Тестування статистичних гіпотез

Перевірка статистичних гіпотез / Hypothesis testing

Перевірка гіпотез – це статистичний метод, який використовується для прийняття статистичних рішень із використанням експериментальних даних. Перевірка гіпотез — це переважно припущення, що ми робимо про параметри генеральної сукупності.

Приклад: ви кажете, що середньому учневі в класі 10 років або що хлопчики в середньому вищі за дівчат.

Ми припускаємо, що ці приклади потребують деякого статистичного способу доказу. Нам потрібно математично довести, що висновок відповідає дійсності.

Адже припустити можна будь-що :)

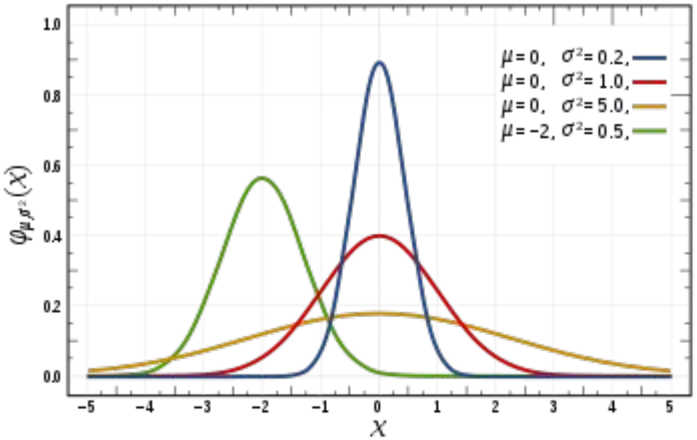
Перевірка статистичних гіпотез

Перевірка гіпотез оцінює **два взаємовиключні твердження про сукупність**, щоб визначити, яке твердження найкраще підтверджується вибіркою даних. Коли говоримо, що результат є **статистично значущим**, це відбувається завдяки перевірці гіпотез.

Перевірка статистичних гіпотез

Розглянемо різні нормальні розподіли. У них різні середні та відхилення.

Ми можемо **математично** дати відповідь на питання, наскільки статистично значуща відмінність між двома розподілами.



Нульова гіпотеза H0

Статистичною називають гіпотезу про властивості генеральної сукупності, що перевіряється на основі вибірки.

Нульова гіпотеза — це певна позиція за замовчанням; твердження, що немає різниці між двома вимірюваними явищами/групами.

Іншими словами, це базове припущення, засноване на знанні предметної галузі чи проблеми.

Приклади:

- компанія виробляє в середньому 50 одиниць продукції на день;
- конверсія перегляду в клік на сайті 10%;
- кожен третій покупець робить повторну покупку на нашому сайті.

Альтернативна гіпотеза H1

Альтернативна гіпотеза — це гіпотеза, яка використовується під час перевірки гіпотез, яка суперечить нульовій гіпотезі. Зазвичай вважається, що спостереження є наслідком реального ефекту (з накладенням певної кількості випадкових варіацій).

Приклади:

- компанія виробляє в середньому понад 50 одиниць продукції на день
- конверсія перегляду в клік на сайті 15%
- кожен другий покупець робить повторну покупку

Статистична значущість

Вказавши нульову та альтернативну гіпотези, ми повинні встановити рівень значущості, на якому ми шукаємо ефект.

Завдання перевірки статистичної значущості полягає в тому, щоб встановити поріг, за межами якого ми вирішуємо, що дані більше не підтримують нульову гіпотезу.

Отже, існує два ризики:

➔ Ми можемо прийняти розбіжність як значущу, коли насправді вона виникла випадковим чином
➔ Ми можемо приписати розбіжність випадковості, коли насправді вона показує справжню розбіжність

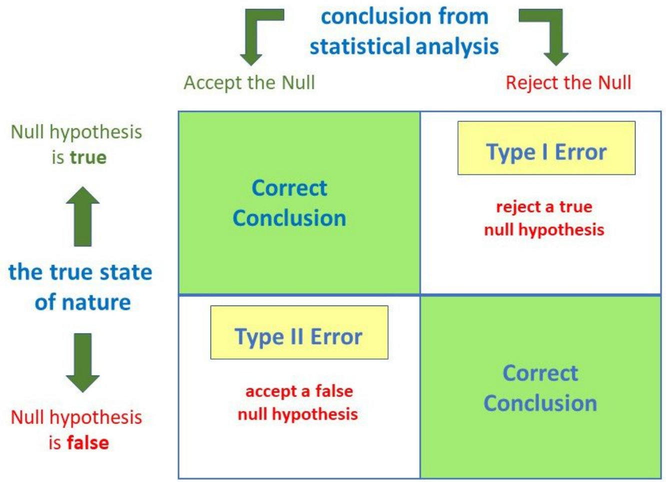
Ці дві можливості називаються відповідно помилками 1-го та 2-го роду.

