

Критерії згоди

Лекція № 10



10.1. Критерій χ^2 (Пірсона) перевірки
непараметричної гіпотези.

10.2. Критерій χ^2 (Пірсона) перевірки параметричної
гіпотези

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

Нехай ξ_1, \dots, ξ_n є вибіркою об'єма n із розподілу \mathcal{F} , який має функцію розподілу F . Припустимо, що перевіряється гіпотеза

$$H_0 : F(x) = F_0(x) \text{ проти альтернативи } H_1 : F(x) \neq F_0(x).$$

Критерії для перевірки гіпотез щодо розподілу генеральної сукупності називаються критеріями згоди (англ. "Goodness of fit criteria").

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

Нехай ймовірнісний простір Ω розбито на події E_1, \dots, E_d такі, що $\bigcap_{k=1}^d E_k = \Omega$ та $E_i \cap E_j = \emptyset$ $i \neq j$.

Зауваження 1

При виборі розбиття ймовірнісного простору на події слід дотримуватись наступного:

- у випадку дискретного розподілу з ймовірностями (p_1, \dots, p_d) , розбиття обирається так, щоб $P\{E_k\} = p_k$, $k = \overline{1, d}$;
- у випадку неперервного розподілу, область визначення ф.р. F_0 розбивається на d інтервалів Δ_k , для яких $P\{\Delta_k\} = p_k$, $k = \overline{1, d}$.

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

Позначимо, через $p_k = F_0(E_k) = P\{\xi_1 \in E_k | H_0\}$, тобто ймовірність попадання в E_k , за умови виконання гіпотези H_0 . Очевидно,

$$p_1 + \dots + p_d = 1.$$

Нехай також

$$v_k = \sum_{j=1}^n 1(\xi_j \in E_k),$$

є кількістю ξ_j , $j = \overline{1, n}$, що попали в E_k , $k = \overline{1, d}$.

Складемо статистику

$$\chi^2 = \sum_{k=1}^d \frac{(v_k - np_k)^2}{np_k} = n \sum_{k=1}^d \left(\frac{v_k}{n} - p_k \right)^2 \frac{1}{p_k}.$$

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

Оскільки частота $\frac{v_k}{n}$ є незміщеною та консистентною оцінкою $F(E_k)$, тобто

$$E \frac{v_k}{n} = F(E_k) = P\{\xi_1 \in E_k\}, \quad \frac{v_k}{n} \xrightarrow{P} F(E_k),$$

то, при виконанні гіпотези H_0 , для всіх $k = \overline{1, d}$

$$F(E_k) = F_0(E_k) = p_k$$

і, як наслідок,

$$\frac{v_k}{n} - p_k = \frac{v_k}{n} - F_0(E_k) \xrightarrow{P} 0, \quad k = \overline{1, d}.$$

При цьому можна показати, що

$$\chi^2 \Rightarrow \chi_{d-1}^2, \quad n \rightarrow \infty.$$

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

Якщо ж H_0 не є вірною, то

$$\frac{v_k}{n} - p_k \xrightarrow{P} F(E_k) - F_0(E_k), \quad k = \overline{1, d},$$

і хоча б для одного $k = \overline{1, d}$, незалежно від розбиття вибіркового простору Ω ,

$$F(E_k) - F_0(E_k) > 0.$$

Останнє означає, що

$$\chi^2 = n \sum_{k=1}^d \left(\frac{v_k}{n} - p_k \right)^2 \frac{1}{p_k} \xrightarrow{P} +\infty, \quad n \rightarrow \infty.$$

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

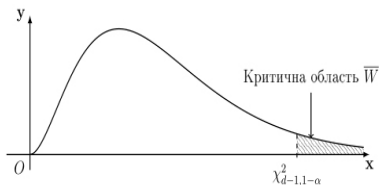
Таким чином, межі, які відокремлюють великі значення статистики χ^2 від малих, знаходяться на підставі того, що при великих n розподіл статистики χ^2 мало відрізняється від розподілу χ^2_{d-1} , за умови вірності гіпотези H_0 , і, у разі вірності альтернативної гіпотези H_1 , χ^2 буде набувати великих значень.

