Статистичні дослідження

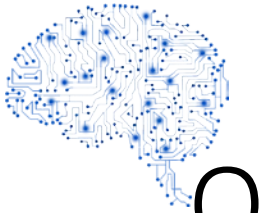
Професор, д.е.н. Ставицький А.В.

Що таке статистика?

Статистика - це вивчення збору, аналізу, інтерпретації, подання та організації даних.

Припущення:

- Спостереження - це значення випадкової величини
- Вибірка представляє сукупність, з якої вона відібрана



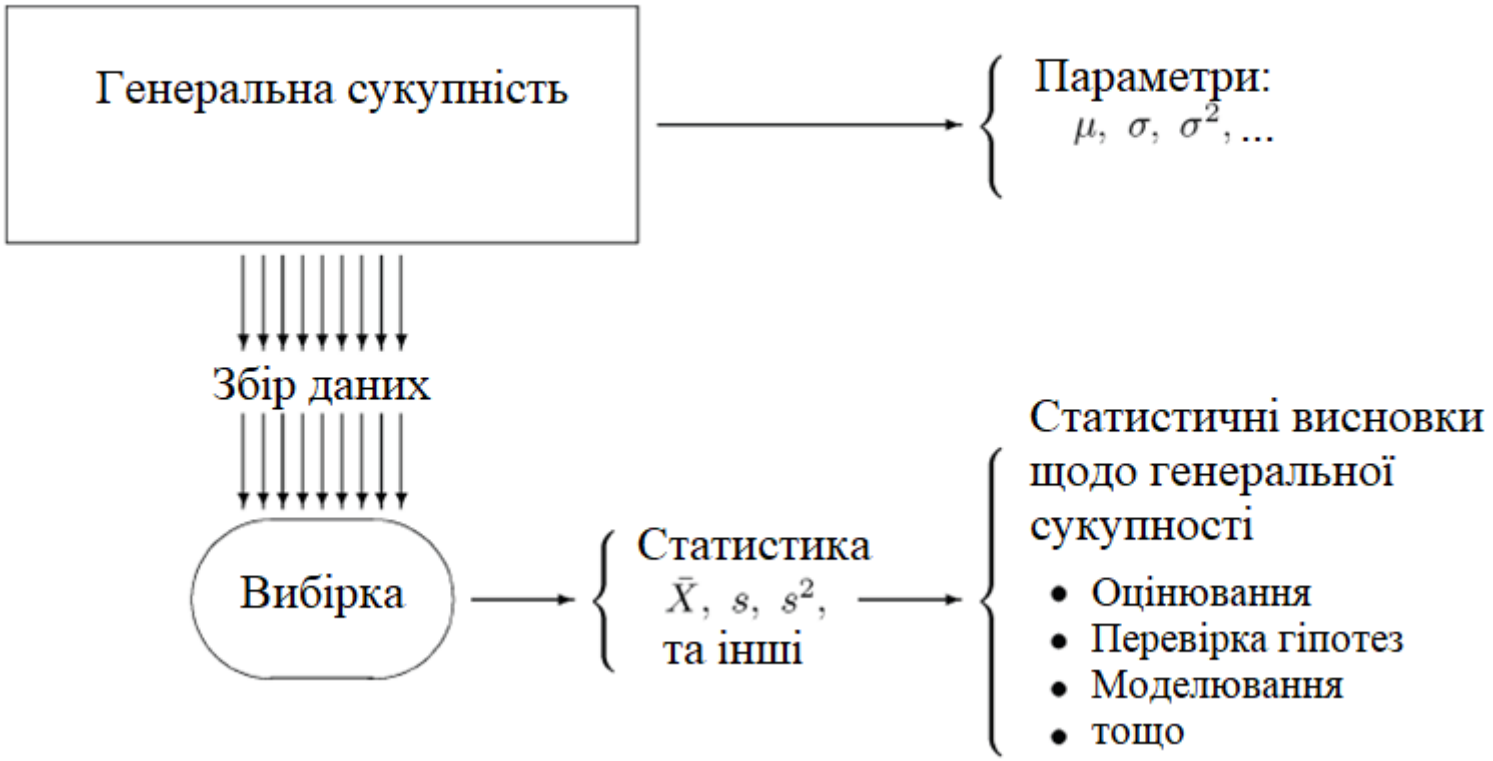
Основні поняття

- Генеральна сукупність: Набір об’єктів щодо яких слід зробити висновок
- Вибірка: репрезентативна частина / підмножина генеральної сукупності
- Випадкова вибірка: елементи сукупності, вибрані випадковим чином і незалежно один від одного

