Mathematical basis of DS

Дані - це колекція фактів, таких як числа, слова, вимірювання, спостереження або просто описи речей.

Якісні та кількісні дані

Дані можуть бути якісними або кількісними.

- Якісні дані - це описова інформація (вони описують щось)
- Кількісні дані - це числова інформація (числа)

Типи даних

Кількісні дані можуть бути дискретними або неперервними:

Дискретні дані можуть приймати лише певні значення (наприклад, цілі числа)
Неперервні дані можуть приймати будь-яке значення (в межах певного діапазону)
Просто кажучи: Дискретні дані підраховуються, неперервні дані вимірюються

Приклад: Що ми знаємо про собаку?

Якісні:

- Він коричневий та чорний
- У нього довга шерсть
- Він має багато енергії

Кількісні:

Дискретні:

- Він має 4 ноги
- У нього є 2 брати

Неперервні:

- Вага - 25.5 кг
- Зріст - 565 мм

Перепис(Census) або вибірка(Sample)

Перепис - це коли ми збираємо дані для кожного члена групи (всієї "популяції").

Вибірка - це коли ми збираємо дані лише для вибраних членів групи.

Приклад: 120 людей у вашому місцевому футбольному клубі
Ви можете запитати кожного (усіх 120), скільки їм років. Це перепис.

Або ви можете просто вибрати людей, які там сьогодні після полудня. Це вибірка.

Перепис є точним, але складним для здійснення. Вибірка не так точна, але може бути достатньою та значно простішою.

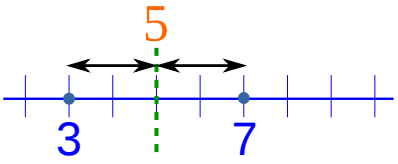
Знаходження центрального значення (середне - avarage)

Коли у вас є два або більше числа, цікаво знайти значення для "центру".

2 числа

З двома числами відповідь проста: візьміть середнє значення.

Приклад: яке центральне значення для чисел 3 і 7?



Відповідь: Середина між ними, тобто 5.

Середнє значення 3 і 7

Ми можемо обчислити його, додаючи 3 і 7, а потім ділимо результат на 2:

(3+7) / 2 = 10/2 = 5

Середнє значення (Avarage)

До цього моменту ми обчислювали Середнє (або Середнє арифметичне):

Середнє: Додайте числа і поділіть на кількість чисел.

Але іноді Середнє значення може вас підвести:

Приклад: День народження

Дядько Боб хоче знати середній вік на вечірці, щоб вибрати розвагу.

Там буде 6 дітей у віці 13 років і 5 малюків у віці 1 рік.

Додайте всі віки і поділіть на 11 (тому що є 11 чисел):

(13+13+13+13+13+13+1+1+1+1+1) / 11 = 7.5...
батут
Середній вік складає близько 7 з половиною років, тому він бере надув

Медіана

Але ви також можете використовувати Медіану: просто впорядкуйте всі числа і оберіть середнє.

Оберіть середнє число:

➡️ 1, 1, 1, 1, 1, 13, 13, 13, 13, 13, 13

Медіана - 13, бо зліва та справа по 5 чисел

Іноді буває два середніх числа. Просто обчислюйте їх середнє значення:

Приклад: Яка медіана для чисел 3, 4, 7, 9, 12, 15?

Є два числа посередині:

➡️ 3, 4, 7, 9, 12, 15

Тому ми обчислюємо їх середнє значення:

(7+9) / 2 = 16/2 = 8
Медіана - 8

Мода

Мода - це значення, яке зустрічається найчастіше:

Приклад: День народження (продовження)

Групуємо числа, щоб їх підрахувати:

➡️ 12, 12, 12, 12, 12, 13, 13, 13, 13, 13, 13

"13" зустрічається 6 разів, "1" зустрічається лише 5 разів, тому мода - 13.

Як запам'ятати? Подумайте про "мода - те, що носить більшість))"

Бімодальність

Але Мода може бути складною, іноді може бути більше однієї Моди.

➡️ 3, 4, 4, 5, 6, 6, 7

Ну ... 4 зустрічається двічі, але 6 також зустрічається двічі.

Тому і 4, і 6 є модами.

Коли є дві Моди, це називається "бімодальним", коли є три або більше Мод, ми називаємо це "мультимодальним".

Бімодальність - це властивість розподілу даних, коли є дві чітко виділені моди або значення, які зустрічаються найбільш часто. В цьому випадку графік або діаграма розподілу матимуть два виражені піки. Наявність бімодальності може вказувати на наявність двох різних підгруп в даних або на різні стани або категорії.

