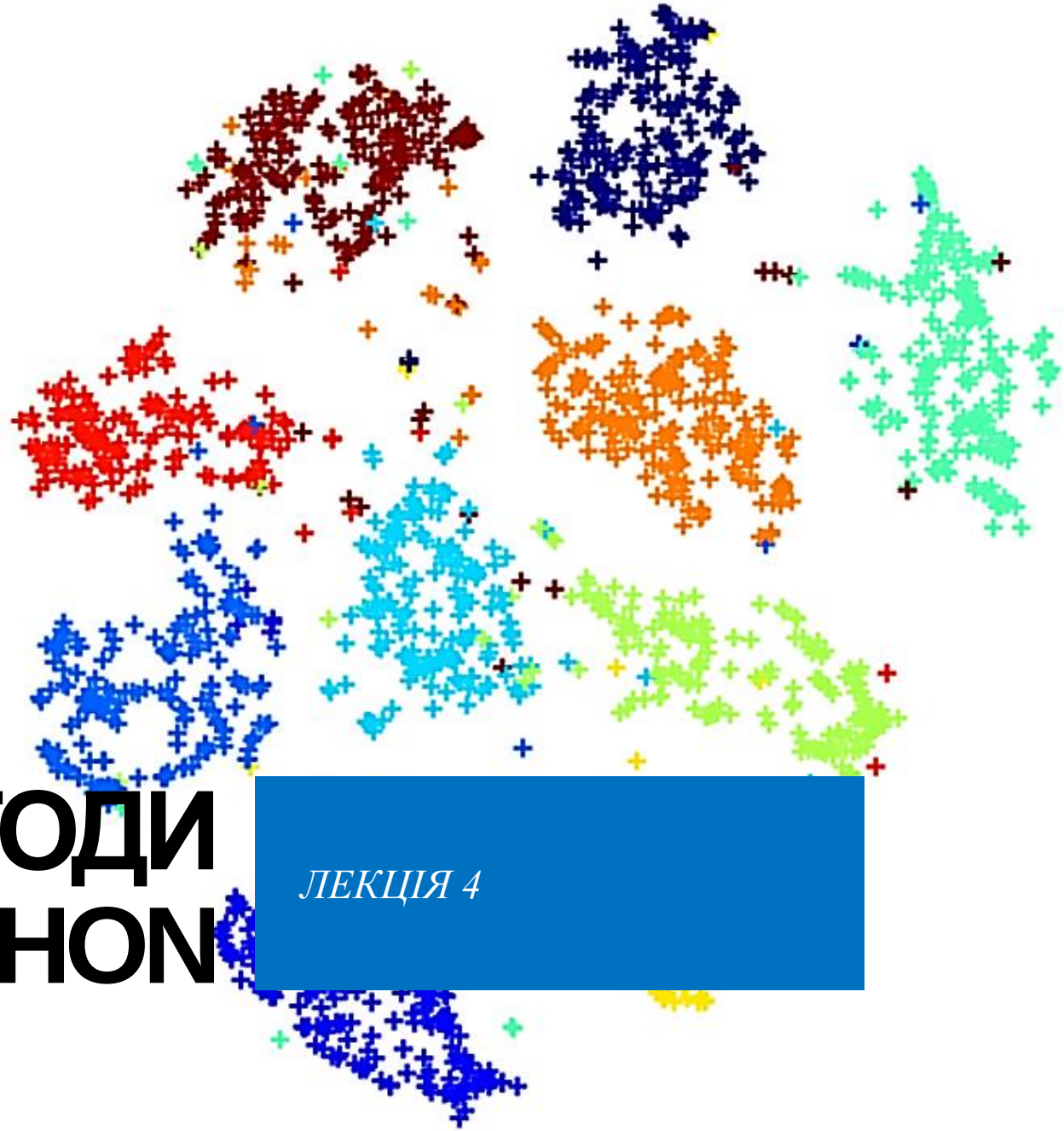
ЖИТОМИРСЬКА ПОЛІТЕХНІКА
ДЕРЖАВНИЙ УНІВЕРСИТЕТ

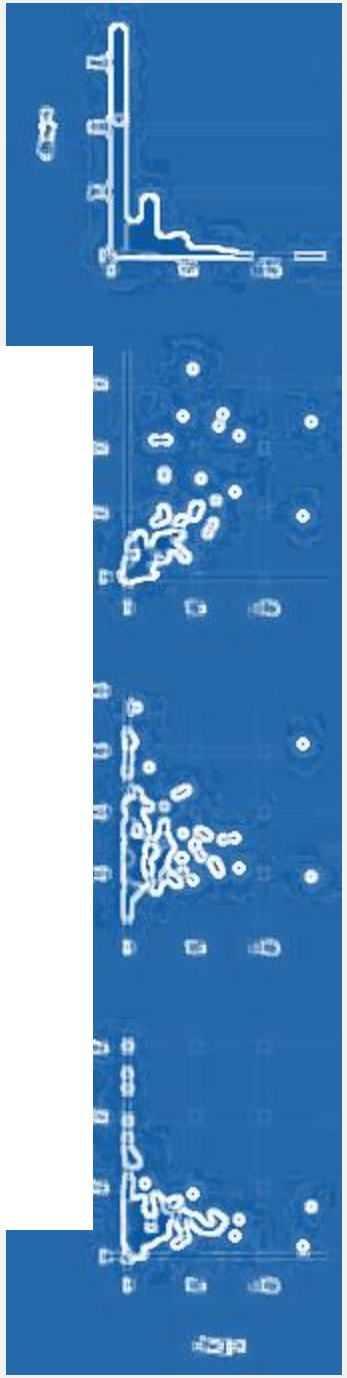


СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ В PYTHON

ЛЕКЦІЯ 4

План

1. Статистичні гіпотези та їх перевірка
2. Розподіли випадкових величин
3. Аналіз розподілів у Python



СТАТИСТИЧНІ ГІПОТЕЗИ ТА ЇХ ПЕРЕВІРКА

Статистична гіпотеза — це формалізоване припущення щодо параметрів генеральної сукупності або розподілу випадкової величини, яке перевіряється на основі вибіркових даних.

Розрізняють:

- **Нульову гіпотезу (H_0)** — це основне припущення, яке стверджує відсутність реального ефекту, різниці або зв'язку між досліджуваними явищами.
- **Альтернативну гіпотезу (H_1)** — припущення про наявність ефекту або різниці.

Приклад :
H_0: новий препарат не впливає на середній рівень артеріального тиску
H_1: новий препарат змінює середній рівень артеріального тиску

КРОКИ СТАТИСТИЧНОЇ ПЕРЕВІРКИ ГІПОТЕЗ

1. ***Формулювання нульової гіпотези***
Нульова гіпотеза як правило розглядається протилежно припущенням.