Статистична значущість

Чим більше ми *зменшуємо* ризик помилок 1-го роду, тим більше ми *збільшуємо* ризик помилок 2-го роду.

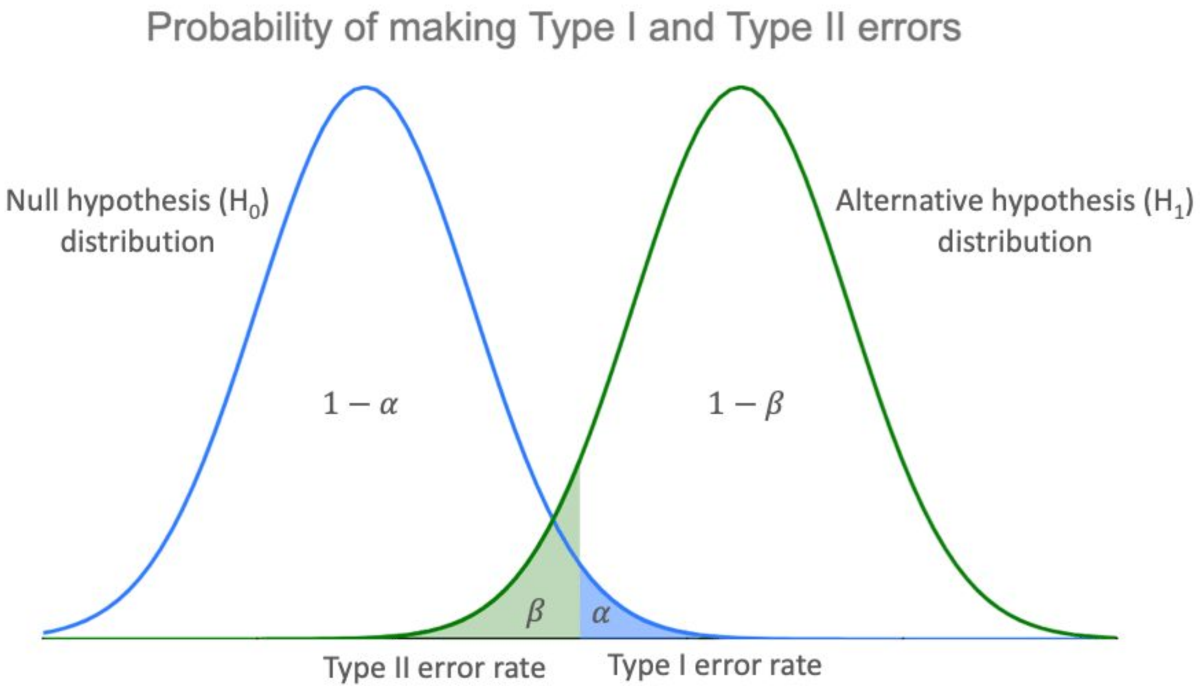
Іншими словами, з чим більшою впевненістю ми хочемо не заявляти про наявність розбіжності, коли її немає, тим більша розбіжність між вибірками нам буде потрібна, щоб заявити про статистичну значущість.

Стат. значимість (позначається літерою “альфа”) - ймовірність хибно позитивного результату (помилки першого роду, що ми відкинули правильну гіпотезу). Результат хибно позитивний щодо альтернативної гіпотези.



Статистична значущість

Так зображують ймовірності помилки першого та другого роду графічно.



Статистичні тести

Статистичні тести

Статистичний тест — це інструмент для перевірки статистичної гіпотези.
Існують наступні найбільш поширені статистичні тести.

Тестування гіпотези	**Тестовая статистика**	**Правило відхилення гіпотези**
z-тести	z-статистика	Якщо тестова статистика ≥ z чи ≤ -z, то відхилити нульову гіпотезу H0.
t-тести	t-статистика	Якщо тестова статистика ≥ t або ≤ -t, то відхилити нульову гіпотезу H0.
Аналіз дисперсії (ANOVA)	F-статистика	Якщо тестова статистика ≥ F, то відхилити нульову гіпотезу H0.
Тести хі-квадрат	Статистика хі-квадрат	Якщо тестова статистика ≥ χ, то відхилити нульову гіпотезу H0.