Викиди

Outliers (викиди) - це значення, які "виходять за межі" інших значень в наборі даних. Вони можуть значно змінювати середнє значення (mean), тому ми можемо виключити їх (і вказати про це) або використовувати медіану (median) або моду (mode) замість них.

Приклад: 3, 4, 4, 5 та 104

Середнє значення (mean): Додаємо всі числа і ділимо на їх кількість (5 у даному випадку):

❗ (3+4+4+5+104) / 5 = 24
24 не добре описує ці числа!

Без числа 104 середнє значення буде:

❗ (3+4+4+5) / 4 = 4
Так на багато краще

Медіана (median): Числа впорядковані, тому просто обираємо середнє число, яким є 4:

➡️ 3, 4, 4, 5, 104

Мода (mode): 4 зустрічається найчастіше, тому мода - це 4

➡️ 3, 4, 4, 5, 104

Особливі методи оцінки середніх значень

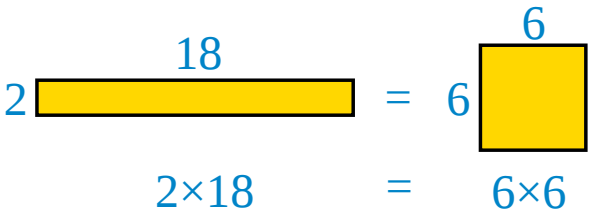
Геометричне середнє

(Geometric Mean) - це особливий тип середнього значення, де ми множимо числа разом, а потім беремо корінь квадратний (для двох чисел), кубічний корінь (для трьох чисел) і т.д.

➡️ **Яке геометричне середнє чисел 2 і 18?**

Спочатку множимо їх: 2 × 18 = 36
Потім (так як є два числа) беремо квадратний корінь: √36 = 6

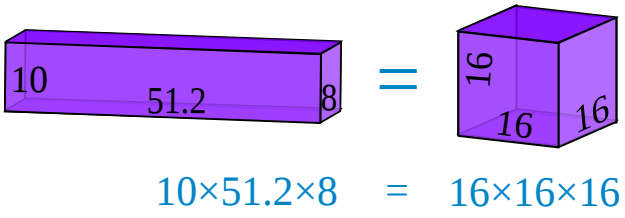
Це, якби площа була однаковою з обох боків



➡️ **Яке геометричне середнє чисел 10, 51.2 і 8?**

Спочатку множимо їх: 10 × 51.2 × 8 = 4096
Потім (так як є три числа) беремо кубічний корінь: 3√4096 = 16

Це, якби об'єм був однаковим з усіх сторін



$$10 \times 51.2 \times 8 = 16 \times 16 \times 16$$

Визначення

Для n чисел: помножте їх всі разом, а потім візьміть n-ий корінь (позначається як n√).

Формально, геометричне середнє n чисел a1 до an вираховується так:

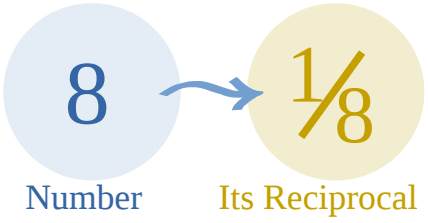
$$n\sqrt{(a1 \times a2 \times ... \times an)}$$

Гармонійне середнє

Гармонійне середнє є: оберненим значенням середнього арифметичного обернених значень.

Так, це багато обернених значень!

Обернене значення означає просто 1/значення.



Формула така:

$$\text{Harmonic Mean} = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots}$$

Гармонійне середнє

Де a, b, c, ... - це значення, а n - кількість значень.

Кроки:

1. Обчисліть обернене значення (1/значення) для кожного значення.
2. Знайдіть середнє арифметичне цих обернених значень (просто додайте їх разом і поділіть на їх кількість).
3. Потім обчисліть обернене значення цього середнього (=1/середнє).

Приклад: Яке гармонійне середнє для 1, 2 і 4?
Обернені значення для 1, 2 і 4:

1/1 = 1, 1/2 = 0.5, 1/4 = 0.25

Тепер додайте їх:

1 + 0.5 + 0.25 = 1.75

Поділіть на їх кількість:

Середнє = 1.753

Обернене значення цього середнього є нашою відповіддю:

$$\text{Гармонійнесереднє} = 3/1.75 = 1.714(\text{до3знаківпіслякоми})$$

Навіщо це нам?

У деяких питаннях, пов'язаних зі співвідношеннями, гармонійне середнє дає правильну відповідь!

