Experimental Design and Data Analysis

designed for HKU Budding Researchers Programme

Instructors:

- Prof. Joshua Ho

Helpers:

- Dr. Chun Hang Eden Ti
- Dr. Ray Hsu
- Ms. Danqing Angela Yin

Table of Contents

Here's a roadmap of our journey into experimental design and data analysis:

By the end of this lesson, you will be able to master the following four skills (**Learning Objectives**):

1. Loading Data

Import excel, Python basics

2. Data Manipulation

Data slicing, indexing, query

3. Basic Statistics

Hypothesis testing, t-test

4. Advance Statistics

Correlation, linear regression

The Mystery: The Secret to Longevity

Today, you are medical **detectives** solve one of humanity's biggest puzzles, i.e. longevity.

Why do people in Japan live to 85 while people in some countries barely reach 55?

1

Your Mission (for the next 2 hours)

- Your Quest: What if we could discover the secrets to a **longer**, **healthier** life?
- Your Clue: WHO Life Expectancy Data
- Your Toolkit: Statistics (and us)!

The World Health Organization tracks life expectancy for 193 countries and the contributing factors below:

- Healthcare Spending
- Education Levels
- GDP
- Disease Rates

By the end of this journey, you'll know how to use Python to uncover life-saving insights from real medical data.

The Detective's Dilemma: How do we find out the TRUTH?

First, Take A Guess?

Can we make assumption on why people in some country live longer?

Second: Find Clues?

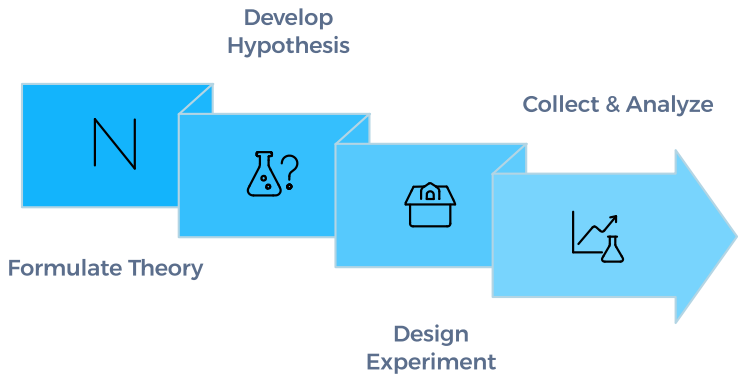
Example: Collect data like WHO to see if there's any patterns in longly lived countries

Third, Do A Study

We need a way to test what actually CAUSES life expectancy to increase

HKU Med

What is Experimental Design? The Researcher's Blueprint



Here is the workflow of conducting biomedical research.

As researchers, we don't just collect data - we design experiments to test our theories.

Experimental Design = The systematic plan for testing cause and effect.

Let's explore the researcher's toolkit: different types of experimental designs for different situations

Detective Tool #1: Randomized Controlled Trials

Have you heard the story on Aspirin, the original wonder drug? How it is being discovered and what is its effect?

The Physicians' Health Study (1980s): The aspirin breakthrough that changed life expectancy

- **Research question:** Can a simple drug like aspirin help people live longer by preventing heart attacks?
- **Method:** A massive, double-blind Randomized Controlled Trial (RCT) with over 22,071 doctors.
  - Some randomly took aspirin; others took a placebo (a dummy pill). **Randomization = Fair Test.**
- **Result:** The aspirin group had **44% fewer heart attacks**.
- **Conclusion:** This RCT provided definitive proof that aspirin can prevent heart attacks.

(credit: New York Times)

Harvard Health

Is aspirin a wonder drug? – Harvar...

...

Detective Tool #2: Looking Back vs. Looking Forward

Retroperspective studies

"Looking Back": The Smoking Detective Story

In the 1950s, researchers noticed lung cancer was rising. They looked backward at medical records and discovered the smoking-cancer link that changed life expectancy forever!

Like examining crime scene evidence after the fact.

Perspective studies

"Looking Forward": The Framingham Life Study

Since 1948, researchers have followed the same families in Framingham, Massachusetts, discovering how cholesterol, blood pressure, and exercise affect how long we live.

Like following suspects to see what they do next.

Both approaches help us solve the life expectancy puzzle.

Detective Tool #3: Statistics

1

Describing the Evidence

Mean, median, variance: Average life expectancy in our WHO dataset: 71.6 years. But some countries reach 84 while others stop at 50. The variance tells us there's a huge mystery to solve!

