Základní pojmy matematické statistiky

Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oborech lidské činnosti. Přitom se řídí principem statistické indukce, tj. na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží učinit závěry o vlastnostech tohoto rozložení.
Ústředním pojmem matematické statistiky je tedy pojem náhodného výběru.

Definice náhodného výběru:

a) Nechť $X_1, ..., X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$. Řekneme, že $X_1, ..., X_n$ je náhodný výběr rozsahu n z rozložení $L(\vartheta)$. (Číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ uspořádané do sloupcového vektoru odpovídají datovému souboru zavedenému v popisné statistice.)
b) Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1), ..., (X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1), ..., (x_n,y_n)$ náhodného výběru $(X_1,Y_1), ..., (X_n,Y_n)$ uspořádané do matice typu $2\times n$ odpovídají dvourozměrnému datovému souboru zavedenému v popisné statistice.)
c) Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\vartheta)$.

Definice statistiky:

Libovolná funkce $T = T(X_1, ..., X_n)$ náhodného výběru $X_1, ..., X_n$ (resp. $T = T(X_1,Y_1, ..., X_n,Y_n)$ náhodného výběru $(X_1,Y_1), ..., (X_n,Y_n)$) se nazývá (výběrová) statistika.

Definice důležitých statistik:

a) Nechť $X_1, ..., X_n$ je náhodný výběr, $n \geq 2$.

Označme $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ … výběrový průměr, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ … výběrový rozptyl, $S = \sqrt{S^2}$ … výběrová směrodatná odchylka

Pro libovolné, ale pevně dané reálné číslo x je statistikou též hodnota výběrové distribuční funkce $F_n(x) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x\}$

b) Nechť je dáno $r \geq 2$ stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2, \ldots, n_r \geq 2$.

Celkový rozsah je $n = \sum_{j=1}^{r} n_j$.

Označme $M_1, \ldots, M_r$ výběrové průměry a $S_1^2, \ldots, S_r^2$ výběrové rozptyly jednotlivých výběrů. Nechť $c_1, \ldots, c_r$ jsou reálné konstanty, aspoň jedna nenulová.

$\sum_{j=1}^{r} c_j M_j$ … lineární kombinace výběrových průměrů, $S_*^2 = \frac{\sum_{j=1}^{r}(n_j - 1)S_j^2}{n - r}$ … vážený průměr výběrových rozptylů.

c) Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení o rozsahu n.

Označme $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$ výběrové průměry, $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2$, $S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2$ výběrové rozptyly.

$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$… výběrová kovariance, $R_{12} = \begin{cases} \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 \neq 0 \\ 0 & \text{jinak} \end{cases}$ … výběrový koeficient korelace.

Pro libovolnou, ale pevně zvolenou dvojici reálných čísel x,y je statistikou též hodnota výběrové simultánní distribuční funkce $F_n(x,y) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x \wedge Y_i \leq y\}$.

Upozornění: Číselné realizace statistik M, S^2, S, S_{12}, R_{12} odpovídají číselným charakteristikám m, s^2, s, s_{12}, r_{12} zavedeným v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n-1}$, nikoliv $\frac{1}{n}$, jak tomu bylo v popisné statistice. Jak uvidíme později, uvedené číselné realizace mohou být považovány za odhady číselných realizací náhodných veličin zavedených v počtu pravděpodobnosti.

Charakteristika vlastnosti	Počet pravděpodobnosti	Matematická statistika	Popisná statistika
poloha	$E(X) = \mu$	M	m
variabilita	$D(X) = \sigma^2$	S^2	$\frac{n-1}{n}s^2$
variabilita	$\sqrt{D(X)} = \sigma$	S	$\sqrt{\frac{n-1}{n}}s$
společná variabilita	$C(X_1, X_2) = \sigma_{12}$	S_{12}	$\frac{n-1}{n}s_{12}$
těsnost vztahu	$R(X_1, X_2) = \rho$	R_{12}	r_{12}
rozložení	$\Phi(x)$	$F_n(x)$	F(x)

Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru $X_1, ..., X_n$ z rozložení $L(\vartheta)$, které závisí na parametru ϑ. Množinu všech přípustných hodnot tohoto parametru označíme Ξ. Tato množina se nazývá parametrický prostor.
Např. je-li $X_1, ..., X_n$ náhodný výběr z rozložení $N(\mu,\sigma^2)$, pak $\vartheta = (\mu, \sigma^2)$ a v tomto případě parametrický prostor $\Xi = (-\infty, \infty) \times \langle 0, \infty)$.

Parametr ϑ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\vartheta)$).

Bodovým odhadem parametrické funkce $h(\vartheta)$ je statistika $T_n = T(X_1, ..., X_n)$, která nabývá hodnot blízkých $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

Intervalovým odhadem parametrické funkce $h(\vartheta)$ rozumíme interval (D, H), jehož meze jsou statistiky
$D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\vartheta)$, ať je hodnota parametru ϑ jakákoliv.

Typy bodových odhadů

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, T, T_1, T_2, ... jsou statistiky.