2. ***Формулювання альтернативної гіпотези.*** Це єдине твердження, що є логічним запереченням нульової гіпотези. H_1 – зв'язок між ознаками є.

3. ***Установка ймовірнісної помилки α (рівня значущості)***

4. ***Вибір відповідного статистичного критерію*** (наприклад, t-критерій Стьюдента, F-критерій Фішера, $\chi 2$ -критерій Пірсона).

5. ***Збір даних***
Визначити, чи збираються дані шляхом експериментального планування або спостереження.

6. ***Розрахунок тестового значення***
7. Порівняння тестового значення з критичною областю p-value з α

8. ***Висновки про H_0***

P-VALUE ТА РІВЕНЬ ЗНАЧУЩОСТІ

p-value — імовірність отримати спостережуваний результат (або більш екстремальний), якщо H_0 істинна.

Рівень значущості (α) — допустима ймовірність помилки I роду (зазвичай 0.05).

Правило прийняття рішення:

- якщо p-value < α → H_0 відхиляється
- якщо p-value ≥ α → немає підстав відхиляти H_0

ОСНОВНІ СТАТИСТИЧНІ ТЕСТИ В PYTHON

У Python статистичні тести реалізовані в бібліотеці ***SciPy (scipy.stats)***.

Тест	Призначення
t-test	Порівняння середніх значень
Mann–Whitney U	Непараметричний аналог t-test
ANOVA	Порівняння більше ніж двох груп
χ^2 (Chi-square)	Аналіз категоріальних змінних
Shapiro–Wilk	Перевірка нормальності

T-TEST (КРИТЕРІЙ СТЬЮДЕНТА)

Перевірка рівності середніх значень **двох незалежних вибірок**.

Гіпотези
H0:μ1=μ2
H1:μ1≠μ2

Статистика тесту

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

де:
$\bar{x}_i$ - середнє
S_i^2 - дисперсія
n_i - обсяг вибірки

Умови застосування
✔ нормальний розподіл
✔ незалежність спостережень
✔ відсутність сильних викидів

📌 **Біомедицина:** порівняння ефекту лікування
📌 **Безпека:** середній час відповіді систем
📌 **КН:** аналіз даних і алгоритмів
📌 **КІ:** продуктивності систем; енергоспоживання; апаратні вимірювання

MANN–WHITNEY U TEST

Непараметричний аналог t-test (для ненормальних розподілів).

Ідея
Порівняння **рангів**, а не середніх.

Статистика

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

де:
R_1 - сума рангів першої вибірки

Гіпотези
H_0: розподіли однакові
H_1: розподіли різні

📌 **Біомедицина:** біомаркери з асиметрією

📌 **Безпека:** затримки мережевого трафіку

📌 **КН:** аналіз даних і алгоритмів

📌 **КІ:** продуктивності систем; енергоспоживання; апаратні вимірювання

ANOVA (ANALYSIS OF VARIANCE)

Порівняння **трьох і більше середніх**.

Гіпотези

H_0:$\mu_1=\mu_2=...=\mu_k$

H_1: хоча б одне середнє відрізняється

F-статистика

$$F = \frac{SS_{between}/(k-1)}{SS_{within}/(N-k)}$$

де:

SS - суми квадратів відхилень

📌 **Критично:** ANOVA **не показує де саме різниця** → потрібні post-hoc тести.

X^2 (CHI-SQUARE TEST)

Аналіз **категоріальних змінних**.

Статистика

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

де:
O_i - спостережені значення
E_i - очікувані значення

Гіпотези
H_0: змінні незалежні
H_1: існує залежність

📌 **Біомедицина:**діагноз vs група пацієнтів

📌 **Безпека:** тип атаки vs результат

X^2 (CHI-SQUARE TEST)

Перевірка ефективності маркетингової кампанії

Припустимо, ми провели рекламну кампанію нового продукту і хочемо перевірити, чи існує зв'язок між **статтю** споживача (чоловіки/жінки) та його **реакцією** на продукт (купив/не купив). Обидві змінні є категоріальними.

Кроки перевірки гіпотези за допомогою $\chi 2$:

Формулювання гіпотез:

$H0$: Між статтю та рішенням про купівлю **немає зв'язку** (вони незалежні).

$H1$: Між статтю та рішенням про купівлю **існує зв'язок** (вони залежні).

Збір даних. Опитуємо 200 людей і зводимо дані в таблицю :

Стать	Купив	Не купив	Всього
Чоловіки	40	60	100
Жінки	70	30	100
Всього	110	90	200

X^2 (CHI-SQUARE TEST)

Перевірка ефективності маркетингової кампанії

Розрахунок: обчислюємо **очікувані частоти** і порівнюємо їх із фактично спостережуваними даними за допомогою формули $\chi 2$:

Висновки:

Якщо розраховане значення $\chi 2$ **велике** (або p-value менше за $\alpha=0.05$), **відхиляємо** $H0$
і робимо висновок, що зв'язок існує.
Якщо значення $\chi 2$ **мале** (p-value більше за $\alpha=0.05$), **приймаємо** $H0$
(або не маємо достатніх доказів її відхилити) і вважаємо, що різниця в покупках між статями випадкова.

РОЗПОДІЛИ

Перевірка **нормальності розподілу**.

