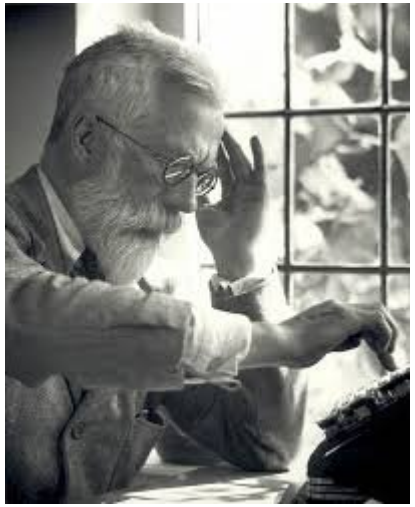
# Дисперсійний аналіз

# ANalisys Of VAriance (ANOVA)

**Змістовий модуль 4**

# Вступ



*Сер Рональд Ейлмер Фішер*

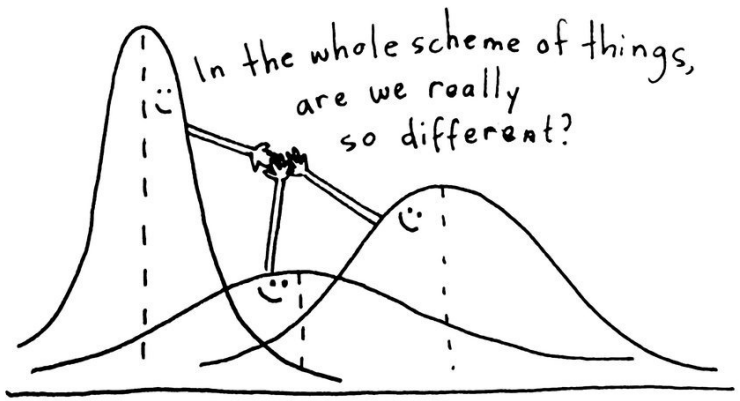
**Дисперсійний аналіз** (**ANOVA - Analysis of Variance**) є сукупністю статистичних методів, призначених для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису, а також для встановлення ступеня впливу факторів та їх взаємодії.

У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations). Вперше цей метод було розроблено Р. Фішером в 1925 р.

**Факторами** називають контрольовані чинники, що впливають на кінцевий результат.

**Рівнем фактора** називають значення, що характеризують конкретний прояв цього фактора. Ці значення зазвичай подають у ***номінальній*** або ***порядковій*** шкалі вимірювань.

Значення вимірюваної ознаки називають **відгуком**.




Здавалося б навіщо потрібен дисперсійний аналіз (ANOVA), якщо існує такий прекрасний і зрозумілий статистичний критерій, як t-критерій Ст'юдента?

Однак, головне обмеження t-критерію перед дисперсійним аналізом полягає в тому, що перший призначений для парних порівнянь, тобто ситуації, коли ми маємо лише дві групи, і він потребує поправок на множинні порівняння у випадку, якщо ми маємо більш ніж дві групи. По-друге, якщо наявні, наприклад, 6 груп і потрібно визначити статистично значимі відмінності між ними, скільки попарних порівнянь у такому разі потрібно зробити?

У такому разі набагато простіше використовуватися критерій, який призначений для ситуацій, коли наявно багато груп і який спроможний дати єдину відповідь на всі досліджувані групи - дисперсійний аналіз.

# Вступ

**Що таке дисперсійний аналіз (ANOVA)?**

- Статистичний метод для оцінки відмінностей між середніми значеннями кількох груп.
- Використовується для перевірки гіпотез, чи є статистично значуща різниця між групами.