2

Testing Our Theories

Null vs. alternative: Hypothesis: 'Countries with higher healthcare spending have longer life expectancy.' We start assuming this is false (null hypothesis) until our data proves otherwise.

3

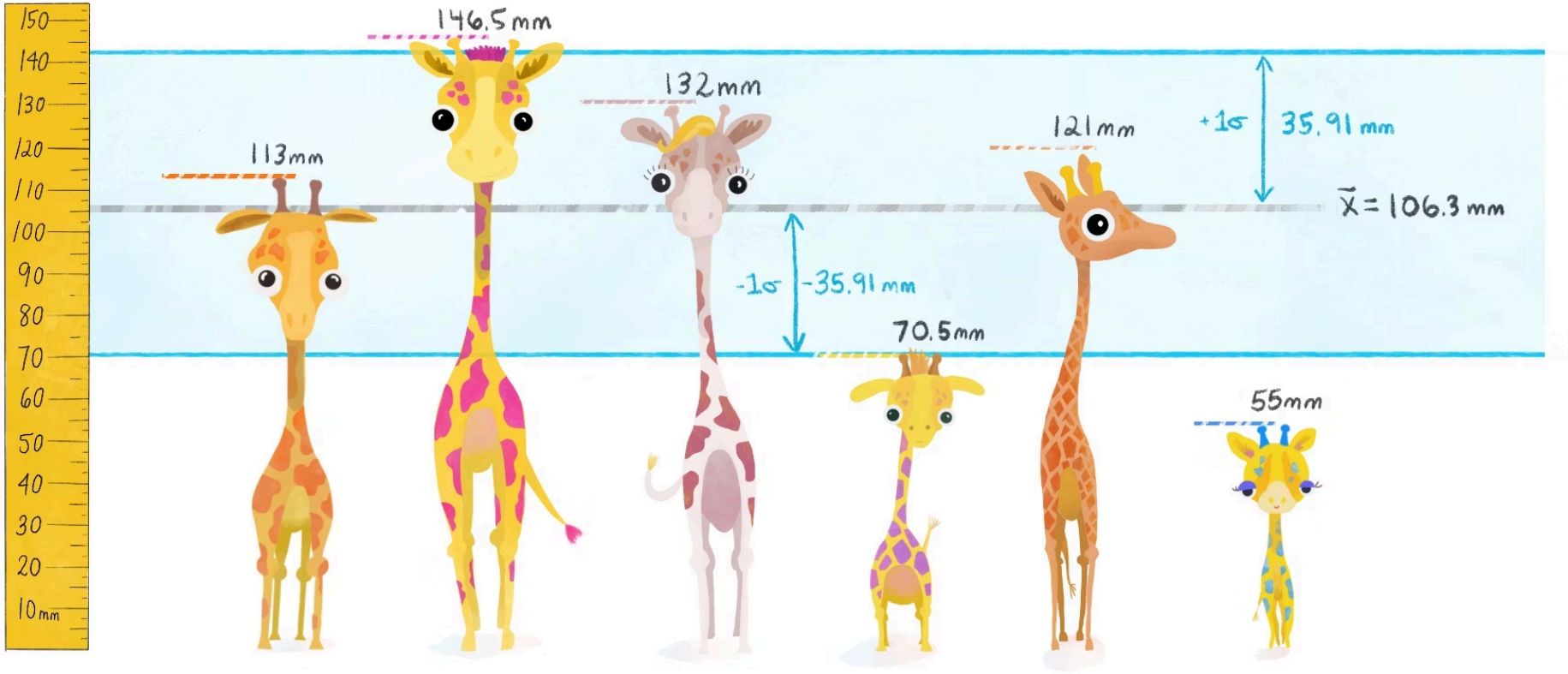
Solving the Case

p-value & alpha (0.05): When p-value < 0.05, we've found a real clue! Like discovering that education level strongly predicts life expectancy ($p < 0.001$) - that's not just coincidence! 🎉

HKU Med

Statistics Made Simple: Understanding Your Data

You bumped into a zoo, let's break down statistics using a visual example e.g. height from giraffe:



credit: https://tinystats.github.io/teacups-giraffes-and-statistics/aboutTheAuthors.html

1

Mean (Average)

Add up all the numbers and divide the sum by the number of values

2

Median (Middle Value)

Line them up from lowest to highest, pick the middle

3

Mode (Most Frequent Value)

The most frequent number

4

Variance (Spread)

5

Standard deviation (Dispersion)

Average squared difference of how far each number is deviated from the mean

Squared root of the variance

Why This Matters: This is called measure of central tendencies. Before we can test theories about what causes

HKU Med

The Detective's Reasoning Framework: Hypothesis Testing

"Imagine you're a detective investigating whether a suspect is guilty...at what level of evidence a suspect is convicted to the crime? This is same as in hypothesis testing in statistics..."