Гіпотези
H0: дані мають нормальний розподіл
H1: розподіл не нормальний

Статистика

$$W = \frac{(\sum a_i x_{(i)})^2}{\sum (x_i - \bar{x})^2}$$

📌 **Ключовий тест перед вибором параметричних методів**

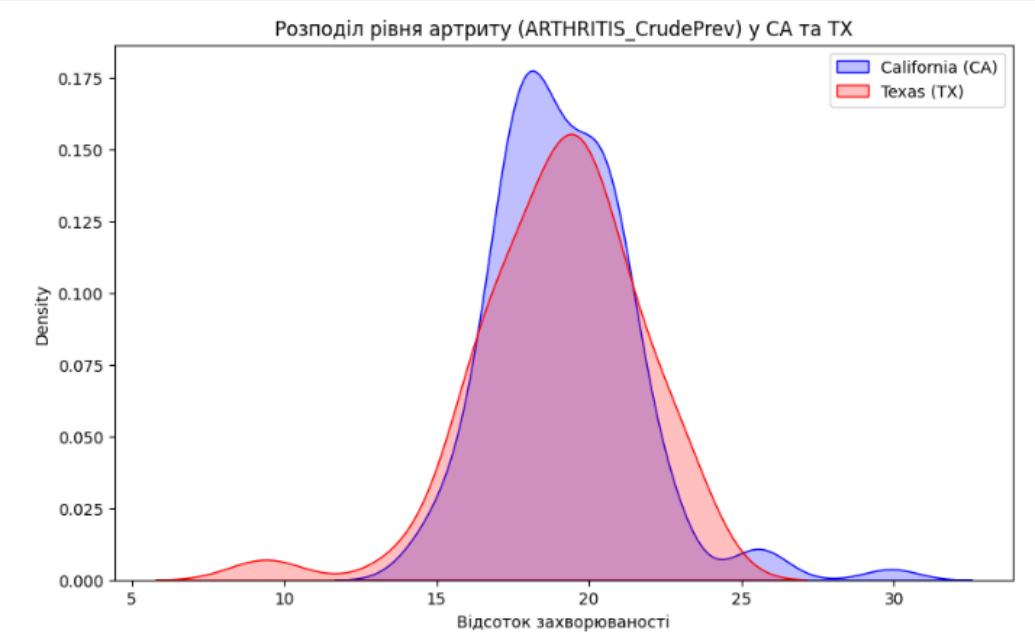
АЛГОРИТМ ВИБОРУ СТАТИСТИЧНОГО ТЕСТУ

СПЕЦІАЛЬНІСТЬ → ТИП ГІПОТЕЗ → ТЕСТ

Спец.	Основна мета гіпотез	Тип H_0	Тип H_1
КН	Порівняння методів	Рівність	Відмінність
КБ	Виявлення аномалій	Нормальність	Відхилення
КІ	Контроль норм	В межах допуску	Порушення
БІ	Ефект впливу	Немає ефекту	Є ефект

ПРИКЛАД 1

датасет 500_Cities_CDC
H_0 стверджує, що ефекту немає (рівень артриту в Каліфорнії та Техасі однаковий).
H_1 стверджує, що різниця є значущою.




```
T-статистика: 0.3686
p-value: 7.1344e-01
```

Висновок.
Немає підстав відхиляти нульову гіпотезу (H_0).
Різниця в рівні артриту не є статистично значущою.

ПРИКЛАД 2

Для порівняння двох штатів використаємо **t-тест Стьюдента** (якщо дані розподілені нормально) або **тест Манна-Вітні** (якщо розподіл відмінний від нормального).

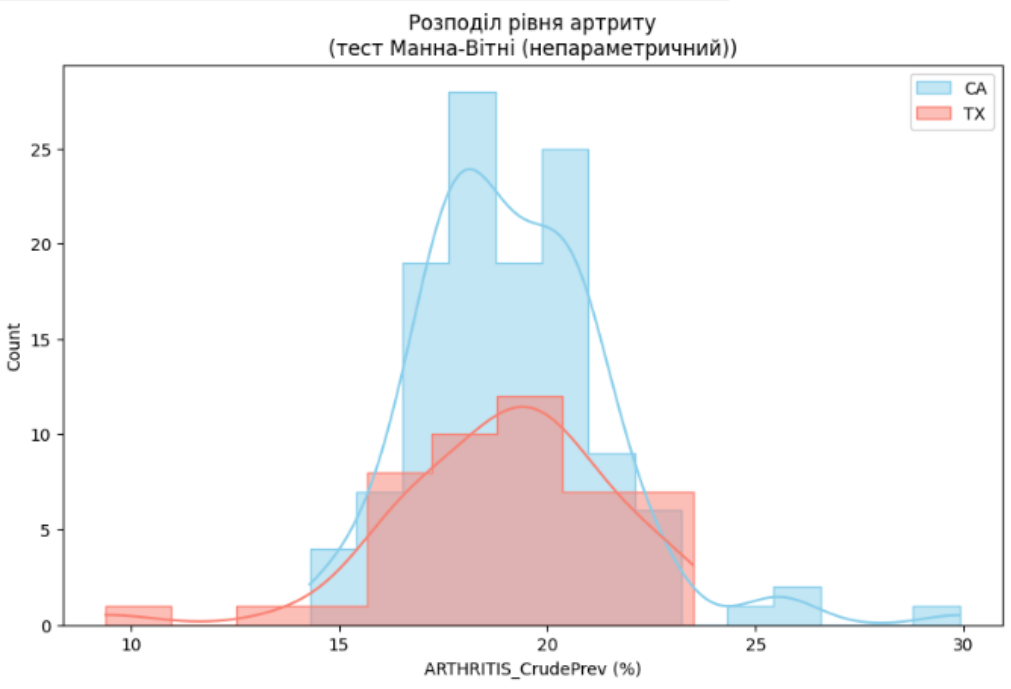
1. Перевірка на нормальність (Тест Шапіро-Вілка)

```
Каліфорнія: W=0.9346, p-value=0.0000
Техас: W=0.9403, p-value=0.0183
```

Дані штату CA НЕ розподілені нормально (p <= 0.05)
Дані штату TX НЕ розподілені нормально (p <= 0.05)

2. Перевірка гіпотези
Використано метод «тест Манна-Вітні» (непараметричний)
Статистика тесту:
`2775.0000 p-value: 0.8101`

Висновок. Немає підстав відхиляти нульову гіпотезу (H_0).
Різниця не є значущою.



СПЕЦИФІКА ЗАСТОСУВАННЯ ГІПОТЕЗ

Біомедичні дані:

- малі вибірки
- часто негаусівські розподіли
- висока ціна помилки I роду

Дані безпеки (кібербезпека, відеоспостереження):

- великі обсяги даних
- дисбаланс класів
- важливість мінімізації помилки II роду (пропуск загрози)

