

Деякі задачі статистики

Вікторія Кнопова

Інститут математики НАНУ

19.09.2026

Зміст

► Історія

Зміст

- ▶ Історія
- ▶ Деякі задачі

Зміст

- ▶ Історія
- ▶ Деякі задачі
- ▶ Деякі парадокси

Трохи історії: XVII століття, Bills of Mortality

Джон Грант (John Graunt, 1620–1674) — англійський торговець і дослідник, який одним із перших систематично застосував кількісний аналіз до даних про населення.

Трохи історії: XVII століття, Bills of Mortality

Джон Грант (John Graunt, 1620–1674) — англійський торговець і дослідник, який одним із перших систематично застосував кількісний аналіз до даних про населення.

У Лондоні з XVI століття публікувалися **Бюлетені смертності (Bills of Mortality)**.

- ▶ щотижневі/щомісячні відомості про кількість померлих;
- ▶ зазначалася причина смерті;
- ▶ дані використовувалися, зокрема, для спостереження за епідеміями;
- ▶ Грант систематизував ці дані та шукав у них закономірності.

Трохи історії: XVII століття, Bills of Mortality

Джон Грант (John Graunt, 1620–1674) — англійський торговець і дослідник, який одним із перших систематично застосував кількісний аналіз до даних про населення.

У Лондоні з XVI століття публікувалися **Бюлетені смертності (Bills of Mortality)**.

- ▶ щотижневі/щомісячні відомості про кількість померлих;
- ▶ зазначалася причина смерті;
- ▶ дані використовувалися, зокрема, для спостереження за епідеміями;
- ▶ Грант систематизував ці дані та шукав у них закономірності.

У 1662 році Джон Грант опублікував

Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality.

Цю працю часто вважають одним із перших прикладів **систематичного статистичного аналізу населення**.

Грант не просто збирав числа, а намагався отримати з них **інформацію про населення**.

Він аналізував:

- ▶ смертність від різних причин;
- ▶ відмінності між чоловіками та жінками;
- ▶ вікову структуру смертності;
- ▶ вплив епідемій;
- ▶ співвідношення народжень і смертей;
- ▶ приблизну чисельність населення Лондона.

Грант не просто збирав числа, а намагався отримати з них **інформацію про населення**.

Він аналізував:

- ▶ смертність від різних причин;
- ▶ відмінності між чоловіками та жінками;
- ▶ вікову структуру смертності;
- ▶ вплив епідемій;
- ▶ співвідношення народжень і смертей;
- ▶ приблизну чисельність населення Лондона.

Важливо: Грант використовував **не повні дані про кожну людину, а агреговані спостереження**, і намагався робити висновки про все населення.

XVIII–XIX століття: статистика стає наукою

Statistik (нім., Gottfried Achenwall, 1748) Status (лат)

XVIII–XIX століття: статистика стає наукою

Statistik (нім., Gottfried Achenwall, 1748) Status (лат)

Після перших досліджень Джона Гранта статистичні методи поступово поширюються на різні галузі.

XVIII–XIX століття: статистика стає наукою

Statistik (нім., Gottfried Achenwall, 1748) Status (лат)

Після перших досліджень Джона Гранта статистичні методи поступово поширюються на різні галузі.

- ▶ **Демографія:** народжуваність, смертність, чисельність населення.
- ▶ **Страхування:** оцінювання ризику смерті та тривалості життя.
- ▶ **Астрономія:** обробка результатів спостережень та оцінювання похибок вимірювань.
- ▶ **Медицина:** порівняння результатів лікування.
- ▶ **Державне управління:** переписи населення та економічна статистика.

XVIII–XIX століття: статистика стає наукою

Statistik (нім., Gottfried Achenwall, 1748) Status (лат)

Після перших досліджень Джона Гранта статистичні методи поступово поширюються на різні галузі.

- ▶ **Демографія:** народжуваність, смертність, чисельність населення.
- ▶ **Страхування:** оцінювання ризику смерті та тривалості життя.
- ▶ **Астрономія:** обробка результатів спостережень та оцінювання похибок вимірювань.
- ▶ **Медицина:** порівняння результатів лікування.
- ▶ **Державне управління:** переписи населення та економічна статистика.

Отже,

Випадковість можна досліджувати!

Закон великих чисел і нормальний розподіл

Якоб Бернуллі (1655–1705): при великій кількості спостережень відносна частота події стабілізується.