Hypothesis: a proposition that researcher wish to verify

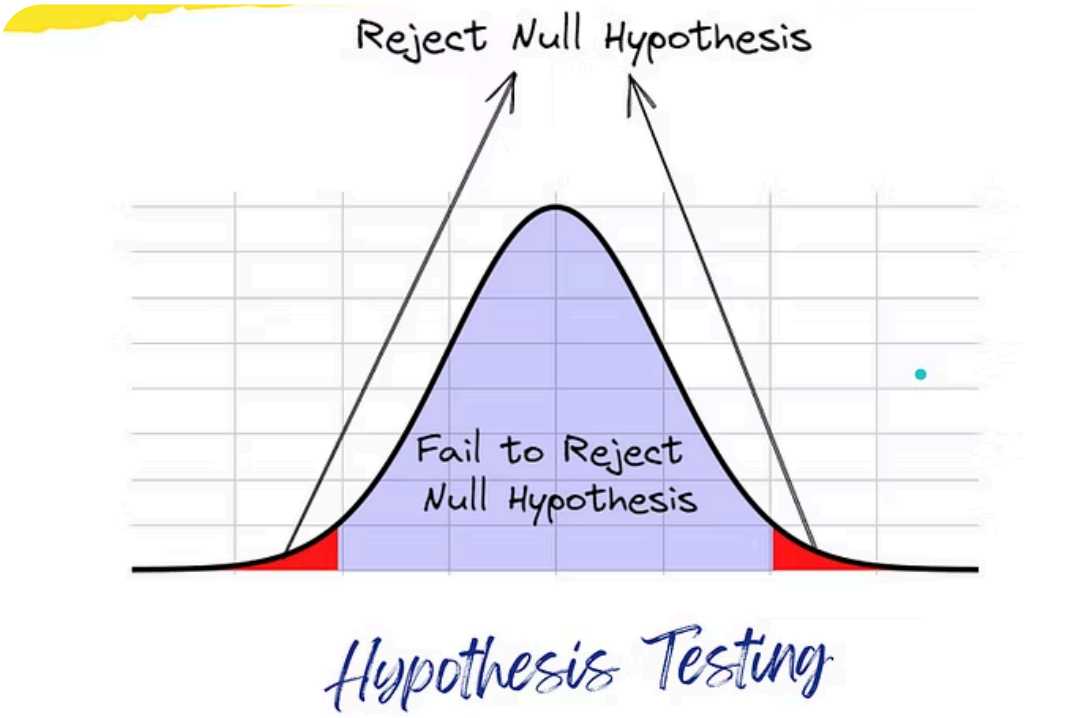
To test our hypothesis, we make following assumptions:

1. **Null Hypothesis (H_0)**:

Assume there is NO effect or difference between two variables

1. **Alternative hypothesis (H_1)**:

Suggest there IS an effect or difference between two variables.



credit: https://medium.com/@dharunasri.na96/hypothesis-testing-0faf0a00ffe3

Let's apply this to Life Expectancy Case

We want to know if there's a significant difference in life expectancy between two distinct groups of countries.

The Evidence Scale: P-value

After we formed our hypothesis, we need to determine at what level whether we accept or reject the hypothesis!"

1

The Bell Curve = Crime Scene Evidence Distribution

"Most evidence points to 'innocent' (the middle of the bell curve). But sometimes evidence is so extreme it suggests 'guilty'!"

2

The 5% Rejection Region = the "Guilty" Zones

"In court, we need 95% certainty to convict. In statistics, we use the same standard: $p < 0.05$ means less than 5% chance it's just coincidence."

Real Detective Example:

"Fingerprints at crime scene: If there's less than 5% chance they got there by accident, we arrest the suspect!"

In Our Life Expectancy Case:

"If there's less than 5% chance that country's development status and life expectancy are related, we conclude level of development DOES matter!"

