Статистичний аналіз даних

До статистичних методів належать наступні:

- Попередній аналіз природи статистичних даних (перевірка гіпотез, оцінка виду функції розподілу, її параметрів і т.п.);
- Виявлення зв'язків і закономірностей (лінійний і нелінійний регресійний аналіз, кореляційний аналіз та ін.);
- Багатовимірний статистичний аналіз (лінійний і нелінійний дискримінантний аналіз, кластерний аналіз, компонентний аналіз, факторний аналіз та ін.);
- Динамічні моделі і прогноз на основі часових рядів.

Статистичний аналіз даних

Випадковою величиною називають таку величину, яка внаслідок випробування може прийняти лише одне числове значення, заздалегідь невідоме і обумовлене випадковими причинами.

Для генерування набору випадкових величин використовується np.random бібліотеки NumPy

Статистичний аналіз даних

Інтегральною функцією розподілу (функцією розподілу) називають ймовірність того, що випадкова величина X прийме значення, менше x.

Функцію розподілу позначають F(x). Таким чином,

$$F(x) = P(X < x)$$

В scipy.stats використовується метод cdf

Cumulative Distribution Function

Наприклад, norm.cdf(0) – значення інтегральної функції розподілу випадкової величини X зі стандартним нормальним розподілом при x=0

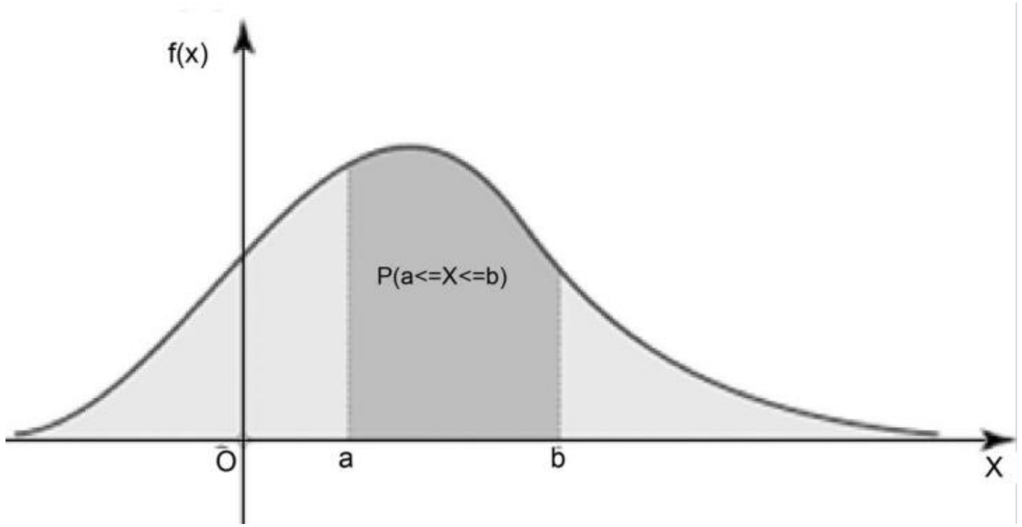
Статистичний аналіз даних

Диференціальною функцією розподілу або **щільністю імовірностей** неперервної випадкової величини називають похідну першого порядку від її інтегральної функції розподілу і позначають

$$f(x) = F'(x).$$

В scipy.stats використовується метод pdf ()

Probability Density Function



Статистичний аналіз даних

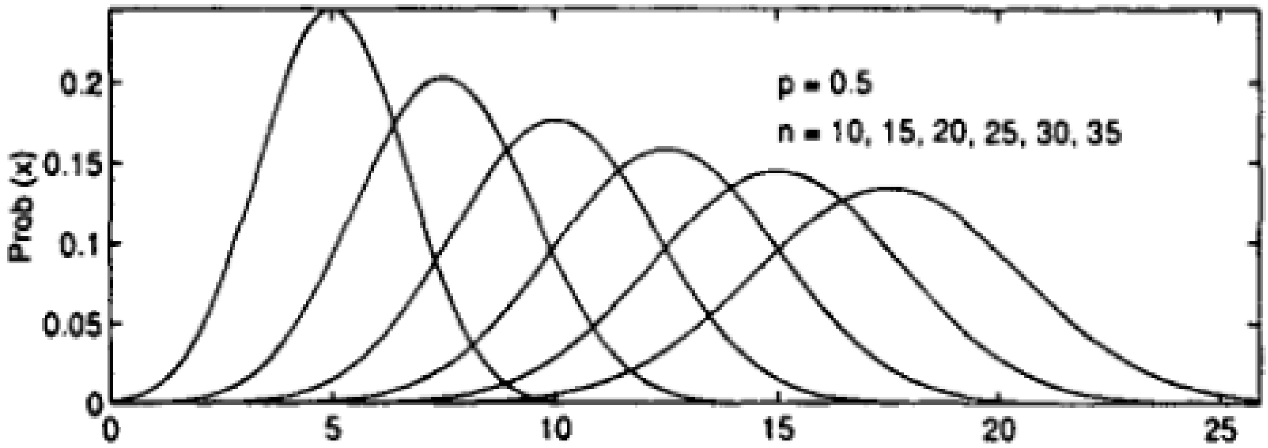
Центральні тенденції розподілу можна охарактеризувати за допомогою математичного сподівання. А розсіювання – за допомогою дисперсії та середньоквадратичного відхилення, яке дорівнює кореню з дисперсії.

Види розподілу

Біноміальний закон розподілу.

$$P_n(x) = C_n^x p^x q^{n-x}$$

використовується у схемі Бернуллі, тобто у випадку *n* незалежних повторних випробувань, в кожному з яких деяка подія з'являється з імовірністю *p*.



np.random.binomial(n, p[, size])

scipy.stats.binom

Види розподілу

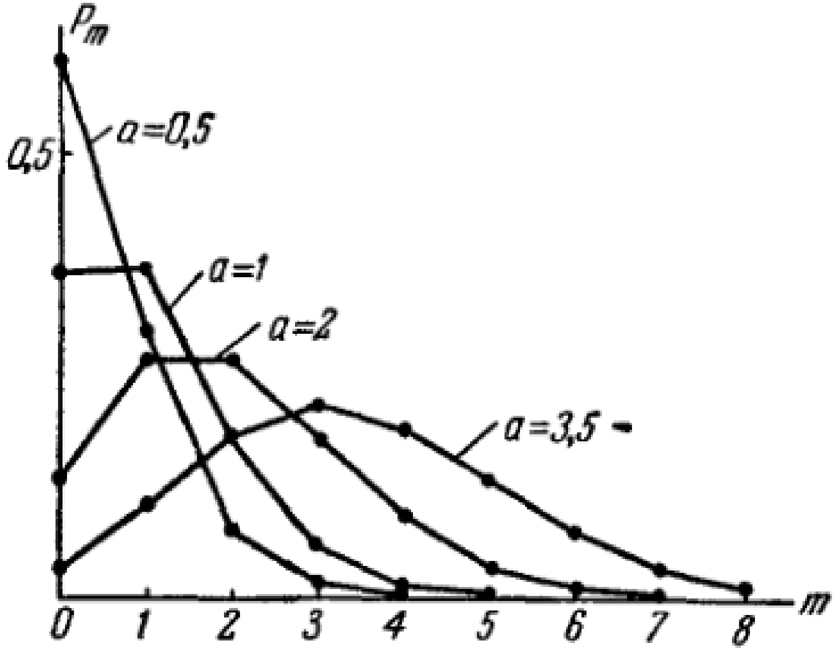
Закон розподілу Пуассона.

