Лекція № 11

11.1. Критерій Смірнова (Колмогорова-Смірнова).
11.2. Критерій знаків.
11.3. Критерій знаків за наявності зв'язків.

Розглянемо задачу з двома вибірками.
Нехай

$$\vec{\xi} = (\xi_1, \ldots, \xi_{n_1})$$

є вибіркою об'єма n_1 із розподілу $\mathcal{F}$, який має функція розподілу F, а

$$\vec{\eta} = (\eta_1, \ldots, \eta_{n_2})$$

є вибіркою об'єма n_2 із розподілу $\mathcal{G}$, який має функція розподілу G.
Припустимо, що перевіряється гіпотеза

$$H_0 \ : \ F = G \text{ проти альтернативи } H_1 \ : \ F \neq G.$$

Критерії для перевірки гіпотез про те, що дві (або більше) вибірки взяті з одного ймовірністного розподілу називаються критеріями однорідності (англ. "Homogeneity Criteria").

11.1. Критерій Смірнова (Колмогорова-Смірнова).

Даний критерій можна використовувати лише у випадку неперервних функцій розподілу.

Нехай $F^*_{n_1}(y)$ та $G^*_{n_2}(y)$ – емпіричні функції розподілу, які побудовані за вибірками $\vec{\xi}$ та $\vec{\eta}$, відповідно.
У якості статистики критерію оберемо

$$\rho\left(\vec{\xi}, \vec{\eta}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{y \in \mathbb{R}} \left| F^*_n(y) - G^*_{n_2}(y) \right|. \tag{1}$$

11.1. Критерій Смірнова (Колмогорова-Смірнова).

Для статистики $\rho\left(\vec{\xi}, \vec{\eta}\right)$ є вірною теорема, аналогічна до теореми Колмогорова.

Теорема 1

Якщо вибірки $\xi_1, \ldots, \xi_{n_1}$ та $\eta_1, \ldots, \eta_{n_2}$ взято з одного розподілу $\mathcal{F}$, який має неперервну функцію розподілу F, то статистика $\rho\left(\vec{\xi}, \vec{\eta}\right)$, яка задається (1), слабко збігається до розподілу Колмогорова, тобто

$$\rho\left(\vec{\xi}, \vec{\eta}\right) \Rightarrow \zeta, \; n_1, n_2 \to \infty,$$

де в.в. ζ має розподіл Колмогорова з неперервною функцією роподілу.

11.1. Критерій Смірнова (Колмогорова-Смірнова).

Зауваження 1

В силу схожести принципів побудови статистики $\rho\left(\vec{\xi}, \vec{\eta}\right)$ критерію однорідності Смірнова та статистики $\mathfrak{D}_n^*$ критерію згоди Колмогорова, а також однакового граничного розподілу при виконанні основної гіпотези, далі ми наведемо лише основні моменти, які стосуються побудови критерію.

11.1. Критерій Смірнова (Колмогорова-Смірнова).

За аналогією, у разі вірності альтернативної гіпотези H_1,

$$\rho\left(\vec{\xi}, \vec{\eta}\right) \xrightarrow{P} \infty, \; n_1, n_2 \to \infty.$$

Знайдемо більш зручну для практичного застосування формулу обчислення статистики $\rho\left(\vec{\xi}, \vec{\eta}\right)$. Нехай $\vec{x} = (x_1, \ldots, x_n)$ є реалізацією вибірки $\vec{\xi}$, а $\vec{y} = (y_1, \ldots, y_n)$ є реалізацією вибірки $\vec{\eta}$. На їх основі будуємо два варіаційних ряди

$$\begin{aligned} -\infty = X_{(0)} &< X_{(1)} < \ldots < X_{(n_1)} < X_{(n_1+1)} = +\infty; \\ -\infty = Y_{(0)} &< Y_{(1)} < \ldots < Y_{(n_2)} < Y_{(n_2+1)} = +\infty. \end{aligned}$$

11.1. Критерій Смірнова (Колмогорова-Смірнова).

Можна показати, що статистика $\rho(\vec{x}, \vec{y})$, побудована за реалізаціями вибірок прийме вигляд

$$\rho(\vec{x}, \vec{y}) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_{1 \le k \le n_1} \left(\left| G^*_{n_2}(X_{(k)}) - \frac{k-1}{n_1} \right|; \left| G^*_{n_2}(X_{(k)}) - \frac{k}{n_1} \right| \right) =$$

$$= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_{1 \le i \le n_2} \left(\left| F^*_{n_1}(Y_{(i)}) - \frac{i-1}{n_2} \right|; \left| F^*_{n_1}(Y_{(i)}) - \frac{i}{n_2} \right| \right).$$

11.1. Критерій Смірнова (Колмогорова-Смірнова).