Z-критерій Фішера / Z-test

Z-тест (z-критерій Фішера) — *клас методів* статистичної перевірки гіпотез, заснованих на нормальному розподілі. Зазвичай застосовується для перевірки рівності середніх значень за відомої дисперсії генеральної сукупності або оцінки вибіркового середнього стандартизованих значень. Z-статистика обчислюється як відношення різниці між випадковою величиною та математичним очікуванням до стандартної помилки цієї випадкової величини:

$$z = \frac{\overline{X} - m}{\mathrm{SE}}$$

де

X з рисочкою — випадкова величина вибіркового середнього,
m - значення математичного очікування,
SE – стандартна помилка цієї величини.

Коли використовуємо z-test

- ➔ Розмір вибірки має бути більшим за 30. В іншому випадку слід використовувати **t-тест**.
- ➔ Вибірки слід відбирати випадковим чином із сукупності.
- ➔ Маємо знати стандартне відхилення генеральної сукупності.
- ➔ Вибірки, взяті із сукупності, мають бути **незалежними** одна від одної.
- ➔ Дані мають бути **нормально розподілені**. Проте у нас завжди є спосіб перейти до нормального розподілу середнього вибірки згенерувавши багато семплів згідно з центральною граничною теоремою *. Щоправда, нам звісно не завжди підійде дослідження середнього замість дослідження оригінально вибірки.

* Тут можна дізнатись про це детальніше https://www.scribbr.com/statistics/central-limit-theorem/

Кроки для виконання z-test

1. Спочатку визначте нульову та альтернативну гіпотези.
2. Визначте рівень значущості альфа.
3. Знайдіть критичне значення z за встановленого рівня стат. значущості альфа, використовуючи таблицю.
4. Знайдіть z значення для нашого тесту за формулою для розрахунку статистики z-тесту.
5. Тепер порівняйте значення з п. (3) та (4).

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Пояснення на наступному слайді

Стандартне відхилення вибіркового середнього

Ми шукаємо випадкове відхилення не випадкової величини, а вибіркового середнього:

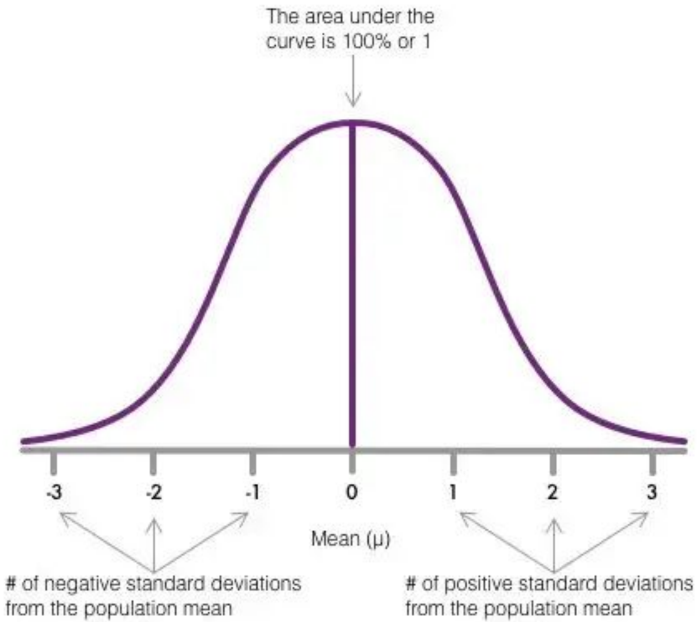
$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}$$

Докладніше про те, чому ми ділимо на sqrt(n), можна почитати тут і тут.

Приклад проведення z-тесту

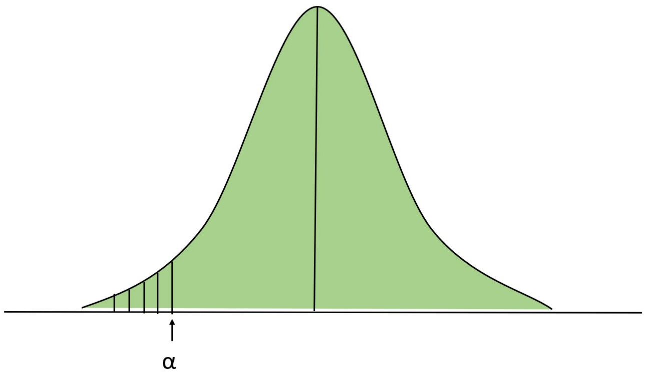
Задача: В одній школі стверджували, що їх учні навчаються краще ніж у середній школі. При підрахунку балів IQ 50-ти учнів середнє значення дорівнювало 110. Середнє значення IQ популяції становить 100, а стандартне відхилення - 15. Вкажіть, чи є твердження директора школи правильним, чи ні, на рівні значущості 5%.