Випадкова величина X, що дорівнює кількості подій у процесі Пуассона, є випадковою змінною Пуассона з параметром λ>0, і

$$P(x) = \frac{e^{-\lambda T} (\lambda T)^x}{x!}$$

де $a = \lambda T$ називається параметром закону Пуассона.

np.random.poisson(a,[, size])



Види розподілу

Рівномірний розподіл. Величина X розподілена рівномірно на проміжку (a, b), якщо усі її можливі значення належать цьому проміжку і щільність її імовірностей на цьому проміжку постійна, тобто

$$f(x) = \begin{cases} C = \dfrac{1}{b-a}, & \text{при } x \in (a,b) \\ 0, & \text{при } x \notin (a,b) \end{cases}$$

np.random.uniform([low, high, size])

Види розподілу

Показниковий розподіл. Випадкова величина X, яка дорівнює відстані між послідовними подіями процесу Пуассона із середньою кількістю подій λ>0 на одиничний інтервал, має показниковий розподіл з параметром λ. Щільність її ймовірностей має вигляд

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{при } x \geq 0; \\ 0, & \text{при } x < 0, \end{cases}$$

де параметр $\lambda > 0$.

np.random.exponential([1/lam, size])

Статистичний аналіз даних

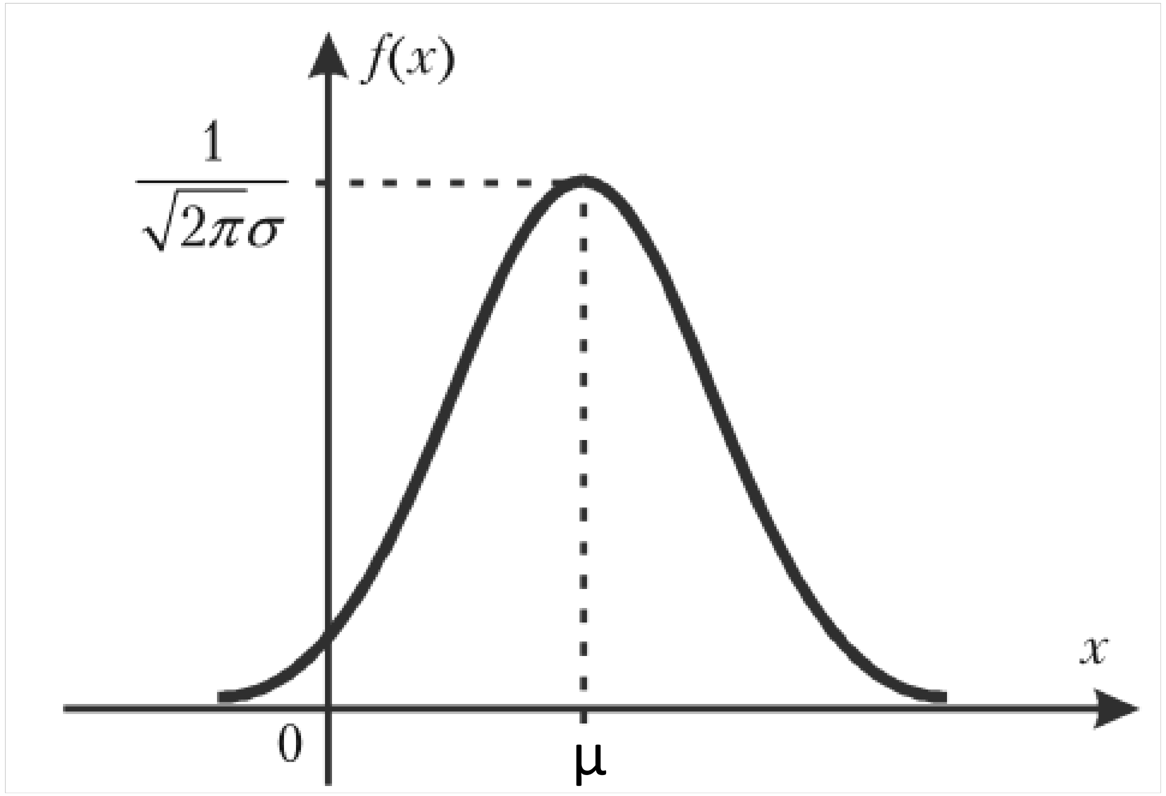
Нормальний розподіл. Випадкову величину X називають розподіленою нормально, якщо щільність її ймовірностей має вигляд

$$f\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(x-\mu\right)^2}{2\sigma^2}}$$

де μ та σ параметри розподілу, що відповідно дорівнюють математичному сподіванню та середньоквадратичному відхиленню величини X.

Види розподілу

Графік цієї функції *f(x)* називають нормальною кривою або **кривою Гаусса.**



Види розподілу

При $\mu = 0$ та $\sigma = 1$ нормальну криву називають нормованою (або стандартною нормальною). Тоді її рівняння буде

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Така функція називається функцією Лапласа.

Види розподілу

Якщо X - нормальна випадкова величина з M(X) = μ і D(X) = σ^2, випадкова величина

$$Z = (X - \mu) / \sigma$$

- це нормальна випадкова величина з M(Z) = 0, а D(Z) = 1. Тобто Z - це стандартна нормальна випадкова величина.

Створення нової випадкової величини за допомогою цього перетворення називається стандартизацією. Випадкова величина Z представляє собою відстань X від її математичного сподівання у терміні середньоквадратичних відхилень.

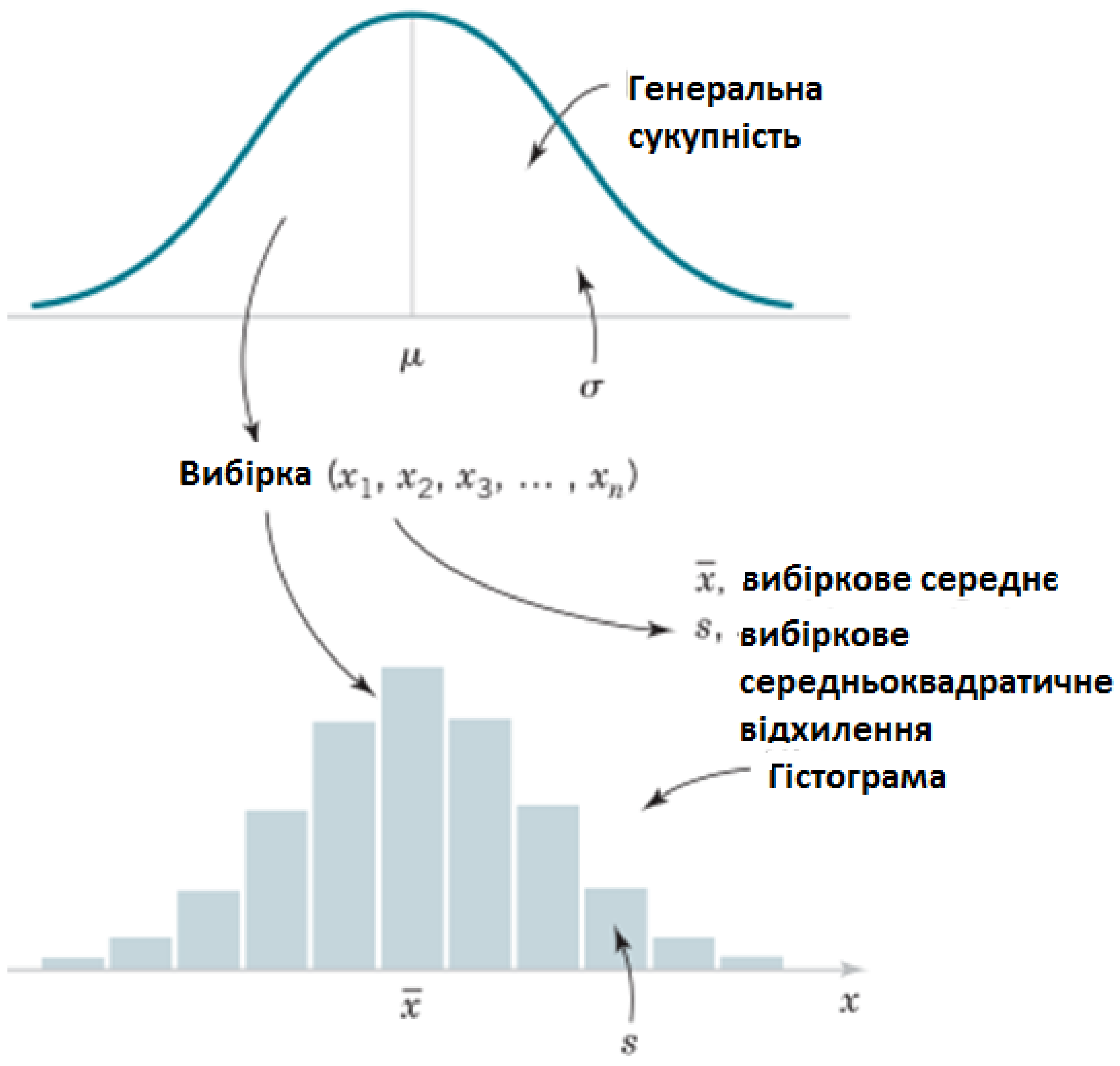
np.random.normal([m, sigma, size])

Статистичний аналіз даних

Вибірковою сукупністю (вибіркою) називають сукупність випадково взятих об’єктів із статистичної сукупності.