❓ Приклад: Ми проїхали 10 км зі швидкістю 60 км/год, потім ще 10 км зі швидкістю 20 км/год. Яка у нас середня швидкість?
Гармонійне середнє = 2/(160 + 120) = 30 км/год

❗ Перевірка: 10 км зі швидкістю 60 км/год зайняло 10 хвилин, 10 км зі швидкістю 20 км/год зайняло 30 хвилин, тому загальною 20 км зайняло 40 хвилин, що становить 30 км за годину.

Гармонійне середнє також добре справляється з великими викидами.

❗ Приклад: 2, 4, 6 і 100
Середнє арифметичне дорівнює 2+4+6+100 = 28
Гармонійне середнє дорівнює 4/(1/2 + 1/4 + 1/6 + 1/100) = 4.32 (до 2 знаків після коми)

Але невеликі викиди погіршують ситуацію!

Інший спосіб розуміння
Ми можемо переставити формулу вигляду, як це:

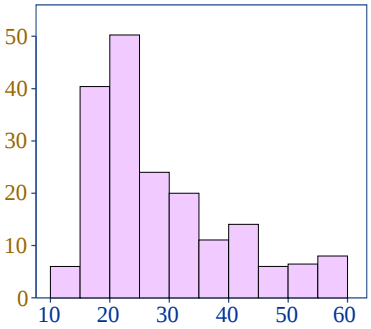
Гармонійне середнє

Використання цього вигляду не є простим, але воно виглядає більш "збалансованим" (n на одній стороні співпадає з n одиницями на іншій стороні, а середнє на одній стороні співпадає зі значеннями на іншій стороні).

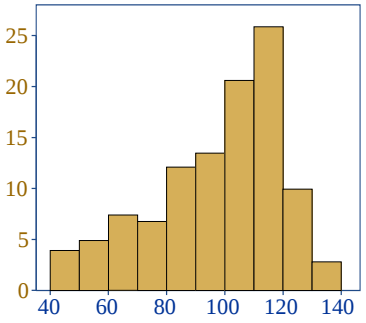
Нормальний розподіл

Дані можуть бути розподілені по-різному. Вони можуть бути розподілені :

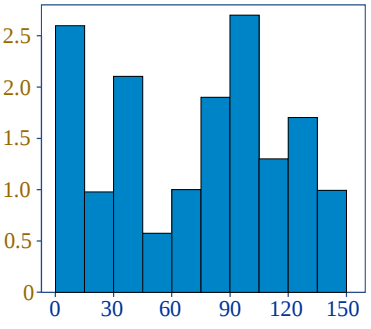
наліво



направо



випадково

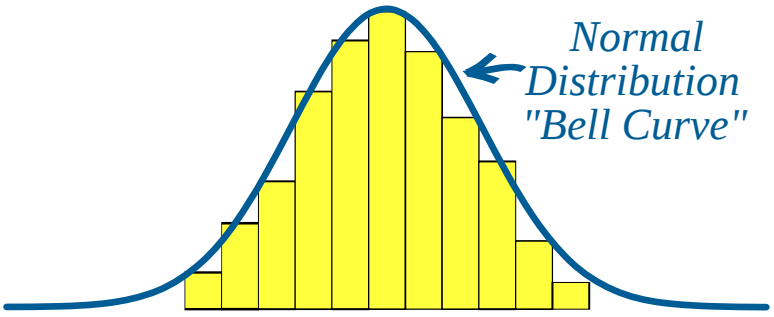


В залежності від такого роду розподілу можна зробити висновки про багато речей та процесів, які відображені у даних.

Випадкові дані

Але є багато випадків, коли дані схильні бути навколо центрального значення без відхилення вліво або вправо, і вони наближаються до "Нормального розподілу" таким чином:

Дзвонова крива



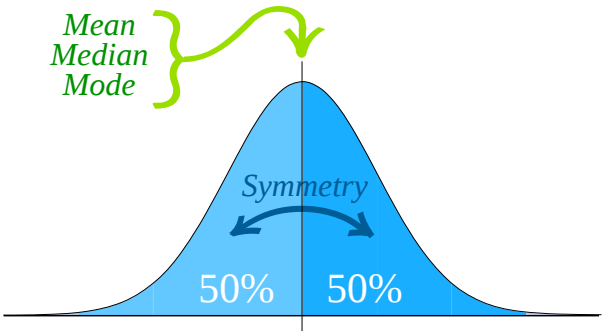
Синя крива - це Нормальний розподіл.
Жовта гістограма показує деякі дані, які слідують йому близько, але не ідеально (це звичайна ситуація).

