Základy matematické statistiky

Problém: máme řadu nějakých pozorování (hodnot), předpokládáme, že vznikla pozorováním nebo měřením nějaké náhodné veličiny.
Úkolem je zjistit něco o té neznámé veličině – jaké má pravděpodobnostní rozdělení nebo alespoň nějaké charakteristiky, např. střední hodnotu, rozptyl.

Náhodná veličina je kvantitativním ohodnocením nějakého náhodného jevu, může nabývat buď konečné nebo spočetné množiny hodnot (např. počet ok na hrací kostce) – takové veličiny nazýváme diskrétní, nebo všech hodnot z nějakého konečného či nekonečného intervalu (např. teplota) – takové veličiny jsou spojité.

Rozdělení náhodné veličiny – charakterizuje se např. distribuční funkcí, která se definuje jako $F(x) = P(X < x)$.
V diskrétním případě, když veličina nabývá hodnot $x_1, x_2, \ldots$, se dá charakterizovat systémem pravděpodobností $p_i = P(X = x_i)$, $i = 1, 2, \ldots$, kde $\sum_{i=1}^{\infty} p_i = 1$. Platí $F(x) = \sum_{\{i : x_i < x\}} p_i$. Pravděpodobnosti p_i se někdy nazývají diskrétní hustota.
Spojité náhodné veličiny jsou charakterizovány hustotou, což je funkce $f(x)$ taková, že platí: $f(x) \geq 0$, $f(x) \leq 1$, $\int_{-\infty}^{\infty} f(x)\,dx = 1$, $F(x) = \int_{-\infty}^{x} f(t)\,dt$.

Další charakteristiky:

<u>Střední hodnota</u> – je hypotetický průměr, kolem kterého kolísají hodnoty náhodné veličiny (je mírou polohy).

<u>diskrétní veličina</u>: $E(X) = \sum_{i=1}^{\infty} x_i \, P(X = x_i)$

<u>spojitá veličina</u>: $E(X) = \int_{-\infty}^{\infty} x \, f(x)\, dx$.

<u>Rozptyl</u> – popisuje vzdálenosti hodnot náhodné veličiny od její střední hodnoty (je mírou měřítka).

značí se σ^2 nebo $var(X)$

definice: $var(X) = E(X - EX)^2$

platí: $var(X) = EX^2 - (EX)^2$

přitom $EX^2 = \sum_i x_i^2 \, P(X = x_i)$ v diskrétním případě

$EX^2 = \int x^2 f(x)\, dx$ ve spojitém případě

<u>Medián</u> $\tilde{x}$ – je "prostřední" hodnota

platí: $F(\tilde{x}) = \frac{1}{2}$, tj. $P(X < \tilde{x}) = P(X \geq \tilde{x}) = \frac{1}{2}$

(pro spojité veličiny to platí přesně, pro diskrétní skoro)

<u>Modus</u> $\hat{x}$ – je nejpravděpodobnější hodnota

platí: $f(\hat{x}) = \max_x \{f(x)\}$ pro spojitá rozdělení

$P(X = \hat{x}) = \max_i P(X = x_i)$

Příklady diskrétních rozdělení:

Alternativní $(0-1)$: $X = 0, 1$, $P(X=1) = p$, $P(X=0) = 1-p$

$$EX = 1 \cdot p + 0 \cdot (1-p) = p$$

$$EX^2 = 1^2 \cdot p + 0^2 \cdot (1-p) = p$$

$$\operatorname{var} X = EX^2 - (EX)^2 = p - p^2 = p(1-p)$$

příklad: hod mincí

$$\left.\begin{array}{ll} X = 1 & \text{když padne "panna"} \\ X = 0 & \quad \text{—"— "orel"} \end{array}\right\} \text{ obojí s pravděpodobností } \frac{1}{2}$$

Binomické: $X = 0, 1, \dots n$; $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

p je parametr rozdělení, $0 < p < 1$

$EX = p$, $\operatorname{var} X = np$

příklad: počet "úspěchů" v posloupnosti n nezávislých pokusů, z nichž každý může skončit úspěchem nebo neúspěchem, $P(\text{úspěchu}) = p$, $P(\text{neúspěchu}) = 1-p$

Poissonovo. $X = 0, 1, \dots$, $P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $\lambda > 0$ je parametr