Генеральною називають сукупність об’єктів, з яких зроблено вибірку. **Об’ємом сукупності** (вибіркової або генеральної) називають кількість об’єктів цієї сукупності. Об’єм генеральної сукупності часто позначається літерою N, а вибіркової - n.

Статистичний аналіз даних



Статистичний аналіз даних

Характеристика, що описує генеральну сукупність та має числове значення, називається **параметром генеральної сукупності**. Будь-яке число, отримане з вибірки з метою оцінки параметру генеральної сукупності, називається **вибірковою статистикою** або коротко **статистикою**. Вибіркова статистика є випадковою величиною, оскільки її значення залежить від того, з якої вибірки її отримано.

Статистичний аналіз даних

Значення x_i вибірки називають варіантами. Послідовність варіант, розміщених у порядку зростання, називають варіаційним рядом. Якщо при цьому x_i повторюється n_i разів, то число n_i називають абсолютною частотою варіанти x_i, а $\frac{n_i}{n}$ – відносною частотою варіанти x_i.

Статистичний аналіз даних

Статистичним розподілом вибірки називають перелік варіант та відповідних частот або відносних частот. Статистичний закон розподілу зручно задавати таблицею, що встановлює зв’язок між значеннями випадкової величини та їх частотами:

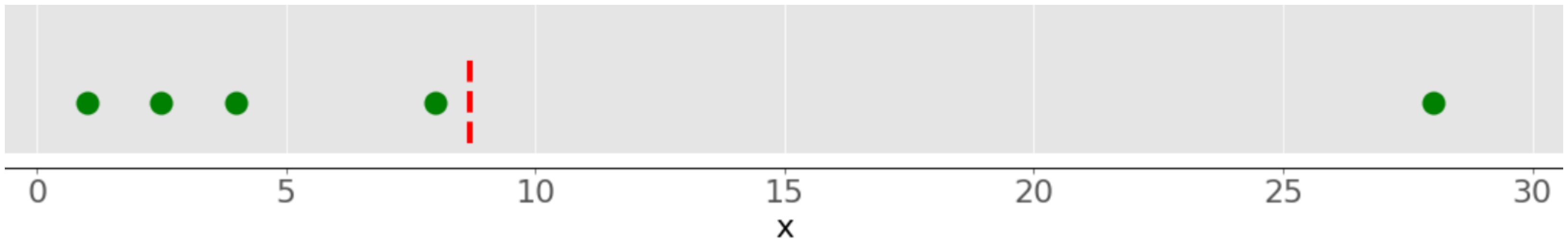
x_i	x_1	x_2	$\ldots$	x_k
n_i	n_1	n_2	$\ldots$	n_k

Статистичний розподіл вибірки, заданий цією таблицею, також називають простим чи незгрупованим статистичним розподілом або розподілом частоти варіанти x_i.

Числові характеристики рядів розподілу

Нехай x_1, x_2, … , x_n – спостереження (значення величини X) у вибірці з об'ємом n. Тоді **вибіркове середнє** можна знайти за формулою:

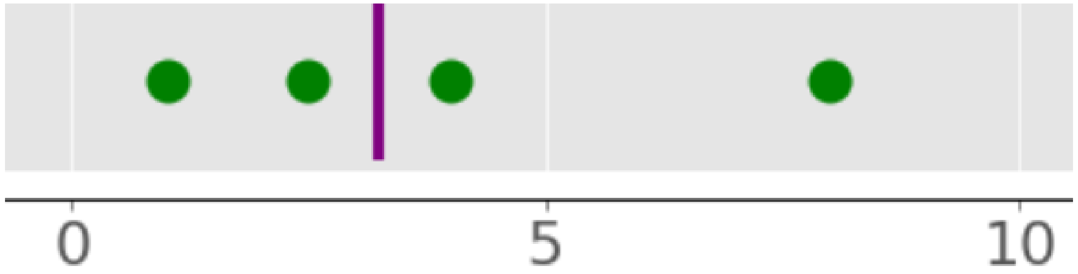
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



NumPy : np.mean()

Числові характеристики рядів розподілу

Медіана вибірки, що має n відсортованих значень, визначається як центральне значення, якщо *n* непарне число, або як середнє значення двох центральних значень, якщо *n* парне число. Різниця між вибірковим середнім та медіаною пропорційно збільшується зі збільшенням асиметрії розподілу даних або у випадку малого об’єму вибірки.



NumPy : np.median()

Числові характеристики рядів розподілу

Мода вибірки визначається як значення, що зустрічається найчастіше у вибірці. Наприклад, вибори, як правило, визначаються модою, тобто обирається той кандидат, який отримає найбільше голосів. Вибірка даних може мати більше однієї моди, і в цьому випадку вона називається мультимодальною вибіркою.

SciPy : scipy.stats.mode()

Числові характеристики рядів розподілу

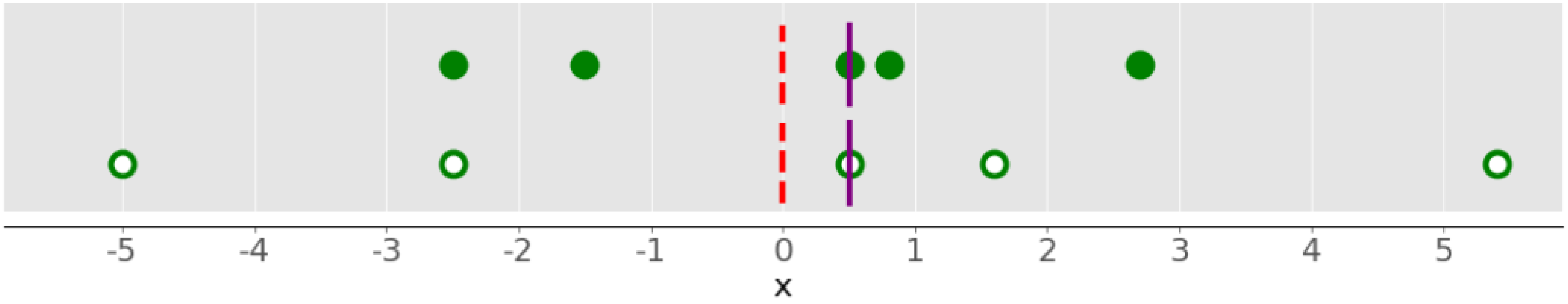
Кількісні параметри є надійними показниками положення даних, якщо їх значення суттєво не змінюється, коли до вибірки потрапляють помилкові дані. Медіана та мода більш стійкі, ніж вибіркове середнє.

На середнє значення вибірки впливають усі значення даних у наборі даних, включаючи так звані **статистичні викиди**, тобто такі значення, що суттєво віддалені від інших значень. Статистичні викиди можуть бути спричинені помилками у вимірюванні або іншими непередбачуваними чинниками.

Числові характеристики рядів розподілу

Якщо x_1 , x_2 , … , x_n є вибіркою з об’ємом n, тоді вибіркова дисперсія (незміщена) розраховується за формулою:

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



Числові характеристики рядів розподілу

Значення $s = \sqrt{s^2}$ називається вибірковим середнім квадратичним відхиленням.

NumPy : np.var(), np.std()

var() та std() має важливий опціональний параметр – ddof – різниця степеней вільності. При обчисленній дисперсії в знаменнику використовується значення N - ddof. За замовчанням ddof=0. Щоб отримати незміщену дисперсію, потрібно задати ddof=1.

