МІРИ МІНЛИВОСТІ
ЛЕКЦІЯ 5

Всі середні показники дають загальну характеристику ряду результатів вимірювань. Кожна міра дає таке значення, яке «представляє» в певному сенсі всі оцінки групи. В цьому випадку нехтують розходженнями, що існують між окремими значеннями. На практиці часто цікавить, наскільки кожен результат відхиляється від середнього значення. Однак легко можна уявити, що дві групи результатів вимірювань мають однакові середні, але різні значення вимірювань.

Приклад:

i	1	2	3	4	5	6	7	8	9	10	11	M_o	M_e	$\overline{X}$
Вибірка 1	3	3	3	4	4	4	4	4	5	5	5	**4**	**4**	**4**
Вибірка 2	2	2	3	4	4	4	4	5	5	5	6	**4**	**4**	**4**

Вибірки різні, але міри центральної тенденції (МЦТ) однакові! Відмінності ж між цими рядами істотні. Тому МЦТ завжди необхідно доповнювати показниками варіації.

Міри мінливості даних – це статистичні показники, що дозволяють судити про ступінь однорідності отриманої множини даних, її компактності.

Найбільш поширені в психологічних дослідженнях показники:

- ✓ розмах
- ✓ дисперсія
- ✓ стандартне відхилення

Розмах (R) – це інтервал між максимальним і мінімальним значеннями ознаки.

Він визначається легко і швидко, але чутливий до випадкових даних особливо при малій кількості досліджуваних. Це найпростіша міра відхилення. Розмах визначається лише крайніми значеннями ознаки

$$\mathbf{R=X_{max}-X_{mjn}}$$

Розмах досить груба міра мінливості, тому що для її обчислення використовуються тільки два елементи вибірки, а розподіл інших елементів не враховується.

Наприклад, розглянемо дві вибірки, кожна з яких складається зі ста елементів:

а) $\underbrace{1......1}_{10\text{ раз}}, \underbrace{2......2}_{10\text{ раз}}, \underbrace{10......10}_{10\text{ раз}}$ $R_1=10-1=9$

б) $1, \underbrace{2\ ....2}_{98\text{ раз}}, 10$ $R_2=10-1=9$

Незважаючи на те, що вибірки істотно розрізняються, розмахи вибірок рівні.

Дисперсія (D) – це міра мінливості, що дозволяє врахувати відмінності між окремими елементами вибірки.

Для обліку цих відмінностей потрібно розглянути суму відхилень кожного елемента вибірки від середнього значення:

$$(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})$$

Але за властивістю середнього значення ця сума завжди дорівнює нулю, тому що відхилення можуть приймати як позитивні, так і від’ємні значення. Для подолання цього недоліку кожне відхилення потрібно звести в квадрат:

$$(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2 = \sum_{i=1}^{n}(x_1 - \overline{x})^2$$

Щоб отримати величину, що характеризує відхилення кожного окремого елемента вибірки від середнього значення, отриману суму необхідно поділити на кількість елементів n. Отримана величина і є дисперсією:

$$D = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n}, \text{ де } n \geq 30$$

Для малих вибірок формула для обчислення дисперсії набуває вигляду:

$$D = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n-1}, \text{ де } n < 30$$

Чим більшою є дисперсія вибірки, тим більше розкидані вихідні значення по числовій осі щодо середнього значення вибірки.

Приклад 1

Обчислити дисперсію для наступної вибірки: 3, 5, 6, 9, 11, 14.

Рішення:

Визначимо обсяг вибірки: n=6.

Обчислимо середнє значення вибірки:

$\overline{x}$=(3+5+6+9+11+14):6=48:6=8

Обчислимо дисперсію з урахуванням того, що n < 30:

$$D = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^2}{n-1} = \frac{(3-8)^2 + (5-8)^2 + (6-8)^2 + (9-8)^2 + (11-8)^2 + (14-8)^2}{6-1} = 16{,}8$$

Важлива характеристика дисперсії полягає в тому, що за її допомогою можна порівнювати вибірки, різні за обсягом.

Чим вища однорідність вибірки, тим нижчим буде значення дисперсії.

Для того щоб наблизити розмірність дисперсії до розмірності вимірюваного ознаки, застосовують операцію вилучення квадратного кореня з дисперсії. Отриману величину називають ***вибірковим стандартним відхиленням***:

Стандартне відхилення або середньоквадратичне відхилення (σ - сигма) – міра мінливості, яка дозволяє охарактеризувати розкид елементів вибірки щодо середнього значення.