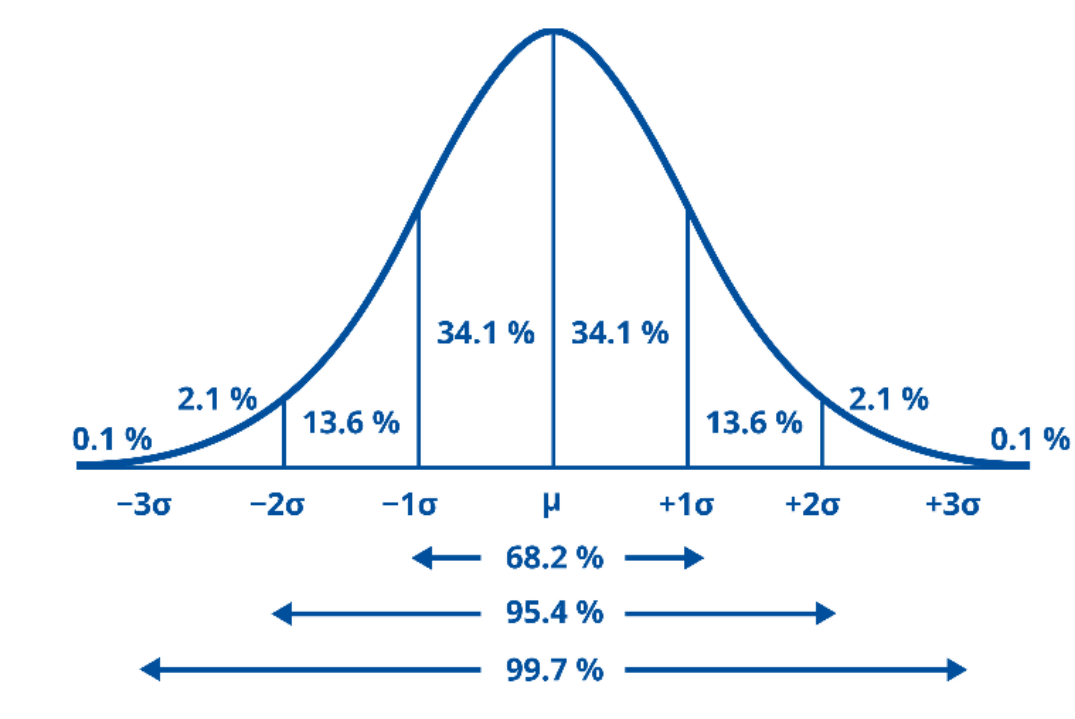
РОЗПОДІЛИ ВИПАДКОВИХ ВЕЛИЧИН

Розподіл описує, як значення випадкової величини розміщуються у вибірці.

Основні характеристики:

- Середнє аріфметичне (Mean)
- Медіана (Median)
- Мода (Mode)
- Дисперсія (Variance)
- Стандартне відхилення (Std)

- Форма розподілу
  - асиметрія (skewness)
  - ексцес (kurtosis)



НАЙПОШИРІНІШИ РОЗПОДІЛИ

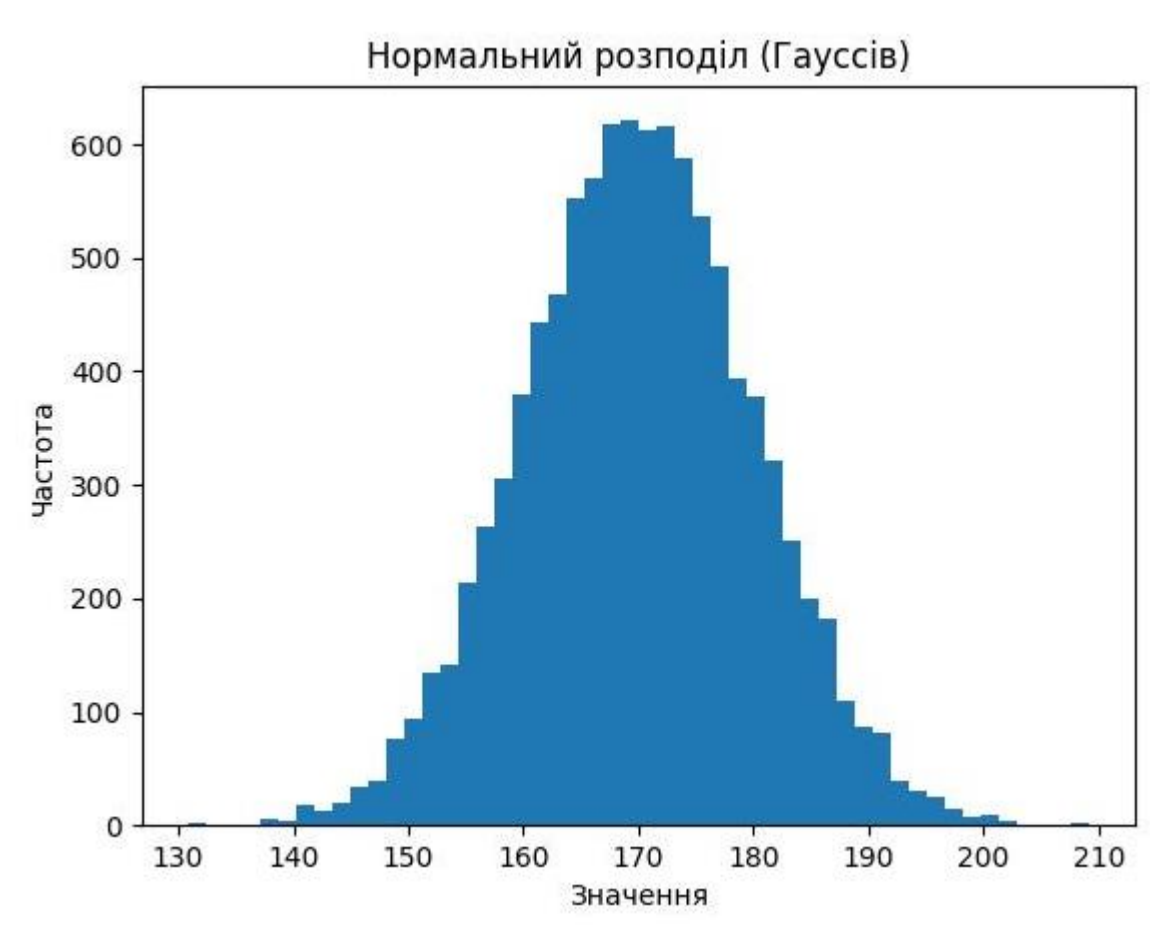
Розподіл	Приклад
Нормальний	Фізіологічні показники
Біноміальний	Подія / не подія
Пуассона	Кількість інцидентів
Експоненційний	Час до події
Логнормальний	Біомаркери

НОРМАЛЬНИЙ РОЗПОДІЛ

Це розподіл, де більшість значень зосереджена навколо середнього, а ймовірність екстремальних значень симетрично спадає в обидва боки.

Де зустрічається: зріст людей, похибки вимірювань, вага товарів.