Різносторонні z-тести



Лівосторонній z-test

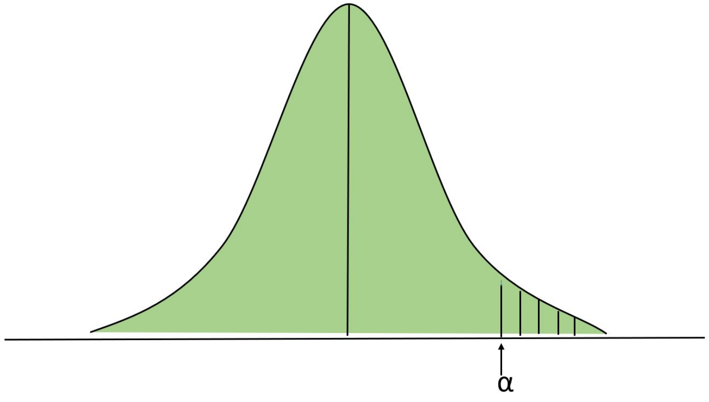
У цьому тесті наша область відхилення нульової гіпотези знаходиться у крайньому **лівому** кутку розподілу. Тут наша нульова гіпотеза полягає в тому, що заявлене значення **більше** середнього значення для генеральної сукупності.



Left-tailed Test

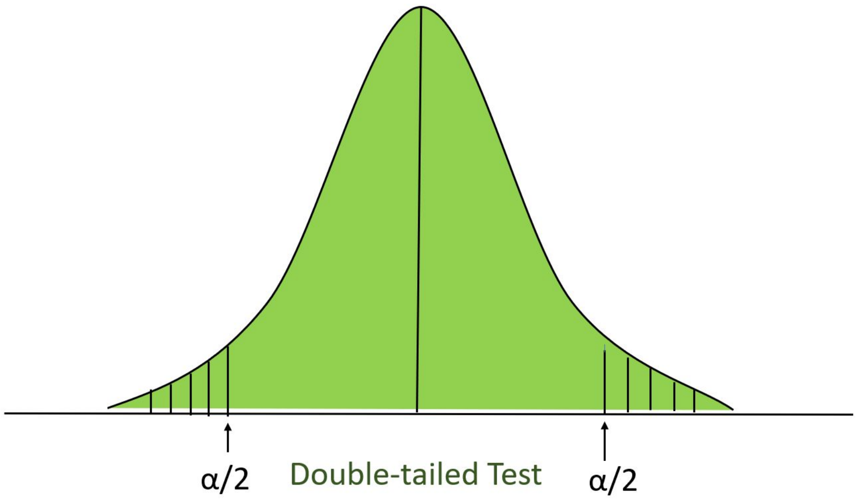
Правосторонній z-test

У цьому тесті наша область відхилення знаходиться у крайньому **правому** кутку розподілу. Тут наша нульова гіпотеза у тому, що заявлене значення **менше або дорівнює** середнього значення для генеральної сукупності.

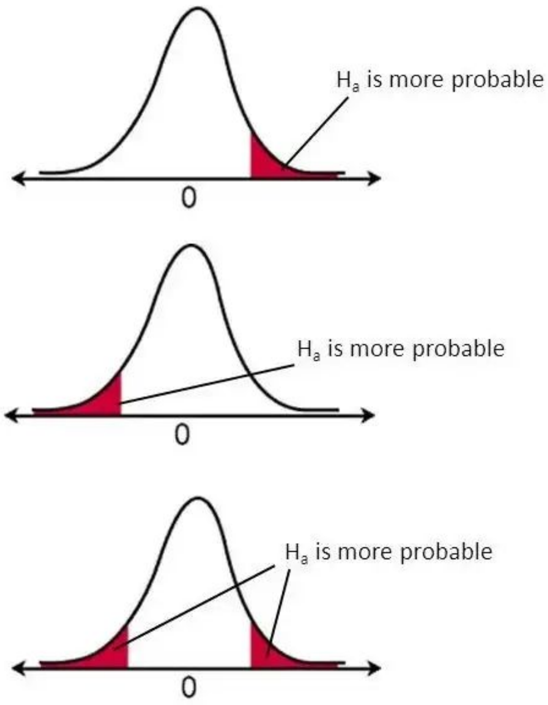


Двосторонній z-test

У цьому тесті наша область відхилення знаходиться в обох точках розподілу. Тут наша нульова гіпотеза полягає в тому, що заявлене значення дорівнює середньому значенню для сукупності.