Стандартне відхилення тісно пов'язане з дисперсією. Оскільки для знаходження дисперсії кожне відхилення ознаки від середнього значення зводили в квадрат, то для отримання величини, порівнянної з середнім значенням, необхідно з дисперсії винести квадратний корінь. Отриману величину і назвали середнім квадратичним або стандартним відхиленням:

$$\boldsymbol{\sigma = \sqrt{D}}$$

Середнє квадратичне відхилення (σ – сигма) показує, наскільки часто відхиляються індивідуальні значення середнього, і вимірюється в тих самих одиницях, що і середнє арифметичне. За даним критерієм ми судимо про щільність вибірки: чим більше значення σ, тим менш щільними є результати вибірки. Чим ближче інваріанти до $\overline{X}$, чим щільнішою є вибірка, тим вищою є її однорідність. Середнє квадратичне відхилення обчислюється за формулою:

Якщо n≥30 $\boldsymbol{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\overline{x})^2}{n}}$

Якщо n < 30 $\boldsymbol{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\overline{x})^2}{n-1}}$

Якщо значення ознаки повторюються, то

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\overline{x})^2 \times n_i}{n-1}}$$

Наприклад 2

Візьмемо два ряди з однаковими середніми арифметичними: X = 4, але тепер їх відмінність один від одного покаже **σ**:

1 група – 2, 8, 2 ⇒

$$\boldsymbol{\sigma} = \sqrt{\frac{(2-4)^2+(8-4)^2+(2-4)^2}{4-1}} = \frac{4+16+4}{3} = \sqrt{\mathbf{8}} = \mathbf{2,83}$$

2 група - 5, 2, 5 ⇒

$$\boldsymbol{\sigma} = \sqrt{\frac{(5-4)^2+(2-4)^2+(5-4)^2}{4-1}} = \frac{1+4+1}{3} = \sqrt{\mathbf{2}} = \mathbf{1,4}$$

тобто друга група більш однорідна.

Приклад 3

Обчислити стандартне відхилення для вибірки з прикладу 2: 3, 5, 6, 9, 11, 14

Розв'язання:

$\sigma=\sqrt{D}$

У попередньому прикладі дисперсія дорівнює 16,8. Тоді

$\sigma=\sqrt{16{,}8} \approx 4{,}1$

Дисперсію і стандартне відхилення можна легко розрахувати за статистичним розподілом вибірки

Приклад 4:

Обчислити дисперсію і стандартне відхилення вибірки за розподілом частот:

X_i	1	2	3	4
n_i	20	15	10	5

Розв'язання:

Визначимо обсяг вибірки: n = 20 + 15 + 10 + 5 = 50.

Обчислимо середнє значення:

$$\overline{X} = \frac{(1*20+2*15+3*10+4*5)}{50} = \frac{100}{50} = 2$$

Обчислимо дисперсію для вибірки обсягом n> 30

$D = \frac{\sum_{i=1}^{n}(x_1-\overline{x})^2}{n}$ =((1-2)^2×20+(2-2)^2×15+(3-2)^2×10+(4-2)^2×5)/50=1

Обчислимо стандартне відхилення

σ=$\sqrt{D}$=$\sqrt{1} = 1$

Невелика величина стандартного відхилення говорить про однорідність вибірки, тобто середнє значення є типовим значенням для варіаційного ряду. При дуже великому стандартному відхиленні середнє значення в меншій мірі характеризує варіаційний ряд, що говорить про значну варіабельності ознаки або неоднорідності досліджуваної групи.

Коефіцієнт варіації (v) – міра мінливості, яка дозволяє оцінити ступінь розсіювання варіант біля середнього значення.

Коефіцієнт варіації обчислюється за формулою:

$$\mathbf{V=\frac{\sigma}{\overline{x}} \times 100\%}$$

Якщо **V<10%,** це говорить про мале розсіювання варіант навколо середнього значення;

якщо **10% <V <20%**, то можна говорити про середнє розсіювання варіант;

якщо **V>20%,** то розсіювання варіант навколо середнього є сильним і середнє не є типовим значенням варіаційного ряду.

Приклад 5

Визначити коефіцієнт варіації за даними прикладу.

Рішення.