Числові характеристики рядів розподілу

Початковим емпіричним моментом порядку k називається вираз

$$\grave{I}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

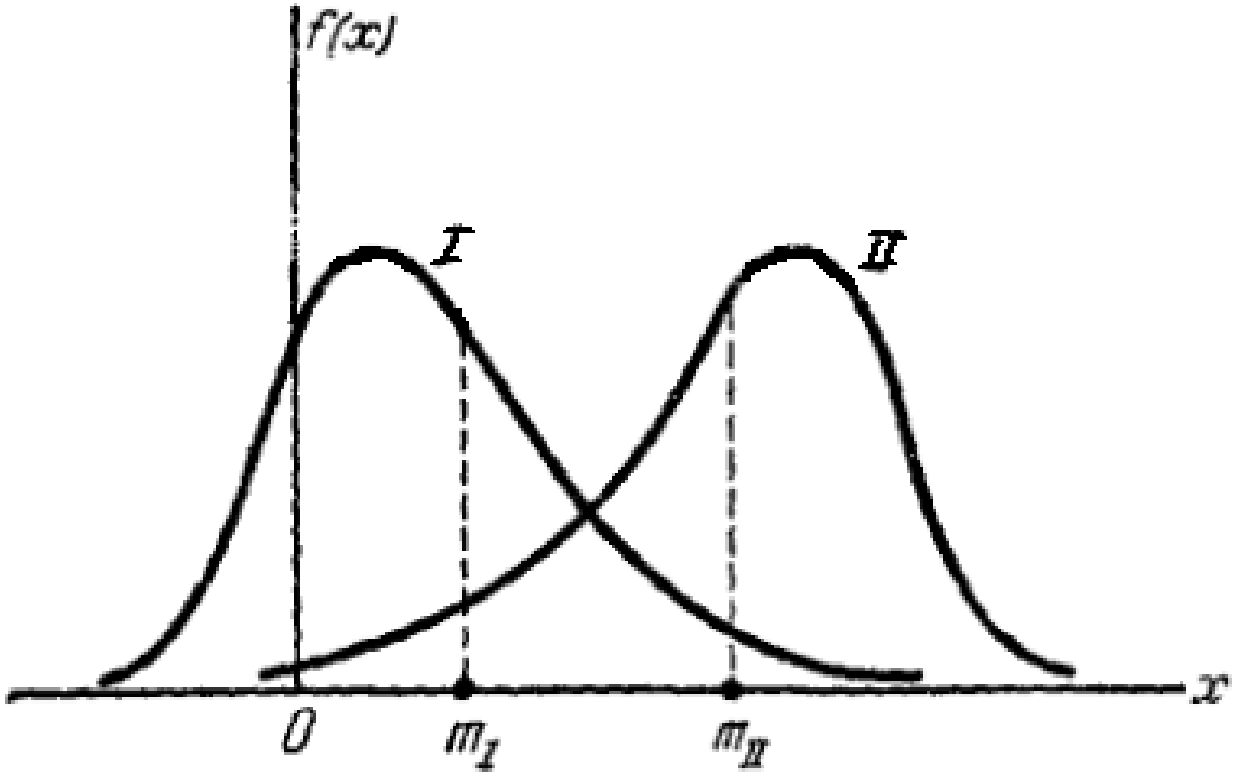
Центральним емпіричним моментом порядку k називається вираз

$$m_k = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^k$$

Числові характеристики рядів розподілу

Коефіцієнтом асиметрії A_s називається відношення центрального емпіричного моменту третього порядку до куба середнього квадратичного відхилення:

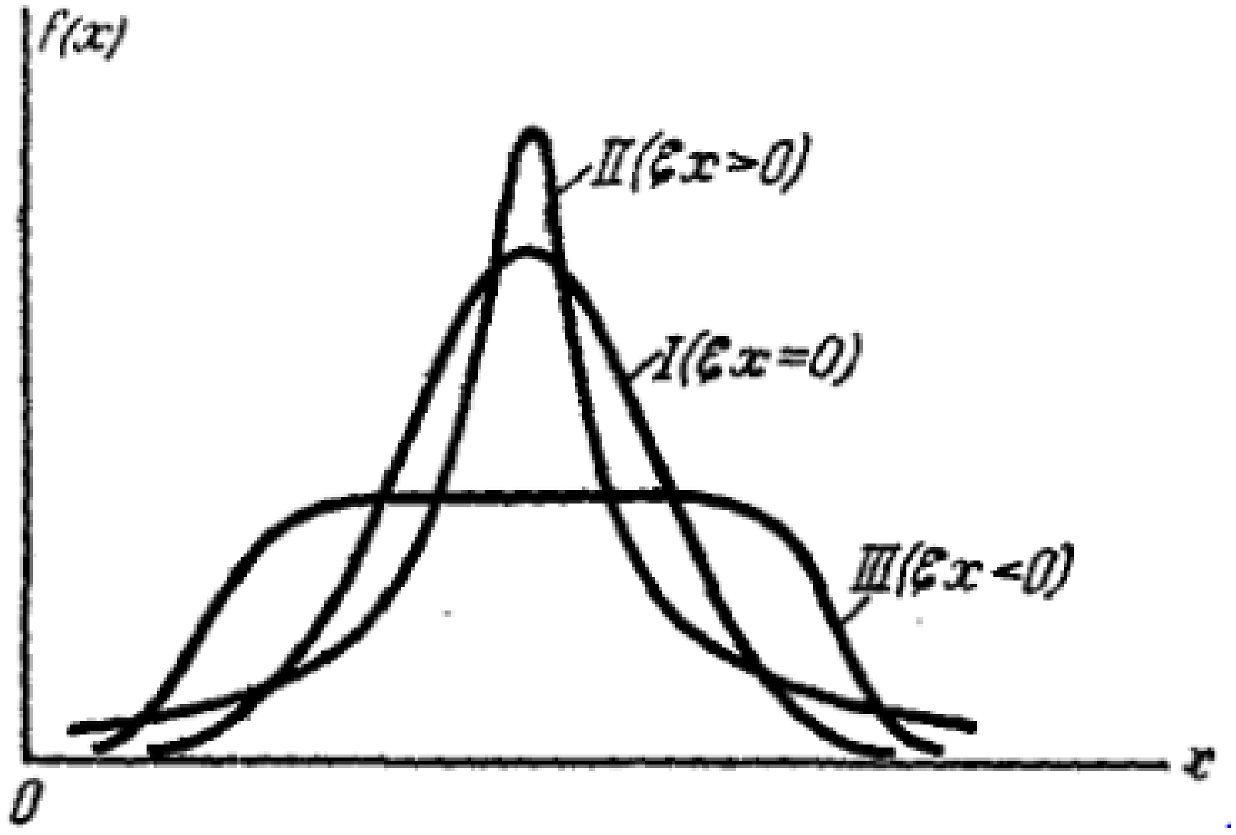
$$A_s = \frac{m_3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$$



Числові характеристики рядів розподілу

Ексцесом Е називається зменшене на три одиниці відношення центрального моменту четвертого порядку до четвертого степеню середнього квадратичного відхилення:

$$E = \frac{m_4}{\sigma^4} - 3$$

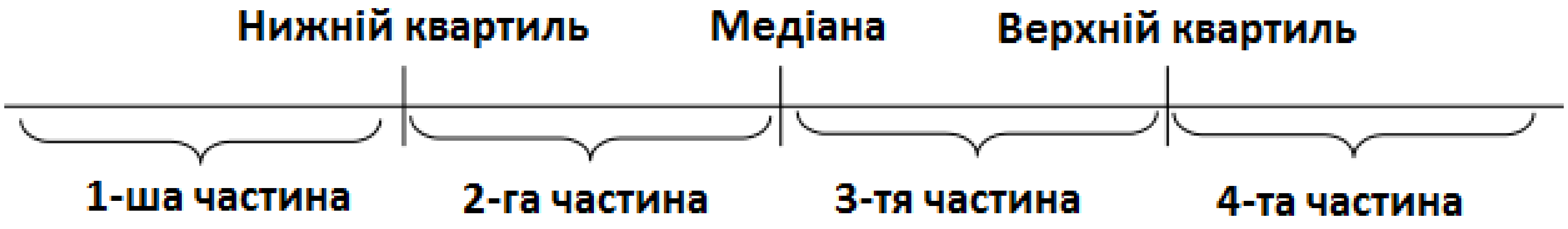


Числові характеристики рядів розподілу

Термін **квантиль** відноситься до числових характеристик, які можуть надати інформацію про розсіювання даних. Точніше, квантиль розділяє відсортовані дані на групи, кожна з яких містить однакову частину набору даних. Якщо група відповідає певному відсотку даних, то квантиль називається **персентілом**.

Числові характеристики рядів розподілу

Найчастіше використовують три квантилі, які називаються **квартилі** Q_L, Q_M, and Q_H, які ділять відсортовані дані на чотири рівні частини.



