Прикладний статистичний аналіз

Ігор Мірошниченко

Юлія Хлевна

2025-03-13

Зміст

# Передмова

# Вступ

# 1. Біноміальний критерій

## 1.1 Генеральна сукупність та вибірка

Ви вирішили створити платформу онлайн-курсів з програмування. Ви записали навчальні відео та запропонували користувачам доступ за передплатою. Вартість курсу для студента становить 1000 рублів, а витрати на підтримку платформи та індивідуальні консультації коштують вам 500 рублів з кожного студента.

Проте ви помічаєте, що деякі люди відмовляються від курсу після першого заняття, якщо матеріал їм здається складним або нецікавим. Інвестори готові підтримати ваш проєкт, якщо рівень відмов буде нижче 50%.

Щоб це перевірити, ви проводите експеримент: залучаєте 30 нових студентів. 20 із них проходять курс й оплачують доступ, а 11 відмовляються. 20 — це більше половини, але чи достатньо цього, щоб довести перспективність проєкту?

Розв’язуючи таку задачу, ми припускаємо, що існує певна аудиторія, яка користуватиметься нашим сервісом. Цю групу називають **генеральною сукупністю**. Якщо запустити сервіс для всіх потенційних користувачів, у ньому буде певна частка успішних випадків, позначимо її як . Це невідомий параметр, який ми не можемо визначити безпосередньо. Натомість ми можемо проводити експерименти та **досліджувати** результати. Оскільки протестувати продукт на всій аудиторії неможливо, ми беремо **вибірку** з генеральної сукупності та аналізуємо частку успішних випадків.

Згідно з результатами нашого експерименту, спостережувана ймовірність оплати становить [[1]](#footnote-22). Це означає, що 67% студентів оплатили доступ. Чи можемо ми зробити висновок, що справжня частка успішних випадків перевищує 50%?

Розгляньмо, чому отримане значення *може не бути* переконливим доказом. Припустимо, що ймовірність успішної оплати дорівнює , і змоделюємо можливі результати для 30 студентів.

Давайте спростимо цю задачу до прикладу з підкиданням монетки та змоделюємо результати для 30 спроб:

* Якщо монетка випаде орлом, студент оплачує доступ.
* Якщо монетка випаде решкою, студент відмовляється від курсу.
* Використаємо метод integers() до класу Generator, яка генерує випадкові цілі числа в заданому діапазоні.
* Підкинемо монетку 30 разів та порахуємо кількість успішних випадків.

rng = np.random.default\_rng(18)  
  
n = 30  
results = rng.integers(0, 1, size = 30, endpoint = True)  
success = np.sum(results) / n  
  
print(f"Кількість успішних випадків: {round(success, 3) \* 100}%")

Рядок 1

Ініціалізуємо генератор випадкових чисел з фіксованим seed.

Рядок 3

Кількість студентів.

Рядок 4

Генеруємо випадкові числа[[2]](#footnote-24) для кожного студента.

Рядок 5

Обчислюємо частку успішних випадків.

Рядок 7

Виводимо результат.

Кількість успішних випадків: 70.0%

Ми бачимо, що в експерименті частка успішних випадків навіть перевищила 63%, тоді як у симуляції була закладена ймовірність 50%.

Тому, на жаль, ми не можемо з абсолютною точністю визначити, яким є справжнє значення у генеральній сукупності та чи перевищує воно 50%, незалежно від того, скільки спостережень ми проводимо. Однак, застосовуючи методи прикладної статистики, ми зможемо використати інструменти, які допоможуть ухвалити правильне рішення, зокрема й у цьому випадку.

## 1.2 Статистичні гіпотези

Ми з’ясували, що навіть за ймовірності можна отримати значну кількість успішних випадків. Насправді ми спеціально підбирали seed для отримання такого результату. Якщо повторити цей експеримент з іншим значенням seed або збільшити кількість спостережень, результат може виявитися іншим.

|  |
| --- |
| Порада |
| Спробуйте змінити seed (наприклад 22) або кількість спостережень та перевірте, як змінюється результат. |

Тож велика кількість успішних випадків може бути результатом випадковості. Щоб вирішити, чи можна вважати результати експерименту **статистично значущими** необхідно отримати відповідь на питання:

Чи можна вважати, що спостережуване значення є більшим від ?

Звернімося до теорії ймовірностей. Факт підписки на наш сервіс для кожного окремого студента можна розглядати як випадкову величину , яка підпорядковується розподілу Бернуллі[[3]](#footnote-29). Параметр цього розподілу, а саме ймовірність успіху, нам невідомий.

де — ймовірність успіху.

Нас цікавить підтвердження того, що . У статистиці для перевірки гіпотез розглядають дві можливості:

* **Нульова гіпотеза** () формулюється як твердження, яке ми прагнемо спростувати.
* **Альтернативна гіпотеза** () висловлює припущення, яке ми хочемо довести.

Скорочено це записують як:

Зауважимо, що якщо в нашому експерименті з 30 студентами можна дивитися не на частку успіхів, а на їх **кількість**.

Тоді питання можна переформулювати так:

За умови вірності наскільки ймовірно отримати 20 або більше успішних випадків з 30?

Якщо ми проводимо незалежних спостережень, то сума цих випадкових величин також підпорядковується біноміальному розподілу[[4]](#footnote-30).

де — випадкова величина, яка показує успіх у -му спостереженні, — кількість успішних випадків у спостереженнях, — кількість спостережень, — ймовірність успіху.

Давайте подивимось, як це виглядає графічно. Для цього побудуємо графік функції щільності ймовірностей для біноміального розподілу з параметрами та .

n = 30  
mu = 0.5  
  
x = np.arange(0, n + 1)  
y = binom.pmf(x, n, mu)  
  
plt.bar(x, y, color=turquoise)  
plt.bar(x[x >= 20], y[x >= 20], color=red\_pink)  
plt.xlabel("Кількість успішних випадків")  
plt.ylabel("Ймовірність")  
plt.show()

|  |
| --- |
| Рисунок 1.1 -  Функція щільності ймовірностей при |

На графіку зображено функцію щільності ймовірностей для біноміального розподілу з параметрами та . Ціановим кольором позначено ймовірності для кожної кількості успішних випадків. Рожевими виділено ймовірності для кількості успішних випадків, яка дорівнює або перевищує 20.

# 2. -критерій Фішера

# 3. -критерій Стьюдента

# 4. Монте-Карло в задачах статистики

# 5. Критерій Колмогорова

# 6. Критерій Пірсона

# 7. U-критерій Манна-Уітні

# 8. Бутстреп

# 9. Підсумки

# Список літератури

1. У статистиці позначається як оцінка параметра . [↑](#footnote-ref-22)
2. Метод integers() генерує випадкові цілі числа в заданому діапазоні. Аргумент endpoint вказує, що верхня межа включається у діапазон. [↑](#footnote-ref-24)
3. Розподіл Бернуллі — це дискретний розподіл ймовірностей, який моделює випадковий експеримент з двома можливими результатами: успіхом або невдачею. [↑](#footnote-ref-29)
4. Біноміальний розподіл моделює кількість успішних випадків у послідовності незалежних випробувань. Сума незалежних випадкових величин з розподілу Бернуллі підпорядковується біноміальному розподілу. [↑](#footnote-ref-30)