Закон великих чисел і нормальний розподіл

Якоб Бернуллі (1655–1705): при великій кількості спостережень відносна частота події стабілізується.

Наприклад, якщо багато разів кидати симетричну монету,

$$\frac{\text{кількість гербів}}{\text{кількість кидань}} \longrightarrow \frac{1}{2}.$$

Закон великих чисел і нормальний розподіл

Якоб Бернуллі (1655–1705): при великій кількості спостережень відносна частота події стабілізується.

Наприклад, якщо багато разів кидати симетричну монету,

$$\frac{\text{кількість гербів}}{\text{кількість кидань}} \longrightarrow \frac{1}{2}.$$

Адольф Кетле (1786–1874): ідеї ймовірності до масових явищ у суспільстві, причини, які впливають на розвиток людини.

Закон великих чисел і нормальний розподіл

Якоб Бернуллі (1655–1705): при великій кількості спостережень відносна частота події стабілізується.

Наприклад, якщо багато разів кидати симетричну монету,

$$\frac{\text{кількість гербів}}{\text{кількість кидань}} \longrightarrow \frac{1}{2}.$$

Адольф Кетле (1786–1874): ідеї ймовірності до масових явищ у суспільстві, причини, які впливають на розвиток людини.

Карл Фрідріх Гаусс (1777-1855): похибки вимірювань і нормальний розподіл.

Закон великих чисел і нормальний розподіл

Якоб Бернуллі (1655–1705): при великій кількості спостережень відносна частота події стабілізується.

Наприклад, якщо багато разів кидати симетричну монету,

$$\frac{\text{кількість гербів}}{\text{кількість кидань}} \longrightarrow \frac{1}{2}.$$

Адольф Кетле (1786–1874): ідеї ймовірності до масових явищ у суспільстві, причини, які впливають на розвиток людини.

Карл Фрідріх Гаусс (1777-1855): похибки вимірювань і нормальний розподіл.

Ідея

Якщо окремі спостереження випадкові, то **великі масиви спостережень можуть демонструвати стійкі закономірності.**

Кінець XIX – початок XX століття: статистичний зв'язок

Статистика починає вивчати не лише окремі величини, а **зв'язки між ними**.

- ▶ **Френсіс Гальтон** (1822–1911): спадковість та ввів ідеї регресії.

Кінець XIX – початок XX століття: статистичний зв'язок

Статистика починає вивчати не лише окремі величини, а **зв'язки між ними**.

- ▶ **Френсіс Гальтон** (1822–1911): спадковість та ввів ідеї регресії.

Наприклад:

”Діти дуже високих або дуже низьких батьків зазвичай мають зріст, ближчий до загального середнього показника, ніж їхні батьки”

Кінець XIX – початок XX століття: статистичний зв'язок

Статистика починає вивчати не лише окремі величини, а **зв'язки між ними**.

- ▶ **Френсіс Гальтон** (1822–1911): спадковість та ввів ідеї регресії.

Наприклад:

”Діти дуже високих або дуже низьких батьків зазвичай мають зріст, ближчий до загального середнього показника, ніж їхні батьки”

- ▶ **Карл Пірсон** (1857–1936) розвинув математичні методи кореляції та перевірки статистичних припущень.

Кінець XIX – початок XX століття: статистичний зв'язок

Статистика починає вивчати не лише окремі величини, а **зв'язки між ними**.

- ▶ **Френсіс Гальтон** (1822–1911): спадковість та ввів ідеї регресії.

Наприклад:

”Діти дуже високих або дуже низьких батьків зазвичай мають зріст, ближчий до загального середнього показника, ніж їхні батьки”

- ▶ **Карл Пірсон** (1857–1936) розвинув математичні методи кореляції та перевірки статистичних припущень.

Наприклад:

зріст батьків $\longleftrightarrow$ зріст дітей.

Кінець XIX – початок XX століття: статистичний зв'язок

Статистика починає вивчати не лише окремі величини, а **зв'язки між ними**.

- ▶ **Френсіс Гальтон** (1822–1911): спадковість та ввів ідеї регресії.

Наприклад:

”Діти дуже високих або дуже низьких батьків зазвичай мають зріст, ближчий до загального середнього показника, ніж їхні батьки”

- ▶ **Карл Пірсон** (1857–1936) розвинув математичні методи кореляції та перевірки статистичних припущень.

