15

Критерій χ-квадрат Пірсона

Нехай $\xi \in F$ і перевіряються гіпотези $H_1: F = F_1$ проти $H_2: F \neq F_1$.

Задача полягає в побудові асимптотичного критерію рівня $1-\varepsilon$. Припустимо, що $\xi \in F_1$, розіб'ємо область можливих значень ξ_1 на деяку кількість проміжків, які не перетинаються:

$$P_1(\xi_1 \in \Delta_1 \cup \ldots \cup \Delta_k) = 1,$$

де Δ_i мають вигляд $\Delta_i = [a_i; b_i)$, $i = 1, \ldots, k$.

Нехай ν_i – кількість спостережень, які потрапили в Δ_i, $i = 1, \ldots, k$, $\nu_1 + \ldots + \nu_k = n$.

Позначимо також

$$p_i = P_1(\xi_1 \in \Delta_i) = F_1(b_i) - F_1(a_i),\ i = 1, \ldots, k.$$

Із закону великих чисел випливає, що

$$\frac{\nu_i}{n} \xrightarrow{P} p_i,\ n \to \infty,$$

за кожного i, якщо правильна гіпотеза H_1.

За міру близькості сукупностей $\left\{\frac{\nu_1}{n}, \ldots, \frac{\nu_k}{n}\right\}$ та $\{p_1, \ldots, p_k\}$ пропонується застосувати величину

$$\Psi_n = n \sum_{i=1}^{k} \frac{1}{p_i}\left(\frac{\nu_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}.$$

<u>Теорема Пірсона.</u> Якщо $0 < p_i < 1$ за всіх $i = 1, \ldots, k$, то для будь-якого $y > 0$ $P_1(\Psi_n < y) \to \chi^2_{k-1}(y)$, $n \to \infty$.

Доведення цієї теореми досить складне, тому ми його не наводимо.

Приступимо до побудови критерію.

Знайдемо число q таке, що $\chi^2_{k-1}(q) = 1-\varepsilon$. Якщо правильна гіпотеза H_1, то з ймовірністю, близькою до $1-\varepsilon$, значення випадкової величини Ψ_n має бути менше q. Тому ми відкидаємо гіпотезу, якщо $\Psi_n \geq q$, і приймаємо її у протилежному випадку. Це значить, що ми приймаємо H_1, якщо відсутні явні протиріччя цієї гіпотези зі значеннями, що спостерігаються.

Критична множина має такий вигляд:

$$K=\{(\xi_1,\dots,\xi_n) : \Psi_n \geq q\}.$$

Для ймовірності помилки першого роду маємо

$$\beta_1 = P_1(\Psi_n \geq q) = 1 - P_1(\Psi_n < q) \simeq$$

$$\simeq 1 - \chi^2_{k-1}(q) = \varepsilon.$$

Зауваження. Наближення $P_1(\Psi_n < q) \simeq \chi^2_{k-1}(q)$ достатньо задовільне для практичного застосування, якщо $np_i \geq 10$ для всіх i. В протилежному випадку слід об'єднати два сусідніх інтервали в один.

Побудова критерію за допомогою довірчого інтервалу.

Припустимо, що $\xi \in F_\theta$, $\theta \in \mathbb{R}$ – невідомий параметр. Задача полягає у перевірці основної гіпотези $H_1 : \theta = \theta_1$ проти $H_2 : \theta \neq \theta_1$.

Якщо існує довірчий інтервал для θ рівня $1-\varepsilon$ (точний або асимптотичний), то за його допомогою можливо побудувати критерій узгодження (також точний або асимптотичний) рівня $1-\varepsilon$. Дійсно, якщо за всіх значень θ

$$P_\theta(A(\xi_1, \dots, \xi_n) < \theta < B(\xi_1, \dots, \xi_n)) \geq 1-\varepsilon,$$

то й коли $\theta = \theta_1$ має бути

$$P_{\theta_1}(A(\xi_1, \dots, \xi_n) < \theta_1 < B(\xi_1, \dots, \xi_n)) \geq 1-\varepsilon.$$