Приклад: Статистика орендної плати для міста Київ

- Населення: всі кімнати, квартири тощо, що здаються в оренду в Києві (занадто багато, щоб аналізувати всі)
- Вибірка: обрана частина; всі квартири з Дарниці
- Випадкова вибірка: дослідження n = 100, 200,… випадкових об'єктів з Києва

Генеральна сукупність та вибірка



Вимірювання «середини» даних

- Число, яке характеризує „центр” даних
- Найважливіші:
  - Вибіркове середнє
  - Медіана
  - Середнє гармонійне

Медіана

- Вибірка: $x_1, x_2, \dots, x_n \rightarrow$ Впорядковуємо за величиною: $x_1 \leq x_2 \leq \cdots \leq x_n$ $\rightarrow$ впорядкована вибірка : $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
- Медіана $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{якщо } n \text{ непарне}, \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{якщо } n \text{ парне} \end{cases}$

Вибірка	ранг
x_1=5	$x_{(1)}$=3
x_2 =9	$x_{(2)}$=4
x_3=3	$x_{(3)}$=5
x_4=8	$x_{(4)}$=6
x_5=19	$x_{(5)}$=8
x_6=4	$x_{(6)}$=9
x_7=6	$x_{(7)}$=19

n = 7 непарне:

$$\tilde{x} = x_{\left(\frac{7+1}{2}\right)} = x_{(4)} = 6$$

Вибірка	ранг
x_1=5	$x_{(1)}$=3
x_2 =9	$x_{(2)}$=4
x_3=3	$x_{(3)}$=5
x_4=8	$x_{(4)}$=6
x_5=19	$x_{(5)}$=7
x_6=4	$x_{(6)}$=8
x_7=6	$x_{(7)}$=9
x_8=7	$x_{(8)}$=19

n = 8 парне:

$$\tilde{x} = \frac{1}{2}\left[x_{\left(\frac{8}{2}\right)} + x_{\left(\frac{8}{2}+1\right)}\right]$$

$$= \frac{1}{2}\left[x_{(4)} + x_{(5)}\right]$$

$$= \frac{1}{2}[6 + 7] = 6.5$$

Медіана для інтервальної вибірки

$$Me = y_i + h_i \frac{\frac{n}{2} - \sum_{k=1}^{i-1} m_k}{m_i}$$

Вибіркове середнє

- Вибіркове середнє
  - Вибірка: $x_1, x_2, \dots, x_n$
  - Розмір вибірки: n

- Вибіркове середнє $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

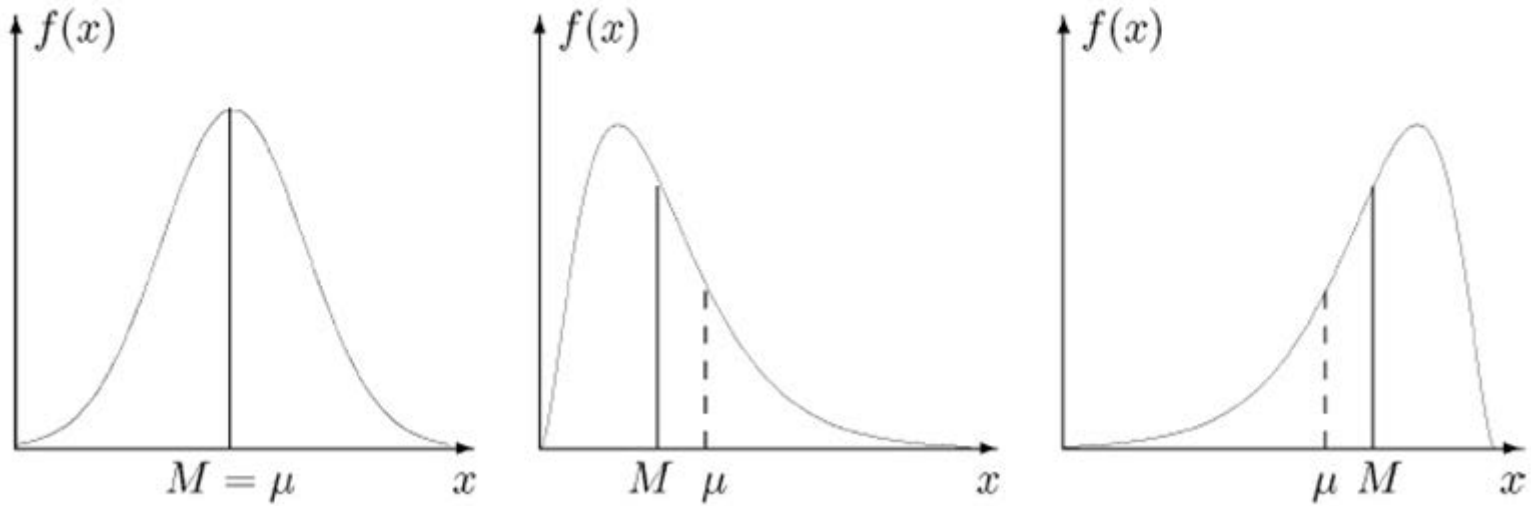
Порівняння медіани та вибіркового середнього