Наприклад:

зріст батьків $\longleftrightarrow$ зріст дітей.

Але виникає принципове питання:

Чи означає статистичний зв'язок причинно-наслідковий зв'язок?

XX століття: сучасна математична статистика

У XX столітті статистика перетворюється на математичну теорію прийняття рішень за спостереженнями.

XX століття: сучасна математична статистика

У XX столітті статистика перетворюється на математичну теорію прийняття рішень за спостереженнями.

- ▶ **Рональд Фішер**(1890–1962)— метод максимальної правдоподібності, планування експериментів, дисперсійний аналіз.

XX століття: сучасна математична статистика

У XX столітті статистика перетворюється на математичну теорію прийняття рішень за спостереженнями.

- ▶ **Рональд Фішер**(1890–1962)— метод максимальної правдоподібності, планування експериментів, дисперсійний аналіз.
- ▶ **Єжі Нейман** (1894-1981), **Егон Пірсон** (1895–1980) — теорія перевірки статистичних гіпотез, похибки першого та другого роду
- ▶ Розвиваються **довірчі інтервали** та статистичне оцінювання.
- ▶ З'являються непараметричні методи та багатовимірна статистика.

Критерій Неймана-Пірсона:

Метод допомагає вирішити завдання:

Критерій Неймана-Пірсона:

Метод допомагає вирішити завдання:

є дві гіпотези

- **нульова** (H_0 , наприклад, «сигналу немає» або «ліки не працюють»)

та

- **альтернативна** (H_1 , наприклад, «сигнал є» або «ліки допомагають»).

Критерій Неймана-Пірсона:

Метод допомагає вирішити завдання:

є дві гіпотези

- **нульова** (H_0 , наприклад, «сигналу немає» або «ліки не працюють»)

та

- **альтернативна** (H_1 , наприклад, «сигнал є» або «ліки допомагають»).

Критерій будує правило прийняття рішення так, щоб обмежити ймовірність хибнопозитивного результату (помилки I роду)

$$P(H_1|H_0)$$

і при цьому максимально збільшити шанс виявити істинний ефект (потужність критерію, або мінімізувати помилку II роду)

$$P(H_1|H_1).$$

Експеримент Бенвеністе: «пам'ять води»

Історія

У 1988 році французький імунолог Жак Бенвеністе та його колеги опублікували в журналі *Nature* результати експериментів із надзвичайно розведеними розчинами антитіл.

Експеримент Бенвеністе: «пам'ять води»

Історія

У 1988 році французький імунолог Жак Бенвеністе та його колеги опублікували в журналі *Nature* результати експериментів із надзвичайно розведеними розчинами антитіл.

Ідея:

навіть якщо молекул антитіла вже не залишилося, розчин нібито зберігає його «інформацію» або «пам'ять».

Експеримент Бенвеністе: «пам'ять води»

Історія

У 1988 році французький імунолог Жак Бенвеністе та його колеги опублікували в журналі *Nature* результати експериментів із надзвичайно розведеними розчинами антитіл.

Ідея:

навіть якщо молекул антитіла вже не залишилося, розчин нібито зберігає його «інформацію» або «пам'ять».

Подальші незалежні та подвійно сліпі перевірки не підтвердили початкового ефекту.....

Експеримент Бенвеністе: «пам'ять води»

Історія

У 1988 році французький імунолог Жак Бенвеністе та його колеги опублікували в журналі *Nature* результати експериментів із надзвичайно розведеними розчинами антитіл.

Ідея:

навіть якщо молекул антитіла вже не залишилося, розчин нібито зберігає його «інформацію» або «пам'ять».

Подальші незалежні та подвійно сліпі перевірки не підтвердили початкового ефекту.....

Тут нульова гіпотеза– ефект є А треба було навпаки:)

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що**
налили першим: молоко чи чай.

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що налили першим**: молоко чи чай.

Експеримент Рональда Фішера (1935):

- ▶ приготували 8 чашок: 4 — молоко спочатку, 4 — чай спочатку;
- ▶ чашки випадково перемішали;
- ▶ леді мала визначити, які 4 чашки належать до першої групи.

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що налили першим**: молоко чи чай.

Експеримент Рональда Фішера (1935):

- ▶ приготували 8 чашок: 4 — молоко спочатку, 4 — чай спочатку;
- ▶ чашки випадково перемішали;
- ▶ леді мала визначити, які 4 чашки належать до першої групи.

