ПРАКТИЧНЕ ЗАНЯТТЯ № 5. (Теорія)
Головні задачі математичної статистики.
Непараметричне оцінювання.

1. Задача оцінювання розподілу генеральної популяції.
2. Емпірична функція розподілу.
3. Формальне визначення емпіричної функції розподілу.

1. *Задача оцінювання розподілу генеральної популяції.*

В найбільш загальній постановці задачу непараметричне оцінювання можна сформулювати наступним чином:

- Припустимо, що функція розподілу $F(x)$ генеральної популяції невідома.
- Маємо в розпорядженні просту випадкову вибірку $\{\xi_1, \xi_2, \dots, \xi_n\}$, що вибрана з цієї генеральної популяції.
- Крім інформації, що містить в собі вибірка $\{\xi_1, \xi_2, \dots, \xi_n\}$, не володіємо ніякими іншими знаннями, що стосуються розподілу досліджуваної ознаки та не пов'язані із статистичним спостереженням.

Задача полягає в тому, щоб на підставі спостережень $\{\xi_1, \xi_2, \dots, \xi_n\}$ оцінити невідому функцію розподілу $F(x)$, тобто побудувати таку функцію $\hat{F}_n(x)$, яку можна було б розглядати, як *статистичне наближення* невідомої функції розподілу $F(x)$: $\hat{F}_n(x) \approx F(x)$, $-\infty < x < +\infty$.

2. *Емпірична функція розподілу.*

Нехай $\{\xi_1, \xi_2, \dots, \xi_n\}$буде простою вибіркою, вибраною з популяції, що має функцію розподілу $F(x)$. Розташуємо її елементи в зростаючому порядку:

$$\xi^*_1 \le \xi^*_2 \le \dots \le \xi^*_n.$$

В математичній статистиці отриманий таким чином вектор $\{\xi^*_1, \xi^*_2, \dots, \xi^*_n\}$ називається *варіаційним рядом*.

- Координата ξ^*_i варіаційного ряду являється i-тою по порядку (за величиною) координатою вектору $\{\xi_1, \xi_2, \dots, \xi_n\}$.

Між іншим: $\xi^*_1 = min(\xi_1, \xi_2, \dots, \xi_n)$; $\xi^*_n = max(\xi_1, \xi_2, \dots, \xi_n)$.

- Координати ξ^*_i, $i = 1, 2, \dots, n$ варіаційного ряду називаються *порядковими* (або *позиційними*) *статистиками*.

Функція розподілу $F(x)$ генеральної популяції визначає розподіл значення ознаки X, що вивчається, серед елементів всієї популяції. Нашим завданням є побудова *наближення* невідомої функції розподілу $F(x)$.

Оскільки проста вибірка *напевно* є репрезентативною, тобто структура розподілу в ній властивості X *несуттєво відрізняється* від структури цього розподілу у всієї популяції, то можна сподіватися, що:

- Розподіл значення ознаки X в вибірці $(\xi_1, \xi_2, \dots, \xi_n)$ можна розглядати, як *наближення* невідомої функції розподілу $F(x)$ генеральної популяції.

Визначення. ***Емпіричною функцією розподілу***, побудованою на підставі простої статистичної вибірки $(\xi_1, \xi_2, \dots, \xi_n)$ називається функція $\hat{F}_n(x)$, $-\infty < x < +\infty$, яка визначає **розподіл значення ознаки X в цій вибірці**.

Іншими словами, для кожного дійсного числа x $\hat{F}_n(x)$ вказує ***частоту появи*** у вибірці $(\xi_1, \xi_2, \dots, \xi_n)$ ***чисел менших від x***.

3. *Формальне визначення емпіричної функції розподілу.*

Запишемо це визначення формальним чином. Позначимо через $\nu_n(x)$ кількість елементів вибірки $(\xi_1, \xi_2, \dots, \xi_n)$ *строго менших* від x. Якщо використати індикатор $I(A)$ випадкової події A, тобто: $\chi[x] = \begin{cases} 1, \; \text{якщо } A \text{ відбулась} \\ 0, \; \text{якщо } A \text{ не відбулась} \end{cases}$, то для підрахунку $\nu_n(x)$ маємо наступну формулу:

$$\nu_n(x) = \sum_{i=1}^{n} \chi[\xi_i < x].$$

Тоді визначення емпіричної функції розподілу приймає вигляд:

$$\hat{F}_n(x) = \frac{\nu_n(x)}{n}, -\infty < x < +\infty.$$

Або

$$\hat{F}_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \chi[\xi_i < x].$$

Використовуючи варіаційний ряд $(\xi^*_1, \xi^*_2, \dots, \xi^*_n)$ і припускаючи, що серед елементів вибірки немає однакових, тобто:

$$\xi^*_1 < \xi^*_2 < \dots < \xi^*_n,$$

можна задати емпіричну функцію розподілу, визначаючи її значення на проміжках $(\xi^*_{i-1}, \xi^*_i]$, $i = 1, 2, \dots, n$, (покладаючи $\xi^*_0 = -\infty$), а саме:

$$\hat{F}_n(x) = \begin{cases} 0, \quad x \le \xi^*_1, \\ \dfrac{k}{n}, \quad \xi^*_k < x \le \xi^*_{k+1}, \\ 1, \quad x > \xi^*_n \,. \end{cases}$$

Іншими словами, на відрізках між двома сусідніми елементами варіаційного ряду емпірична функція розподілу $\hat{F}_n(x)$ зберігає постійні значення, які є кратними величині $1/n$. При цьому:

- На відрізку $(\xi^*_{i-1}, \xi^*_i]$ $\hat{F}_n(x)$ зберігає значення: $(i-1)/n$, $i = 1, 2, \dots, n$.
- Якщо в якійсь точці x значення $\hat{F}_n(x)$ дорівнює m/n, то це означає, що у вибірці $(\xi_1, \xi_2, \dots, \xi_n)$ **є *точно m* елементів**, значення яких ***строго менших від x***.
- Функція $\hat{F}_n(x)$ розривна, в точках ξ^*_i, $i = 1, \dots, n$, **має *стрибки* величиною $1/n$.**
- Оскільки $[\xi^*_i, i = 1, 2, \dots, n]$ та $[\xi_i, i = 1, 2, \dots, n]$ – це та сама множина дійсних чисел, то функція $\hat{F}_n(x)$ має *стрибки* в точках $(\xi_1, \xi_2, \dots, \xi_n)$.

Легко переконатися, що для функції $\hat{F}_n(x)$ виконуються всі характеристичні властивості функції розподілу, тобто:

- $\hat{F}_n(x)$ неспадна функція.
- $\hat{F}_n(-\infty) = 0$, $\hat{F}_n(+\infty) = 1$.
- $\hat{F}_n(x)$ лівостороннє-неперервна.