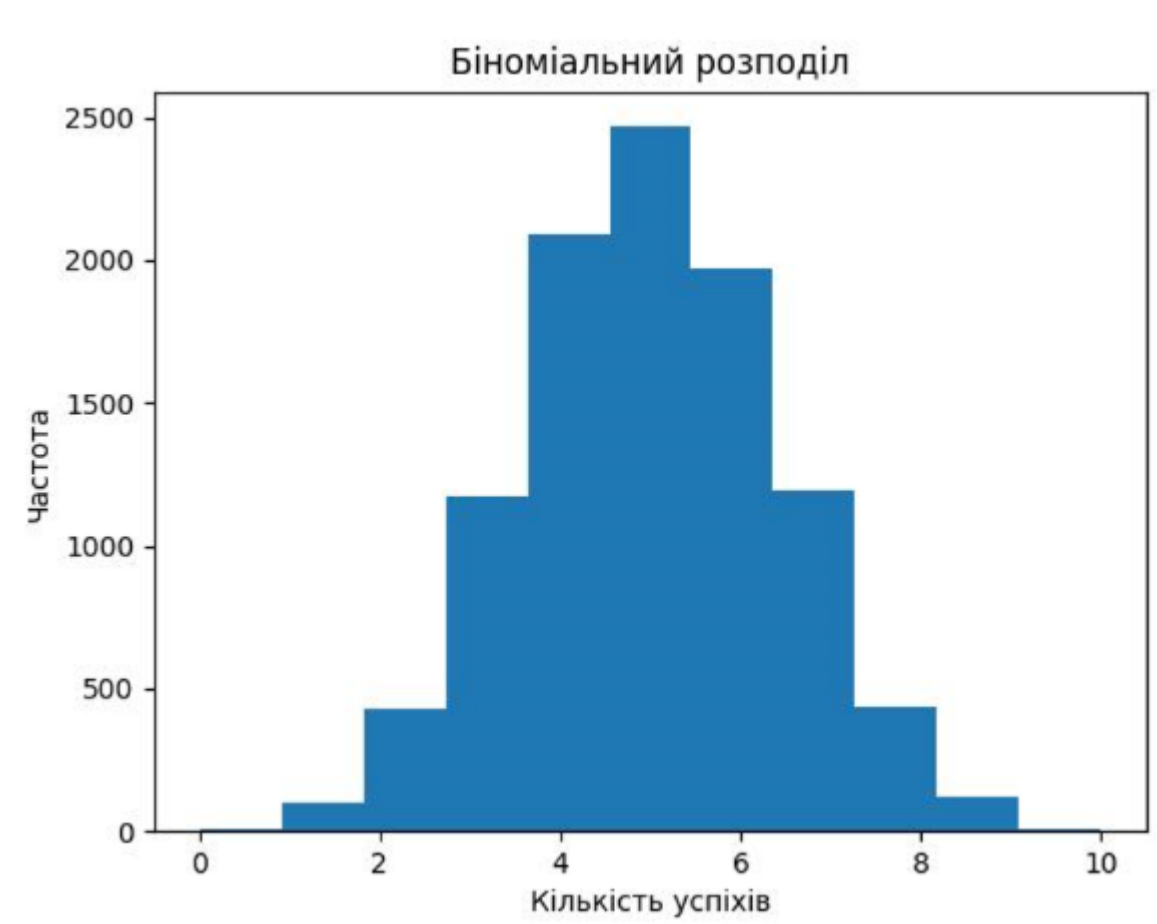
Чому важливий? Багато статистичних тестів (наприклад, t-тест) та алгоритмів (Linear Regression, LDA) працюють найкраще саме на таких даних.



РОЗПОДІЛ БЕРНУЛЛІ ТА БІНОМІАЛЬНИЙ РОЗПОДІЛ

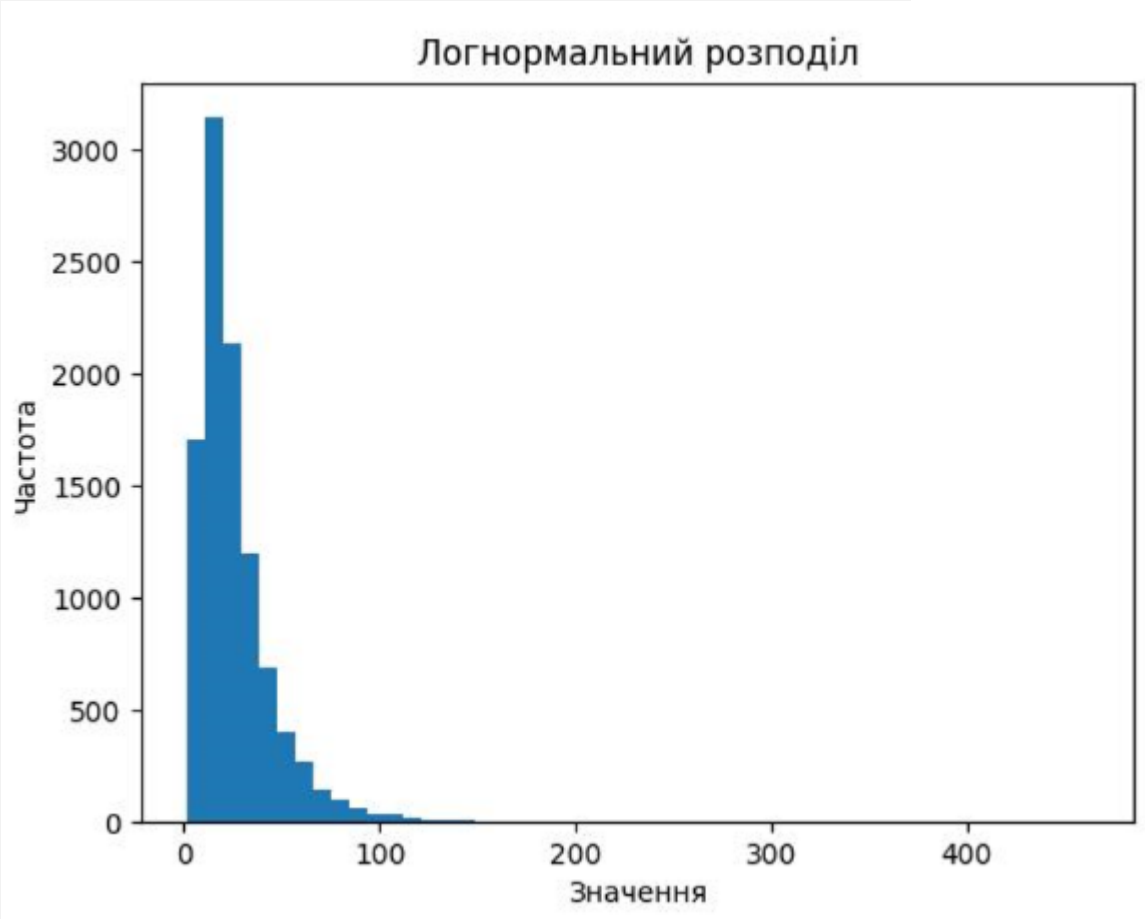
Розподіл Бернуллі має лише два результати (успіх/невдача, 0/1). Це основа для логістичної регресії.

Біноміальний розподіл описує кількість успіхів у серії незалежних випробувань (наприклад, скільки людей зі 100 клікнуть на рекламу).