H_0 : леді не може розрізнити ці два способи приготування.

Тобто за H_0 її правильні відповіді пояснюються лише випадковістю.

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що налили першим**: молоко чи чай.

Експеримент Рональда Фішера (1935):

- ▶ приготували 8 чашок: 4 — молоко спочатку, 4 — чай спочатку;
- ▶ чашки випадково перемішали;
- ▶ леді мала визначити, які 4 чашки належать до першої групи.

H_0 : леді не може розрізнити ці два способи приготування.

Тобто за H_0 її правильні відповіді пояснюються лише випадковістю.

Усього можливих способів вибрати 4 чашки з 8: $C_8^4 = 70$.

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що налили першим**: молоко чи чай.

Експеримент Рональда Фішера (1935):

- ▶ приготували 8 чашок: 4 — молоко спочатку, 4 — чай спочатку;
- ▶ чашки випадково перемішали;
- ▶ леді мала визначити, які 4 чашки належать до першої групи.

H_0 : леді не може розрізнити ці два способи приготування.

Тобто за H_0 її правильні відповіді пояснюються лише випадковістю.

Усього можливих способів вибрати 4 чашки з 8: $C_8^4 = 70$.

Якщо леді правильно визначила **всі чашки**, то

$$P(\text{усі 8 правильні} \mid H_0) = \frac{1}{70} \approx 0,014.$$

Леді та чай з молоком: експеримент Фішера

Леді Муріель Брістоль стверджувала, що може визначити, **що налили першим**: молоко чи чай.

Експеримент Рональда Фішера (1935):

- ▶ приготували 8 чашок: 4 — молоко спочатку, 4 — чай спочатку;
- ▶ чашки випадково перемішали;
- ▶ леді мала визначити, які 4 чашки належать до першої групи.

H_0 : леді не може розрізнити ці два способи приготування.

Тобто за H_0 її правильні відповіді пояснюються лише випадковістю.

Усього можливих способів вибрати 4 чашки з 8: $C_8^4 = 70$.

Якщо леді правильно визначила **всі чашки**, то

$$P(\text{усі 8 правильні} \mid H_0) = \frac{1}{70} \approx 0,014.$$

Висновок Фішера:

Таке спостереження дуже малоімовірне за H_0 .

Біноміальний розподіл

Біноміальний розподіл описує кількість успіхів у n незалежних випробуваннях, якщо в кожному випробуванні є лише два результати: **успіх** або **невдача**.

Біноміальний розподіл

Біноміальний розподіл описує кількість успіхів у n незалежних випробуваннях, якщо в кожному випробуванні є лише два результати: **успіх** або **невдача**.

Приклади:

- ▶ скільки разів випаде “орел” при 10 підкиданнях монети;
- ▶ скільки правильних відповідей буде у тесті з 20 запитань;
- ▶ скільки виробів із 100 виявляться бракованими.

Біноміальний розподіл

Біноміальний розподіл описує кількість успіхів у n незалежних випробуваннях, якщо в кожному випробуванні є лише два результати: **успіх** або **невдача**.

Приклади:

- ▶ скільки разів випаде “орел” при 10 підкиданнях монети;
- ▶ скільки правильних відповідей буде у тесті з 20 запитань;
- ▶ скільки виробів із 100 виявляться бракованими.

X = кількість успіхів, n = кількість випробувань, $p = P(\text{успіх})$.

$$X \sim \text{Bin}(n, p).$$

Біноміальний розподіл

Біноміальний розподіл описує кількість успіхів у n незалежних випробуваннях, якщо в кожному випробуванні є лише два результати: **успіх** або **невдача**.

Приклади:

- ▶ скільки разів випаде “орел” при 10 підкиданнях монети;
- ▶ скільки правильних відповідей буде у тесті з 20 запитань;
- ▶ скільки виробів із 100 виявляться бракованими.

X = кількість успіхів, n = кількість випробувань, $p = P(\text{успіх})$.

$$X \sim \text{Bin}(n, p).$$

Ймовірність отримати рівно k успіхів:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Нормальний розподіл

Нормальний розподіл — це модель для величин, значення яких

- ▶ коливаються навколо певного середнього значення;
- ▶ мають приблизно симетричний розподіл;
- ▶ найчастіше трапляються близькі до середнього значення.