Для попереднього прикладу $\overline{X} = 2$, σ= 1

$\mathbf{V=\frac{\sigma}{\overline{x}} \times 100\% = \frac{1}{2} \times 100\% = 50\%}$ **-** сильне розсіювання варіант біля середнього значення

Асиметрія та ексцес

Існують характеристики вибірки, які показують, чи спостерігаються в ній переважні значення ознаки. Вони також визначають форму графічного представлення даних і називаються мірами форми. До мір форми відносяться асиметрія і ексцес.

Асиметрією (А) називають ступінь відхилення графічного представлення даних (гістограми або полігону частот) від симетричного виду щодо середнього значення.

Для обчислення асиметрії використовується формула:

$$A = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^3}{n \times \sigma^3}$$

.

де n - об'єм вибірки, σ – стандартне відхилення, $\overline{X}$ — середнє значення

Асиметрія показує, які значення переважають у вибірці: більше середнього значення або менше за нього.

Властивості асиметрії:

1) якщо **A>0**, то говорять про позитивну (лівосторонню) асиметрію, тобто у вибірці найчастіше зустрічаються значення менше середнього значення

2) якщо **A <0**, то говорять про від'ємну (правосторонню) асиметрію, тобто у вибірці найчастіше зустрічаються значення більші за середнє значення;

3) якщо **A=0**, то це означає, що вихідна вибірка (її гістограма) є симетричною відносно прямої X $=\overline{x}$

На практиці повністю симетричні гістограми або полігони частот зустрічаються досить рідко.

Ексцес (E) – це міра, що характеризує відносну опуклість або згладженість розподілу вибірки порівняно з нормальним розподілом.

Для обчислення ексцесу використовується формула:

$$E = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^4}{n \times \sigma^4} - 3$$

Позитивний ексцес позначає відносно загострений розподіл, негативний – відносно згладжений.

Ексцес дозволяє визначити, наскільки кучно основна маса даних групується навколо середнього значення.

Властивості ексцесу:

1) якщо **E>0**, то говорят, що вихідні дані відповідають гостровершинному розподілу, тобто купчасто розташовані навколо середнього значення;

2) якщо **-3<E<0**, то говорять, що вихідні дані відповідають плосковершинному розподілу, тобто розсіяні відносно середнього значення;

3) якщо **E=0**, то говорять, що вихідні дані відповідають середньовершинному розподілу.

Приклад 6:

За даними прикладу обчислимо асиметрію та ексцес

X_i	1	2	3	4
n_i	20	15	10	5

В результаті попередніх розрахунків було встановлено, що

n = 50 $\overline{X} = 2$ σ$= 1$

Скористаємося формулою для обчислення асиметрії:

$$A = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^3}{n \times \sigma^3} = \frac{(1-2)^3 \times 20 + (2-2)^3 \times 15 + (3-2)^3 \times 10 + (4-2)^3 \times 5}{50 \times 1^3}$$
$$= \frac{-20 + 0 + 10 + 40}{50} = \frac{30}{50} = 0{,}6$$

лівостороння асиметрія

Скористаємося формулою для обчислення ексцесу:

$$\boldsymbol{E = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})^4}{n \times \sigma^4} - 3}$$

$$E = \frac{(1-2)^4 \times 20 + (2-2)^4 \times 15 + (3-2)^4 \times 10 + (4-2)^4 \times 5}{50 \times 1^4} - 3$$
$$= \frac{20 + 0 + 10 + 80}{50} - 3 = \frac{110}{50} - 3 = -0{,}8$$

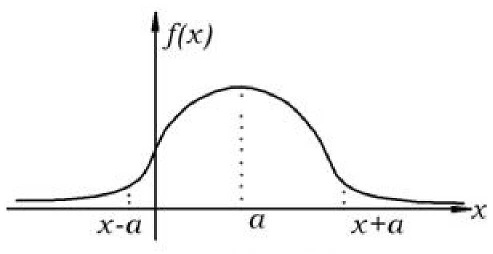
данні відповідають плосковершинному розподілу

Вибір числових характеристик для опису вибірки залежить від того, в якій шкалі проведено вимірювання ознаки. Для даних, представлених в номінальній шкалі, підходящою мірою є мода. Дані, виміряні в рангової шкалою, можна описати за допомогою медіани. Кількісні дані можна представити за допомогою середнього значення і стандартного відхилення.

Розподілом ознаки називається закономірність появи різних її значень. Для його опису використовуються параметри розподілу.

Параметри розподілу – це числові характеристики, що представляють розподіл. Одні з них – міри центральної тенденції – вказують, де в основному розташовані значення ознаки; інші – міри розкиду – наскільки значення ознаки мінливі; треті – міри форми – чи спостерігається переважна поява певних значень ознаки.