NumPy : np.percentile(), np.quantile()

```
np.percentile(a, [25, 50, 75])
np.quantile(a, [0.25, 0.5, 0.75])
```

Числові характеристики рядів розподілу

Основні статистичні характеристики можна отримати за допомогою функції SciPy: scipy.stats.describe ()

За замовчання ddof=1

Повертає:

- кількість елементів
- мінімальне та максимальне значення
- вибіркове середнє
- дисперсію
- коефіцієнт асиметрії (skewness)
- ексцес (kurtosis)

Статистичні гіпотези

Часто необхідно знати закон розподілу генеральної сукупності. Якщо закон розподілу невідомий, але є міркування для припущення його певного вигляду А, наприклад, розподіл рівномірний, показниковий або нормальний, тоді висувають гіпотезу:

генеральна сукупність розподілена за законом А.

Статистичними називають гіпотези про вигляд розподілу генеральної сукупності або про параметри відомих розподілів.

Статистичні гіпотези

Основною (нульовою) називають висунуту гіпотезу і позначають H_0.

Вона завжди є твердженням виду: генеральна сукупність розподілена за законом А або параметр розподілу генеральної сукупності дорівнює θ.

Альтернативною (конкурентною) називають гіпотезу, що суперечить основній, її позначають H_1. Альтернативна гіпотеза може бути **двосторонньою**:

генеральна сукупність не розподілена за законом А або параметр розподілу генеральної сукупності не дорівнює θ.

або **одностронньою**:

параметр розподілу генеральної сукупності менше θ або параметр розподілу генеральної сукупності більше θ.

Статистичні гіпотези

Якщо за висновком буде відкинута правильна гіпотеза, то кажуть, що це **помилка першого роду**.

Якщо за висновком буде прийнята неправильна гіпотеза, то кажуть, що це **помилка другого роду**.

Рішення	**H_0 правильна**	**H_0 хибна**
Не вдалося відкинути H_0	Немає помилки	Помилка другого роду
Відкинути H_0	Помилка першого роду	Немає помилки

Критерії для перевірки гіпотез

Статистичним критерієм узгодження перевірки гіпотези (або просто критерієм) називають випадкову величину K, розподіл якої (точний або наближений) відомий і яка застосовується для перевірки основної гіпотези.

Спостереженим значенням критерію узгодження називають значення відповідного критерію, обчислене за даними вибірки.

Критерії для перевірки гіпотез

Після обрання певного критерію узгодження, множину усіх його можливих значень поділяють на дві підмножини, що не перетинаються: одна з них містить значення критерію, при яких основна гіпотеза відхиляється, а друга – при яких вона приймається.

Критерії для перевірки гіпотез

Критичною областю називають сукупність значень критерію, при яких основна гіпотеза відхиляється.

Областю прийняття гіпотези (областю допустимих значень) називають множину значень критерію, при яких гіпотезу приймають.

Критерії для перевірки гіпотез

Критерій узгодження K – одновимірна випадкова величина, усі її можливі значення належать деякому інтервалу. Тому критична область та область прийняття гіпотези також будуть інтервалами, а це означає, що існують точки, які ці інтервали відокремлюють.

Критичними точками (межами) критерію K називають точки $k_{кр}$, які відокремлюють критичну область від області прийняття гіпотези.

Критерії для перевірки гіпотез

Правобічною називають критичну область, що визначається нерівністю $K>k_{kp}$. Відповідає односторонній альтернативній гіпотезі виду $\theta> \theta_0$

Лівобічною називають критичну область, що визначається нерівністю $K < k_{kp}$. Відповідає односторонній альтернативній гіпотезі виду $\theta < \theta_0$

Критерії для перевірки гіпотез

Щоб знайти однобічну критичну область, треба знайти критичну точку $k_{кр}$. Для цього задають достатньо малу ймовірність – рівень значущості α, а потім шукають критичну точку з врахуванням вимоги

$$P(K>k_{кр})=\alpha$$

у випадку правобічної критичної області або

$$P(K<k_{кр})=\alpha$$

у випадку лівобічної критичної області.

Критерії для перевірки гіпотез

У випадку двобічної критичної області повинно виконуватись тотожність

$$P(K< k_1)+P(K>k_2)= \alpha$$

Алгоритм перевірки гіпотез

1) Визначити параметр, стосовно якого потрібно перевірити гіпотезу
2) Визначити основну гіпотезу H_0
3) Визначити гіпотезу H_1 альтернативну до гіпотези H_0
4) Обрати статистичний критерій для перевірки
5) Визначити критерій відхилення основної гіпотези, наприклад, рівень значущості α
6) Розрахувати числові характеристики вибірки, потрібні для отримання значення статистичного критерію. Обчислити значення статистичного критерію
7) Зробити висновки: чи потрібно відхилити основну гіпотезу.

Критерії для перевірки гіпотез

P-значення - це найменший рівень значущості, який би призвів до відкидання нульової гіпотези H_0 з наведеними даними.

P-значення надає інформацію про вагу доказів проти H_0, і тому можна зробити висновок на будь-якому визначеному рівні значущості.

Зазвичай прийнято вважати значення статистичного критерію (і даних) суттєвими, коли нульова гіпотеза H_0 відкидається; отже, *P-значення* можна вважати найменшим рівнем α, на якому дані є значущими. Іншими словами, *P-значення* - це спостережуваний рівень значущості. Після того, як P-значення стане відомим, особа, яка приймає рішення, може визначити, наскільки важливими є дані, не зважаючи на попередньо вибраний рівень значущості.

Критерії для перевірки гіпотез

При тестуванні гіпотез часто використовують ймовірність помилки першого роду або рівень значущості $\alpha = 0{,}05$. Таке значення є результатом досвіду і не є універсальним.

Таким чином, нульова гіпотеза відхиляється, якщо значення статистичного критерію потрапляє в критичну область або р-значення менше 0,05.

Гіпотези про математичне сподівання

Розглянемо тестування гіпотез щодо математичного сподівання генеральної сукупності з нормальним розподілом і з невідомою дисперсією.

Наприклад, перевірити чи середня довжина пелюстки півника дорівнює 1,5 см.

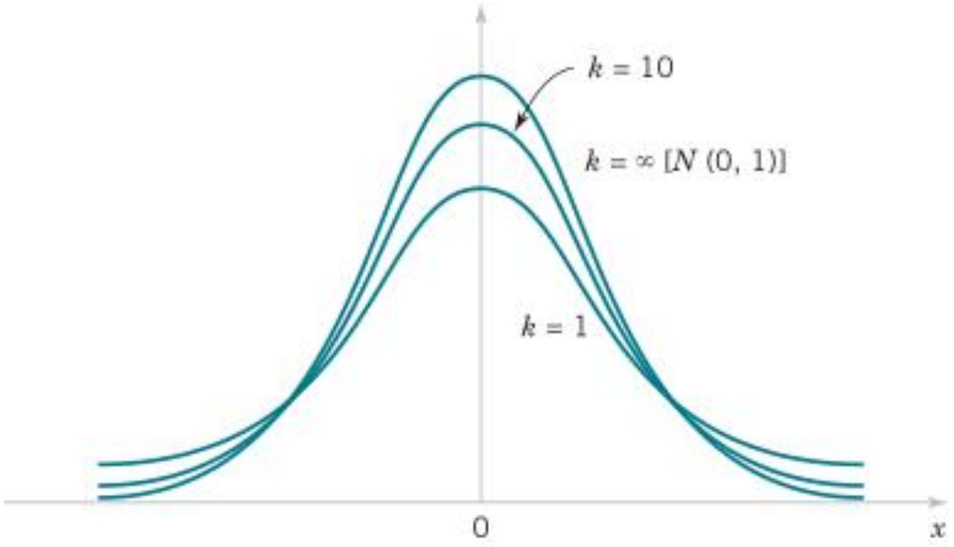
Гіпотези будуть мати вигляд:

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

Статистичний критерій:

$$T_0 = \left(\bar{X} - \mu_0\right) / \left(S / \sqrt{n}\right)$$

`scipy.stats.ttest_1samp()`



Гіпотези про математичне сподівання

```
import numpy as np
from scipy import stats
a = np.random.normal(size=100)
stats.ttest_1samp(a,5.0)
```

Результат:

Ttest_1sampResult(statistic=-44.927784285791816, pvalue=1.1717539782314286e-67)

```
stats.ttest_1samp(a, 0.0)
```

Результат:

Ttest_1sampResult(statistic=0.12155047853769656, pvalue=0.9035014128541538)

Гіпотези про математичне сподівання

Можливо також перевірити чи має генеральна сукупність нормальний розподіл: **`normaltest()`**

```
import numpy as np
from scipy import stats
a = np.random.normal(size=100)
stats.normaltest(a)
```

Результат:

NormaltestResult(statistic=1.6336291460900272, pvalue=0.44183685426247565)

Статистичні гіпотези

Перевірка гіпотези про рівність математичних сподівань нормальних генеральних сукупностей при невідомих середньоквадратичних відхиленнях.

