Прикладний статистичний аналіз

Ігор Мірошниченко

Юлія Хлевна

2025-03-14

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 гривень, а витрати на підтримку платформи та індивідуальні консультації коштують вам 500 гривень 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

### 1.2.1 Постановка задачі

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 |

[Рисунок 1.1](#fig-binom-pmf) демонструє функцію щільності ймовірностей для біноміального розподілу з параметрами та . Ціановим[[5]](#footnote-35) кольором позначено ймовірності для кожної кількості успішних випадків. Рожевими виділено ймовірності для кількості успішних випадків, яка перевищує або дорівнює 20.

### 1.2.2 Критерій

Щойно ми розробили алгоритм, який на основі вибірки або визнає наявність доказів на користь , або повідомляє, що таких доказів немає. Відповідно, він або відхиляє , або не відхиляє її.

Такий алгоритм називається **критерієм**. Його можна подати у вигляді функції , яка приймає реалізацію вибірки та повертає , якщо слід відхилити , та в іншому випадку.

Давайте припустимо, що ми вирішили відхилити , якщо кількість успішних випадків перевищує або дорівнює 21. Тоді критерій набуде вигляду:

Зазвичай скорочують запис і пишуть просто правило, за яким відхиляємо

Позначимо , , тоді критерій набуває вигляду:

Так влаштована більшість класичних критеріїв у прикладній статистиці, тому величинам у ньому дано спеціальні назви. називається **статистикою критерію**, — **критичним значенням**.

може бути будь-якою функцією від вибірки, яку ви вважаєте логічною для перевірки гіпотези. У нашому випадку це кількість успіхів, або сума всіх . Але ви можете вибрати й інші: максимальне значення, суму перших 5 значень або навіть просто перший елемент.

### 1.2.3 Критична область

Знову перепишемо наше основне запитання, тільки тепер з використанням нашого критерію :

Наскільки часто може бути таке, що за справедливості критерій відхиляє гіпотезу?

Відповідь на це запитання залежить від критичного значення. Зараз ми взяли його рівним 21, побачивши на картинці, що великі відхилення відбуваються при рідко. Але що означає рідко й наскільки рідко, не сказали. Тепер наша мета зрозуміти, як вибрати критичне значення , виходячи з **частоти помилок** нашого критерію.

Вибираючи , ми можемо або часто відхиляти нульову гіпотезу, коли мале, або можемо робити це рідше, коли велике. Щоб вибрати правильне значення, потрібно визначитися, коли наш критерій помиляється.

* . Якщо відхиляти гіпотезу при отриманні хоча б 16 успішних підписок із 30, то це навряд чи влаштує інвесторів. Так, успіхів більше половини. Але якщо в генеральній сукупності ймовірність 0.5, то майже в половині випадків ми будемо відхиляти гіпотезу. Критерій помилково повертає , тобто це помилка **хибно позитивна** (false positive, **FP**).
* . У такому разі будемо відхиляти гіпотезу тільки за 29 або 30 успіхів. Ці значення, звісно, говорять про те, що відхилення від 50% успіхів сильне. Але якщо в генеральній сукупності ймовірність, наприклад, 60%, то такі значення будуть виходити рідко. Але ж такі ймовірності теж влаштували б інвесторів, й ми б змогли відкрити стартап! А з таким критерієм ми навряд чи доб’ємося цього. Не відхилити гіпотезу , коли вона неправильна — це теж помилка. Вона називається **хибно негативна** (false negative, **FN**), оскільки критерій повернув 0 помилково.

У нашому завданні інвесторам важливіше хибно позитивна помилка. Їм дуже не хочеться потрапити в ситуацію, коли їм показали доказ успішності бізнесу, а виявилося, більшість користувачів відмовляється оформлювати підписку й компанія не отримує прибуток. Це призведе до збитків. Хибно негативна помилка призведе до того, що ви втратите успішний бізнес, але інвестори грошей не втратять.

Тому виберемо поріг, щоб ймовірність хибно позитивної помилки була задовільною, або ж **частота хибнопозитивних спрацьовувань** (False Positive Rate, FPR). Для цього треба зрозуміти, як часто ми будемо відхиляти гіпотезу, за умови вірності .

Тепер знову переформулюємо основне питання, повністю з використанням нових термінів, й врешті-решт відповімо на нього.

Який FPR у критерію для перевірки гіпотези проти ?

Коли є вірною, щоб порахувати кількість успіхів ми проводили 30 разів підкидання монетки з ймовірністю орла . Кількість орлів (тобто успіхів) у такому експерименті має розподіл, який називається біноміальним, тобто при наша статистика має біноміальний розподіл .

Обчислимо FPR для

Це вже ймовірність події за конкретного розподілу випадкової величини. Його можна подивитися за таблицею або, що зручніше, обчислити з використанням мов програмування.

### 1.2.4 Обчислення FPR

Давайте порахуємо суму ймовірностей для кількостей успіхів від 21 до 30 включно. Покажемо графічно, як це виглядає на [Рисунку 1.2](#fig-binom-pmf-fpr).

x = np.arange(0, n + 1)  
y = binom.pmf(x, n, 0.5)  
  
plt.bar(x, y, color=turquoise)  
plt.bar(x[x >= crit\_subs], y[x >= crit\_subs], color=red\_pink)  
for i in range(crit\_subs - 2, crit\_subs + 4):  
 plt.text(i + 0.5, y[i] + 0.001, f"{round(y[i] \* 100, 1)}%",  
 ha='center', va='bottom', size=8, rotation = 30)  
plt.xlabel("Кількість успішних випадків")  
plt.ylabel("Ймовірність")  
plt.show()

|  |
| --- |
| Рисунок 1.2 -  Ймовірність хибно відхилити за умови її вірності |

Залишається лише обчислити суму ймовірностей для кількостей успіхів від 21 до 30 включно. Це і буде нашим FPR.

У нашому випадку це буде 2.1%. Якщо FPR не перевищує деякої константи , то критерій називається критерієм **рівня значущості** . Статистичний критерій з = 100% створити тривіально — достатньо завжди відхиляти — тому така постановка не має сенсу.

Рівень значущості зазвичай обирають на основі бізнес-міркувань. Він позначає те, який ризик неправильного прийняття позитивного рішення ми вважаємо прийнятним. Зазвичай беруть , але якщо потрібне більш точне ухвалення рішення, можуть вибрати , , . Якщо ж рішення не таке критичне, можуть вибрати .

Припустимо, вибрали значення , скористаємося критерієм : тобто якщо кількість успішних випадків перевищує або дорівнює 21, то відхиляємо .

Якщо уважно подивитись на [Рисунок 1.2](#fig-binom-pmf-fpr), то можна помітити, що ми можемо відхиляти при кількості успіхів від 20, а не 21, оскільки такий все ще буде відповідати :

Якщо ж обрати 19, то FPR буде більше :

## 1.3 Статистичні функції в Python

# 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

# 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)
5. Англ. *cyan*, від грец. κυανoῦς — “блакитний”, “лазуровий”. [↑](#footnote-ref-35)