$EX = \lambda$, $\operatorname{var} X = \lambda$

je limitním případem binomického, když $\lim_{n \to \infty} np = \lambda$

<u>geometrické</u>: $X = 0, 1, \ldots$, $P(X=k) = p(1-p)^k$

$EX = \frac{1-p}{p}$, $\operatorname{var} X = \frac{1-p}{p^2}$

příklad: rozdělení "doby čekání" na první výskyt "úspěchu"

<u>rovnoměrné</u>: $X = 1, 2, \ldots, n$; $P(X=k) = \frac{1}{n}$

$EX = \frac{n+1}{2}$, $\operatorname{var} X = \frac{n^2-1}{12}$

příklad: hod kostkou $X = 1, 2, \ldots, 6$, $P(X=k) = \frac{1}{6}$

<u>Příklady spojitých rozdělení</u>

<u>rovnoměrné $R(a,b)$</u>: $X \in (a,b)$, $f(x) = \frac{1}{b-a}$

$EX = \frac{a+b}{2}$, $\operatorname{var} X = \frac{(b-a)^2}{12}$

<u>exponenciální</u>: $X \in (0, \infty)$, $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0$ je parametr

$EX = \frac{1}{\lambda}$, $\operatorname{var} X = \frac{1}{\lambda^2}$

používá se v teorii spolehlivosti, teorii hromadné obsluhy

My se budeme nejčastěji setkávat s následujícími rozděleními:

<u>normální $N(\mu, \sigma^2)$</u>: $X \in (-\infty, \infty)$, $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

μ, σ jsou parametry, $\sigma > 0$

$EX = \mu$, $\operatorname{var} X = \sigma^2$

<u>normované normální rozdělení</u> : $X \in (-\infty, \infty)$, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$EX = 0$, $\operatorname{var} X = 1$

<u>chí-kvadrát rozdělení χ^2_n</u> (s n stupni volnosti) : $X \in (0, \infty)$

$$f(x) = \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} x^{n/2 - 1} e^{-x/2}$$

kde $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x}\, dx$ pro $a > 0$ je Γ-funkce

plati: $\Gamma(a+1) = a\Gamma(a)$, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$EX = n$, $\operatorname{var} X = 2n$

<u>Platí</u> (v důsledku věty o transformaci náhodných veličin):

jsou-li $X_1, X_2, \dots, X_n$ nezávislé náhodné veličiny a každá má rozdělení $N(0,1)$, pak $U = X_1^2 + \dots + X_n^2$ má rozdělení χ^2_n

<u>Studentovo rozdělení t_n</u> (t-rozdělení s n stupni volnosti):

$X \in (-\infty, \infty)$

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$EX = 0$, $\operatorname{var} X = \frac{n}{n-2}$ pro $n > 2$

<u>Platí</u>: jsou-li X a U nezávislé náhodné veličiny, $X \sim N(0,1)$, $U \sim \chi^2_n$, pak veličina $T = \frac{X}{\sqrt{\frac{U}{n}}}$ má rozdělení t_n.

<u>Fisherovo rozdělení F_{n_1,n_2}</u> (F-rozdělení s n_1 a n_2 stupni volnosti).

$X \in (0,\infty)$

$$f_{n_1,n_2}(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1}\left(1+\frac{n_1}{n_2}\right)^{-\frac{n_1+n_2}{2}}$$

$$EX = \frac{n_2}{n_2-2} \quad \text{pro } n_2 > 2$$

$$\operatorname{var} X = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)(n_2-4)} \quad \text{pro } n_2 > 4$$

<u>Platí</u>: Jsou-li U a V nezávislé náhodné veličiny, $U \sim \chi^2_m$, $V \sim \chi^2_n$, pak $F = \dfrac{U/m}{V/n}$ má rozdělení $F_{m,n}$.

Tvary hustot těchto rozdělení není třeba si pamatovat, dříve byly uváděny ve statistických tabulkách, dnes se konkrétní hodnoty vypočtou na počítači.

Důležité je poznat, kde se dá které rozdělení použít.

Budeme potřebovat hlavně kvantily nebo kritické hodnoty těchto rozdělení (opět jsou v tabulkách).

<u>Kvantil</u> rozdělení je hodnota u_α, pro kterou platí $P(U < u_\alpha) = \alpha$.

(U je náhodná veličina s tímto rozdělením).

<u>Kritická hodnota</u> rozdělení je hodnota x_α, pro kterou $P(X > x_\alpha) = \alpha$

(X je náh. veličina s tímto rozdělením).