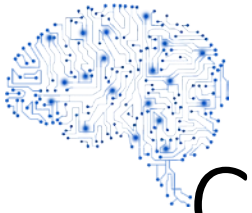
- Обидві вибірки мають медіану 2500
- $\bar{x} = 3000$ та $\bar{x'} = 5000$ - середні значення
- На середнє може сильно впливати одне значення

- Медіана є більш стійкою проти екстремальних значень („викидів”)
- Тим не менш, вибіркове середнє частіше використовується на практиці, оскільки має інші бажані властивості.

i	x_i	x'_i	відсортовані x_i та x'_i
1	2000	2000	1500
2	5000	15000	2000
3	4000	4000	2500
4	1500	1500	4000
5	2500	2500	5000 / 15000

Виібркове середнє vs. медіана





Середнє гармонійне

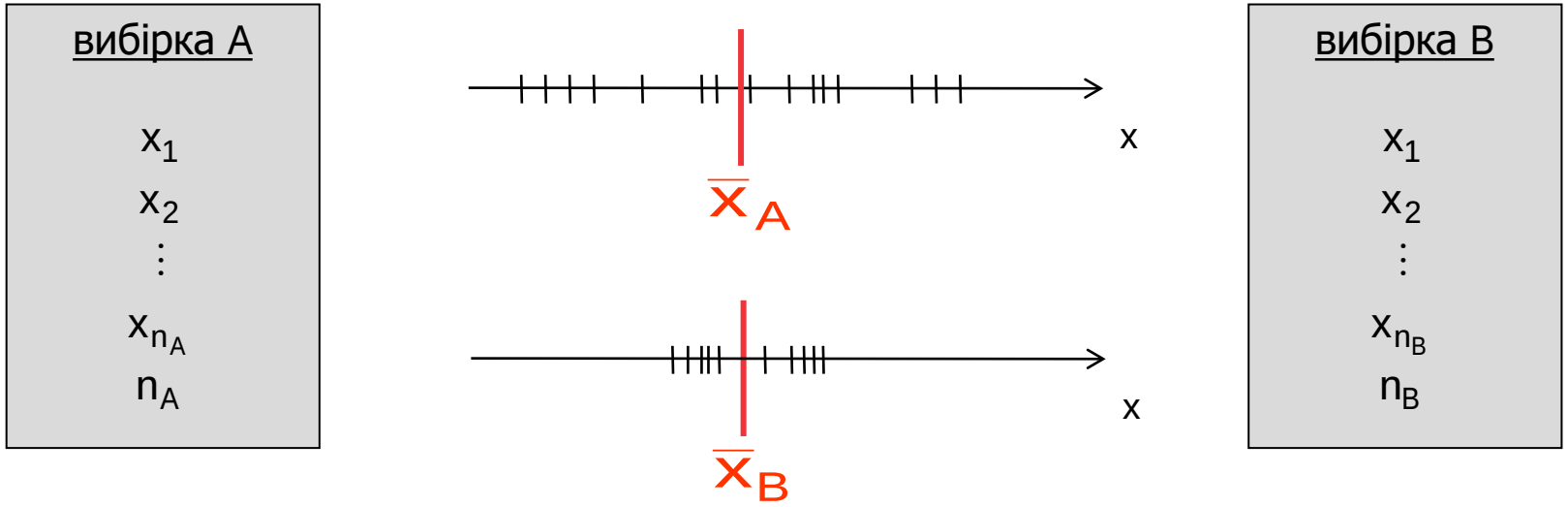
- Гармонійне середнє значення є меншим за арифметичне середнє і має тенденцію підкреслювати вплив малих викидів, при цьому мінімізуючи вплив великих викидів.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

- Цікаво! Гармонійне середнє використовується для розрахунку оцінки F, що є мірою точності тестів. Оцінка F визначається як середньозважене гармонічне часток успіхів (p) та невдач (r) при тестуванні:

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Волатильність даних



→ Вибіркове середнє (або медіана) недостатнє для опису вибірки

Вимір розкиду змінних

- Найважливіші:
  - Мінімум, максимум, діапазон
  - Вибіркова дисперсія
  - Вибіркове стандартне відхилення

Дисперсія і стандартне відхилення

- Міра виразити розкид навколо центру (середнє значення) єдиним значенням:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Емпіричне стандартне відхилення s - просто квадратний корінь дисперсії,

$$\mathrm{s} = \sqrt{s^2}$$

Навіщо ділити на $n - 1$ замість n?