Тому ми відхиляємо H_1, якщо $\theta_1 \notin (A(\xi_1, \dots, \xi_n), B(\xi_1, \dots, \xi_n))$, через те, що така подія має малу ймовірність (не більше ε) за умови справедливості H_1. Критична множина має такий вигляд:

$$K = \{(\xi_1, \dots, \xi_n) : \theta_1 \notin (A(\xi_1, \dots, \xi_n), B(\xi_1, \dots, \xi_n))\}.$$

Перевірка гіпотез у випадку двох вибірок

Припустимо, що здійснені дві серії незалежних випробувань, за результатами яких отримані дві незалежні вибірки

$$\xi = (\xi_1, \dots, \xi_n) \in F$$

і

$$\eta = (\eta_1, \dots, \eta_n) \in G.$$

Найчастіше перевіряється основна гіпотеза про співпадання розподілів $F = G$. В цьому випадку критерії називаються критеріями однорідності. В інших ситуаціях перевіряється гіпотеза про співпадання лише деяких параметрів розподілів F і G. Розглянемо такі задачі.

Попередньо зауважимо, що тепер ми маємо $n+m$ спостережень, отже, вибірковим простором буде $\mathbb{R}^{n+m}$ і критична множина K буде $n+m$ вимірною.

Нехай спочатку

$$\xi = (\xi_1, \dots, \xi_n) \in \Phi_{a_1, \sigma_1^2},$$

$$\eta = (\eta_1, \dots, \eta_m) \in \Phi_{a_2, \sigma_2^2}.$$

Всі чотири параметри невідомі.

1. **Перевірка гіпотез про співпадання дисперсій.**

Перевіряємо основну гіпотезу $H_1 : \sigma_1^2 = \sigma_2^2$ проти $H_2 : \sigma_1^2 \neq \sigma_2^2$. Виберемо мале число $\varepsilon > 0$, і нехай

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i, \quad S_\xi^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2,$$

$$\bar{\eta} = \frac{1}{m}\sum_{i=1}^{n} \eta_i, \quad S_\eta^2 = \frac{1}{m}\sum_{i=1}^{m}(\eta_i - \bar{\eta})^2.$$

На основі теореми про властивості вибірок із нормального розподілу

$$\frac{n S_\xi^2}{\sigma_1^2} \in \chi^2_{n-1}, \quad \frac{m S_\eta^2}{\sigma_2^2} \in \chi^2_{m-1},$$

ці випадкові величини незалежні, тому, що побудовані із незалежних вибірок. З цих випадкових величин можливо побудувати випадкову величину, що має розподіл Фішера:

$$\frac{1}{n-1}\frac{n S_\xi^2}{\sigma_1^2} : \frac{1}{m-1}\frac{m S_\eta^2}{\sigma_2^2} = \frac{n(m-1)\sigma_2^2 S_\xi^2}{m(n-1)\sigma_1^2 S_\eta^2} \in F_{n-1,m-1}.$$

Якщо правильна гіпотеза H_1, тобто $\sigma_1^2 = \sigma_2^2$, то

$$\eta = \frac{n(m-1) S_\xi^2}{m(n-1) S_\eta^2} \in F_{n-1,m-1}.$$

За допомогою таблиць розподілу $F_{n-1,m-1}$ і ε можливо знайти числа q_1 та q_2 такі, що

$$F_{n-1,m-1}(q_1) = \frac{\varepsilon}{2}, \quad F_{n-1,m-1}(q_2) = 1 - \frac{\varepsilon}{2}.$$

Тоді маємо

$$P_1(q_1<\xi<q_2)=F_{n-1,m-1}(q_2)-F_{n-1,m-1}(q_1)=$$
$$=1-\varepsilon.$$

Тому логічно відкидати H_1, якщо $\xi\notin(q_1,q_2)$, ймовірність цієї події точно дорівнює ε, якщо правильна гіпотеза H_1. Маємо таку критичну множину

$$K=\{(\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_m):\xi\notin(q_1,q_2)\}.$$

г. <u>Перевірка гіпотези про співпадання середніх.</u>

Перевірку гіпотези будемо проводити припускаючи, що дисперсії співпадають тобто:

$$\sigma_1^2=\sigma_2^2=\sigma^2. \quad \sigma^2 - \text{невідоме}.$$

Перевіряється гіпотеза $H_1: \alpha_1=\alpha_2$ проти гіпотези $H_2: \alpha_1\neq\alpha_2$.