Багато речей тісно слідують за Нормальним розподілом:

- рост людей
- розмір речей, виготовлених машинами
- похибки вимірювання
- тиск крові
- оцінки на тесті
- Ми кажемо, що дані "нормально розподілені":

нормальний розподіл зі середньою, медіаною і модою в центрі

У Нормального розподілу є:



середнє значення = медіана = мода
симетрія стосовно центру
50% значень менше за середнє
і 50% більше за середнє

Quincunx

Ви можете побачити, як нормальний розподіл створюється випадково!

https://www.youtube.com/watch?v=AwEaHCjgeXk

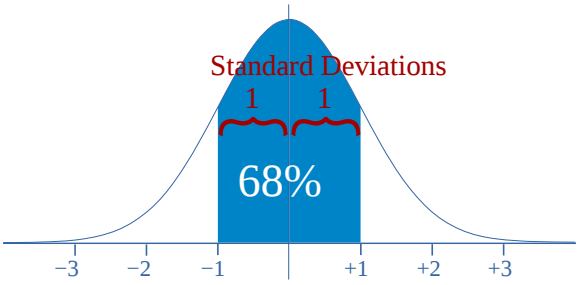
Це називається Квінкункс, і це дивовижно, бо це фізичний процес, який з хауса будує гармонію.

Стандартне відхилення

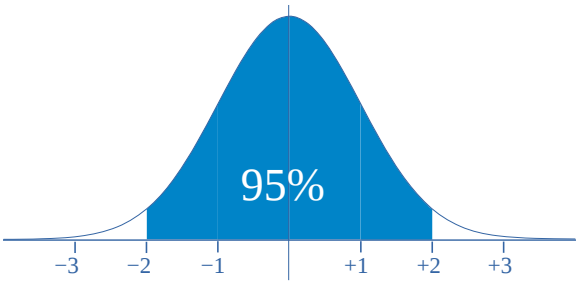
Стандартне відхилення - це міра того, як розподілені числа на деякій множені значень.

Коли ми розраховуємо стандартне відхилення, ми зазвичай встановлюємо наступне:

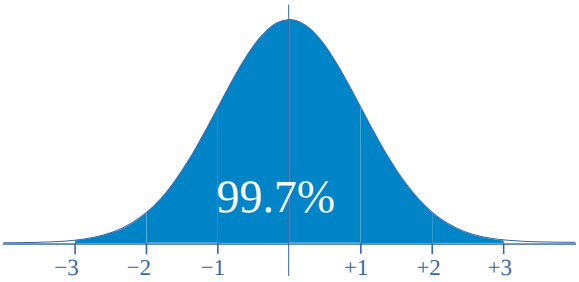
нормальний розподіл 68%, 95%, 99.7%



68% значень знаходяться в межах 1 стандартного відхилення від середнього



95% значень знаходяться в межах 2 стандартних відхилення від середнього



99.7% значень знаходяться в межах
3 стандартних відхилення від середнього

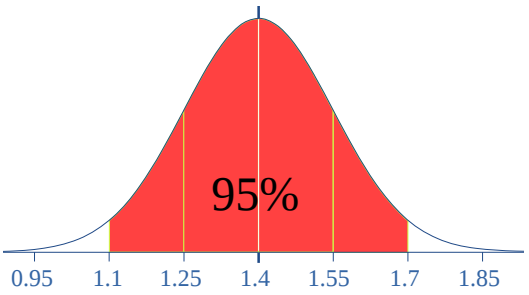
? 95% учнів у школі знаходяться у діапазоні від 1.1 м до 1.7 м зросту. Припускаючи, що ці дані нормально розподілені, ви можете розрахувати середнє значення та стандартне відхилення?

Середнє значення знаходиться посередині між 1.1 м та 1.7 м:

Середнє = (1.1 м + 1.7 м) / 2 = 1.4 м

95% - це 2 стандартних відхилення від середнього (всього 4 стандартних відхилення), тому:

1 стандартне відхилення = (1.7 м - 1.1 м) / 4 = 0.6 м / 4 = 0.15 м
І ось результат: нормальний розподіл 95%



Корисно знати стандартне відхилення, оскільки ми можемо сказати, що будь-яке значення:

- майже завжди знаходиться в межах 1 стандартного відхилення (з 100, 68 повинні бути у межах)
- дуже ймовірно знаходиться в межах 2 стандартних відхилень (з 100, 95 повинні бути у межах)
- майже точно знаходиться в межах 3 стандартних відхилень (з 1000, 997 повинні бути у межах)

Стандартні показники

Кількість стандартних відхилень від середнього також називається "Стандартним показником", "сигма" або "z-показником". Звикайте до цих слів!