Різносторонні z-тести: підсумуємо



Right-tail test

H_a: $\mu >$ value

Left-tail test

H_a: $\mu <$ value

Two-tail test

H_a: $\mu \neq$ value

[Розділ](#) з американського підручника зі статистики по тестуванню гіпотез для глибшого ознайомлення

p-value

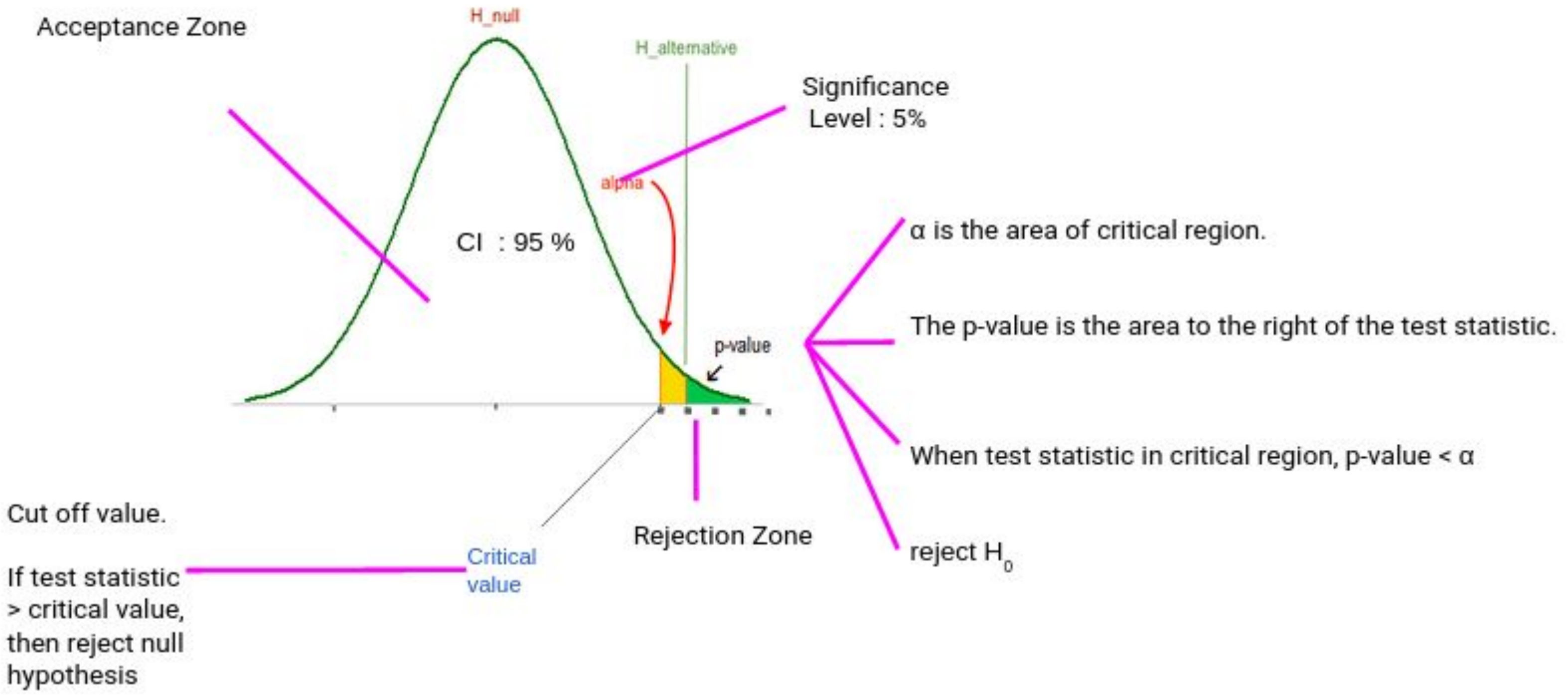
Р-значення — це можливість отримати таке ж або більше "екстремальне значення" за умови правдивості нульової гіпотези. Грубо кажучи р-значення — це можливість того, що H0 - правдива.