10.1. Критерій χ^2 (Пірсона) перевірки непараметричної гіпотези.

Нехай задано рівень значущості α . Тоді критична область критерію χ^2

$$(\chi^2)^* > \chi_{d-1, 1-\alpha}^2,$$

де $\chi_{d-1, 1-\alpha}^2$ – квантиль розподілу χ_{d-1}^2 рівня $1 - \alpha$, при попаданні в яку відхиляється нульова гіпотеза H_0 , і не відхиляється в протилежному випадку.



10.2. Критерій χ^2 (Пірсона) перевірки параметричної гіпотези

Нехай ξ_1, \dots, ξ_n є вибіркою об'єма n із невідомого розподілу \mathcal{F} , який має функцію розподілу F . Нехай перевіряється гіпотеза

$$H_0 : F(x) = F_{\theta}(x), \theta \in \Theta \subset \mathbb{R}^q,$$

проти альтернативи

$$H_1 : F(x) \neq F_{\theta}(x),$$

для всіх $\theta \in \Theta$, де θ – невідомий параметр, q – його розмірність.

10.2. Критерій χ^2 (Пірсона) перевірки параметричної гіпотези

Нехай ймовірнісний простір Ω розбито на події

$$E_1, \dots, E_d,$$

як це вказано у попередньому пункті з $d > q + 1$.

На відміну від попереднього пункту

$$p_k = P\{E_k|H_0\} = p_k(\theta)$$

залежить від невідомого параметра, а це означає, що від параметра залежить також

$$\chi^2(\vec{\xi}, \theta) = \sum_{k=1}^d \frac{(v_k - np_k(\theta))^2}{np_k(\theta)}.$$

10.2. Критерій χ^2 (Пірсона) перевірки параметричної гіпотези

В цьому випадку природно діяти наступним чином: у якості значень невідомих параметрів

$$\theta = (\theta_1, \dots, \theta_q)$$

взяти їх оцінки

$$\theta^* = (\theta_1^*, \dots, \theta_q^*),$$

які знаходяться за вибіркою ξ_1, \dots, ξ_n , а у якості гіпотетичного розподілу розглянути $F_{\theta^*}(x)$. Статистика критерію при цьому, буде мати вигляд

$$\chi^2(\vec{\xi}, \theta^*) = \sum_{k=1}^d \frac{(v_k - np_k(\theta^*))^2}{np_k(\theta^*)}.$$

10.2. Критерій χ^2 (Пірсона) перевірки параметричної гіпотези

Р.Фішером було встановлено, що у разі вірності нульової гіпотези H_0 і виборі у якості оцінки $\theta^* = \hat{\theta}$ – оцінки максимальної вірогідності,

$$\chi^2(\vec{\xi}, \theta^*) \Rightarrow \chi^2_{d-q-1}, n \rightarrow \infty.$$

Таким чином, при оцінюванні параметрів методом максимальної вірогідності, критерій Пірсона можна застосовувати у наступному вигляді.

Нехай задано рівень значущості α . Тоді критична область критерію χ^2

$$\chi^2(\vec{x}, \theta^*) > \chi^2_{d-q-1, 1-\alpha},$$

де $\vec{x} = (x_1, \dots, x_n)$ – реалізація вибірки, $\chi^2_{d-q-1, 1-\alpha}$ – квантиль розподілу χ^2_{d-q-1} рівня $1 - \alpha$, при попаданні в яку відхиляється нульова гіпотеза H_0 , і не відхиляється в протилежному випадку.

10.2. Критерій χ^2 (Пірсона) перевірки параметричної гіпотези

Зауваження 2

Для того, щоб критерій Пірсона можна було коректно застосовувати необхідно вимагати, щоб для всіх d подій E_k розбиття виконувалась нерівність:

$$np_k \geq 10, \quad k = \overline{1, d}. \quad (1)$$

Якщо ж для деяких E_k значення

$$np_k < 10,$$

то такі E_k об'єднують з іншими подіями так, щоб для нового розбиття \tilde{E}_k , $k = \overline{1, d}$, нерівності (1) мали місце.