❓ В тій самій школі один з ваших друзів має зріст 1.85 м

нормальний розподіл 95%
На діаграмі дзвону видно, що 1.85 м відхиляється на 3 стандартні відхилення від середнього 1.4, тому:

Зріст вашого друга має "z-показник" 3.0 (3-я ланка графіку, а також можна знайти математично)

Також можливо розрахувати, на скільки стандартних відхилень 1.85 відхиляється від середнього

Наскільки 1.85 відхиляється від середнього?

Це 1.85 - 1.4 = 0.45 м від середнього

Скільки це становить стандартних відхилень? Стандартне відхилення становить 0.15 м, тому:

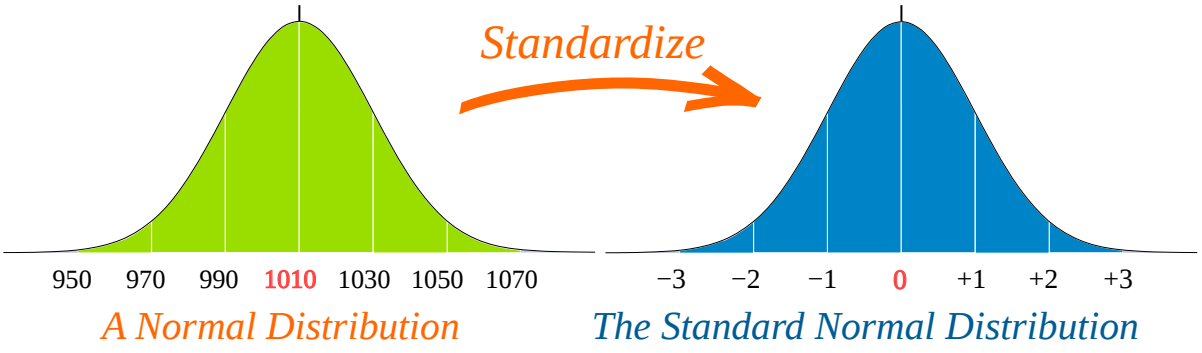
0.45 м / 0.15 м = 3 стандартних відхилення

Тому для перетворення значення в Стандартний показник ("z-показник"):

- спочатку відняти середнє значення,
- потім поділити на стандартне відхилення. І це називається "стандартизацією":

Стандартизація

Ми можемо взяти будь-який нормальний розподіл і перетворити його на Стандартний нормальний розподіл.



Приклад: Час подорожі
Дослідження щоденного часу подорожі мало такі результати (у хвилинах):

26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34

Середнє значення - 38.8 хвилини, а стандартне відхилення - 11.4 хвилини.

Перетворіть значення на Стандартні показники ("стандартні оцінки").

Щоб перетворити 26:

спочатку відняти середнє значення:

$$26 - 38.8 = -12.8$$

потім поділити на стандартне відхилення

$$-12.8/11.4 = -1.12$$

Тому зовнішнє зміщення для 26 становить -1.12.

Аналогічно обчисліть всі інші зміщення.

Потім ці значення можна використовувати для обчислення інтервалів довіри, розрахунку ймовірностей тощо.

Це основа багатьох статистичних тестів і аналізів даних!

> **Що може означати такі показники**: якщо ви отримаєте стандартний нормальний розподіл у межах від - 1 до 1, це означає що ваш отриманий результат та наведені дані відповідають високий імовірності того, що вони є компетентними. Якщо ви виходити за ці межі то це означає, що що ці дані слабо відповідають дійсності.

z-score formula

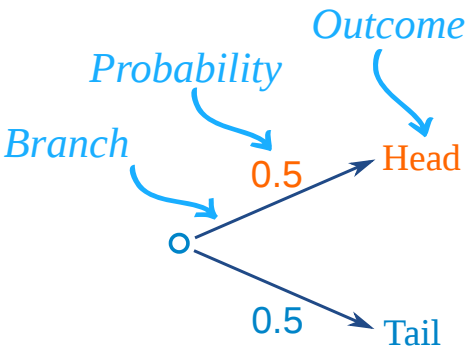
$$z = (x - \mu)/\sigma$$

- **z** is the "z-score" (Standard Score)
- **x** is the value to be standardised
- **μ** ('mu") is the mean
- **σ** ("sigma") is the standard deviation