Приклад:

x_1
x_2
⋮
x_n
n
$\bar{x}$

$x_1 = 75$

$x_2 = 2$

$x_3 = 270$

$x_4 = 4 \cdot 100 - 75 - 2 - 270 = 53$

$n = 4$

$\bar{x} = 100$

x_4 =53 не вільне, але задане іншими значеннями, коли середнє значення відоме.

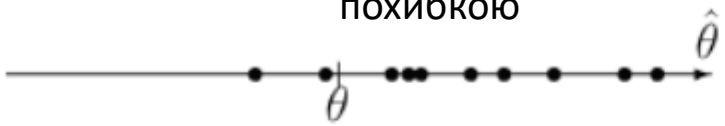
→ s^2 має (n-1) ступенів свободи (f)

→ $s^2 = \frac{1}{f} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$

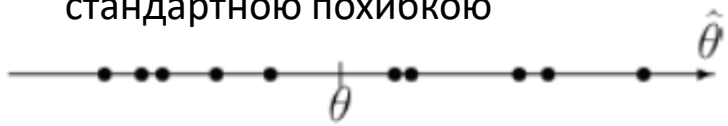
Стандартні помилки оцінок

Для оцінювача T для параметра θ його стандартна похибка - Std (T), і це вказує на точність і надійність T

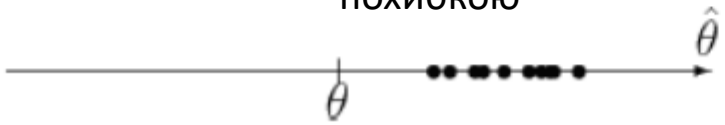
Зміщена оцінка з великою стандартною похибкою



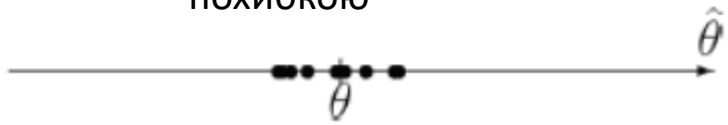
Незміщена оцінка з великою стандартною похибкою

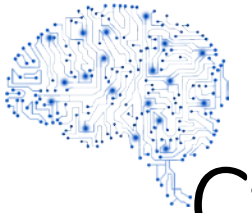


Зміщена оцінка з малою стандартною похибкою



Незміщена оцінка з малою стандартною похибкою





Статистичні гіпотези

Статистичний тест використовує дані з вибірки для оцінки параметра генеральної сукупності

Статистичні тести містять дві конкуруючі гіпотези:

- Нульова гіпотеза (H_0): Стверджує, що впливу певного фактору немає.
- Альтернативна гіпотеза (H_a): доповнююча гіпотеза, за якої вплив суттєвий.

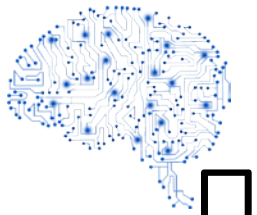
Гіпотези завжди стосуються параметрів генеральної сукупності

Приклад: Сон проти кофеїну - 1

Студентам давали запам’ятати слова, а потім випадковим чином їх дали або спати 90 хвилин, або таблетку кофеїну. Через 2 з половиною години їх перевіряли на здатність до згадування слів, які вони вчили.

- Пояснювальна змінна: сон або кофеїн
- Залежна змінна: кількість згаданих слів

Що краще для пам’яті: сон чи кофеїн?

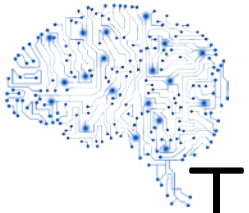


Приклад: Сон проти кофеїну - 2

- Нехай μ_s та μ_c – вибіркове середнє кількості згаданих слів після сну та кофеїну відповідно.
- Чи є різниця в середньому згадуванні слова між сном та кофеїном?
- Які нульові та альтернативні гіпотези?