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

Нехай $\vec{\xi} = (\xi_1, \dots, \xi_n)$ є вибіркою об'єма n із розподілу \mathcal{F} , який має неперервну ф.р. F . Перевіряється проста гіпотеза

$H_0 : F(x) = F_0(x)$ проти складної альтернативи $H_1 : F(x) \neq F_0(x)$.

Побудуємо статистику

$$\mathfrak{D}_n = \sqrt{n} \sup_{y \in \mathbb{R}} |F_n^*(y) - F_0(y)|,$$

де $F_n^*(y)$ – емпірична функція розподілу, яку побудовано за вибіркою $\vec{\xi}$.

У разі вірності гіпотези H_0 , згідно з теоремою Колмогорова

$$\mathfrak{D}_n \Rightarrow \eta, \quad n \rightarrow \infty,$$

де η має розподіл Колмогорова.

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

Зауважимо, що у разі невірності гіпотези H_0

$$\sup_{y \in \mathbb{R}} |F_n^*(y) - F_0(y)| \xrightarrow{P} 0, \quad n \rightarrow \infty,$$

а тому

$$\mathfrak{D}_n \xrightarrow{P} \infty, \quad n \rightarrow \infty.$$

Таким чином, межі, які відокремлюють великі значення статистики \mathfrak{D}_n від малих, знаходяться на підставі того, що при великих n розподіл статистики \mathfrak{D}_n мало відрізняється від розподілу Колмогорова, за умови вірності гіпотези H_0 , і, у разі вірності альтернативної гіпотези H_1 , \mathfrak{D}_n буде набувати великих значень.

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

Знайдемо більш зручну для практичного застосування формулу обчислення статистики \mathfrak{D}_n . Нехай $\vec{x} = (x_1, \dots, x_n)$ є реалізацією вибірки. На її основі будуюмо варіаційний ряд

$$X_{(1)}, \dots, X_{(n)},$$

який доповнюємо $X_{(0)} = -\infty$ та $X_{(n+1)} = +\infty$. Тоді емпірична ф.р. можна записати як

$$F_n^*(y) = \frac{k}{n}, \quad X_{(k)} < y \leq X_{(k+1)}, \quad k = \overline{0, n},$$

причому

$$F_n^*(X_{(k)}) = \frac{k-1}{n}, \quad k = \overline{0, n+1}.$$

Помітимо, що тоді

$$\mathfrak{D}_n^* = \sqrt{n} \max_{1 \leq k \leq n} \left(\left| F_0(X_{(k)}) - \frac{k-1}{n} \right|; \left| F_0(X_{(k)}) - \frac{k}{n} \right| \right).$$

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

Нехай задано рівень значущості α . Тоді критичну область критерію Колмогорова можна визначати наступним чином:

- Якщо $n < 100$, тоді у якості критичного рівня критерію доцільно обирати значення $\sqrt{n}\varepsilon_{\alpha,n}$, де $\varepsilon_{\alpha,n}$ визначається як верхня межа розподілу відхилень $\sup_{y \in \mathbb{R}} |F_n^*(y) - F(y)|$ рівня α , тобто найменше значення $\varepsilon > 0$, для якого

$$P \left\{ \sup_{y \in \mathbb{R}} |F_n^*(y) - F(y)| \geq \varepsilon \right\} \leq \alpha.$$

В цьому випадку критична область критерію Колмогорова буде мати вигляд

$$\mathfrak{D}_n^* > \sqrt{n}\varepsilon_{\alpha,n}.$$

10.3. Критерій Колмогорова (Колмогорова-Смірнова)

- Якщо значення $n \geq 100$, тоді у якості критичного рівня критерію доцільно обирати значення λ_α – квантиль розподілу Колмогорова рівня $1 - \alpha$, тобто корінь рівняння

$$1 - K(\lambda) = \alpha,$$

де $K(\lambda)$ – ф.р. розподілу Колмогорова. В цьому випадку критична область критерію Колмогорова буде мати вигляд

$$\mathfrak{D}_n^* > \lambda_\alpha.$$

Дякуємо за увагу!