HKU Med

The P-Value (continued)

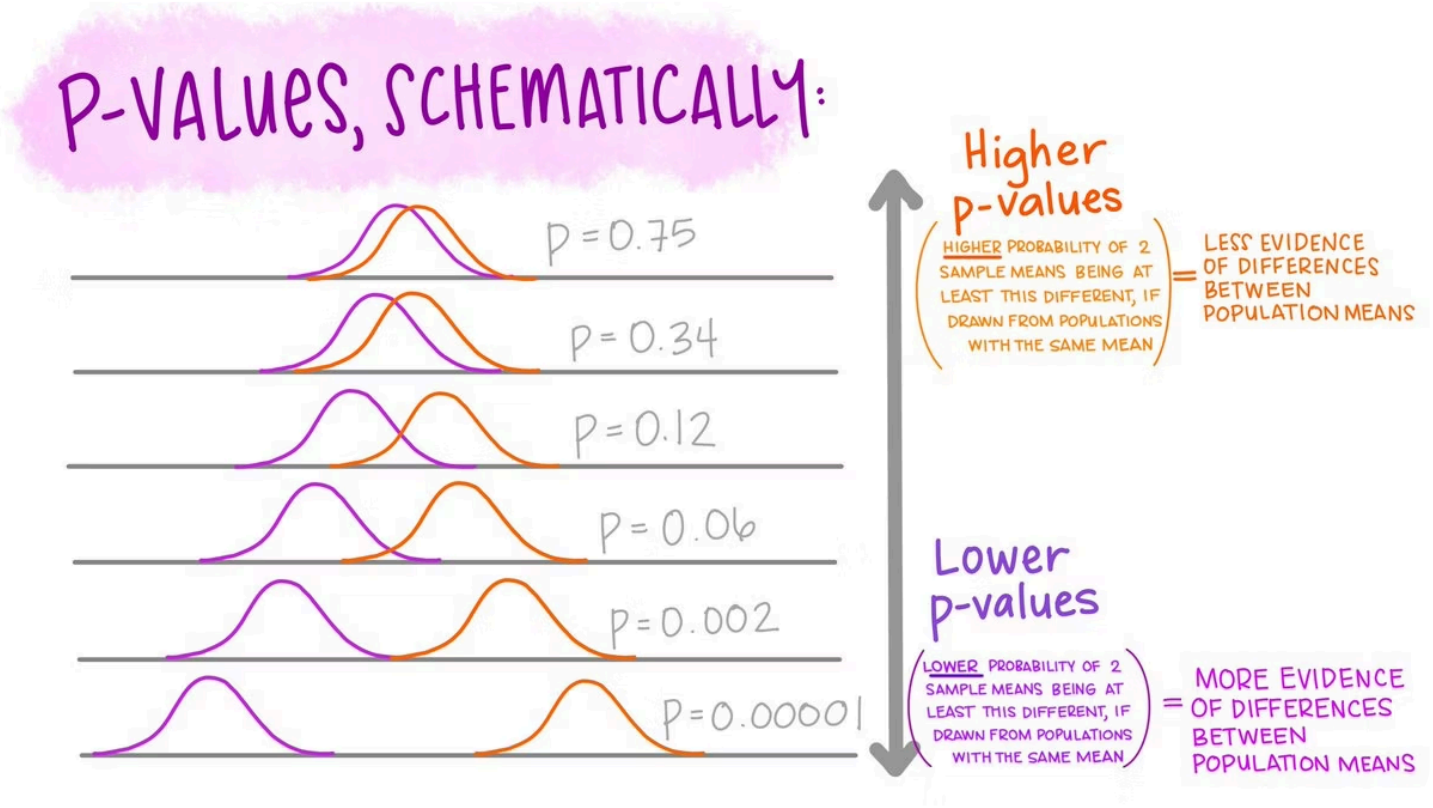
"Think of p-value as your 'Evidence Strength Meter' - like a detective's confidence scale!"

p > 0.10

"Weak evidence - not enough to make an arrest"

p = 0.05

"Strong evidence - just enough to convict (our threshold!)"



credit: Alison Horst

The Detective's Rule:

"If $p < 0.05$, we have enough evidence to 'convict' - reject the null hypothesis and say the relationship is REAL!"

Memory Trick:

"Lower p-value = Stronger evidence. Think: 'p' stands for 'probably just chance' - the lower it is, the less likely it's just chance!"

The Detective's Advanced Toolkit: T-Tests

"As a super-sleuth, you don't just look for clues – you compare them! A T-test is your go-to tool when you need to compare two groups of evidence and see if they're truly different, or if it's just a trick of the light."