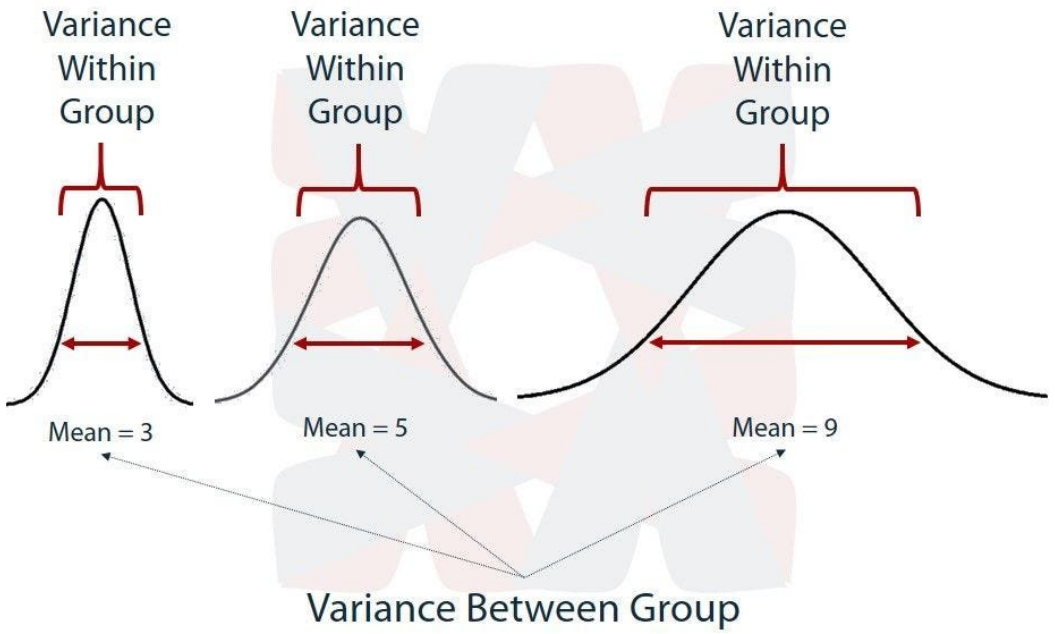
**Приклади**:

- **Вплив різних дієт на масу тіла**: три групи піддослідних тварин були розподілені на три різні дієти. Через місяць були зібрані дані про масу тіла кожної тварини. Визначте, чи є статистично значуща різниця між середніми масами тіла в цих трьох групах. **Дані для аналізу**: маса тіла для кожної тварини в трьох групах. **Мета**: перевірити, чи дієти вплинули на масу тіла.
- **Вплив різних доз препарату на рівень глюкози**: дослідники досліджують вплив трьох різних доз препарату на рівень глюкози в крові. Піддослідні тварини були випадковим чином розподілені на три групи, кожній групі давали різну дозу препарату. Після експерименту виміряли рівень глюкози. **Дані для аналізу**: рівень глюкози в крові для кожної тварини в трьох групах. **Мета**: визначити, чи є відмінність між середніми рівнями глюкози при різних дозах.
- **Вплив умов освітлення на швидкість росту рослин**: три групи рослин вирощували за різних умов освітлення: природне світло, штучне світло та комбіноване світло. Після 4 тижнів виміряли висоту рослин у кожній групі. **Дані для аналізу**: висота рослин у кожній з трьох груп. **Мета**: визначити, чи умови освітлення впливають на швидкість росту.

# Основні концепції ANOVA

**Дисперсія** - міра того, наскільки спостереження відрізняються від середнього значення.

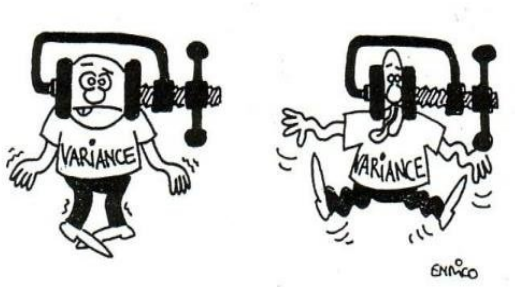
- Міжгрупова дисперсія: варіативність між групами.
- Внутрішньогрупова дисперсія: варіативність всередині груп.



# Мета дисперсійного аналізу

Визначити, чи середні значення вибірок (груп) суттєво відрізняються одне від одного.