Нормальний розподіл

Нормальний розподіл — це модель для величин, значення яких

- ▶ коливаються навколо певного середнього значення;
- ▶ мають приблизно симетричний розподіл;
- ▶ найчастіше трапляються близькі до середнього значення.

Приклади: зріст людей, похибки вимірювань, результати тестів (за певних умов).

μ = середнє, σ = стандартне відхилення.

- ▶ μ визначає, де знаходиться центр розподілу;
- ▶ σ визначає, наскільки широко розкидані значення.

Нормальний розподіл

Нормальний розподіл — це модель для величин, значення яких

- ▶ коливаються навколо певного середнього значення;
- ▶ мають приблизно симетричний розподіл;
- ▶ найчастіше трапляються близькі до середнього значення.

Приклади: зріст людей, похибки вимірювань, результати тестів (за певних умов).

μ = середнє, σ = стандартне відхилення.

- ▶ μ визначає, де знаходиться центр розподілу;
- ▶ σ визначає, наскільки широко розкидані значення.

Правило 68–95–99,7

Приблизно

$68\% \in [\mu - \sigma, \mu + \sigma]$, $95\% \in [\mu - 2\sigma, \mu + 2\sigma]$, $99,7\% \in [\mu - 3\sigma, \mu + 3\sigma]$.

Типові задачі: A-level, або німецька гімназія

1. Вибірка та випадкові величини

З 1000 виробів випадково вибирають 20. Імовірність того, що виріб має дефект, дорівнює 0.03.

$$X \sim \text{Bin}(20, 0.03).$$

Знайти $P(X = 0)$ та $P(X \geq 2)$.

Типові задачі: A-level, або німецька гімназія

1. Вибірка та випадкові величини

З 1000 виробів випадково вибирають 20. Імовірність того, що виріб має дефект, дорівнює 0.03.

$$X \sim \text{Bin}(20, 0.03).$$

Знайти $P(X = 0)$ та $P(X \geq 2)$.

2. Нормальний розподіл

Результати тесту мають розподіл

$$X \sim N(68, 10^2).$$

Яка частка учнів отримує більше ніж 80 балів?

Типові задачі: A-level, або німецька гімназія

1. Вибірка та випадкові величини

З 1000 виробів випадково вибирають 20. Імовірність того, що виріб має дефект, дорівнює 0.03.

$$X \sim \text{Bin}(20, 0.03).$$

Знайти $P(X = 0)$ та $P(X \geq 2)$.

2. Нормальний розподіл

Результати тесту мають розподіл

$$X \sim N(68, 10^2).$$

Яка частка учнів отримує більше ніж 80 балів?

3. Перевірка гіпотези

Виробник стверджує, що частка дефектних виробів дорівнює 5%. У вибірці з 100 виробів виявлено 9 дефектних.

$$H_0 : p = 0.05, \quad H_1 : p > 0.05.$$

Чи є достатньо статистичних підстав відкинути твердження виробника?

Типова задача:

Задано таблицю або діаграму.

- ▶ обчислити середнє значення та стандартне відхилення;
- ▶ побудувати або інтерпретувати Boxplot;
- ▶ (порівняти дві групи);
- ▶ зробити висновок про дані.

реальна ситуація



математична модель



обчислення



статистична інтерпретація результату

Порядковий критерій: дві вибірки

Нехай маємо дві незалежні вибірки:

X : 2, 4, 5, 7, 9,

Y : 1, 3, 6, 8, 10.

Перевіряємо гіпотезу

H_0 : X, Y з однієї популяції.

Проти альтернативи

H_1 : X, Y з різних популяцій.

Об'єднаємо всі 10 спостережень і впорядкуємо їх:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Присвоїмо їм ранги:

значення	1	2	3	4	5	6	7	8	9	10
ранг	1	2	3	4	5	6	7	8	9	10
вибірка	Y	X	Y	X	X	Y	X	Y	X	Y

Порядковий критерій: суми рангів

Сума рангів першої вибірки:

$$R_X = 2 + 4 + 5 + 7 + 9 = 27.$$

Сума рангів другої вибірки:

$$R_Y = 1 + 3 + 6 + 8 + 10 = 28.$$

Порядковий критерій: суми рангів

Сума рангів першої вибірки:

$$R_X = 2 + 4 + 5 + 7 + 9 = 27.$$

Сума рангів другої вибірки:

$$R_Y = 1 + 3 + 6 + 8 + 10 = 28.$$

За нульової гіпотези ранги мають випадково розподілятися між двома вибірками.