- $H_0: \mu_s \neq \mu_c$, $H_a: \mu_s = \mu_c$
- $H_0: \mu_s = \mu_c$, $H_a: \mu_s \neq \mu_c$
- $H_0: \mu_s \neq \mu_c$, $H_a: \mu_s > \mu_c$
- $H_0: \mu_s = \mu_c$, $H_a: \mu_s > \mu_c$
- $H_0: \mu_s = \mu_c$, $H_a: \mu_s < \mu_c$

Нульова гіпотеза - «немає різниці» або про те, що засоби рівні. Альтернативна гіпотеза полягає в тому, що є різниця.



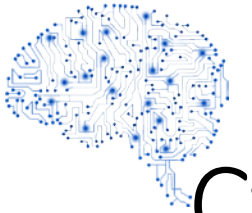
Тестування гіпотез

- Процес прийняття суджень про велику групу (сукупність) на основі невеликої підмножини цієї групи (вибірки) відомий як статистичний висновок.
- Тестування гіпотез, одне з двох полів статистичного висновку, дозволяє нам об'єктивно оцінити ймовірність того, що твердження про сукупність є істинними. Оскільки ці твердження мають ймовірнісний характер, ми ніколи не можемо бути впевнені в їх правдивості.
- Етапи тестування гіпотез
  - Формування гіпотези
  - Визначення відповідної статистики тесту та його розподілу ймовірностей.
  - Визначення рівня значущості.
  - Встановлення правила прийняття рішення.
  - Збір даних та обчислення статистики тесту.
  - Прийняття статистичного рішення.
  - Прийняття економічного чи інвестиційного рішення.

Помилки в тестах гіпотези

- Помилки типу I трапляються, коли ми відкидаємо нульову гіпотезу, яка насправді є правдою.
- Помилки типу II трапляються, коли ми не відкидаємо нульову гіпотезу, що є помилковою.

Рішення	**Реальна ситуація**	
	H_0: True	**H_0: False**
Не відхиляти	Вірне рішення	Помилка II роду
Відхиляти	Помилка I роду	Вірне рішення



Статистична значимість

- Коли параметр для перевірки суттєво відрізняється результатів у самій вибірці, ми вважаємо, що результати вибірки є статистично значущими
- Якщо наш результат є статистично значущим, ми маємо переконливі докази проти H_0 на користь H_a
- Якщо наш вибірка не є статистично значущим, наш тест є непереконливим

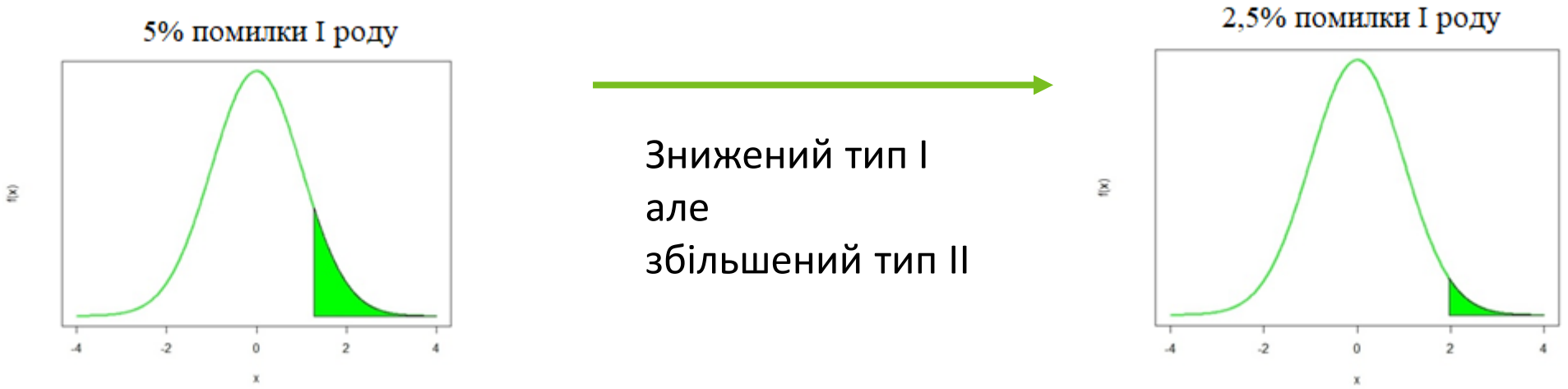
Визначення рівня значущості

- Рівень значущості ідентичний рівню помилки типу I, його часто називають "альфа".
- Рівень довіри до статистичних результатів безпосередньо пов'язаний з рівнем значущості тесту і, таким чином, з ймовірністю помилки типу I.