- **Основна [нульова] гіпотеза ($H_0$): Середні значення у всіх групах рівні**
- **Альтернативна гіпотеза ($H_1$): Є хоча б одна пара середніх значень, що відрізняються**




**Основною метою однофакторного аналізу** зазвичай є оцінка величини впливу конкретного фактора на досліджуваний відгук. Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають контрастом факторів. Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик. У випадку, коли нульова гіпотеза не може бути відкинутою, зазвичай її приймають і роблять висновок про відсутність впливу. Але, якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об'єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

# Типи дисперсійного аналізу

Однофакторний дисперсійний аналіз (ANOVA):

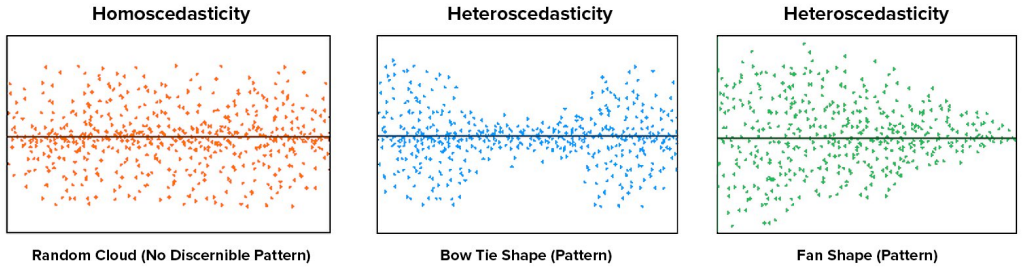
- Оцінює вплив одного фактора на залежну змінну.
- **Приклад**: Вплив різних доз препарату на рівень глюкози.

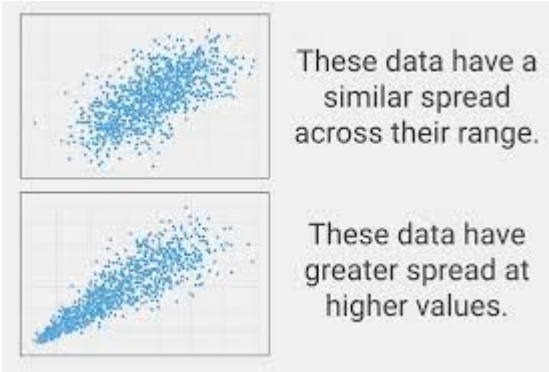
Двофакторний дисперсійний аналіз:

- Оцінює вплив на залежну змінну двох факторів одночасно.
- **Приклад**: Вплив дози препарату і типу харчування на рівень глюкози.

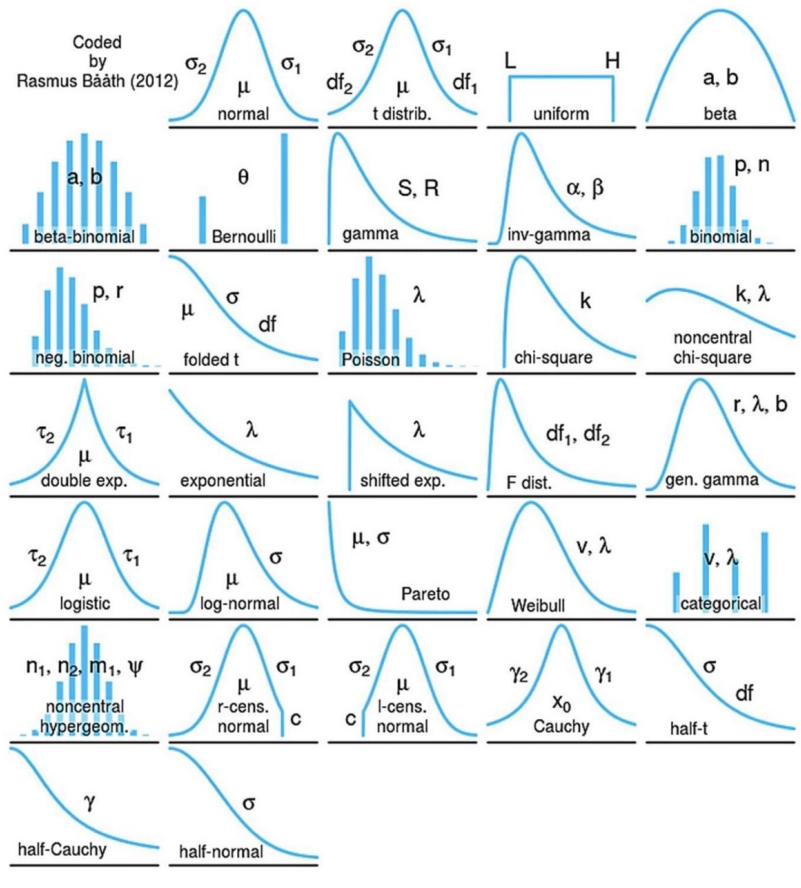
# Припущення дисперсійного аналізу

- Дані мають нормальний розподіл.
- Однорідність дисперсій (гомоскедастичність) між групами.
- Незалежність спостережень.