Vztah mezi kritickou hodnotou a kvantilem téhož rozdělení je tedy

$$x_\alpha = u_{1-\alpha}.$$

Je ale třeba dát pozor:

1) pro některá rozdělení se veličiny v předchozích definicích berou v absolutní hodnotě
2) kvantily i kritické hodnoty se ve statistické literatuře obvykle značí stejnými symboly
3) v některých tabulkách se uvádějí kvantily, v jiných kritické hodnoty

Je tedy třeba vždy důkladně číst celý text, nejen vzorečky!

Nezávislost náhodných veličin.

Veličiny $X_1, \dots, X_n$ jsou nezávislé, jestliže platí $F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n)$, kde $F(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n)$ je sdružená distribuční fce, $F_i(x_i) = P(X_i < x_i)$ jsou marginální distribuční fce jednotlivých veličin (musí to platit pro každé n a všechny možné hodnoty $x_1, \dots, x_n$).

Jinak: $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$

Důsledky: $E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n)$

$$var(X_1 + \dots + X_n) = var(X_1) + \dots + var(X_n)$$

$$cov(X_i, X_j) = E(X_i - EX_i)(X_j - EX_j) = 0 \text{ pro všechna } i, j$$

Náhodný výběr je posloupnost navzájem nezávislých náhodných veličin se stejným rozdělením.

Použití mnoha statistických metod předpokládá normální rozdělení, ale tento předpoklad nebývá vždy v praxi splněn. Přesto jsou některé z nich použitelné. Proč? Platí tzv. Centrální limitní věta:

Nechť $X_1, X_2, \dots, X_n, \dots$ jsou nezávislé stejně rozdělené náhodné veličiny, $E X_i = \mu < \infty$, $\operatorname{var} X_i = \sigma^2 < \infty$ pro všechna i. Položme $Y_n = \sum_{k=1}^{n} X_k$. Pak pro distribuční fci $F_n(x)$ náhodné veličiny $Y_n^* = \frac{Y_n - E(Y_n)}{\sqrt{\operatorname{var} Y_n}}$ platí

$\lim_{n\to\infty} F_n(x) = \Phi(x)$, kde $\Phi(x)$ je distr. fce rozdělení $N(0,1)$.

neboli: $$\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n\sigma^2}} \xrightarrow[n\to\infty]{} Z \sim N(0,1) \text{ v distribuci}$$

když zlomek vydělíme n: $$\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{} Z \sim N(0,1)$$

Převod rozdělení $N(0,1)$ na $N(\mu, \sigma^2)$ a naopak:

nechť $X \sim N(0,1)$, pak $Y = \sigma X + \mu \sim N(\mu, \sigma^2)$

nechť $X \sim N(\mu, \sigma^2)$, pak $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$

Centrální limitní věta pak vlastně říká, že aritmetické průměry

$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ mají asymptoticky normální rozdělení $N(\mu, \frac{\sigma^2}{n})$,

i když veličiny X_k normální rozdělení nemají!

Další trik, který nám někdy dovoluje ověřit předpoklady, je zákon velkých čísel. Existuje ve 2 podobách:

Slabý zákon velkých čísel: Nechť $X_1, X_2, \ldots$ jsou po dvou nezávislé náhodné veličiny, $EX_i = \mu < \infty$, $\operatorname{var} X_i = \sigma^2 < \infty$ pro všechna i. Pak

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

(konvergují podle pravděpodobnosti).

Znamená to, že pro libovolné $\varepsilon > 0$ platí $\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$.

Silný zákon velkých čísel: Nechť $X_1, X_2, \ldots$ jsou nezávislé náhodné veličiny s týmž rozdělením, $EX_i = \mu < \infty$, $\operatorname{var} X_i = \sigma^2 < \infty$. Pak

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu$ skoro jistě, tj. $P(\lim_{n \to \infty} \bar{X}_n = \mu) = 1$.