Рівень значущості	Пропонований опис
0.10	"деякий доказ"
0.05	“сильний доказ”
0.01	“дуже сильний доказ”

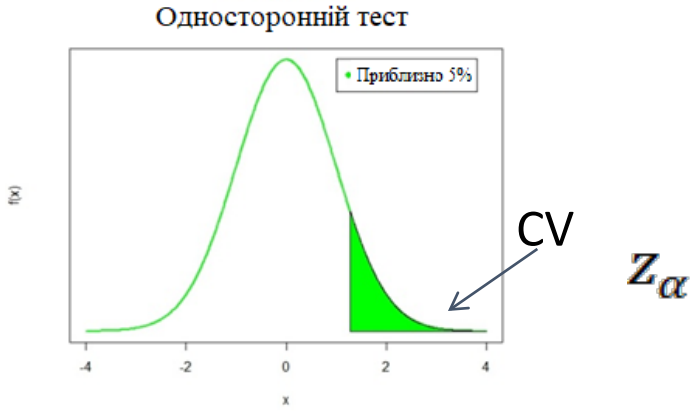
Обернений зв’язок при тестуванні гіпотез

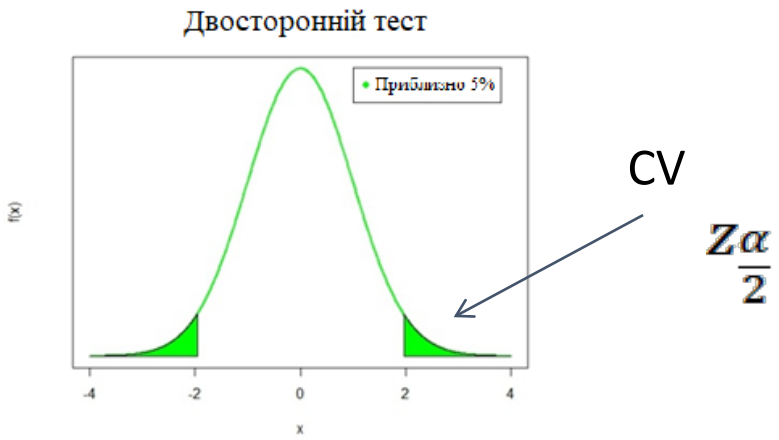
- Якщо ми зменшимо ймовірність помилки типу I, вказавши менший рівень значущості, збільшуємо ймовірність помилки типу II.
- Єдиний спосіб зменшити ймовірність обох помилок одночасно - збільшити розмір вибірки.

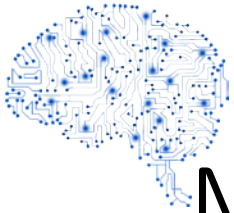


Встановлення правила прийняття рішення

- Правило прийняття рішення використовує рівень значущості та розподіл ймовірності тестової статистики для визначення значення, вище (нижче) якого нульова гіпотеза відхиляється.
- Критичне значення (CV) тестової статистики - це значення вище (нижче), яке нульова гіпотеза відкидається.
- Односторонні тести позначаються підписом α.
- Двосторонні тести позначаються підписом α / 2.







Можливі гіпотези

- Тестування вибіркового середнього

$$H_0: \mu=\mu_0$$
$$H_1: \mu\neq\mu_0$$

- Тестування різниці середніх

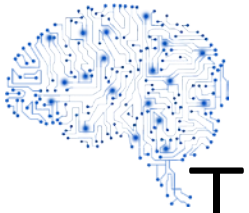
$$H_0: \mu_1-\mu_2=\mu_0$$
$$H_1: \mu_1-\mu_2\neq\mu_0$$

- Тестування дисперсії

$$H_0: \sigma^2=\sigma_0$$
$$H_1: \sigma^2\neq\sigma_0$$

- Тестування рівності дисперсій

$$H_0: \sigma_1^2=\sigma_2^2$$
$$H_1: \sigma_1^2\neq \sigma_2^2$$

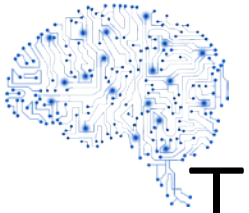


Тестування вибіркового середнього – 1

H_0: $\mu=\mu_0$

H_1: $\mu\neq\mu_0$

- Ми майже ніколи не знаємо дисперсію генеральної сукупності, і в таких випадках тести середнього проводяться через t-тести або z-тести.