Порядковий критерій: суми рангів

Сума рангів першої вибірки:

$$R_X = 2 + 4 + 5 + 7 + 9 = 27.$$

Сума рангів другої вибірки:

$$R_Y = 1 + 3 + 6 + 8 + 10 = 28.$$

За нульової гіпотези ранги мають випадково розподілятися між двома вибірками.

Для $n_X = n_Y = 5$

$$\text{середнє} = \frac{n_X(n_X + n_Y + 1)}{2} = \frac{5 \cdot 11}{2} = 27.5.$$

Порядковий критерій: суми рангів

Сума рангів першої вибірки:

$$R_X = 2 + 4 + 5 + 7 + 9 = 27.$$

Сума рангів другої вибірки:

$$R_Y = 1 + 3 + 6 + 8 + 10 = 28.$$

За нульової гіпотези ранги мають випадково розподілятися між двома вибірками.

Для $n_X = n_Y = 5$

$$\text{середнє} = \frac{n_X(n_X + n_Y + 1)}{2} = \frac{5 \cdot 11}{2} = 27.5.$$

Ми отримали

$$R_X = 27,$$

що дуже близько до очікуваного значення 27.5.

Парадокс Сімпсона: вступ до університету

Розглянемо вступ до двох факультетів:

природничі науки, соціальні науки.

	Подали документи		Прийняті	
	хлопці	дівчата	хлопці	дівчата
Природничі	300	10	138	8
Соціальні	50	110	2	22

Парадокс Сімпсона: вступ до університету

Розглянемо вступ до двох факультетів:

природничі науки, соціальні науки.

	Подали документи		Прийняті	
	хлопці	дівчата	хлопці	дівчата
Природничі	300	10	138	8
Соціальні	50	110	2	22

Рівень прийому по факультетам:

	хлопці	дівчата
Природничі	$138/300 = 46\%$	$8/10 = 80\%$
Соціальні	$2/50 = 4\%$	$22/110 = 20\%$

У кожній групі частка прийнятих дівчат вища.

Але що відбувається в цілому?

Якщо об'єднати два факультети, отримаємо:

	Подали документи	Прийняті
хлопці	350	140
дівчата	120	30

Тому

$$P(\text{прийняті} \mid \text{хлопці}) = \frac{140}{350} = 40\%,$$

тоді як

$$P(\text{прийняті} \mid \text{дівчата}) = \frac{30}{120} = 25\%.$$

Але!!! У кожному окремому факультеті:

$$80\% > 46\%, \quad 20\% > 4\%.$$

Напрямок залежності змінився!

Чому виникає парадокс Сімпсона?

Ключове питання:

Чи однаково розподілені хлопці та дівчата між факультетами?

Чому виникає парадокс Сімпсона?

Ключове питання:

Чи однаково розподілені хлопці та дівчата між факультетами?

Ні.

Чому виникає парадокс Сімпсона?

Ключове питання:

Чи однаково розподілені хлопці та дівчата між факультетами?

Ні.

Частка хлопців, які подали на точні науки:

$$\frac{300}{350} \approx 86\%$$

також: на точних науках високий рівень прийому.

Чому виникає парадокс Сімпсона?

Ключове питання:

Чи однаково розподілені хлопці та дівчата між факультетами?

Ні.

Частка хлопців, які подали на точні науки:

$$\frac{300}{350} \approx 86\%$$

також: на точних науках високий рівень прийому.

Частка дівчат, які подали на соціальні науки:

$$\frac{110}{120} \approx 92\%$$

На соціальних науках низький рівень прийому.

Чому виникає парадокс Сімпсона?

Ключове питання:

Чи однаково розподілені хлопці та дівчата між факультетами?

Ні.

Частка хлопців, які подали на точні науки:

$$\frac{300}{350} \approx 86\%$$

також: на точних науках високий рівень прийому.

Частка дівчат, які подали на соціальні науки:

$$\frac{110}{120} \approx 92\%$$

На соціальних науках низький рівень прийому.

Отже, ми змішуємо дві різні групи з дуже різними рівнями прийому.

Сімпсон/(Пірсон, Юл): результат може змінитися після об'єднання груп.

Парадокс Сімпсона: альтруїсти та егоїсти

Уявімо дві групи в популяції:

A = альтруїсти, E = егоїсти.

