Непараметричні методи в бізнес-аналізі

Розглянемо скінченну послідовність незалежних однаково розподілених числових випадкових величин, тобто вибірку $\xi_1, \xi_2, \ldots, \xi_n$, де n – обсяг вибірки, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} X(x - \xi_i)$ - функція розподілу елементів вибірки,

де $X(u) = 1$ при $x > 0$ та $X(u) = 0$ при $x \leq 0$. Емпірична функція розподілу $F_n(x)$ є обґрунтованою оцінкою функції розподілу $F(x)$, тобто функції розподілу елементів вибірки.

Справедливою є теорема В.І. Глівенко:

для будь-якої неперервної функції розподілу *F(x)* при n->∞

емпірична функція розподілу $F_n(x)$ рівномірно сходиться до неї:

$$\sup_x \left| F_n(x) - F_n \right| \to 0$$

(сходимість за ймовірністю)

Статистичні тести дозволяють зробити вибір між двома конкуруючими гіпотезами:

H0 (нульова гіпотеза) полягає в тому, що нічого не відбулося (ви отримали очікувану відповідь, середнє не змінилось, модель не покращилась тощо);

H1 (альтернативна гіпотеза) полягає в тому, що щось відбулося (ви отримали неочікувану відповідь, середнє значення зросло, модель підходить краще тощо).

Алгоритм проведення статистичного тесту:

1. припустимо, що нульова гіпотеза є вірною;
2. підрахуємо тестову статистику (наприклад, середнє значення вибірки);
3. виходячи зі статистики та її розподілу, можемо розрахувати p-значення, ймовірність значення тестової статистики як екстремального або більш екстремального, ніж те що спостерігається, за умови вірності нульової гіпотези;
4. якщо p-значення занадто мале, то маємо докази проти нульової гіпотези (***спростування нульової гіпотези***);
5. Якщо p-значення велике, то не маємо таких доказів (***неможливість спростувати нульову гіпотезу***).

Зазвичай p-значення менше 0,05 вказує на те, що змінні, ймовірно, не є незалежними, у той час як p-значення, яке перевищує 0,05, не може надати таких доказів. Це лаконічний спосіб сказати:

- нульова гіпотеза полягає в тому, що змінні є незалежними;
- альтернативна гіпотеза полягає в тому, що змінні не є незалежними;
- у випадку рівня значущості $\alpha=0,05$, якщо $p<0,05$, то спростовується нульова гіпотеза, даючи докази того, що змінні не є незалежними; якщо $p>0,05$, то не можливо її спростувати;
- можна обрати власне значення α, в цьому випадку рішення про спростування або неможливість зробити це будуть відрізнятися.

Критерії згоди: використовуються для оцінки близькості емпіричного розподілу до теоретичного нормального розподілу.

Критерії згоди:

- критерій згоди Пірсона;
- критерій В. І. Романовського;
- критерій А. М. Колмогорова;
- критерій Б. С. Ястремського.

Критерій згоди Пірсона:

ґрунтується на визначенні величини χ^2, яка визначається як сума квадратів різниць емпіричних та теоретичних частот, віднесених до теоретичних частот, тобто

$$\chi^2 = \sum \frac{(m - m')^2}{m'},$$

m - емпіричні частоти;

m' - теоретичні частоти.

Для оцінки того, наскільки даний емпіричний розподіл відображається нормальним розподілом, розраховують за розподілом Пірсона ймовірності досягнення χ^2 даного значення $P(\chi^2)$.

$P(\chi^2)$ - табличні значення.

У таблиці по рядках значення χ^2 , а по стовпцях значення ступенів волі (k) варіювання емпіричного розподілу ($k=n-s-1$).

n- число груп у сукупності;

s – число статистичних характеристик, які використовуються для розрахунку теоретичного розподілу.

Критерій згоди В. І. Романовського:

ґрунтується на обчисленні відношення

$$\frac{\chi^2 - k}{\sqrt{2k}},$$

k - число ступенів волі.

якщо абсолютне значення відношення менше 3, то пропонується розбіжність між теоретичним та емпіричним розподілами вважати несуттєвим (можливість прийняти нормальний розподіл як емпіричний);

Якщо відношення більше 3, то розбіжність суттєва.

Критерій згоди А. М. Колмогорова:

Встановлює близькість теоретичних та емпіричних розподілів шляхом порівняння інтегральних розподілів та розраховується як

$$\lambda = \frac{D}{\sqrt{N}},$$

D - максимальна межа різниці: накопичених теоретичних та накопичених емпіричних частот;.

Критерій згоди А. М. Колмогорова:

$p(\lambda)$ – ймовірність того, що λ досягне даної величини.

Якщо знайденому значенню λ відповідає дуже мала ймовірність $p(\lambda)$, то розбіжність між емпіричним та теоретичним розподілом неможливо вважати випадковою, й таким чином, перше мало відображає друге. Навпаки, якщо $p(\lambda)$ – значна величина (більше 0,05), то розбіжність між частотами може бути випадковою й розподіли добре відповідають один одному.

Критерій згоди Б. С. Ястремського:

У загальному вигляді можна записати нерівністю

$$I \le 3\sqrt{2n + 4\Theta},$$

$$I = \left| C - n \right|,$$

$$C = \sum \frac{(m - m')^2}{m'q},$$

m - емпіричні частоти;

m' - теоретичні частоти;

n - число груп.

Критерій згоди Б. С. Ястремського:

Для числа груп, менших 20:

$$\Theta = 0,6,$$

$$q = 1 - p$$

$I \le 3\sqrt{2n + 4\Theta}$ - показує несуттєвість розбіжності між емпіричними та теоретичними частотами;

$I > 3\sqrt{2n + 4\Theta}$ - розбіжності між емпіричними та теоретичними частотами суттєві

Елементарні прийоми визначення “нормальності” розподілу:

- числа Вестергарда (0,3; 0,7; 1,1; 3);
- порівняння середньої арифметичної, моди, медіани.

Елементарні прийоми визначення “нормальності” розподілу: числа Вестергарда

- визначення основних характеристик (середнє арифметичне $\bar{x}$, середнє квадратичне відхилення σ.
- емпіричний розподіл відповідає нормальному розподілу, якщо:

$\bar{x}-0,3\sigma$	1/4 всієї сукупності	$\bar{x}+0,3\sigma$
$\bar{x}-0,7\sigma$	1/2 всієї сукупності	$\bar{x}+0,7\sigma$
$\bar{x}-1,1\sigma$	¾ всієї сукупності	$\bar{x}+1,1\sigma$
–3	0,998 всієї сукупності	3