Імовірнісна деревовидна діаграма

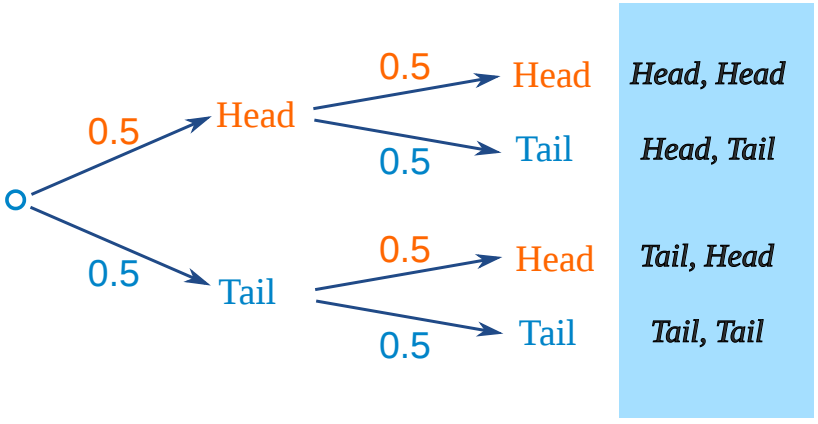
Розрахунок ймовірностей може бути складним завданням, іноді їх додають, іноді множать, і часто важко зрозуміти, що робити ... деревовидні діаграми на допомогу!

Ось деревовидна діаграма для підкидання монети:
Є дві "гілки" (Голови і Решки)

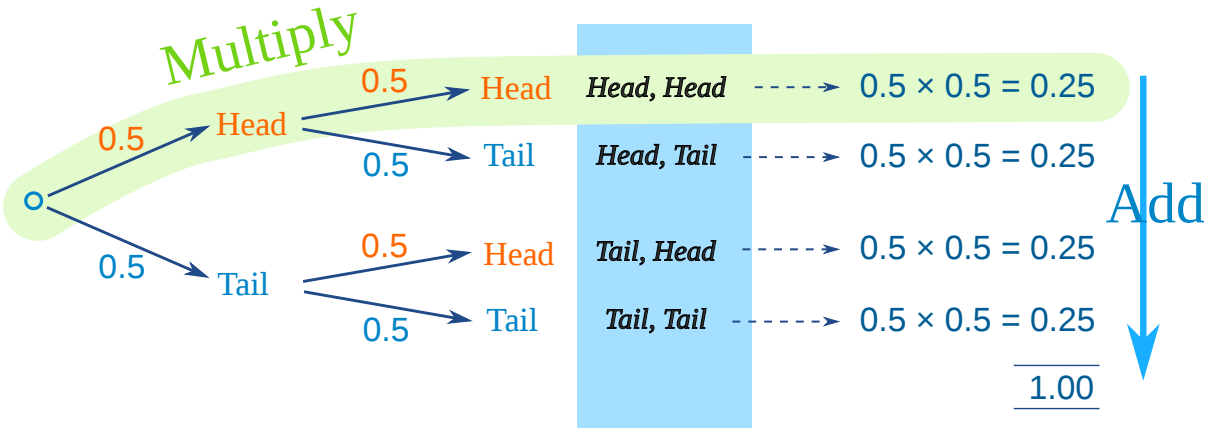


Ймовірність кожної гілки записана на гілці. При 1-му підкидання він дорівнює 0.5. Насправді це не так, бо імперично доведено, що така ймовірність становить 0.49999, бо в середньому на кожні 10000 підкидання є 1 Решка
Результат записаний на кінці гілки
Ми можемо розширити деревовидну діаграму на два підкидання монети:

деревовидна діаграма 2 підкидання монети



Як ми розраховуємо загальні ймовірності?



Ми множимо ймовірності вздовж гілок
Ми додаємо ймовірності по стовпцях
розрахунки ймовірностей за деревовидною діаграмою (множення та додавання)

Тепер ми можемо бачити такі речі, як:

Ймовірність "Голова, Голова" дорівнює 0,5 × 0,5 = 0,25
Всі ймовірності додаються до 1,0 (що завжди є хорошою перевіркою)
Ймовірність отримати принаймні одну Голову з двох підкидань дорівнює 0,25 + 0,25 + 0,25 = 0,75
... та інше

Розподіл Бернуллі (Біномний розподіл)

Розподіл Бернуллі - це статистичний розподіл, який моделює випадковий експеримент з двома можливими результатами: успіхом (позитивним результатом) або невдачею (негативним результатом). Цей розподіл названий на честь швейцарського математика Жака Бернуллі.

У розподілі Бернуллі є один параметр - ймовірність успіху, позначена як p. Вона вказує на ймовірність того, що випадок успіху відбудеться під час одного експерименту.