Нехай дві нормально розподілені генеральні сукупності мають рівні дисперсії, а математичні сподівання можуть бути різними.

З сукупностей зробили вибірку об'єму n_1 і n_2 і знайшли вибіркові

середні $\overline{x}_1$ та $\overline{x}_2$, а також вибіркові середньоквадратичні відхилення s_1 та s_2 відповідно.

Наприклад, перевірити чи однакова довжина пелюстки двох видів півників?

Статистичні гіпотези

Перевіримо гіпотезу про те, що різниця математичних сподівань двох генеральних сукупностей дорівнює певному числу (яке також може дорівнювати нулю). Тоді

H_0: $\mu_1 - \mu_2 = c_0$

Альтернативна гіпотеза буде

H_1: $\mu_1 - \mu_2 \neq c_0$

Статистичний критерій: $$T_0 = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - c_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Де $S_p^2 = \dfrac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}$ - об'єднана дисперсія

Статистичні гіпотези

Якщо нульова гіпотеза вірна, то статистичний критерій приблизно буде мати розподіл Стьюдента з $n_1 + n_2 - 2$ степенями вільності.

Якщо дві вибірки є незалежними, тоді використовується:

`scipy.stats.ttest_ind()`

Якщо дві вибірки є залежними, тоді використовується:

`scipy.stats.ttest_rel()`

Наприклад, перевірити чи змінилась довжина пелюстки півників за сто років?

Статистичні гіпотези

```
import numpy as np
from scipy import stats
a = np.random.normal(size=100)
b = np.random.normal(size=90)
stats.ttest_ind(a,b)
```

Результат:

Ttest_indResult(statistic=-1.052724075942592, pvalue=0.2938186373247703)

Статистичні гіпотези

Якщо достатньо вивести лише р-значення:

```
stats.ttest_ind(a,b).pvalue
```

0.4595654995289018

За замовчанням перевіряється двостороння альтернативна гіпотеза, для односторонньої потрібно змінити параметр alternative на ‘less’ або ‘greater’

```
c=stats.ttest_ind(a,b).pvalue
d=stats.ttest_ind(a,b,alternative='less').pvalue
c,d
```

(0.6464796938532716, 0.3232398469266358)

Наприклад, перевірити чи довжина пелюстки півника першого виду більша, ніж довжина пелюстки другого виду?

Гіпотези про розподіл

Критерій узгодження Пірсона (хі-квадрат) ефективно використовують для перевірки гіпотези про розподіл генеральної сукупності, тобто гіпотези що розподіл випадкової величини має певний функціональний вираз.

Гіпотези про розподіл

Основна гіпотеза H_0: генеральна сукупність має даний розподіл.

Альтернативна гіпотеза H_1: генеральна сукупність не має даного розподілу.

Процедура перевірки вимагає випадкової вибірки об’єму n з генеральної сукупності, розподіл якої невідомий. Ці n спостережень можна поділити на k класів (як для побудови гістограми). Нехай O_i - спостережувана частота в i-ому класі. Виходячи з теоретичного розподілу ймовірностей обчислюємо очікувану частоту в i-ому класі E_i. Статистичний критерій:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

Гіпотези про розподіл

χ_0^2 приблизно має розподіл хі-квадрат з k−p−1 степенями вільності, де p являє собою кількість параметрів теоретичного розподілу, оцінених за статистикою вибірки. Ми повинні відкинути нульову гіпотезу, якщо значення статистичного критерію занадто велике. Тому P-значення буде дорівнювати $P = P(\chi_{k-p-1}^2 > \chi_0^2)$

Для тесту з фіксованим рівнем значущості ми відкидаємо нульову гіпотезу, якщо обчислене значення статистичного критерію

$$\chi_0^2 > \chi_{\alpha,k-p-1}^2$$

Гіпотези про розподіл

```
scipy.stats.chisquare(f_obs, f_exp=None, ddof=0, axis=0)
```

f_obs – спостережувані частоти

f_exp – очікувані частоти

ddof – різниця ступенів вільності; кількість ступенів вільності k - 1 - ddof

Гіпотези про розподіл

Приклад. Існує припущення, що кількість дефектів на друкованих платах має розподіл Пуассона. Було зібрано випадкову вибірку з n = 60 друкованих плат та виявлено наступну кількість дефектів.

Кількість дефектів	Частота
0	32
1	15
2	9
3	4

Перевірити правильність припущення з рівнем значущості 0,05.

Гіпотези про розподіл

<u>Приклад</u>. Математичне сподівання генеральної сукупності невідоме. Його оцінка, тобто вибіркове середнє дорівнює 0,75. Для розподілу Пуассона з параметром 0,75 можна обчислити p_i, теоретичну ймовірність і-го класу.

$$p_1 = P(X = 0) = \frac{e^{-0,75}(0,75)^0}{0!} = 0,472$$

$$p_2 = P(X = 1) = \frac{e^{-0,75}(0,75)^1}{1!} = 0,354$$

$$p_3 = P(X = 2) = \frac{e^{-0,75}(0,75)^2}{2!} = 0,133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0,041$$

Гіпотези про розподіл

Приклад. Очікувані частоти обчислюються шляхом множення об’єму вибірки $n = 60$ в на ймовірності p_i. Тобто $E_i = np_i$. Очікувані частоти наступні:

Кількість дефектів	Ймовірності	Очікувані частоти
0	0,472	28,32
1	0,354	21,24
2	0,133	7,98
3 (або більше)	0,041	2,46

Гіпотези про розподіл

Приклад. Оскільки очікувана частота в останній комірці менше 3, ми об’єднуємо дві останні комірки:

Кількість дефектів	Частоти	Очікувані частоти
0	32	28,32
1	15	21,24
2	13	10,44

Гіпотези про розподіл

Приклад.

1. Параметр, що цікавить: розподіл генеральної сукупності.
2. Основна гіпотеза: H_0: генеральна сукупність має розподіл Пуассона.
3. Альтернативна гіпотеза: H_1 : розподіл генеральної сукупності – це не розподіл Пуассона.
4. Статистичний критерій:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

Гіпотези про розподіл

<u>Приклад</u>.

5. Статистичний критерій хі-квадрат матиме $k-p-1 = 3-1-1 = 1$ степенів вільності. Відхилити нульову гіпотезу, якщо P-значення менше, ніж 0,05.

6. Розрахунки:

$$\chi_0^2 = \frac{(32-28,32)^2}{28,32} + \frac{(15-21,24)^2}{21,24} + \frac{(13-10,44)^2}{10,44} = 2,94$$

Гіпотези про розподіл

Приклад.

7. Висновки: З таблиці розподілу хі-квадрат можна знайти, що

$\chi^2_{0.10,1} = 2,71$ і $\chi^2_{0.05,1} = 3,84$

Оскільки $\chi^2_0 = 2,94$ лежить між цими значеннями, ми робимо висновок, що P-значення становить від 0,05 до 0,1. Тому, оскільки P-значення 0,05, ми не можемо відкинути нульову гіпотезу про те, що розподіл дефектів на друкованих платах є розподілом Пуассона.