У психології найчастіше посилаються на нормальний розподіл. Він характеризується тим, що крайні значення ознаки зустрічаються досить рідко, а середні – досить часто. Графік нормального розподілу являє собою колоколообразну средньовершинну криву, симетричну відносно прямої X =$\overline{x}$.



Графік нормального розподілу

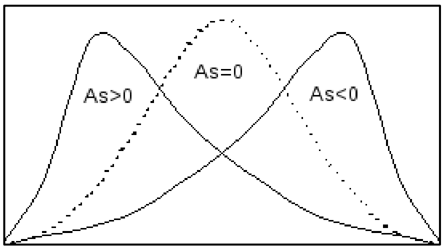
Нормальний розподіл характеризується наступними параметрами:

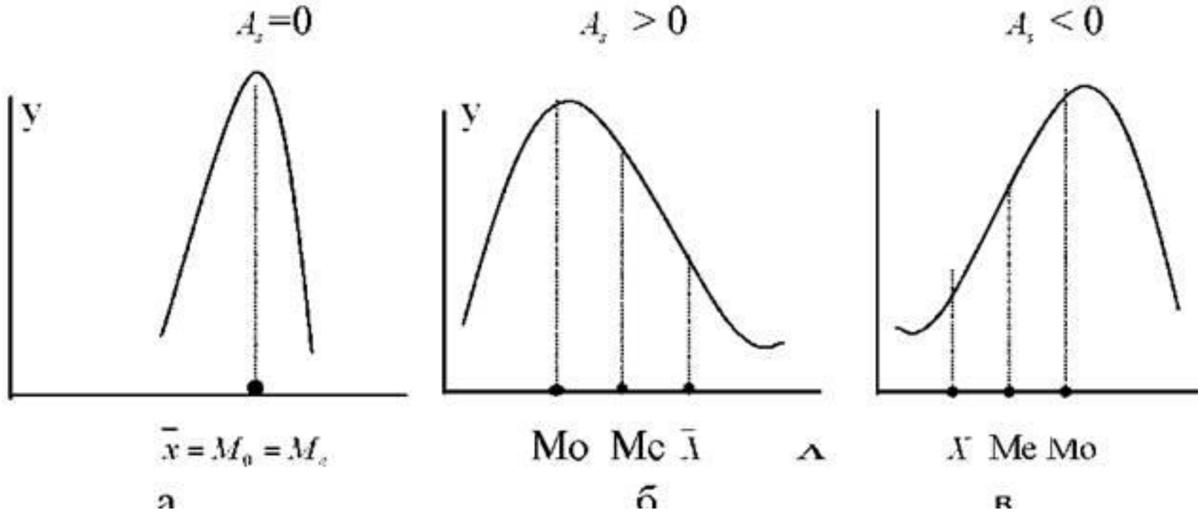
1) мода, медіана і середнє значення вибірки збігаються: Mo≈Me ≈$\overline{x}$

2) дисперсія і стандартне відхилення невеликі;

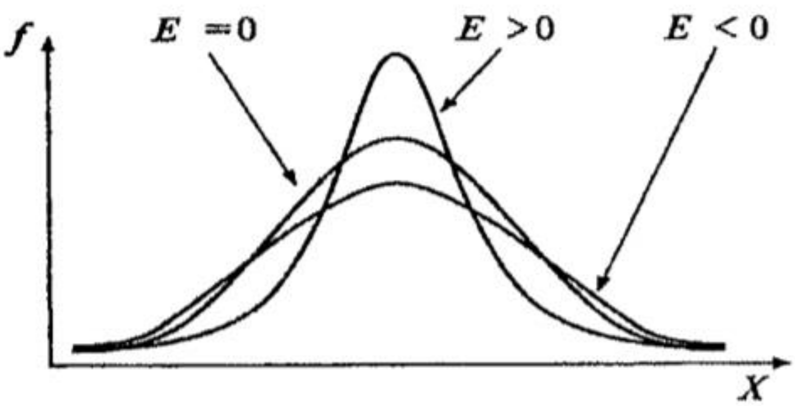
3) асиметрія і ексцес дорівнюють нулю.

Якщо **асиметрія** відмінна від нуля, то розподіл ознаки є асиметричним. Графічно асиметричні розподілу можна представити таким чином





Розподілу ознак з різними ексцесами графічно виглядають наступним чином



Таким чином, розподіл ознаки дозволяє в графічному вигляді представити варіативність цієї ознаки.