Функція ймовірності розподілу Бернуллі визначає ймовірності успіху та невдачі. Ця функція може бути використана для обчислення ймовірностей конкретних подій в експерименті з двома можливими результатами.

Розподіл Бернуллі є основою для більш складних розподілів, таких як біноміальний розподіл, де проводиться серія незалежних експериментів з розподілом Бернуллі.

❓ **Підкидання монети**:
Чи випав герб (Heads - H) чи решка (Tails - T)
Ми кажемо, що ймовірність того, що монета впаде гербом, становить ½
А ймовірність того, що монета впаде решкою, також ½

❓ **Кидання кубика:**
Чи випала четвірка...?... чи ні?
Ми кажемо, що ймовірність випадіння четвірки становить 1/6 (одна з шести граней - четвірка)
А ймовірність не випадіння четвірки становить 5/6 (п'ять з шести граней - не четвірка)

Зверніть увагу, що у кубика є 6 граней, але тут ми розглядаємо лише два випадки: "четвірка: так" або "четвірка: ні"

Підкинемо “чесну”* монету три рази... Яка ймовірність отримати рівно два «голови»?

Використовуючи H для «голови» та T для «хвоста», ми можемо отримати будь-який з цих 8 результатів:

*- притримується всіх природніх законів

Які результати нас цікавлять?
"Two Heads" можуть бути в будь-якому порядку: "HHT", "THH" і "HTH" мають дві «голови» (і один «хвіст»).

Таким чином, 3 з 8 результатів дають "Two Heads".

Яка ймовірність кожного результату?
Кожен результат має однакову ймовірність, і їх всього 8, тому кожен результат має ймовірність 1/8.

Отже, ймовірність події "Two Heads" дорівнює:

$$3 \times 1/8 = 3/8$$

Отже, ймовірність отримати дві «голови» становить 3/8.

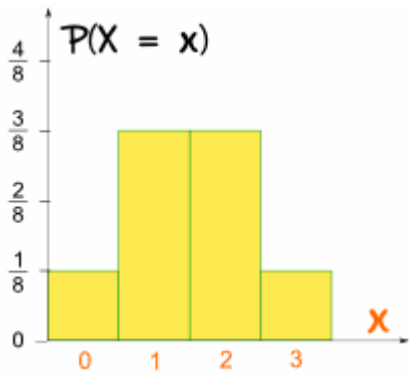
Ми використовували спеціальні терміни

- Результат: будь-який результат підкидання монети тричі (8 різних можливостей)
- Подія: "Two Heads" з трьох підкидань монети (3 результати мають це)

ЕКСПЕРЕМЕНТУЄМО

Розрахунки виглядають наступним чином (P означає "ймовірність"):

P(Three Heads) = P(HHH) = 1/8
P(Two Heads) = P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8
P(One Head) = P(HTT) + P(THT) + P(TTH) = 1/8 + 1/8 + 1/8 = 3/8
P(Zero Heads) = P(TTT) = 1/8
Це виглядає на графіку таким чином: Він симетричний!



Обчислення

Уявіть, що ми хочемо визначити ймовірність отримання 5 «голів» після 9 підкидань монети: перерахувати всі 512 результатів займе багато часу! Для такого можна використати фомулу

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Ця формула може виглядати страшно, але її легко використовувати. Нам потрібно лише два числа:

n - загальна кількість
k - кількість, яку ми хочемо отримати
Символ "!" означає "факторіал", наприклад, 4! = 1×2×3×4 = 24

Примітка: часто це називають "n обрати k"

Давайте спробуємо це:

❓ **При 3 підкиданнях, яка ймовірність отримати 2 «голови»?**
У нас є n = 3 і k = 2:
n!/k!(n-k)! = 3!/2!(3-2)!
= 3×2×1/2×1 × 1
= 3
Отже, існує 3 результати з "2 головами".

Давайте використаємо її для складнішого питання:

Приклад: при 9 підкиданнях, яка ймовірність отримати 5 «голів»?
У нас є n = 9 і k = 5:

n!/k!(n-k)! = 9!5!(9-5)!
= 9×8×7×6×5×4×3×2×1/5×4×3×2×1×4×3×2×1
= 126
Таким чином, 126 з результатів будуть мати 5 «голів».

А для 9 підкидань всього є 2^9 = 512 результатів, а ймовірність при цьому вийде наступна:

Кількість потрібних результатів	множимо	Ймовірність кожного результату	отримуємо	P(X=5)
126	**×**	**1 / 512**	**=**	**126 / 512**

P(X=5) = 126/512 = 0.24609375

Приблизно 25% шанс.