What does T-tests do?

It helps us **compare two groups** to see if their **average values are significantly different**. For example, is the average life expectancy in Country Group A truly higher than in Country Group B?

Our Detective Analogy

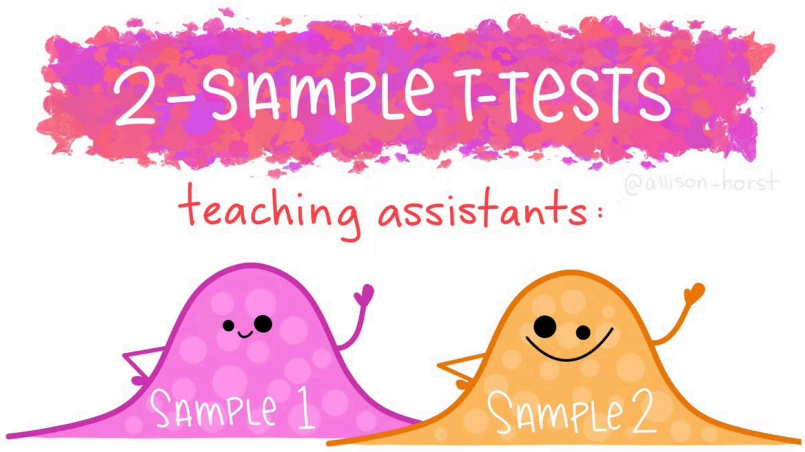
Imagine you have two suspects. A T-test helps you compare a specific characteristic (like their height, or their details) to see if there's a significant difference between them, or if they're just typical variations.

Types of T-Tests in Our Case:

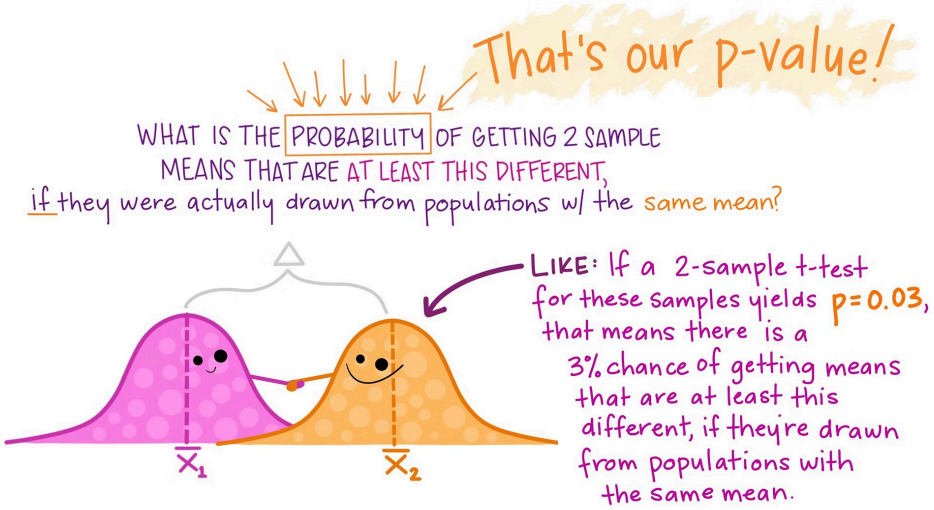
One-Sample T-Test

Two-Sample T-Test

"Is the average life expectancy of a specific country (e.g., USA) different from the *global average*?" You compare one group to a known standard or target.



"Is the average life expectancy of *developed nations* different from *developing nations*?" You compare two completely separate groups to each other.



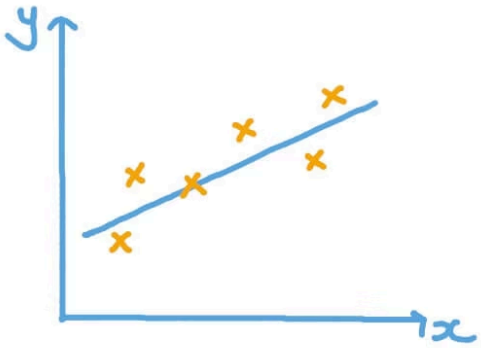
In Our Life Expectancy Case:

We might use a T-test to compare the average life expectancy of countries with high healthcare spending versus countries with low healthcare spending. If the T-statistic is large enough, and the p-value is small (e.g., < 0.05), we can conclude that the difference in life expectancy between these groups is statistically significant.