ЛОГНОРМАЛЬНИЙ РОЗПОДІЛ

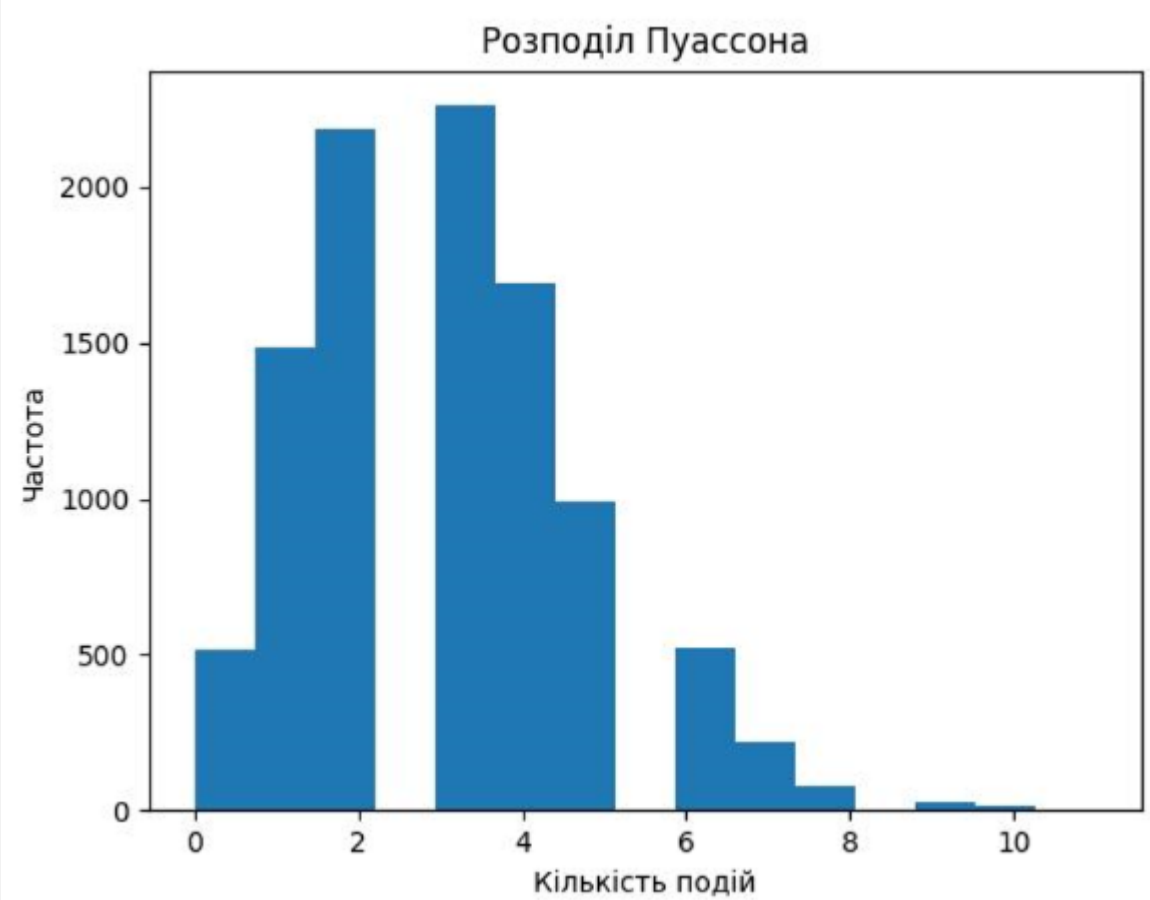
Це розподіл випадкової величини, логарифм якої розподілений нормально. Він має довгий "хвіст" праворуч.
Де зустрічається: доходи населення (більшість отримує середню зарплату, але є невелика кількість мільйонерів), кількість жителів у містах (як у датасеті CDC).
📌 Такі дані часто вимагають логарифмування перед подачею в модель, щоб "стиснути" хвіст і наблизити розподіл до нормального.



РОЗПОДІЛ ПУАССОНА

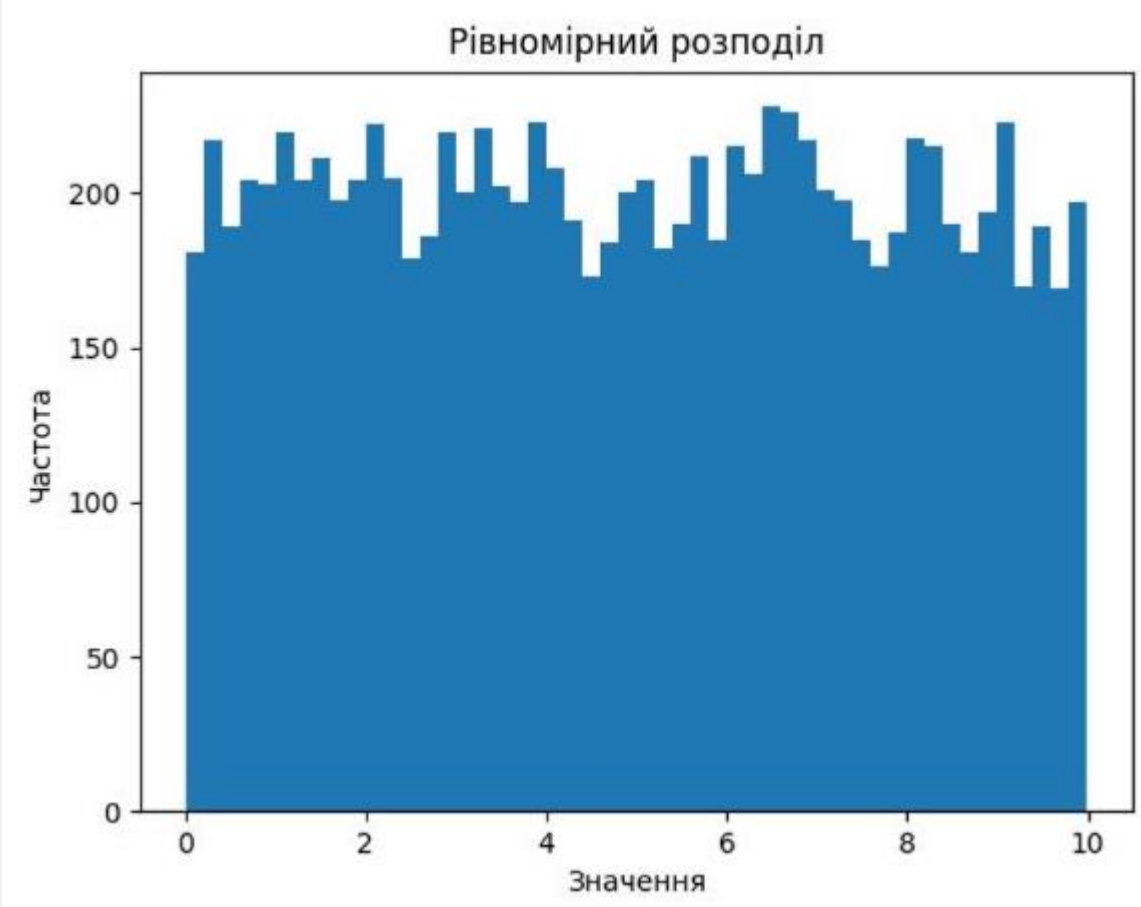
Описує кількість подій, що відбуваються за певний проміжок часу або в певному просторі.