Гіпотези про розподіл

```
import numpy as np
from scipy import stats
stats.chisquare([32,15,13],f_exp=[28.32,21.24,10.44],ddof=1)
```

Результат:

Power_divergenceResult(statistic=2.939151892980063, pvalue=0.0864561164314485)

Очікувані частоти можна було розрахувати як:

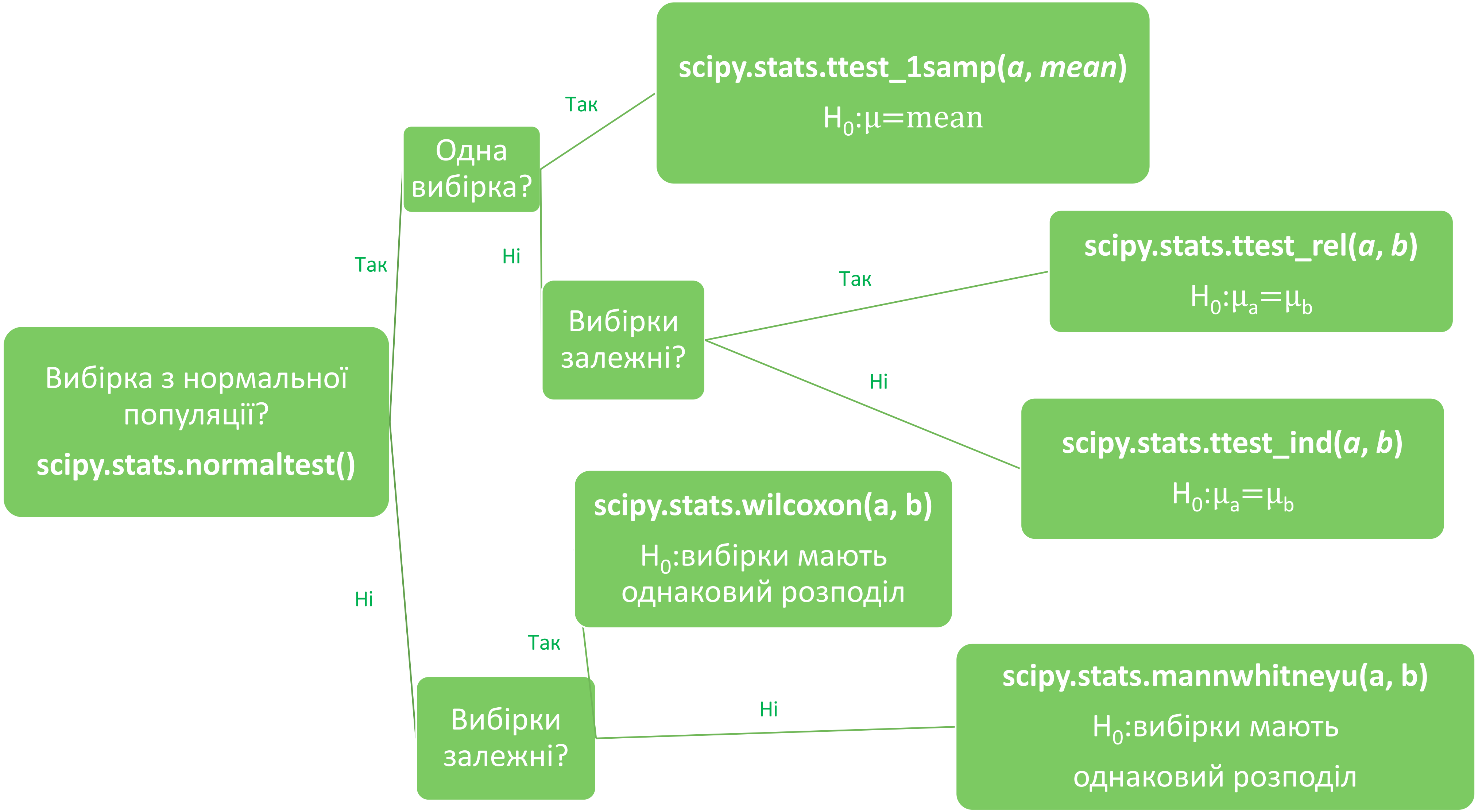
```
f_exp = 60*stats.poisson.pmf([0,1,2], 0.75)
```

Статистичні гіпотези

Гіпотези або тести, які потребують попередніх умов, наприклад, що розподіл генеральної сукупності повинен бути нормальним, називаються параметричними.

Гіпотези, які не потребують таких умов, є непараметричними.

Статистичні гіпотези



Кореляційний аналіз

Дві випадкові величини **незалежні**, якщо закон розподілу однієї з них не залежить від того, які можливі значення прийняла друга величина.

Випадкові величини **залежні**, якщо закон розподілу однієї величини залежить від того, які значення прийняла друга величина.

Кореляційний аналіз

Для вимірювання одночасно розсіювання двох величин від їхніх математичних сподівань використовується коваріація K_{XY}.

Або, для більшої зручності, коефіцієнт кореляції:

$$r_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}$$

Коефіцієнт кореляції є кількісною характеристикою лінійності залежності випадкових величин X та Y.

Випадкові величини X та Y називають **некорельованими**, якщо їх кореляційний момент або коефіцієнт кореляції дорівнює нулеві.

Кореляційний аналіз

Властивості коефіцієнта кореляції:

1) $|r_{XY}| \le 1$

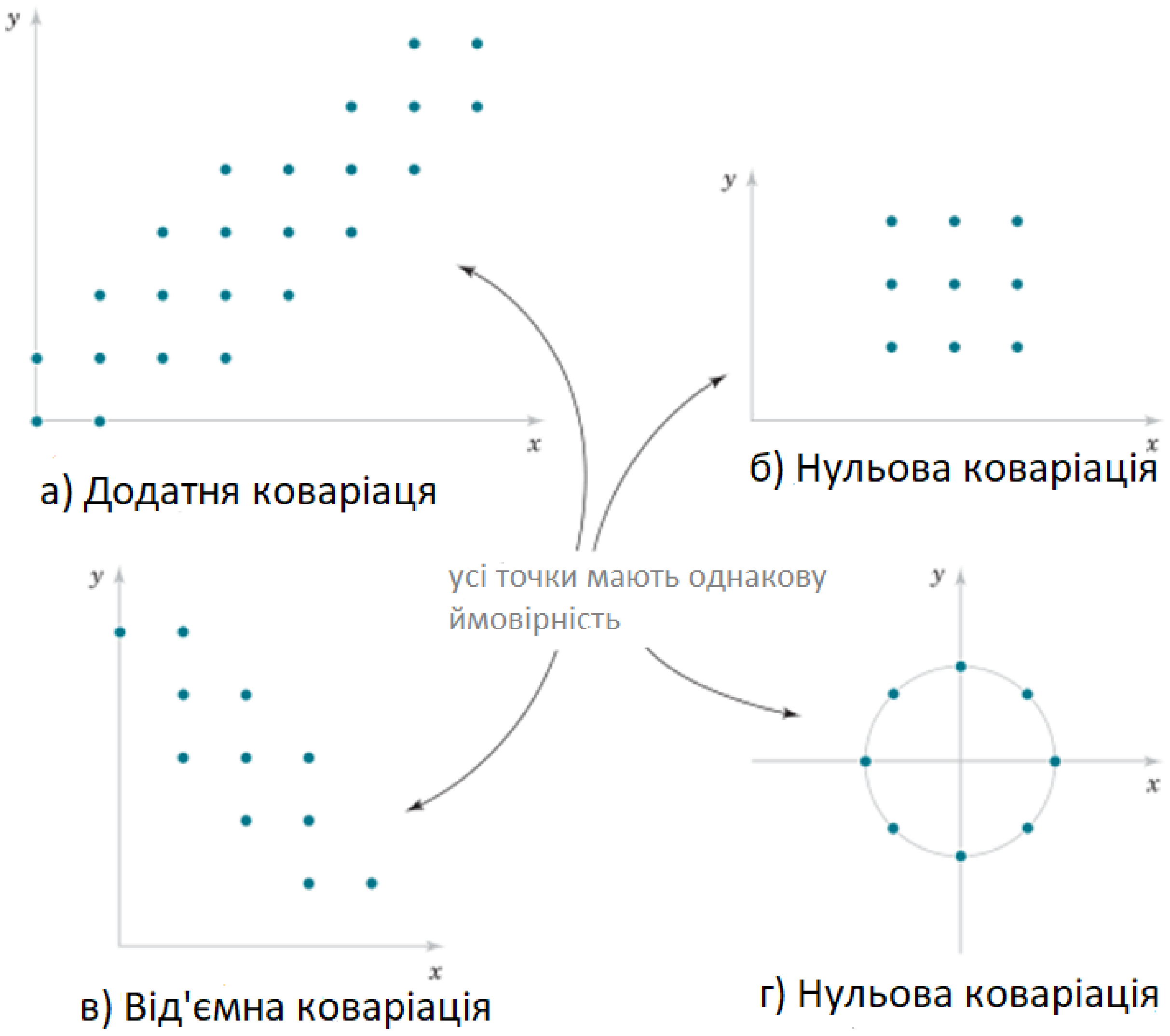
2) якщо X та Y незалежні, то $r_{XY} = 0$

3) якщо між X та Y є лінійна залежність Y = aX + b, де a та b – постійні, то $r_{XY} = 1$

Кореляційний аналіз

Із незалежності двох величин випливає їх некорельованість, але із некорельованості ще не випливає незалежність цих величин. У випадку нормального розподілу величин із некорельованості випадкових величин випливає їх незалежність.

