Описова статистика (Descriptive Statistics)

Лекція № 1

1.1. Основні поняття математичної статистики
1.2. Методи представлення інформації про вибірку

Означення 1

Математична статистика – це розділ математики, в якому вивчаються методи збору, систематизації та обробки інформації з метою виявлення існуючих закономірностей.

Припустимо, що ми повторюємо один й той самий випадковий експеримент в однакових умовах і отримуємо певний набір даних (числових чи деяких інших). При цьому виникають наступні питання.

1) Якщо спостерігається одна випадкова величина (в.в.) – як з набору її значень в кількох експериментах зробити як умога точніший висновок про розподіл в.в. : функція розподілу (ф.р.), щільність або закон розподілу, числові характеристики такі як математичне сподівання, дисперсія, моменти в.в., тощо.
2) Якщо ми спостерігаємо декілька в.в. одночасно, що можна сказати про їх залежність або сумісний розподіл?

Означення 2

Генеральна сукупність (г.с.) - це деякий (як правило невідомий) ймовірностний розподіл $\mathcal{F}$.

Означення 3

Вибірка – це набір незалежних в.в. $\xi_1, \xi_2, \ldots, \xi_n$ або $\vec{\xi} = (\xi_1, \ldots, \xi_n)$, кожна з яких має розподіл $\mathcal{F}$. При цьому n називається об'ємом вибірки.

Означення 4

Реалізація вибірки – це значення $x_1,\ x_2,\ \ldots,\ x_n$, які набулися в.в. $\xi_1,\ \xi_2,\ \ldots,\ \xi_n$ в результаті конкретного стохастичного експерименту. Тобто

$$x_1 = \xi_1(\omega_0),\ x_2 = \xi_2(\omega_0),\ \ldots,\ x_n = \xi_n(\omega_0).$$

При цьому x_k називається варіантою.

Множина всіх реалізацій S вибірки $x_1,\ x_2,\ \ldots,\ x_n$ називається вибірковим простором. Пара $(S,\ \mathcal{F})$ називається статистичною моделлю опису серії спостережень, які породжують вибірку.

1.2. Методи представлення інформації про вибірку

- 1.2.1. Первинна обробка інформації
- 1.2.2. Графічні методи представлення інформації
- 1.2.3. Аналітичні методи представлення інформації

Статистика має спеціальні означення для понять важливих для статистичної аргументації. В описовій статистиці ви збираєте дані та описуєте їх. Якщо ви аналізуєте та робите інтерпретацію даних, то ви використовуєте статистичні дані.

1.2.1. Первинна обробка інформації

Нехай маємо реалізацію вибірки $x_1,\ x_2,\ \ldots,\ x_n$.

Означення 5

Варіаційним рядом $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ називаються елементи вибірки, які впорядковані за зростанням. При цьому

$$x_{(1)} = \min_{1\le k\le n} x_k,\ \ x_{(n)} = \max_{1\le k\le n} x_k.$$

$x_{(k)}$ називається порядковою статистикою порядку k.

Процес впорядкування вибірки називається "ранжуванням" .

Статистичні розподіли.

Нехай розподіл г.с. $\mathcal{F}$ є дискретним. Тоді нехай $x_1^*, \ldots, x_m^*$ – елементи вибірки, впорядковані за зростанням, причому кожне значення вказується лише один раз, n_k – число разів появи x_k^* в реалізації вибірки. n_k називається частотою появи x_k^*.
Зауважимо, що

$$n_1 + \ldots + n_m = n.$$

Величина $\nu_k = \frac{n_k}{n}$, відношення частоти n_k до об'єму вибірки n, називається відносною частотою.
Сума відносних частот елементів $x_1^*, \ldots, x_k^*$ називається накопичувальною частотою ν_k^* елементу x_k^*, тобто

$$\nu_k^* = \nu_1 + \ldots + \nu_k.$$

1.2. Методи представлення інформації про вибірку

На основі вказаних вище характеристик можна побудувати статистичний розподіл г.с.

Значення	x_1^*	x_2^*	...	x_m^*
Частоти	n_1	n_2	...	n_m
Відносні частоти	$\nu_1 = \frac{n_1}{n}$	$\nu_2 = \frac{n_2}{n}$	...	$\nu_m = \frac{n_m}{n}$
Накопичувальні чатоти	$\nu_1^* = \nu_1$	$\nu_2^* = \nu_1 + \nu_2$	...	$\nu_m^* = \nu_1 + \ldots + \nu_m = 1$

Якщо розподіл г.с. є неперервним або дискретним з великою кількістю значень, тоді використовують інтервальний статистичний розподіл.