Тестування вибіркового середнього – 2

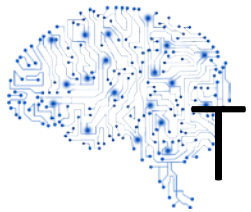
- Тести, що порівнюють виібркове середнє з заданим значенням:
- Використовуйте t-тест з df = n - 1, коли
  - Дисперсія генеральної сукупнсоті невідома або
  - Вибірка велика або
  - Вибірка мала, але приблизно нормально розподілена.

$$t_{n-1} = \frac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}$$

- Використовуйте z-тест, якщо
  - Вибірка велика або
  - Генеральна сукупність має нормальний розподіл

$$z = \frac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}$$

- Зауважимо, що два тести використовують вибіркове стандартне відхилення як оцінку стандартного відхилення генеральної сукупності.



Тестування вибіркового середнього: Приклад – 1

- Ви зібрали дані про щомісячну дохідність власного капіталу та визначили, що середня доходність протягом 48-місячного періоду, який ви перевіряєте, становила 12,94% при стандартному відхиленні 15,21%. Ви хочете перевірити, чи дорівнює ця середня віддача 15-відсотковій доходності, яку досягли Ваші конкуренти. Ви хочете бути на 95% впевнені у своїх результатах.

Тестування вибіркового середнього: Приклад – 2

- Формулюємо гіпотезу
  - H_0: θ = 15%
  - H_a: θ ≠ 15% (двосторонній тест).
- Визначаємо відповідну статистику тесту
  - t-тест, оскільки невідома дисперсія генеральної сукупності.
- Вказуємо рівень значущості
  - 0.05 за умовою задачі
- Обраховуємо критичне значення
  - t=2.01174.
- Збираємо дані з здачі та обраховуємо статистику:
  - $t_{n-1} = \frac{\bar{X} - \mu_0}{\hat{s}/\sqrt{n}}$,

$$t_{48-1} = \frac{0.1294 - 0.15}{0.1521/\sqrt{48}} = -0.938$$

- Робимо статистичний висновок
  - Не відхиляти нульову гіпотезу.

Тестування різниці середніх: Незалежні вибірки

H_0: $\mu_1-\mu_2=\mu_0$

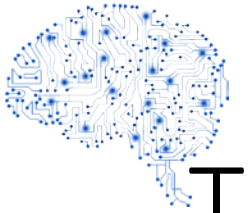
H_1: $\mu_1-\mu_2\neq\mu_0$

- Нормально розподілені, рівні, але невідомі відхилення
  - Використовуємо спеціальну оцінку дисперсії, s_p^2, яка є зваженою середньою вибіркових дисперсій.

$$t=\frac{(\bar{X}_1-\bar{X}_2)-\mu_0}{\sqrt{\left(\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}\right)}} \qquad s_p^2=\frac{(n_1-1)\hat{s}_1^2+(n_2-1)\hat{s}_2^2}{n_1+n_2-2} \qquad \mathrm{df}=n_1+n_2-2$$

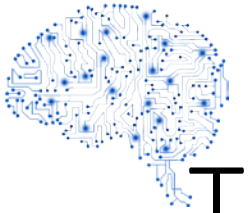
- Нормально розподілені, неоднакові та невідомі дисперсії
  - Використовуємо іншу оцінку дисперсії

$$t=\frac{(\bar{X}_1-\bar{X}_2)-\mu_0}{\sqrt{\left(\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}\right)}} \qquad \mathrm{df}=n_1+n_2-2$$



Тестування різниці середніх: Приклад – 1

- Ви вирішили дослідити, чи буде покращена прибутковість пенсійного портфеля вашого клієнта шляхом додавання іноземних акцій. Відповідно, спершу ви хочете перевірити, чи мають іноземні акції такий же прибуток, як внутрішні акції, перш ніж продовжувати аналіз.
- Американські акції принесли 12,94% із стандартним відхиленням 15,21% за попередні 48 місяців. Ви визначили, що за той же період європейські акції принесли 17,67% зі стандартним відхиленням 16,08%. Ви хочете того ж рівня впевненості в цьому результаті (5%).
- Ви готові припустити, що зараз дві вибірки є незалежними, приблизно нормально розподіленими та взятими з сукупності з однаковою базовою дисперсією.