Де зустрічається:
кількість дзвінків у кол-центр за годину; кількість відвідувачів сайту за хвилину.



РІВНОМІРНИЙ РОЗРОДІЛ

Кожне значення в певному діапазоні має однакову ймовірність випадання.
Де зустрічається: кидання кубика, генерація випадкових чисел у криптографії.
Візуально виглядає як прямокутник.



АНАЛІЗ РОЗПОДІЛІВ У PYTHON

Основні бібліотеки:

- numpy
- scipy.stats
- matplotlib, seaborn

Методи аналізу:

- гістограми
- KDE-графіки
- Q–Q plots

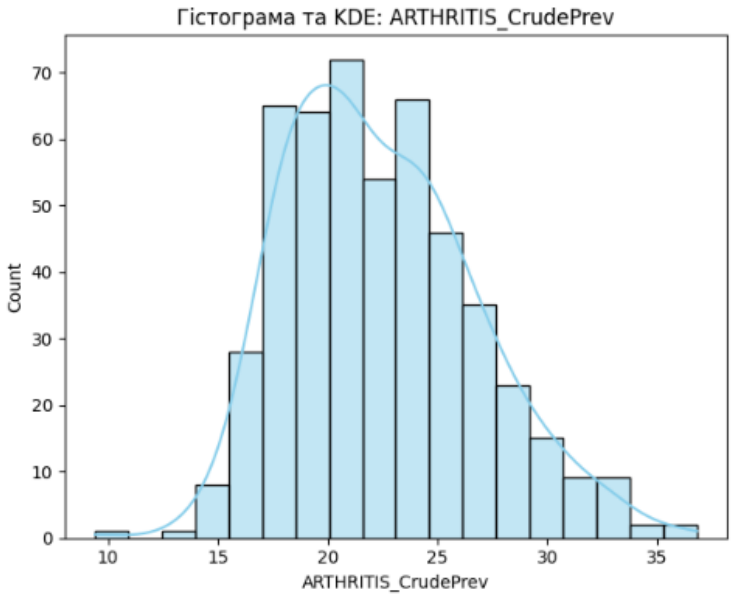
Як визначити тип розподілу в Python?

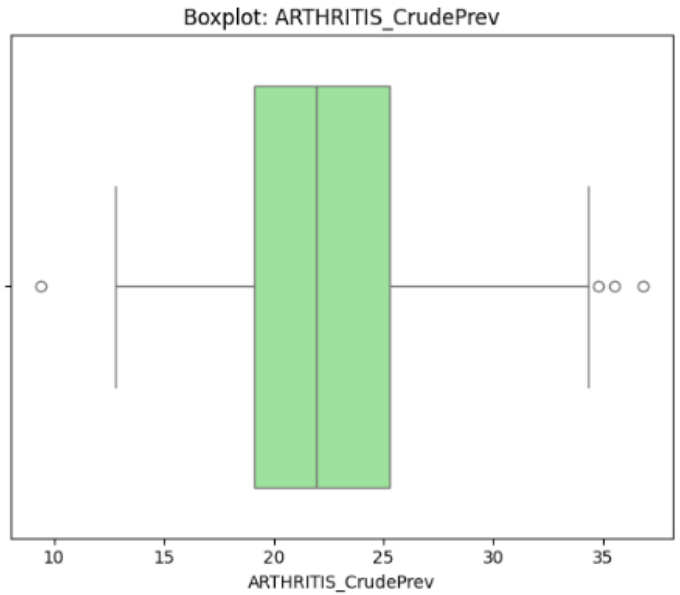
На етапі EDA використовують:
Гістограми та KDE
sns.histplot(data, kde=True) — візуальна перевірка.

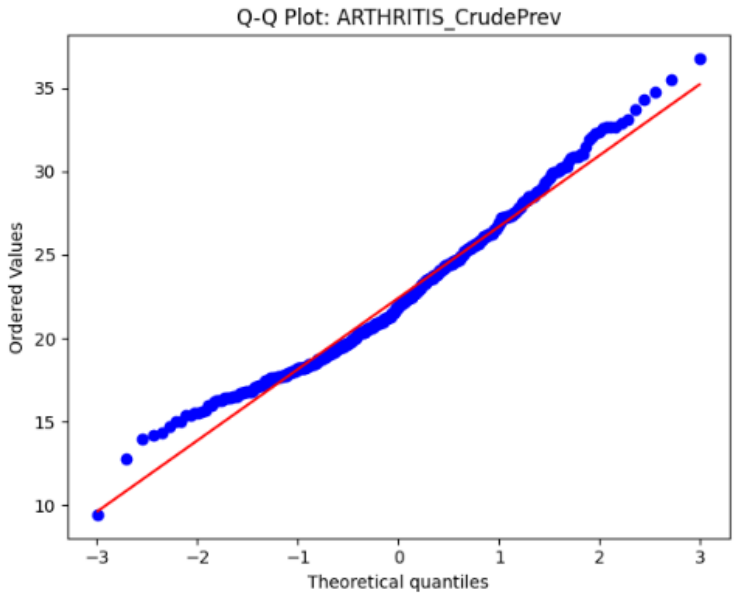
Q-Q Plot (Квантиль-квантиль) дозволяє побачити відхилення від нормальності вздовж діагоналі.

Статистичні тести
Тест Шапіро-Вілка або Д’Агостіно на нормальність.

ПРИКЛАД АНАЛІЗУ РОЗПОДІЛІВ У PYTHON







Інтерпретація:

1. Якщо пік зміщений вліво, а хвіст тягнеться вправо - це **позитивна асиметрія** (характерно для доходів або рідкісних хвороб).
2. Будь-які точки за межами "вусів" є статистичними викидами. Це сигнал для етапу *Data Preparation*, що дані потребують очищення або використання *RobustScaler*.
3. Якщо сині точки лежать на червоній лінії - дані розподілені **нормально**. Якщо кінці точок відхиляються вгору або вниз - у даних є **"важкі хвости"** (багато викидів/екстремальних значень).

ПЕРЕВІРКА НОРМАЛЬНОСТІ

Ключовий етап перед вибором статистичного тесту.