Область, в якому лежать всі значення реалізації вибірки розбивають на m інтервалів $\Delta_1, \ldots, \Delta_m$. Кількість інтервалів, на які слід розбити область можна знайти за формулою Стреджеса

$$m = 1 + [\log_2 n] = 1 + [3.322 \cdot \lg n].$$

Довжини інтервалів при цьому знаходяться за формулою

$$h = \frac{x_{(n)} - x_{(1)}}{m}.$$

Через n_k позначимо кількість елементів вибірки, які попали в інтервал Δ_k. ν_k та ν_k^* задаються так само, як і в попередньому випадку.

На основі вказаних вище характеристик будується інтервальний статистичний розподіл г.с.

Значення	Δ_1	Δ_2	...	Δ_m
Частоти	n_1	n_2	...	n_m
Відносні частоти	ν_1	ν_2	...	ν_m
Накопичувальні частоти	ν_1^*	ν_2^*	...	$\nu_m^* = 1$

Зауваження 1

В деяких випадках спостереження можуть надаватись не як індивідуальні значення, а вже розподілені по інтервалах тобто у вигляді інтервального статистичного розподілу. В цьому випадку дані називаються груповаными.

1.2.2. Графічні методи представлення інформації

Полігон частот. Використовується у випадку, коли розподіл г.с. є дискретним, на основі статистичного розподілу г.с. Будують систему координат таку, що на осі абсцис будуть відображатися елементи вибірки $x_1^*, \ldots, x_m^*$, а на осі ординат відповідні відносні частоти. Далі у вказаній системі координат будують точки $M_k\left(x_k^*, \nu_k\right)$, $k = \overline{1,m}$, які з'єднують між собою у ламану $M_1M_2 \ldots M_m$.

Гістограма. Використовується у випадку, коли розподіл г.с. є неперервним, на основі інтервального статистичного розподілу г.с. Будують систему координат таку, що на осі абсцис будуть відображатися інтервали $\Delta_1, \ldots, \Delta_m$, а на осі ординат відповідні скориговані відносні частоти. Далі будують у вказаній системі координат прямокутники з основами Δ_k, $k = \overline{1,m}$, та відповідними висотами

$$h_k = \frac{\nu_k}{l(\Delta_k)},$$

де $l(\Delta_k)$ – довжина інтервалу Δ_k. Зауважимо, що в цьому випадку площа отриманої фігури буде дорівнювати

$$S_{\text{гіст.}} = \sum_{k=1}^{m} l(\Delta_k) \cdot h_k = \sum_{k=1}^{m} l(\Delta_k) \cdot \frac{\nu_k}{l(\Delta_k)} = 1.$$

Емпірична ф.р. Нехай $F(y) = P\{\xi_1 < y\}$ – теоретична ф.р. Тоді у якості оцінки теоретичної ф.р. можна взяти так звану емпіричну ф.р.

Для дискретної г.с., на основі статистичного розподілу, вона будується наступним чином:

$$F_n^*(y) = \frac{\text{кількість } x_k \text{ менших за } y}{n}.$$

У разі, коли розподіл г.с. є неперервним або дискретний з великою кількістю значень, емпіричну ф.р. можна будувати на основі інтервального статистичного розподілу

$$F_n^*(y) = \frac{\text{кількість } x_k \text{ менших за праву границю інтервала, що містить } y}{n}.$$

1.2.3. Аналітичні методи представлення інформації

Графічні методи є чудовим способом для отримання швидкого огляду вибіркових даних, але вони не є точними та не призводять самі по собі до наступних досліджень. Для цього нам потрібно введення числових параметрів таких як, наприклад, середнє.
Існують різні шляхи, за допомогою яких ми можемо спробувати описати розподіл. Розглянемо деякі з них, які корисні при описі гістограми або полігона частот.

Нехай $x_1, \ldots, x_n$ реалізація вибірки з г.с. $\mathcal{F}$

Параметри зсуву (measure of location).

Вибіркове середнє $(\overline{x})$ є одним із найбільш відомих параметрів зсуву і знаходиться за формулою: $\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$.

Зауваження 2

Вказана формула може бути використана лише у випадку, коли всі індивідуальні значення x_j є відомими.

Якщо ж відносно вибірки є відомим лише розбиття спостережень на класи (Δ_i) та частота попадань елементів вибірки у кожен з цих інтервалів (n_i). В цьому випадку групованих даних можна використовувати наступну формулу: $\overline{x} = \frac{1}{n} \sum_{i=1}^{m} n_i x_i = \sum_{i=1}^{m} \nu_i x_i$, де n_i – частота попадання в інтервал Δ_i, x_i – точка, що є серединою інтервалу Δ_i, $i = \overline{1, m}$.

Вибіркова медіана.

Однією із вад вибіркового середнього воно дає нерепрезентативні результати, оскільки є чутливим до викидів у вибірці (значення які значно відрізняються від інших спостережень вибірки)та симетрії. Тому іноді є більш привабливим використання інших параметрів зсуву, які є більш стійкими до таких екстремальних значень. Одним із таких параметрів є вибіркова медіана. Вона знаходиться за наступною процедурою.