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\vartheta)$, jestliže
$\forall \vartheta \in \Xi : E(T) = h(\vartheta)$.
(Význam nestrannosti spočívá v tom, že odhad T nesmí parametrickou funkci $h(\vartheta)$ systematicky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li T_1, T_2 nestranné odhady téže parametrické funkce $h(\vartheta)$, pak řekneme, že T_1 je lepší odhad než T_2, jestliže
$\forall \vartheta \in \Xi : D(T_1) < D(T_2)$.

c) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost asymptoticky nestranných odhadů parametrické funkce $h(\vartheta)$, jestliže
$$\forall \vartheta \in \Xi : \lim_{n\to\infty} E(T_n) = h(\vartheta).$$
(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu.)

d) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost konzistentních odhadů parametrické funkce $h(\vartheta)$, jestliže
$$\forall \vartheta \in \Xi \forall \varepsilon > 0 : \lim_{n\to\infty} P\left(\left|T_n - h(\vartheta)\right| > \varepsilon\right) = 0.$$
(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat „daleko“ od parametrické funkce $h(\vartheta)$.)

Lze dokázat, že z nestrannosti odhadu vyplývá jeho asymptotická nestrannost a z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů odhadu konverguje k nule.

Vlastnosti důležitých statistik

a) Případ jednoho náhodného výběru: Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou μ, rozptylem σ^2 a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme M_n výběrový průměr, S_n^2 výběrový rozptyl a pro libovolné, ale pevně dané $x \in R$ označme $F_n(x)$ hodnotu výběrové distribuční funkce. Pak pro libovolné hodnoty parametrů μ , σ^2 a libovolné, ale pevně dané reálné číslo x platí:

$E(M_n) = \mu$,

$D(M_n) = \frac{\sigma^2}{n}$,

$E(S_n^2) = \sigma^2$,

$D(S_n^2) = \frac{\gamma_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, kde γ_4 je 4. centrální moment,

$E(F_n(x)) = \Phi(x)$,

$D(F_n(x)) = \frac{\Phi(x)[1-\Phi(x)]}{n}$

Znamená to, že M_n je nestranným odhadem μ, S_n^2 je nestranným odhadem σ^2, pro libovolné, ale pevně dané $x \in R$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$.

Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů μ,

$\{S_n^2\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů σ^2,

pro libovolné, ale pevně dané $x \in R$ je $\{F_n(x)\}_{n=1}^{\infty}$ posloupnost konzistentních odhadů $\Phi(x)$.

b) Případ $r \geq 2$ stochasticky nezávislých náhodných výběrů: Nechť $X_{11},\ldots,X_{1n_1}$, ..., $X_{r1},\ldots,X_{rn_r}$ je r stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2$, ..., $n_r \geq 2$ z rozložení se středními hodnotami μ_1, ..., μ_r a rozptylem σ^2. Celkový rozsah je $n = \sum_{j=1}^{r} n_j$. Nechť c_1, ..., c_r jsou reálné konstanty, aspoň jedna nenulová. Pak pro libovolné hodnoty parametrů μ_1, …, μ_r a σ^2 platí:

$$E\left(\sum_{j=1}^{r} c_j M_j\right) = \sum_{j=1}^{r} c_j \mu_j ,$$

$E(S_*^2) = \sigma^2$.

Znamená to, že lineární kombinace výběrových průměrů $\sum_{j=1}^{r} c_j M_j$ je nestranným odhadem lineární kombinace středních hodnot $\sum_{j=1}^{r} c_j \mu_j$ a vážený průměr výběrových rozptylů $S_*^2 = \dfrac{\sum_{j=1}^{r} (n_j - 1) S_j^2}{n - r}$ je nestranným odhadem rozptylu σ^2.

c) Případ jednoho náhodného výběru z dvourozměrného rozložení: Nechť (X_1,Y_1), ..., (X_n,Y_n) je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Pak pro libovolné hodnoty parametrů σ_{12} a ρ platí:
$E(S_{12}) = \sigma_{12}$,
$E(R_{12}) \approx \rho$ (shoda je vyhovující pro $n \geq 30$).

Znamená to, že výběrová kovariance S_{12} je nestranným odhadem kovariance σ_{12}, avšak výběrový koeficient korelace R_{12} je vychýleným odhadem koeficientu korelace ρ.

Číselné charakteristiky diskrétních a spojitých náhodných veličin aspoň intervalového typu

Charakteristika polohy: střední hodnota E(X) – číslo, které charakterizuje polohu realizací náhodné veličiny na číselné ose s přihlédnutím k jejich pravděpodobnostem.

Diskrétní případ: náhodná veličina X má pravděpodobnostní funkci $\pi(x)$.

Střední hodnota $E(X) = \sum_{x=-\infty}^{\infty} x\pi(x)$, pokud je suma vpravo konečná.

Fyzikální význam: střední hodnota je těžiště soustavy hmotných bodů, jejichž celková hmotnost je 1 a bod o souřadnici x má hmotnost $\pi(x)$.

Spojitý případ: náhodná veličina X má hustotu pravděpodobnosti $\varphi(x)$.