Náhodný výběr z normálního rozdělení $N(\mu,\sigma^2)$

posloupnost $X_1, \dots, X_n$ navzájem nezávislých náhodných veličin, každá má rozdělení $N(\mu,\sigma^2)$
(v praxi to budou napozorované hodnoty)

Bodové odhady parametrů:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

odhad je <u>nestranný</u>, tj: $E\bar{X} = \mu$ $\quad \left(\text{dk. } E\bar{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu\right)$

<u>konzistentní</u>, tj: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu$ podle zákona velkých čísel

<u>maximálně věrohodný</u> (zatím nebudeme definovat)

má rozdělení $N\left(\mu, \frac{\sigma^2}{n}\right)$ $\quad \left(\operatorname{var}\bar{X} = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}X_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\right)$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \left(= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2\right)\right)$$

odhad je <u>nestranný</u>, tj: $ES^2 = \sigma^2$

<u>konzistentní</u>, tj: $S^2 \longrightarrow \sigma^2$

není maximálně věrohodný

veličina $\frac{(n-1)S^2}{\sigma^2}$ má rozdělení χ^2_{n-1}

platí $$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\underbrace{\left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{\text{tyto veličiny mají rozdělení } N(0,1)}$$

nejsou ale všechny nezávislé, proto je počet stupňů volnosti $n-1$ a ne n.

$\bar{X}$ a S^2 jsou <u>nezávislé</u>

Skutečné bodové odhady parametrů se budou nacházet s velkou pravděpodobností v nějakém intervalu kolem svých bodových odhadů. Meze se určí podle toho, jak velkou psl chceme.

Nejjednodušší případ: <u>Interval spolehlivosti pro μ (neznámé) při známém σ^2.</u>

zvolíme nějaké $\alpha \in (0,1)$. – mělo by být malé, např. $\alpha = 0{,}05$

hledáme a, b tak, aby platilo

$$\left.\begin{array}{ll} \text{buď} & P(a < \mu) = 1-\alpha \\ \text{nebo} & P(\mu < b) = 1-\alpha \end{array}\right\} \text{ jednostranné intervaly spolehlivosti}$$

$$\text{nebo} \quad P(a < \mu < b) = 1-\alpha \qquad \text{oboustranný} \ \text{–''–}$$

$1-\alpha$ je koeficient spolehlivosti (konfidence)

vyjdeme z odhadu $\bar{X}$, znormujeme ho na $\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

platí $P\left(\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\alpha}\right) = 1-\alpha$, kde $u_{1-\alpha}$ je kvantil rozdělení $N(0,1)$

pak $1-\alpha = P\left(\bar{X}-\mu < \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right) = P\left(\underbrace{\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}}_{a} < \mu\right)$

Definice kvantilu: Nechť Y je náhodná veličina s distribuční fcí F. Hodnota x_α, pro kterou $F(x_\alpha) = P(Y < x_\alpha) = \alpha$, se nazývá α-kvantilem příslušného rozdělení.

Kvantily známých rozdělení se najdou v tabulkách.

podobně: $P\left(\mu < \overbrace{\bar{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}}^{b}\right) = P\left(-u_{1-\alpha} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu\right) =$

$$= P\left(-u_{1-\alpha} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < -u_{1-\alpha}\right) =$$

$$= 1 - \Phi(-u_{1-\alpha}) = 1 - [1 - \Phi(u_{1-\alpha})] = \Phi(u_{1-\alpha}) = 1-\alpha$$

Φ je distribuční fce $N(0,1)$

platí pro ni $\Phi(-u) = 1 - \Phi(u)$, protože rozdělení je symetrické

oboustranný interval bude mít meze $\left(\bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}},\ \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$

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

$$= P\left(-u_{1-\alpha/2} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\alpha/2}\right) = \Phi(u_{1-\alpha/2}) - \Phi(-u_{1-\alpha/2}) =$$

$$= 1 - \alpha/2 - (1 - (1 - \alpha/2)) = 1 - \alpha$$

Interval spolehlivosti pro μ při neznámém σ^2

Když σ^2 neznáme, nahradíme ho jeho odhadem S^2 a použijeme výraz $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$, ten už ale nebude mít normální rozdělení.

platí: $$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{Y}{\sqrt{\frac{Z}{n-1}}}, \quad \text{kde } Y \sim N(0,1),\ Z \sim \chi^2_{n-1}$$

takže $$\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Intervaly spolehlivosti budou vypadat stejně, pouze místo kvantilu u_α normálního rozdělení se použijí kvantily $t_{n-1,\alpha}$ Studentova t-rozdělení:

<u>Interval spolehlivosti pro σ^2</u>

vyjdeme z odhadu S^2 a z toho, že víme, že $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

pak $P\left(\frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha}\right) = 1-\alpha$, když $\chi^2_{n-1,1-\alpha}$ je kvantil

tj. tedy $P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}} < \sigma^2\right) = 1-\alpha$

podobně $P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}\right) = 1-\alpha$

a oboustranně $P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha.$