Методи:

- Shapiro–Wilk
- Kolmogorov–Smirnov
- Anderson–Darling

```
k2, p = stats.normaltest(df['ARTHRITIS_CrudePrev'].dropna())
alpha = 0.05
print(f"p-value: {p:.5f}")
if p < alpha:
    print("розподіл НЕ є нормальним")
else:
    print("Розподіл  нормальний")
```

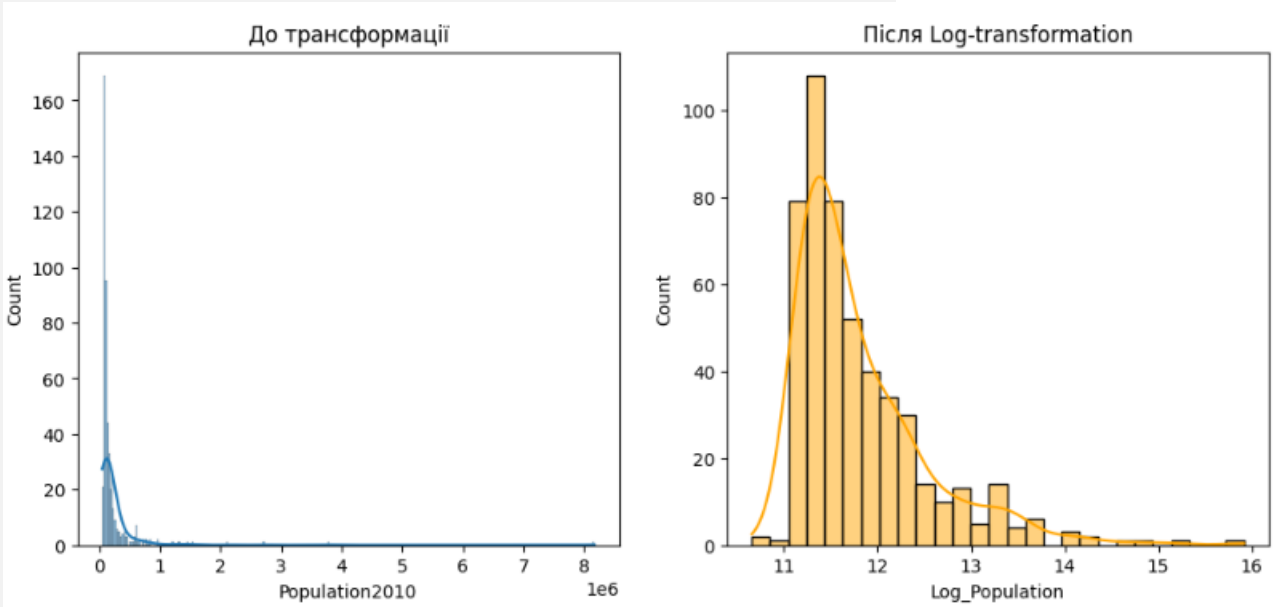
```
p-value: 0.00007
розподіл НЕ є нормальним
```

ТРАНСФОРМАЦІЯ ДАНИХ (LOG-TRANSFORMATION)

Якщо ваші графіки (KDE або Q-Q Plot) показали значну позитивну асиметрію (довгий хвіст праворуч), це може «заплутати» багато моделей Data Mining. Найпоширеніший спосіб це виправити - логарифмування.

```
# Створюємо логарифмовану ознаку
df['Log_Population'] =
np.log1p(df['Population2010'])

fig, axes = plt.subplots(1, 2,
figsize=(12, 5))
sns.histplot(df['Population2010'],
kde=True, ax=axes[0]).set_title('До
трансформації')
sns.histplot(df['Log_Population'],
kde=True, ax=axes[1],
color='orange').set_title('Після Log-
transformation')
plt.show()
```



ВИЗНАЧЕННЯ КАТЕГОРІАЛЬНИХ ЗМІННИХ

Важливо розрізняти **номінальні** (назви штатів) та **порядкові** (розмір міста: Small < Medium < Large) змінні.

```
def analyze_categories(df):
    results = []
    for col in df.columns:
        unique_count = df[col].nunique()
        total_count = len(df)
        ratio = (unique_count / total_count) * 100
# Критерій: якщо унікальних значень мало (< 5%) або тип 'object'
        is_categorical = "Так" if (ratio < 5 or df[col].dtype == 'object') else "Ні"
        results.append({
            'змінна': col,
            'Тип': df[col].dtype,
            'Унікальних': unique_count,
            '% унікальних': round(ratio, 2),
            'Категоріальна?': is_categorical
        })
    return pd.DataFrame(results)
print(analyze_categories(df))
```

ВИЗНАЧЕННЯ КАТЕГОРІАЛЬНИХ ЗМІННИХ

Результат

...	змінна	Тип	Унікальних	% унікальних	Категоріальна?
0	StateAbbr	object	51	10.2	Так
1	PlaceName	object	474	94.8	Так
2	PlaceFIPS	int64	500	100.0	Ні
3	Population2010	int64	497	99.4	Ні
4	ACCESS2_CrudePrev	float64	229	45.8	Ні
..	...	...	...	...	...
113	TEETHLOST_Crude95CI	object	472	94.4	Так
114	TEETHLOST_AdjPrev	float64	180	36.0	Ні
115	TEETHLOST_Adj95CI	object	465	93.0	Так
116	Geolocation	object	500	100.0	Так
117	Log_Population	float64	497	99.4	Ні

[118 rows x 5 columns]

МЕТОДИ ОБРОБКИ ДАНИХ ЗАЛЕЖНО ВІД ТИПУ РОЗПОДІЛУ

Тип розподілу	Візуальні ознаки	Рекомендований Scaler	Рекомендована трансформація	Підходящі алгоритми
Нормальний (Gaussian)	Симметричний "дзвін", точки на Q-Q Plot лежать на лінії.	StandardScaler (Z-score)	Не потрібна	Лінійна/Логістична регресія, LDA, Gaussian NB
Логнормальний (Skewed)	Довгий хвіст праворуч (позитивна асиметрія).	MinMaxScaler	Log Transformation або Box-Cox	Нейронні мережі, KNN (після трансформації)
З важкими хвостами	Багато точок за межами "вусів" Boxplot.	RobustScaler (на основі медіани та IQR)	Winsorization (обрізка викидів)	Дерева рішень, Random Forest (стійкі до викидів)
Бімодальний	Два або більше піків на KDE-графіку.	MinMaxScaler	Створення нових ознак (Binning)	Кластеризація (K-Means), Gaussian Mixture Models
Категоріальний	Дискретні стовпчики, низька кількість унікальних значень.	Не застосовується	One-Hot Encoding або Label Encoding	CatBoost, XGBoost, Decision Trees

ВИСНОВКИ

1. Статистичні методи — основа інтелектуального аналізу даних.

2. Python надає потужний інструментарій для перевірки гіпотез, аналізу кореляцій та розподілів.

3. Коректне застосування статистики потребує урахування предметної області.

4. Інтерпретація результатів важливіша за самі обчислення.