> ! The General Binomial Probability Formula
>
> P(k out of n) = *n!* / **k!(n-k)!** * p^k(1-p)^(n-k)

Skewed Data

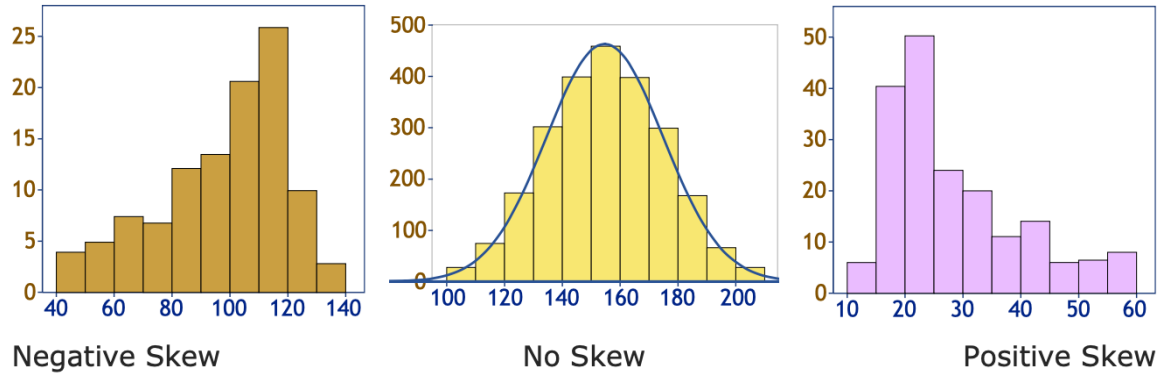
Наскільки перекреслені (сківні) дані

У статистиці, коли розглядаються дані, може виникнути ситуація, коли розподіл значень не є симетричним навколо середнього значення. Це називається сківністю або перекресленістю даних.

Перекреслені дані можуть мати розподіл, де хвіст з одного боку більш виражений або довший, ніж з іншого боку. Це може вказувати на наявність викидів або неоднорідність у даних.

Для дослідження сківності даних можна використовувати різні статистичні методи і графічні зображення, такі як гістограми, діаграми розсіювання та коштовні діаграми.

Граючись зі сківністю даних, ви можете спостерігати, як змінюється форма розподілу та його параметри, такі як середнє значення та стандартне відхилення.



Negative Skew | No Skew | Positive Skew

Сківність називають **негативною**, оскільки довгий "хвіст" знаходиться з боку від'ємних значень відносно вершини графіку розподілу.

Деякі люди вважають це "перекресленням ліворуч" (довгий хвіст знаходиться зліва від вершини).

Середнє значення також розташоване ліворуч від вершини графіку.

> **Нормальний розподіл не має скосості.**

Позитивна косість відбувається, коли довгий хвіст знаходиться на позитивній стороні піку, вона "схилена вправо".

Середнє значення знаходиться праворуч від значення піку.

Негативна косість (negative skew) відбувається, коли розподіл даних має довгий хвіст на негативній стороні піку, тобто ліворуч від середнього значення. Ось декілька прикладів, коли може спостерігатися негативна косість:

1. Дохід населення: В багатьох країнах розподіл доходів може мати негативну косість, оскільки більшість людей має низький дохід, але деякі особи мають дуже високі доходи, що створює довгий хвіст уліво від середнього доходу.
2. Розмір компаній: У бізнесі розмір компаній також може мати негативну косість. Більшість компаній є невеликими або середнього розміру, але деякі гіганти займають значну частку ринку, створюючи довгий хвіст уліво від середнього розміру компаній.

Позитивна косість (positive skew) відбувається, коли розподіл даних має довгий хвіст на позитивній стороні піку, тобто праворуч від середнього значення. Ось декілька прикладів, коли може спостерігатися позитивна косість:

1. Вартість нерухомості: У ринку нерухомості може бути позитивна косість. Більшість нерухомості має помірну вартість, але деякі розкішні властивості можуть мати значно вищу вартість, що створює довгий хвіст управо від середньої вартості.
2. Вік пенсіонерів: Розподіл віку пенсіонерів також може мати позитивну косість. Більшість пенсіонерів мають типовий вік пенсійного віку, але деякі люди можуть жити дуже довго, що призводить до позитивної косості та довгого хвоста управо від середнього віку пенсіонерів.

Звичайно, це лише приклади, і негативна або позитивна косість можуть спостерігатися у різних контекстах і ситуаціях. Важливо зазначити, що косість вказує на зсунення розподілу даних відносно його піку і може мати значення для аналізу та розуміння даних.