Парадокс Сімпсона: альтруїсти та егоїсти

Уявімо дві групи в популяції:

A = альтруїсти, E = егоїсти.

Спочатку:

	A	E	Разом
Група 1	80	20	100
Група 2	20	80	100
Разом	100	100	200

Частка альтруїстів у кожній групі:

80% та 20%.

У цілому: $\frac{100}{200} = 50\%$.

Припустимо, що всередині кожної групи егоїсти мають перевагу:
частка альтруїстів після розмноження зменшується.

Наприклад,

80% $\longrightarrow$ 70%

у першій групі та

Припустимо, що всередині кожної групи егоїсти мають перевагу:
частка альтруїстів після розмноження зменшується.

Наприклад,

80% $\longrightarrow$ 70%

у першій групі та

20% $\longrightarrow$ 10%

у другій групі.

Припустимо, що всередині кожної групи егоїсти мають перевагу:
частка альтруїстів після розмноження зменшується.

Наприклад,

$$80\% \longrightarrow 70\%$$

у першій групі та

$$20\% \longrightarrow 10\%$$

у другій групі.

Отже, всередині кожної групи альтруїсти програють егоїстам:

$$70\% < 80\%, \quad 10\% < 20\%.$$

На перший погляд, це означає, що альтруїстична поведінка повинна зникати з популяції.

Альтруїсти програють у кожній групі...

Однак припустимо, що альтруїзм підвищує ефективність усієї групи. Групи, в яких більше альтруїстів, краще співпрацюють, ефективніше використовують ресурси і тому залишають більше нащадків.

Після розмноження:

	A	E	Разом
Група 1	280	120	400
Група 2	10	90	100
Разом	290	210	500

Альтруїсти програють у кожній групі...

Однак припустимо, що альтруїзм підвищує ефективність усієї групи. Групи, в яких більше альтруїстів, краще співпрацюють, ефективніше використовують ресурси і тому залишають більше нащадків.

Після розмноження:

	A	E	Разом
Група 1	280	120	400
Група 2	10	90	100
Разом	290	210	500

Отже, загальна частка альтруїстів:

$$\frac{290}{500} = 58\%.$$

Альтруїсти програють у кожній групі...

Однак припустимо, що альтруїзм підвищує ефективність усієї групи. Групи, в яких більше альтруїстів, краще співпрацюють, ефективніше використовують ресурси і тому залишають більше нащадків.

Після розмноження:

	A	E	Разом
Група 1	280	120	400
Група 2	10	90	100
Разом	290	210	500

Отже, загальна частка альтруїстів:

$$\frac{290}{500} = 58\%.$$

50% → 58%

Альтруїсти програють у кожній групі...

Однак припустимо, що альтруїзм підвищує ефективність усієї групи. Групи, в яких більше альтруїстів, краще співпрацюють, ефективніше використовують ресурси і тому залишають більше нащадків.

Після розмноження:

	A	E	Разом
Група 1	280	120	400
Група 2	10	90	100
Разом	290	210	500

Отже, загальна частка альтруїстів:

$$\frac{290}{500} = 58\%.$$

50% $\rightarrow$ 58%

Альтруїстів стало більше в популяції, хоча в кожній окремій групі їх стало менше.

Що тут “парадоксального”?

Маємо два рівні відбору.

Всередині групи:

егоїсти мають перевагу $\Rightarrow$ частка альтруїстів зменшується.

Між групами:

група з більшою часткою альтруїстів виявилася значно продуктивнішою:

група 1 зростає до 400 особин, група 2 — 100 як і було

Тому внесок групи 1 у загальну популяцію стає набагато більшим.

Внутрішньогруповий відбір vs. міжгруповий відбір

Це приклад того, як **різні рівні агрегації** можуть давати протилежні висновки.

Парадокс Берксона: приклад із COVID-19

Розглянемо просту математичну модель.

$A = \text{COVID-19}$, $B = \text{серцево-судинне захворювання}$.

Припустимо, що в загальній популяції ці події незалежні:

$$A \perp B.$$

Парадокс Берксона: приклад із COVID-19

Розглянемо просту математичну модель.

$A = \text{COVID-19}$, $B = \text{серцево-судинне захворювання}$.

Припустимо, що в загальній популяції ці події незалежні:

$$A \perp B.$$

Нехай $P(A) = P(B) = 0.1$.