Střední hodnota $E(X) = \int_{-\infty}^{\infty} x\varphi(x)dx$, pokud je integrál vpravo konečný.

Fyzikální význam: střední hodnota je těžiště hmotné přímky, jejíž celková hmotnost je 1 a hmota je na přímce rozprostřena podle předpisu φ (x).

Centrovaná náhodná veličina: Y = X - E(X).
(Pro náhodnou veličinu Y platí: E(Y) = 0.)

Charakteristika variability: rozptyl D(X) - číslo, které charakterizuje proměnlivost realizací náhodné veličiny kolem její střední hodnoty s přihlédnutím k jejich pravděpodobnostem.

Definiční vzorec: $D(X) = E\left([X - E(X)]^2\right)$ (rozptyl je střední hodnota kvadrátu centrované náhodné veličiny).

Výpočetní vzorec: $D(X) = E(X^2) - [E(X)]^2$ (rozptyl je střední hodnota kvadrátu mínus kvadrát středních hodnot).

$$D(X) = \begin{cases} \sum_{x=-\infty}^{\infty} x^2 \pi(x) - \left[\sum_{x=-\infty}^{\infty} x\pi(x)\right]^2 \\ \int_{-\infty}^{\infty} x^2 \varphi(x)dx - \left[\int_{-\infty}^{\infty} x\varphi(x)dx\right]^2 \end{cases}$$

Směrodatná odchylka $\sqrt{D(X)}$ - vyjadřuje průměrnou variabilitu realizací náhodné veličiny X kolem její střední hodnoty.

Standardizovaná náhodná veličina: $Z = \frac{X - E(X)}{\sqrt{D(X)}}$

(Pro náhodnou veličinu Z platí: E(Z) = 0, D(Z) = 1.)

Charakteristika společné variability: kovariance $C(X_1, X_2)$ – číslo, které charakterizuje variabilitu realizací dvou náhodných veličin X_1, X_2 kolem jejich středních hodnot s přihlédnutím k pravděpodobnostem těchto realizací.

Definiční vzorec: $C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)])$ (kovariance je střední hodnota součinu centrovaných náhodných veličin).

Výpočetní vzorec: $C(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$ (kovariance je střední hodnota součinu mínus součin středních hodnot).

$$C(X_1, X_2) = \begin{cases} \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1 x_2 \pi(x_1, x_2) - \sum_{x_1=-\infty}^{\infty} x_1 \pi_1(x_1) \sum_{x_2=-\infty}^{\infty} x_2 \pi_2(x_2) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 \varphi(x_1, x_2) dx_1 dx_2 - \int_{-\infty}^{\infty} x_1 \varphi_1(x_1) dx_1 \int_{-\infty}^{\infty} x_2 \varphi_2(x_2) dx_2 \end{cases}$$

Význam kovariance: Je-li kovariance kladná (záporná), pak to svědčí o existenci jistého stupně přímé (nepřímé) lineární závislosti mezi realizacemi náhodných veličin X_1, X_2. Je-li kovariance nulová, pak říkáme, že náhodné veličiny X_1, X_2 jsou nekorelované a znamená to, že mezi jejich realizacemi není žádný lineární vztah. Pozor – z nekorelovanosti nevyplývá stochastická nezávislost, zatímco ze stochastické nezávislosti plyne nekorelovanost.

Vlastnosti střední hodnoty

a) $E(a) = a$

b) $E(a + bX) = a + bE(X)$

c) $E(X - E(X)) = 0$

d) $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$

e) Jsou-li náhodné veličiny $X_1, ..., X_n$ stochasticky nezávislé, pak $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$

Vlastnosti kovariance

a) $C(a_1, X_2) = C(X_1, a_2) = C(a_1, a_2) = 0$

b) $C(a_1 + b_1X_1, a_2 + b_2X_2) = b_1b_2C(X_1, X_2)$

c) $C(X, X) = D(X)$

d) $C(X_1, X_2) = C(X_2, X_1)$

e) $C(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$

f) $C\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} C(X_i, Y_j)$

Vlastnosti rozptylu

a) $D(a) = 0$

b) $D(a + bX) = b^2 D(X)$

c) $D(X) = E(X^2) - [E(X)]^2$

d) $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C(X_i, X_j)$ (jsou-li náhodné veličiny X_1, ..., X_n nekorelované, pak $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$)

Vlastnosti koeficientu korelace

a) $R(a_1, X_2) = R(X_1, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X_1, a_2 + b_2X_2) = \mathrm{sgn}(b_1b_2)\, R(X_1, X_2)$

c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X_1, X_2) = R(X_2, X_1)$

e) $R(X_1, X_2) = \begin{cases} \dfrac{C(X_1, X_2]}{\sqrt{D(X_1)}\sqrt{D(X_2)}} & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0 \\ 0 \text{ jinak} & \end{cases}$

f) $|R(X_1, X_2)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X_1, X_2 existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a_1, a_2 tak, že $P(X_2 = a_1 + a_2X_1) = 1$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)