(x) = P(Z^2 < x) = P(-\sqrt{x} < Z < \sqrt{x}) =$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \int_0^y \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} e^{-\frac{t}{2}} dt$$

Hustotu dostaneme derivací distribuční funkce

$$f(x) = \frac{d}{dx} F(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

Toto je gamma rozdělení $\Gamma(\frac{1}{2}, \frac{1}{2}) = \chi^2(1)$.

Věta Nechť $X_1, X_2, \dots, X_n$ jsou nezávislé náhodné veličiny s rozděleními $\Gamma(a_i, b)$. Pak $Y = X_1 + X_2 + \dots + X_n$ má rozdělení $\Gamma(a_1 + a_2 + \dots + a_n, b)$.

Důkaz plyne z výpočtu momentů

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \frac{b^{a_1}}{(b-t)^{a_1}} \cdot \frac{b^{a_2}}{(b-t)^{a_2}} \cdots \frac{b^{a_n}}{(b-t)^{a_n}}$$

$$= \left(\frac{b}{(b-t)}\right)^{a_1 + a_2 + \dots + a_n}.$$

■

Rozdělení $\chi^2(n)$ – chí kvadrát s n stupni volnosti

Nechť $Y = Z_1^2 + Z_2^2 + \dots + Z_n^2$, kde $Z_1, \dots, Z_n$ jsou nezávislé náhodné veličiny

s rozdělením $N(0,1)$. Pak Z_i^2 mají rozdělení $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$ a Y má rozdělení

$$\Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

To se nazývá chí kvadrát s n stupni volnosti a značí se $\chi^2(n)$. ■