Кореляційний аналіз



Кореляційний аналіз

Відповідно, можно знайти коваріацію та кореляційний момент для вибірок.

Для розрахунку коваріації можна використати функцію

np.cov()

```
import numpy as np
a = np.random.normal(0,5,100)
b = np.random.normal(0,5,100)
np.cov(a,b), np.cov(a), np.var(b,ddof=1)
```

Результат:

```
array([[24.66813   ,  4.1147734 ],
       [ 4.1147734 , 28.14296067]]),
 array(24.66813),
 28.14296066642728
```

Кореляційний аналіз

Для розрахунку коефіцієнта кореляції можна використати функцію

`np.corrcoef()`

`np.corrcoef(a,b)`

Результат:

```
array([[1.        , 0.15616831],
       [0.15616831, 1.        ]])
```

Кореляційний аналіз

Знайти коефіцієнт кореляції (Пірсона), а також перевірити гіпотезу про те, що коефіцієнт кореляції дорівнює нулю можна використавши функцію:

```
scipy.stats.pearsonr()
```

```
from scipy import stats
stats.pearsonr(a,b)
```

Результат:

(0.15616831244485752, 0.1207615031832596)

Кореляційний аналіз

Коефіцієнт кореляції рангу Спірмена є непараметричним показником монотонності зв'язку між двома наборами даних. На відміну від кореляції Пірсона, кореляція Спірмена не передбачає нормального розподілу обох наборів даних. Наступна функція також перевіряє гіпотезу про те, що коефіцієнт кореляції дорівнює нулю. Цей коефіцієнт приймає значення від -1 до 1.

```
scipy.stats.spearmanr()
```

```
from scipy import stats
a=np.random.randint(0,10,100)
b=np.random.randint(0,10,100)
stats.spearmanr(a,b)
```

Результат:

SpearmanrResult(correlation=0.1157619680397407, pvalue=0.251412112736803)

Кореляційний аналіз

Перевірити, чи є зв’язок між шириною і довжиною пелюстки півників.

```
from scipy import stats
petal_l = np.array(data['petal_length'])
petal_w = np.array(data['petal_width’])
stats.spearmanr(petal_l,petal_w)
```

Результат:

SpearmanrResult(correlation=0.9360033509355781, pvalue=5.383649646073561e-69)

Кореляційний аналіз

Вибірки з нормальної популяції?

scipy.stats.normaltest()

Так

Лінійний зв’язок:

scipy.stats.pearsonr(*x*, *y*)

H_0:r=0

Ні

Монотонний зв’язок:

scipy.stats.spearmanr(a, *b)*

H_0:r=0

Лінійна регресія

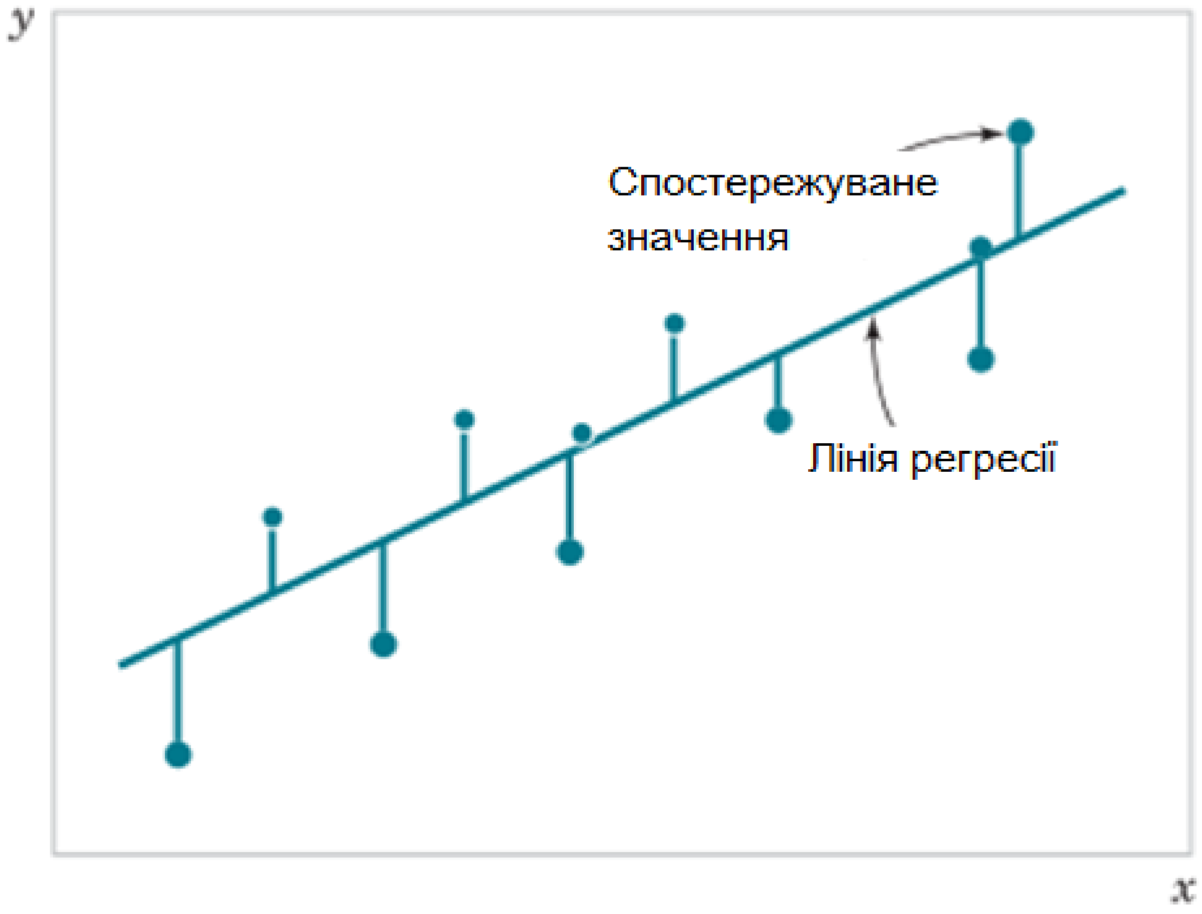
Набір статистичних інструментів, які використовуються для моделювання та дослідження взаємозв'язків між змінними, пов'язаними недетермінованим чином, називається **регресійним аналізом**.

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

де ε - випадкова помилка. Ця модель називається простою **лінійною регресійною моделлю**, оскільки вона має лише одну незалежну змінну або **регресор**. Регресійна модель є емпіричною моделлю.

Лінійна регресія

Припустимо, існує *n* пар спостережень (x_1,y_1), (x_2,y_2), … ,(x_n,y_n). Оцінки β_0 та β_1 повинні описувати лінію, яка найкраще підходить до даних.



Лінійна регресія

Для знаходження лінійної регресії можна використати наступну функцію:

`scipy.stats.linregress ()`

Вона повертає:

slope – нахил лінії регресії, β_1
intercept – перетин лінії регресії, β_0
rvalue – коефіцієнт кореляції
pvalue – p-значення тесту про те, що нахил дорівнює нулю
stderr, intercept_stderr – середнє квадратичне відхилення середнього арифметичного нахилу та перетину.

Лінійна регресія

```
scipy.stats.linregress ()

import numpy as np
from scipy import stats
a = np.random.random(10)
b = 2*a+5
stats.linregress(a,b)
```

Результат:
LinregressResult(slope=2.0000000000000004, intercept=5.000000000000001, rvalue=1.0, pvalue=4.375000000000076e-80, stderr=0.0, intercept_stderr=0.0)

Лінійна регресія

```
import numpy as np
from scipy import stats
a = np.random.random(10)
b = 2*a+np.random.random(7)
stats.linregress(a,b)
```

Результат:

LinregressResult(slope=2.2294269436640164, intercept=0.4245031492174052, rvalue=0.9510331726885244, pvalue=0.0009924760424436887, stderr=0.3240379584817232, intercept_stderr=0.19174282728240075)