Для 10000 людей маємо:

	B	$\overline{B}$	Разом
A	100	900	1000
$\overline{A}$	900	8100	9000
Разом	1000	9000	10000

Парадокс Берксона: приклад із COVID-19

Розглянемо просту математичну модель.

$A = \text{COVID-19}$, $B = \text{серцево-судинне захворювання}$.

Припустимо, що в загальній популяції ці події незалежні:

$$A \perp B.$$

Нехай $P(A) = P(B) = 0.1$.

Для 10000 людей маємо:

	B	$\bar{B}$	Разом
A	100	900	1000
$\bar{A}$	900	8100	9000
Разом	1000	9000	10000

Справді, $P(B | A) = 0.1$

Отже, $P(A \cap B) = P(A)P(B)$

$A \perp B.$

Що змінює відбір?

Тепер введемо подію

H = людина потрапила до нашої вибірки.

Що змінює відбір?

Тепер введемо подію

H = людина потрапила до нашої вибірки.

Для ілюстрації припустимо:

$$H = A \cup B.$$

Тобто люди без жодного з двох факторів не потрапляють до вибірки.

Що змінює відбір?

Тепер введемо подію

H = людина потрапила до нашої вибірки.

Для ілюстрації припустимо:

$$H = A \cup B.$$

Тобто люди без жодного з двох факторів не потрапляють до вибірки.
Вилучаємо 8100 людей із

$$\overline{A} \cap \overline{B}.$$

Залишається:

	B	$\overline{B}$	Разом
A	100	900	1000
$\overline{A}$	900	0	900
Разом	1000	900	1900

	B	$\overline{B}$	Разом
A	100	900	1000
$\overline{A}$	900	0	900
Разом	1000	900	1900

	B	$\bar{B}$	Разом
A	100	900	1000
$\bar{A}$	900	0	900
Разом	1000	900	1900

Тепер:

$$P(B | A, H) = \frac{100}{1000} = 0.1,$$

але

$$P(B | \bar{A}, H) = \frac{900}{900} = 1.$$

	B	$\bar{B}$	Разом
A	100	900	1000
$\bar{A}$	900	0	900
Разом	1000	900	1900

Тепер:

$$P(B | A, H) = \frac{100}{1000} = 0.1,$$

але

$$P(B | \bar{A}, H) = \frac{900}{900} = 1.$$

Отже,

$$P(B | A, H) < P(B | \bar{A}, H).$$

Виникла **негативна статистична залежність**.

В межах клініки,

Ймовірність мати хворе серце, за умови, що є COVID, менша, ніж у тих, у кого COVID не виявлено.

Парадокс Берксона: головна ідея

Схематично:

$$A \longrightarrow H \longleftarrow B.$$

Обидві змінні впливають на те, хто потрапляє до відібраної групи.
У початковій популяції:

$$A \perp B.$$

Але після відбору за H :

$$A \not\perp B \mid H.$$

Інтуїція:

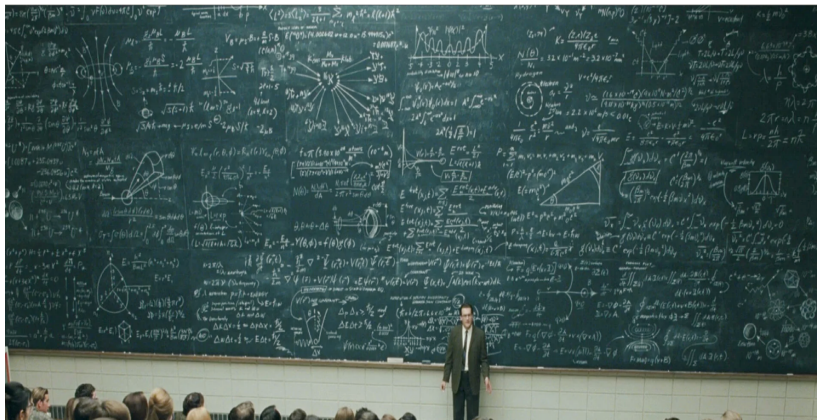
якщо людина вже має A , їй “не потрібне” B , щоб задовольнити умову потрапляння до вибірки.

Якщо ж A немає, то для потрапляння до вибірки потрібне B .

Тому в обраній групі A і B стають негативно пов’язаними.

Негативна залежність $\neq$ негативний причинний вплив.

Насправді, ми нічого не знаємо (с)...



(фото з фільму "Серйозна людина")

Дякую за увагу!