

ŠTATISTIKA

Základné pojmy

Definícia 1.

Náhodným výberom rozsahu n zo základného súboru s rozdelením pravdepodobnosti náhodnej premennej X budeme nazývať náhodný vektor $(X_1, X_2, \dots, X_n)$ s vlastnosťami

1. $X_1, X_2, \dots, X_n$ sú nezávislé náhodné premenné,
2. pre všetky $i = 1, 2, \dots, n$ X_i má rovnaké rozdelenie ako X .

Definícia 2.

Ľubovoľnú náhodnú premennú, ktorá je transformáciou náhodného výberu nazývame *štatistikou*.

Podľa účelu použitia (v teórii odhadu resp. v testovaní hypotéz) hovoríme o *odhadovej štatistike*, resp. o *testovacej štatistike*.

Definícia 3.

Nech $T_n = T(X_1, X_2, \dots, X_n)$ je odhadová štatistika pre funkciu $\tau(\theta)$ parametra θ . Hovoríme, že odhadová štatistika T_n je *nevychýlená*, ak $E(T_n) = \tau(\theta)$. Ak $\lim_{n \rightarrow \infty} E(T_n) = \tau(\theta)$, nazývame štatistiku T_n *asymptoticky nevychýlenou*.

Číslo

$$MSE(T_n) = E((T_n - \tau(\theta))^2)$$

nazývame *strednou kvadratickou chybou* odhadovej štatistiky T_n .

Hovoríme, že odhadová štatistika T_n je *konzistentná*, ak

$$T_n \xrightarrow{P} \tau(\theta)$$

Veta 1.

Ak T_n je nevychýlená odhadová štatistika a $\lim_{n \rightarrow \infty} \text{var}(T_n) = 0$, potom odhadová štatistika T_n je konzistentná.

Testovanie hypotéz

Definícia 4.

Akkoľvek tvrdenie o rozdelení pravdepodobnosti sa nazýva *štatistická hypotéza*. Praktický problém sa formuluje ako dvojica štatistických hypotéz, ktoré nemôžu byť súčasne pravdivé. Jedna z nich sa nazýva *nulová* (H_0) a druhá *alternatívna* (H_1).

Definícia 5.

Test hypotézy je pravidlo, ktoré určuje, kedy nulovú hypotézu zamietneme. Spravidla je založený na *kritickej oblasti*, t. j. na podmnožine množiny $\mathbb{R}^n$ definovanej prostredníctvom testovacej štatistiky. Ak realizácia náhodného výberu patrí do kritickej oblasti, zamietneme nulovú hypotézu.

Definícia 6.

Ak zamietneme nulovú hypotézu a nulová hypotéza platí, hovoríme o *chybe prvého druhu*. Pravdepodobnosť chyby prvého druhu nazývame *hladina významnosti testu* a označujeme α . Ak nulovú hypotézu prijmeme, a ona neplatí, hovoríme o *chybe druhého druhu* a jej pravdepodobnosť označujeme β . Hodnotu $1 - \beta$ nazývame *sila testu*. Ak sila testu je funkciou $1 - \beta(\theta)$ testovaného parametra θ rozdelenia pravdepodobnosti, nazývame ju *silofunkcia* testu.

Niektoré dôležité neparametrické testy

Chí–kvadrát test dobrej zhody rozdelenia pravdepodobnosti

H_0 : X má rozdelenie pravdepodobnosti
určené distribučnou funkciou F

H_1 : X nemá rozdelenie pravdepodobnosti
určené distribučnou funkciou F

Testovacia štatistika:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

kde f_i sú výberové početnosti a e_i sú hypotetické početnosti vypočítané pomocou parametrov odhadnutých na základe momentových štatistík náhodného výberu.

Kritická oblasť na hladine významnosti α :

$$W = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n: \chi^2 > q(1-\alpha, k-p-1)\}$$

kde p je počet parametrov hypotetického rozdelenia odhadovaných z realizácie náhodného výberu a $q(1-\alpha, k-p-1)$ je $(1-\alpha)$ -kvantil chí-kvadrát rozdelenia s $k-p-1$ stupňami voľnosti.

Chí-kvadrát test dobrej zhody nezávislosti zložiek náhodného vektora (X, Y)

H_0 : X, Y sú nezávislé

H_1 : X, Y sú závislé

Testovacia štatistika:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$$

kde $f_{i,j}$ sú výberové početnosti a $e_{i,j}$ sú hypotetické početnosti vypočítané pomocou vzorca $e_{i,j} = f_{i,X} f_{Y,j} / n$, kde $f_{Y,j} = \sum_{i=1}^k f_{i,j}$ a $f_{i,X} = \sum_{j=1}^l f_{i,j}$.

Kritická oblasť na hladine významnosti α :

$$W = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n: \chi^2 > q(1-\alpha, (k-1)(l-1))\}$$

kde $q(1-\alpha, (k-1)(l-1))$ je $(1-\alpha)$ -kvantil chí-kvadrát rozdelenia s $(k-1)(l-1)$ stupňami voľnosti.

Kolmogorovov test zhody rozdelenia pravdepodobnosti

H_0 : X má rozdelenie pravdepodobnosti
určené distribučnou funkciou F

H_1 : X nemá rozdelenie pravdepodobnosti
určené distribučnou funkciou F

Testovacia štatistika:

$$D = \max_{i=1,2,\dots,n} \{\max\{|F(i) - i/n|, |F(i) - (i-1)/n|\}\}$$

Kritická oblasť na hladine významnosti α :

$$W = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n: D > d(1-\alpha, n)\}$$

kde $d(1-\alpha, n)$ je $(1-\alpha)$ -kvantil testovacej štatistiky D (tabuľková hodnota).

Tabuľka najpoužívanějších kvantilov testovacej štatistiky D .

$n \backslash 1-\alpha$	0,90	0,95	0,99
10	0,37	0,41	0,49
20	0,26	0,29	0,36
30	0,22	0,24	0,29
40	0,19	0,21	0,25
50	0,17	0,19	0,23
$< n$	$1,22/\sqrt{n}$	$1,36/\sqrt{n}$	$1,63/\sqrt{n}$