# Probability Distributions



При однофакторному дисперсійному аналізі вихідні дані подають у вигляді таблиць, у яких:

- кількість стовпчиків дорівнює кількості рівнів фактора;
- кількість значень у кожному стовпчику - кількості спостережень при відповідному рівні фактора.

Для різних рівнів фактора кількість спостережень може бути різною. При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів. Завданням аналізу є перевірка нульо вої гіпотези про рівність середніх значень сукупностей, що розглядаються.

| Результати вимірювань | Рівні фактора | | | |
|---|---|---|---|---|
| | 1 | 2 | ... | k |
| 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1k}$ |
| 2 | $x_{21}$ | $x_{22}$ | ... | $x_{2k}$ |
| ... | ... | ... | ... | ... |
| $n_i$ | $x_{ni1}$ | $x_{ni2}$ | ... | $x_{nik}$ |

***Форма таблиці спостережень при проведенні однофакторного дисперсійного аналізу***

# Формула дисперсійного аналізу

F-статистика:

***F = Міжгрупова дисперсія / Внутрішньогрупова дисперсія***

Якщо отримане значення F велике, це свідчить про те, що середні груп суттєво відрізняються.

# Однофакторний дисперсійний аналіз (ANOVA) – Приклад

Дані з дослідження про вплив трьох різних дієт на масу тіла щурів.

Етапи:

- Обчислення середніх для кожної групи.
- Розрахунок дисперсій.
- Обчислення F-статистики та її порівняння з критичним значенням.

# Інтерпретація результатів

Якщо **F-статистика > критичне значення**: Відхиляємо нульову гіпотезу (середні відрізняються).

Якщо **F-статистика < критичне значення**: Не відхиляємо нульову гіпотезу (середні не відрізняються).

## *Приклад: Вплив дієти (фактор) на зміну маси тіла (відгук)*

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Изменение веса** | | | | | | | | | |
| 2 | **Диета 1** | **Диета** | **Диета** | *Дисперсионный анализ: однофакторный* | | | | | | |
| 3 | 3,8 | 0 | 7 | | | | | | | |
| 4 | 6 | 0 | 5,6 | *ИТОГО* | | | | | | |
| 5 | 0,7 | -2,1 | 3,4 | *Группы* | *Количество* | *Сумма* | *Среднее* | *Дисперсия* | | |
| 6 | 2,9 | 2 | 6,8 | *Диета 1* | 24 | 79,2 | 3,300 | 5,018 | | |
| 7 | 2,8 | 1,7 | 7,8 | *Диета 2* | 27 | 81,7 | 3,026 | 6,367 | | |
| 8 | 2 | 4,3 | 5,4 | *Диета 3* | 27 | 139 | 5,148 | 5,739 | | |
| 9 | 2 | 7 | 6,8 | | | | | | | |
| 10 | 8,5 | 0,6 | 7,2 | | | | | | | |
| 11 | 1,9 | 2,7 | 7 | *Дисперсионный анализ* | | | | | | |
| 12 | 3,1 | 3,6 | 7,3 | *Источник дисперсии* | *Сумма квадратов* | *степень свободы* | *Квадрат среднего* | *F* | *Значение P* | *F критическое* |
| 13 | 1,5 | 3 | 0,9 | *Между группами* | 71,094 | 2,000 | 35,547 | 6,197 | 0,003 | 3,119 |
| 14 | 3 | 2 | 7,6 | *В группах* | 430,179 | 75,000 | 5,736 | | | |
| 15 | 3,6 | 4,2 | 4,1 | *Всего* | 501,273 | 77,000 | | | | |
| 16 | 0,9 | 4,7 | 6,3 | | | | | | | |
| 17 | -0,6 | 3,3 | 5 | | | | | | | |
| 18 | 1,1 | -0,5 | 2,5 | | | | | | | |
| 19 | 4,5 | 4,2 | 0,9 | | | | | | | |
| 20 | 4,1 | 2,4 | 3,5 | | | | | | | |
| 21 | 9 | 5,8 | 0,5 | | | | | | | |
| 22 | 2,4 | 3,5 | 2,8 | | | | | | | |
| 23 | 3,9 | 5,3 | 8,6 | | | | | | | |
| 24 | 3,5 | 1,7 | 4,5 | | | | | | | |
| 25 | 5,1 | 5,4 | 2,8 | | | | | | | |
| 26 | 3,5 | 6,1 | 4,1 | | | | | | | |
| 27 | | 7,9 | 5,3 | | | | | | | |
| 28 | | -1,4 | 9,2 | | | | | | | |
| 29 | | 4,3 | 6,1 | | | | | | | |
| 30 | 3,3 | 3,0 | 5,1 | | | | | | | |
| 31 | 2,24 | 2,56 | 2,44 | | | | | | | |