Якщо p-значення невелике, це означає, що таке або більш екстремальне значення отримати при нульовій гіпотезі **малоймовірно**.

В A/B тестах чим **більша** різниця між значеннями, що спостерігаються, тим **менше** p-значення.

Порівнюють p-значення зазвичай з рівнем значущості 0.05. Якщо воно менше, то ми відхиляємо H0. Інакше ми НЕ можемо відхилити H0 (fail to reject пишуть у статтях).

P-value: графічне зображення



[Hypothesis testing]

Приклад проведення z-testy

Розв'язок задачі з прикладу:

1. Визначаємо гіпотези: $H_0 : \mu = 100 \quad H_A : \mu > 100$
2. Задаємо рівень стат. значущості як 5%.
3. Тепер подивимось на z-таблицю. Для значення альфа = 0.05 z-оцінка правостороннього тесту становить 1.645. І тут ми можемо стверджувати, що 95% середніх студентів мають IQ нижче 110.
4. Обчислюємо нашу Z-статистику за формулою, де
   - ➔ X = 110
   - ➔ Середнє (mu) = 100
   - ➔ Стандартне відхилення (сигма) = 15
   - ➔ Рівень значущості (альфа) = 0.05
   - ➔ n = 50
5. Тут 4,71 > 1,645, тому ми **відкидаємо** нульову гіпотезу, бо занадто велика різниця між передбачуваним базовим розподілом та спостереженнями в експерименті. Якщо статистика z-критерію менша, ніж z-score, ми не відкидатимемо нульову гіпотезу.

Розрахунок z-статистики:

$$\frac{(110-100)}{15/\sqrt{50}}$$

$$\frac{10}{(15/sqrt(50))}$$

$$\frac{10}{2.12}$$

$$4.71$$

Приклад z-таблиці

Для пошуку Z-значення на певному рівні значущості нам треба знайти в стовпці, де рівень значущості рівний заданому, рядок, де ймовірність більша за 0.95.

Або, якщо ми знаємо Z-значення отримане на вибірці, ми можемо зрозуміти, який відсоток популяції мав значення величини, що спостерігається, менше, ніж задане. У попередньому прикладі ми могли поцікавитися, який відсоток популяції має IQ менше 110. Враховуючи, що z-значення вийшло у нас 4.71, то ймовірність того, що в середній школі IQ менше 110 за стандартного відхилення 15 - вище 0.9997, виходячи з таблиці збоку.

0 Z

Hundredths Digits

Tenths Digits	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Подивимося, як це можна реалізувати в Python

Зімітуємо експеримент подібний до наведеного зі студентами та їх рівнем IQ, та проведемо процедуру тестування гіпотези.

Ноутбук: **Lecture 7. Statistics.ipynb**

Тести на двох вибірках

Двохвибірковий z-test

У цьому тесті нам надано 2 нормально розподілені та незалежні популяції, і ми довільним чином відібрали вибірки з обох популяцій. Тут ми вважаємо mu1 і mu2 середніми генеральної сукупності, X1 і X2 - середніми, що спостерігаються за вибіркою. Тут наші гіпотези можуть бути такими:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Двохвибірковий z-test

Формула Z-статистики

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

де sigma1 і sigma2 – стандартні відхилення, а n1 та n2 – розміри вибірок популяції, mu1 та mu2 – значення математичних очікувань.

t-тест та z-тест

Порівнюючи середні двох незалежних вибірок ми використовуємо:

- T-критерій Стьюдента, якщо справжня дисперсія популяцій, з яких було вилучено зразки, невідома (зазвичай ми з таким і працюємо);
- Z-тест, чи відома справжня дисперсія s^2 генеральної сукупності.

Z-тест використовується, коли розмір вибірки **великий** (n > 30) чи **відома** дисперсія генеральної сукупності. t-критерій використовується, коли розмір вибірки **невеликий** (n<30) і дисперсія генеральної сукупності **невідома**.

Формула T-тесту така сама, як у Z-тесту, крім того, що ми беремо стандартне відхилення **не** генеральної сукупності, а вибіркове (пораховане за вибіркою - s):

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Обидва ці тести передбачають, що дані розподілені нормально та з гомогенною дисперсією. І обидва тести є параметричними. Якщо ми не знаємо розподіл вибірки — можна високристовувати і ці тести, але краще скористатись непараметричним тестом: гайд по стат. тестам.