Тестування різниці середніх: Приклад – 2

- Формулюємо гіпотезу
  - H_0: $\mu_{USAEq} - \mu_{EUEq} = 0$
  - H_a: $\mu_{USAEq} - \mu_{EUEq} \neq 0$
- Визначаємо відповідну статистику тесту та його розподіл ймовірностей
  - t-тест для нерівних середнії з нормальним розподілом і невідомими, але рівними дисперсіями
- Визначаємо рівень значущості та критичне значення
  - α=0.05, CV = –1.986
- Вказуємо правило прийняття рішення
  - Відхилити нульову гіпотезу, якщо |TS| > 1.986

Тестування різниці середніх: Приклад – 3

- Збираємо дані та обчислюємо статистику тесту →

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \mu}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$

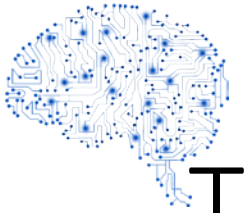
$$s_p^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{(47)0.1521^2 + (47)0.1608^2}{48 + 48 - 2}$$

$$t = \frac{(0.1294 - 0.1767)}{\sqrt{\left(\frac{0.0242^2}{48} + \frac{0.0242^2}{48}\right)}} = -1.4806$$

$$\mathrm{df} = 48 + 48 - 2$$

- Робимо висновок→ Не відхиляти нульову гіпотезу



Тестування дисперсії

H_0: $\sigma^2=\sigma_0$

H_1: $\sigma^2\neq\sigma_0$

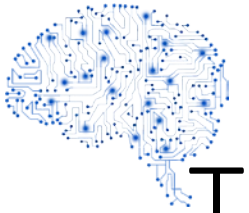
- Тести щодо дисперсії
  - Нормальний розподіл генеральної сукупності
  - Хі-квадрат тест з df = n - 1
  - Дуже чутливі до основних припущень
  - Статистика тесту $$\chi^2_{n-1}=\frac{(n-1)\hat{s}^2}{\sigma_0^2}$$

Тестування дисперсії: приклад

- Чи відрізняється статистично дохідність внутрішнього капіталу у нашому попередньому прикладі (15,21%) від 10%?

$$\chi^2 = \frac{(n-1)\hat{s}^2}{\sigma_0^2} \qquad \chi^2 = \frac{47 \cdot 15.21^2}{10^2} = 108.73$$

- Критичне значення для α = 5% 64.0011 → Відхилити нульову гіпотезу

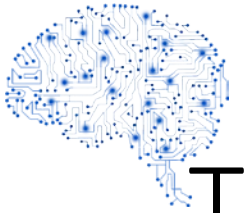


Тестування рівності дисперсій

H_0: $\sigma_1^2=\sigma_2^2$

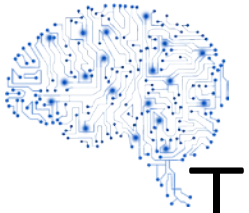
H1: $\sigma_1^2 \neq \sigma_2^2$

- Якщо тестуються дві нормально розподілені сукупності, то тест співвідношення двох дисперсій буде відповідати F- розподілу Фішера.
  - $F(\mathrm{df}_1, \mathrm{df}_2) = \frac{\hat{s}_1^2}{\hat{s}_2^2}$
  - $\mathrm{df}_i = n_i - 1$
- Якщо статистика тесту більше критичного значення для F-розподілу з df_1 і df_2 ступенями свободи, нульова гіпотеза відхиляється.



Тестування рівності дисперсій: приклад

- Повернемося до нашого попереднього прикладу, порівнюючи дохідність європейського та американського капіталу. У прикладі ми припускали, що дисперсії були рівними. Проведіть необхідний тест для оцінки обґрунтованості цього припущення. У нас було 48 спостережень за кожним портфелем акцій, прибутковість європейського капіталу мала стандартне відхилення 16,08%, а американського - 15,21%.



Тестування рівності дисперсій: приклад

- Формулюємо гіпотезу
  - H_0: $\sigma_{USAEq} = \sigma_{EUEq}$
  - H_a: $\sigma_{USAEq} \neq \sigma_{EUEq}$
- Визначаємо відповідну статистику тесту та розподіл ймовірностей
  - F-тест для відношення дисперсій
- Визначаємо рівень значущості
  - CV = 1.6238
- Визначаємо правило прийняття рішення
  - Відхилити нульову гіпотезу, якщо TS > 1.6238
- Збираємо дані та обраховуємо статистику

$$F(\mathrm{df}_1, \mathrm{df}_2) = \frac{\hat{s}_1^2}{\hat{s}_2^2} \qquad \mathrm{df}_i = n_i - 1 \qquad F\left(47,47\right) = \frac{0.1608^2}{0.1521^2} = 1.1177$$

- Робимо висновок → не можемо відхилити нульову гіпотезу

Приклад для MS Excel

- Надайте описову статистику для двох вібирок
- Проведіть тестування гіпотез

- Відео

Але…

- В усіх формулах в знаменнику є кількість елементів.
- Якщо n<100, 200, … 1000, формули працюють коректно
- Але якщо n>10000000, 10000000000, … ?

Питання?