Непараметричним аналогом однофакторного дисперсійного аналізу є **ранговий однофакторний аналіз Краскела - Уолліса**. Він розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р.

Цей критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві. Нульова гіпотеза полягає в тому, що k вибірок обсягами n1 , n2 , ..., nk отримані з однієї і тієї самої генеральної сукупності. Критерій Краскела - Уолліса є узагальненням U-критерію Манна - Уїтні на випадок, коли кількість вибірок k > 2 .

***! Рангові методи, у тому числі й метод Краскела - Уолліса, не передбачають нормальності розподілу результатів спостережень і можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для***

Критерій **Джонкхієра (Джонкхієра - Терпстра)** запропонований незалежно один від одного нідерландським математиком Т. Дж. Терпстрою в 1952 р. й британським психологом Е. Р. Джонкхієром в 1954 р.

Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактора, який вимірюють у порядковій шкалі. Таблиця даних має такий самий вигляд, як і в попередньому випадку. Будемо вважати, що її перший стовпчик відповідає найменшому рівню фактора, другий - наступному за величиною тощо, останній стовпчик відповідає найбільшому рівню. При виконанні таких припущень критерій Джонкхієра є більш потужним, ніж критерій Краскела - Уолліса, стосовно гіпотези про монотонний вплив фактора.

# Пост-гок ( *post-hoc* ) аналіз

Якщо виявлено суттєві відмінності між групами, використовують пост-гок аналіз (наприклад, метод Тьюкі) для визначення, які саме групи відрізняються.

Приклад: Порівняння всіх пар середніх для точного визначення відмінностей.

# Двофакторний дисперсійний аналіз

**Двофакторний дисперсійний аналіз** застосовують для ***пов'язаних нормально розподілених*** вибірок. Дані подають у вигляді таблиці, у стовпчиках якої наводять дані, що відповідають певному рівню першого фактора, а в рядках - дані, що відповідають рівням другого. Таблиця даних має розмірність n x k , де n і k - кількість рівнів першого та другого факторів, відповідно.

Основною відмінністю від таблиці однофакторного дисперсійного аналізу є можлива неоднорідність даних у стовпцях, якщо вплив одного з факторів є більш суттєвим, ніж іншого.

На практиці часто використовують і складніші таблиці двофакторного дисперсійного аналізу, зокрема такі, у яких кожна комірка містить набір даних (повторні вимірювання), що відповідають фіксованим значенням рівнів обох факторів.

| Рівні | Рівні фактора B | | | |
|---|---|---|---|---|
| фактора A | 1 | 2 | ... | k |
| 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1k}$ |
| 2 | $x_{21}$ | $x_{22}$ | ... | $x_{2k}$ |
| ... | ... | ... | ... | ... |
| $n_i$ | $x_{ni1}$ | $x_{ni2}$ | ... | $x_{nik}$ |

# Застосування дисперсійного аналізу у біології

- Оцінка ефективності різних методів лікування.
- Аналіз впливу факторів навколишнього середовища на ріст і розвиток організмів.
- Дослідження поведінкових або фізіологічних реакцій тварин у різних умовах.

# Переваги та обмеження дисперсійного аналізу

**Переваги**:

- Можливість порівняння більше ніж двох груп одночасно.
- Простота застосування.

**Обмеження**:

- Вимагає дотримання припущень про нормальний розподіл і однорідність дисперсій.
- Не визначає, які саме групи відрізняються (потребує пост-гок аналізу).