Застосуємо розподіл Стьюдента. Через те, що $\bar\xi$ і $\bar\eta$ незалежні і

$$\bar\xi\in\Phi_{\alpha_1,\frac{\sigma^2}{n}},\quad \bar\eta\in\Phi_{\alpha_2,\frac{\sigma^2}{n}},$$

маємо
$$\bar\xi-\bar\eta\in\Phi_{\alpha_1-\alpha_2,\,\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)},$$

а після стандартизації

$$\frac{\bar\xi-\bar\eta-(\alpha_1-\alpha_2)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}}\in\Phi_{0,1}.$$

7

За властивістю розподілу χ-квадрат

$$\frac{nS_\xi^2}{\sigma^2}+\frac{mS_\eta^2}{\sigma^2}\in\chi^2_{n+m-2};$$

ця випадкова величина не залежить від різниці $\bar\xi-\bar\eta$. Таким чином,

$$\frac{\bar\xi-\bar\eta-(\alpha_1-\alpha_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}:\sqrt{\frac{1}{n+m-2}\frac{nS_\xi^2+mS_\eta^2}{\sigma^2}}\in T_{n+m-2}.$$

Якщо правильна гіпотеза H_1, то $\alpha_1-\alpha_2=0$ і

$$\psi=\frac{\bar\xi-\bar\eta}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{nS_\xi^2+mS_\eta^2}{n+m-2}}}\in T_{n+m-2}.$$

Із таблиць розподілу T_{n+m-2} знаходимо число q таке, що $T_{n+m-2}(-q)=\frac{\varepsilon}{2}$. Тоді маємо

$$P_1(-q<\psi<q)=T_{n+m-2}(q)-T_{n+m-2}(-q)=1-\varepsilon.$$

Отже, вибравши

$$K=\{(\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_m):|\psi|\ge q\},$$

матимемо

$$\beta_1=P_1((\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_m)\in K)=\varepsilon.$$

3. Критерій Колмогорова–Смирнова однорідності двох вибірок.

Нехай

$$\xi = (\xi_1, \ldots, \xi_n) \in F,$$

$$\eta = (\eta_1, \ldots, \eta_m) \in G,$$

де F і G – неперервні функції розподілу. Перевіряється гіпотеза $H_1: F = G$ проти гіпотези $H_2: F \neq G$. Побудуємо асимптотичний критерій рівня $1-\varepsilon$.

Нехай F_n^* і G_m^* – емпіричні функції розподілу, побудовані за вибірками ξ і η відповідно. Введемо

$$D_{n,m} = \sup_y \left| F_n^*(y) - G_m^*(y) \right|.$$

Якщо правильна гіпотеза H_1, то збільшенням об'ємів вибірок емпіричні функції розподілу збігаються за ймовірністю до загальної границі, тобто $D_{n,m} \xrightarrow{P} 0$.

Наступна теорема, визначає швидкість з якою це відбувається.

Теорема Колмогорова–Смирнова. Нехай правильною є гіпотеза H_1 і загальна функція розподілу вибірок неперервна. Тоді для будь-якого $y > 0$, коли $n \to \infty$, $m \to \infty$

$$P_1\left(\sqrt{\frac{nm}{n+m}}\, D_{n,m} < y\right) \to \mathcal{K}(y) =$$

$$= \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2y^2}.$$

Нехай q таке, що $\mathcal{K}(q) = 1 - \varepsilon$.

Визначимо

$$K = \left\{ (\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_m) : \sqrt{\frac{nm}{n+m}}\, D_{n,m} \geq q \right\},$$

тобто ми відкидаємо гіпотезу про однорідність, якщо розбіжність між двома емпіричними функціями розподілу досить велика.

Тоді за великих n

$$\beta_1 \geq P_1\left(\sqrt{\frac{nm}{n+m}}\, D_{n,m} \geq q\right) \simeq 1 - \mathcal{K}(q) = \varepsilon,$$