Якщо у вибірці відомі всі індивідуальні спостереження. Будуємо варіаційних ряд $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ та знаходимо середній елемент цього варіаційного ряду M_e^*, який і називається вибірковою медіаною, причому,

$$M_e^* = \begin{cases} x_{(k+1)}, & \text{якщо } n = 2k+1; \\ \frac{x_{(k)}+x_{(k+1)}}{2}, & \text{якщо } n = 2k. \end{cases}$$

Якщо вибірка представлена групованими даними, тоді вибіркова медіана розраховується у два етапи : спочатку ми маємо визначити інтервал у якому міститься елемент, який відповідає медіані, а потім маємо розрахувати де в інтервалі цей елемент буде міститися. Для цього можна використати формулу

$$M_e^* = x_L + (x_U - x_L) \cdot \frac{\frac{n+1}{2} - N_{me}}{n_{me}},$$

де x_L, x_U – відповідно нижня та верхня межі інтервала, який містить медіану, N_{me} – накопичувальна частота інтервалів розбиття до інтервалу, що містить медіану (але не включаючи сам інтервал), n_{me} – частота інтервалу, який містить медіану.
Зазначимо, що дріб у правій частині останньої рівності вказує нам, яку частину інтервалу ми маємо пройти, щоб дійти до медіани.

Вибіркова мода

Вибіркова мода $\mathcal{M}_0^*$ – це елемент вибірки, який зустрічається частіше за все.

У випадку групованих даних визначення більш складне. Інтервал, який має найбільшу частоту попадання в нього, називається модальним, причому при розбиті на інтервали, вони мають бути однакової ширини (в протилежному випадку розгляд більш широких інтервалів є нечесним порівняно з інтервалами меншої довжини). Значення моди знаходиться за формулою

$$\mathcal{M}_0^* = x_L + (x_U - x_L) \cdot \frac{n_{mo} - n_{mo-1}}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})},$$

де x_L, x_U – відповідно нижня та верхня межі модального інтервала n_{mo}, n_{mo-1}, n_{mo+1} – частоти відповідно модального, передмодального та післямодального інтервалів.

Зауваження 3

Вказані три параметри зсуву дають різну інформацію. Але якщо розподіл симетричний, то вони будуть давати приблизно однакові значення.

Параметри розсіювання (Measure of dispersion)

При розгляді двох різних розподілів, вони можуть мати однакові або дуже близькі середні, але при цьому суттєво відрізнятись. Наприклад, при порівнянні рівня добробуту двох країн: у них може бути дуже близькими середній добробут по країні, але в одній країні всі отримують приблизно однаковий рівень достатку, а в інший можуть екстремальні значення великого достатку та злиденного. Параметри розсієвання як раз і потрібні для відокремлення таких випадків.
Найпростішим параметром розсіювання є розмах вибірки, який є різницею між найбільшим та найменшим спостереженням даної вибірки.

Вибіркова дисперсія

Найбільш корисним параметром розсіювання є вибіркова дисперсія. Вона є середнім квадратів відхилень елементів вибірки від середнього значення.
Якщо відомо математичне сподівання розподілу $\mathcal{F}$, $E\xi_j = \mu$, тоді вибіркова дисперсія розраховується наступним чином:

- якщо дані не груповані, то

$$\sigma_*^2 = \mathfrak{D}_\xi^* = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \mu\right)^2;$$

- якщо дані груповані, то

$$\sigma_*^2 = \mathfrak{D}_\xi^* = \frac{1}{n}\sum_{i=1}^{m} n_i \left(x_i - \mu\right)^2,$$

де x_i – середини інтервалів Δ_i.

Якщо ж математичне сподівання розподілу $\mathcal{F}$ не відомо, тоді розглядають або вибіркову дисперсію

- для не групованих даних

$$s^2 = \mathfrak{D}_\xi^{**} = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2;$$

- для групованих

$$s^2 = \mathfrak{D}_\xi^{**} = \frac{1}{n}\sum_{i=1}^{m} n_i\left(x_i - \overline{x}\right)^2;$$

або виправлену (незміщену) вибіркову дисперсію

- для не групованих даних

$$s_0^2 = \mathfrak{D}_\xi^{***} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2;$$

- для групованих

$$s_0^2 = \mathfrak{D}_\xi^{***} = \frac{1}{n-1}\sum_{i=1}^{m} n_i\left(x_i - \overline{x}\right)^2.$$

Стандартне відхилення.

Оскільки при знаходженні дисперсії ми підносили відхилення до квадрату, то не зовсім зрозумілим є одиниці вимірювання дисперсії (Наприклад, при вимірі добробуту ми користувались грн, тоді в дисперсії ми отримуємо грн.^{2}). Тому природньо взяти від цього виразу корінь квадратний. Таким чином, ми отримуємо величину, рівну корню квадратному від вибіркової дисперсії, яка називається стандартним відхиленням.

Дякуємо за увагу!