Нехай задано рівень значущості α. Тоді критичну область критерію Смірнова будується наступним чином:

- Якщо $n = n_1 + n_2 < 100$, критична область доцільно обрати

$$\rho(\vec{x}, \vec{y}) > \sqrt{n}\varepsilon_{\alpha,n},$$

де $\varepsilon_{\alpha,n}$ визначається так само, як у попередньому пункті.

- Якщо $n \geq 100$, тоді критична область матиме вигляд

$$\rho(\vec{x}, \vec{y}) > \lambda_\alpha,$$

де λ_α – квантиль розподілу Колмогорова рівня $1 - \alpha$.

Нехай маємо n пар спостережень

$$(\xi_1, \eta_1), \ \ldots, \ (\xi_n, \eta_n),$$

причому

$$\vec{\xi} = (\xi_1, \ldots, \xi_{n_1})$$

і

$$\vec{\eta} = (\eta_1, \ldots, \eta_{n_2})$$

– дві вибірки однакового об'єму n із неперервних розподілів $\mathcal{F}$ та $\mathcal{G}$, відповідно.
Гіпотезу щодо рівності розподілів сформулюємо у наступному вигляді:
H_0 : при кожному $i = \overline{1, n}$ обидва результати спостережень ξ_i та η_i є незалежними, однаково розподіленими випадковими величинами

Розглянемо різниці

$$\zeta_i = \xi_i - \eta_i, \ \ i = \overline{1, n},$$

та позначимо через

$$\mathcal{K} = \sum_{i=1}^{n} 1\{\zeta_i > 0\}.$$

кількість додатних різниць ζ_i, а $n - \mathcal{K}$ тоді кількість від'ємних різниць.
Помітимо, що $\zeta_1, \ldots, \zeta_n$ є незалежними однаково розподіленими випадковими величинами.

Якщо гіпотеза H_0 є вірною, тоді для кожного $i = \overline{1,n}$

$$P\{\xi_i - \eta_i > 0\} = P\{\zeta_i > 0\} = P\{\zeta_i < 0\} = P\{\xi_i - \eta_i < 0\} = \frac{1}{2},$$

оскільки, в силу неперервності розподілів $P\{\xi_i = \eta_i\} = P\{\zeta_i = 0\} = 0$.
Тоді $\mathcal{K}$ матиме біноміальний розподіл з $p = \frac{1}{2}$.

Зазначимо, що ймовірність, що серед $\zeta_1, \ldots, \zeta_n$ кількість додатних величин буде більшим за m, тобто

$$P\{\mathcal{K} > m\} = \left(C_n^{m+1} + \cdots + C_n^n\right) \cdot \frac{1}{2^n},$$

Зафіксуємо рівень значущості α та позначимо через $m_{\alpha,n}$ найменше m при якому

$$P\{\mathcal{K} > m\} \leq \alpha.$$

Тоді критична область двосторонньою критерію буде мати вигляд

$$\mathcal{K}^* > m_{\frac{\alpha}{2},n} \text{ або } n - \mathcal{K}^* > m_{\frac{\alpha}{2},n}.$$

Припустимо, що ф.р. F та G зв'язані співвідношенням

$$G(x) = F(x - \theta),$$

причому ні розподіл F, ні параметр θ невідомі. Тоді основну гіпотезу можна переписати у більш спрощеному вигляді: $H_0 : \theta = 0$, і розглядати,окрім двосторонньою критерія, ще й односторонні. При цьому критичні області для односторонніх критеріїв будуть мати вигляд:

- для правостороннього критерію

$$\mathcal{K}^* > m_{\alpha,n};$$

- для лівостороннього критерію

$$n - \mathcal{K}^* > m_{\alpha,n}.$$

10.3. Критерій знаків за наявності зв'язків.

Якщо зняти вимогу про неперервність розподілів в.в. ξ_i і η_i, то різниці

$$\zeta_i = \xi_i - \eta_i, \ \ i = \overline{1, n},$$

можуть набувати нульових значень з ненульовою ймовірністю (в цьому випадку кажуть, що є зв'язки у вибірці). В цьому випадку можна також використовувати критерій знаків, але попередньо слід відкинути всі рівні нулю різниці і застосовувати критерій лише до ненульових різниць, які залишаться.

10.3. Критерій знаків за наявності зв'язків.

Нехай серед n спостережень рівно s з ненульовими різницями. Тоді, якщо $\mathcal{K}$ – кількість додатних різниць, то кількість від'ємних буде $S - \mathcal{K}$, а критичні області критерію будуть мати вигляд:

- для двостороннього критерію

$$\mathcal{K}^* > m_{\frac{\alpha}{2},s} \text{ або } n - \mathcal{K}^* > m_{\frac{\alpha}{2},s}.$$

- для правостороннього критерію

$$\mathcal{K}^* > m_{\alpha,s};$$

- для лівостороннього критерію

$$n - \mathcal{K}^* > m_{\alpha,s}.$$

Дякуємо за увагу!