The Detective's Magnifying Glass: Finding Connected Clues (Correlation)

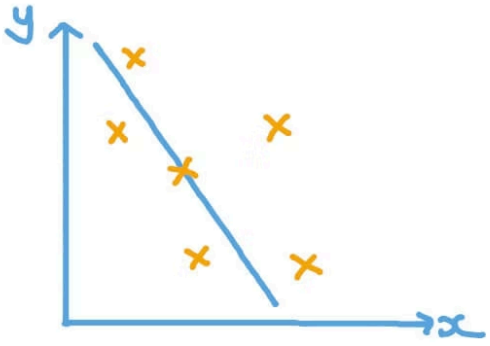
As detectives, we often find two clues that seem linked. Correlation is simply finding if two things change together. Does one go up when the other goes up? Or does one go down when the other goes up?

What type of linear correlation might exist between the mass loaded onto a spring and its extension?



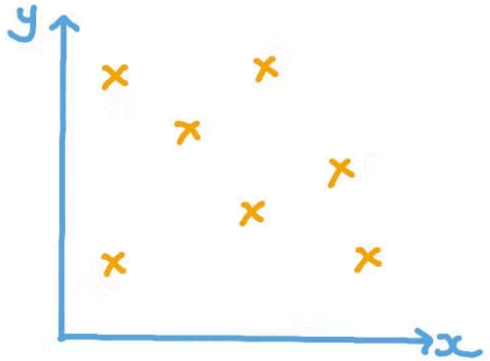
Positive correlation

✓ In general, as x increases, y increases

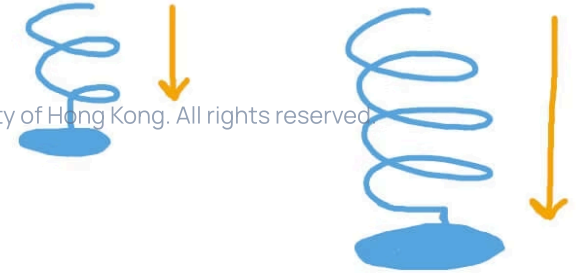


Negative correlation

In general, as x increases, y decreases



NO linear correlation



As mass increases, extension also increases.

A positive correlation

HKU Med

The Suspects: Variables in Our Life Expectancy Case

Independent Variables

"What We Control"

Healthcare spending per person, education years, GDP per capita - these are the factors we suspect might cause longer life.

Think: 'What might be causing some countries to live longer?'

Dependent Variables

"What We Measure"

Life expectancy in years - this is what we're trying to explain and predict.

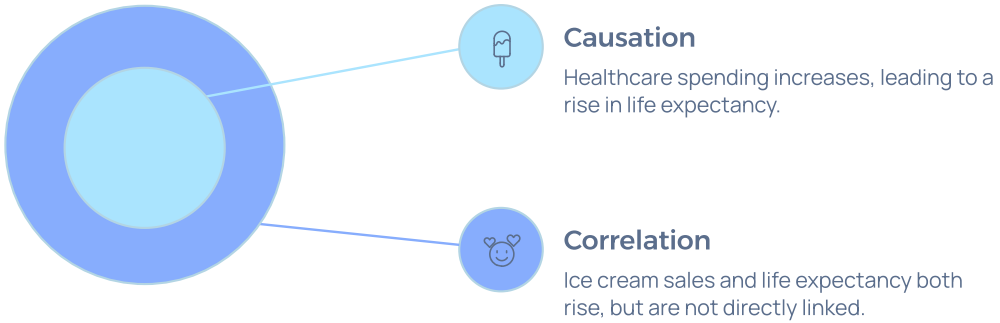
Think: 'This is what changes when our suspects are involved'

The Controls

What We Keep Constant

Geographic region, data collection year - we control these so they don't confuse our investigation.

Think: 'What could mislead us if we don't account for it?'



Correlation vs. Causation: The Detective's Crucial Distinction

"As keen data detectives, understanding the difference between correlation and causation is paramount. Just because two clues appear together, doesn't mean one caused the other! This is a common trap in any investigation."

1

🔍 Correlation

What it is: Two variables move together. When one changes, the other tends to change in a predictable way.

- Positive: Both increase or decrease together (e.g., more study, higher grades).
- Negative: One increases as the other decreases (e.g., more exercise, lower weight).

2

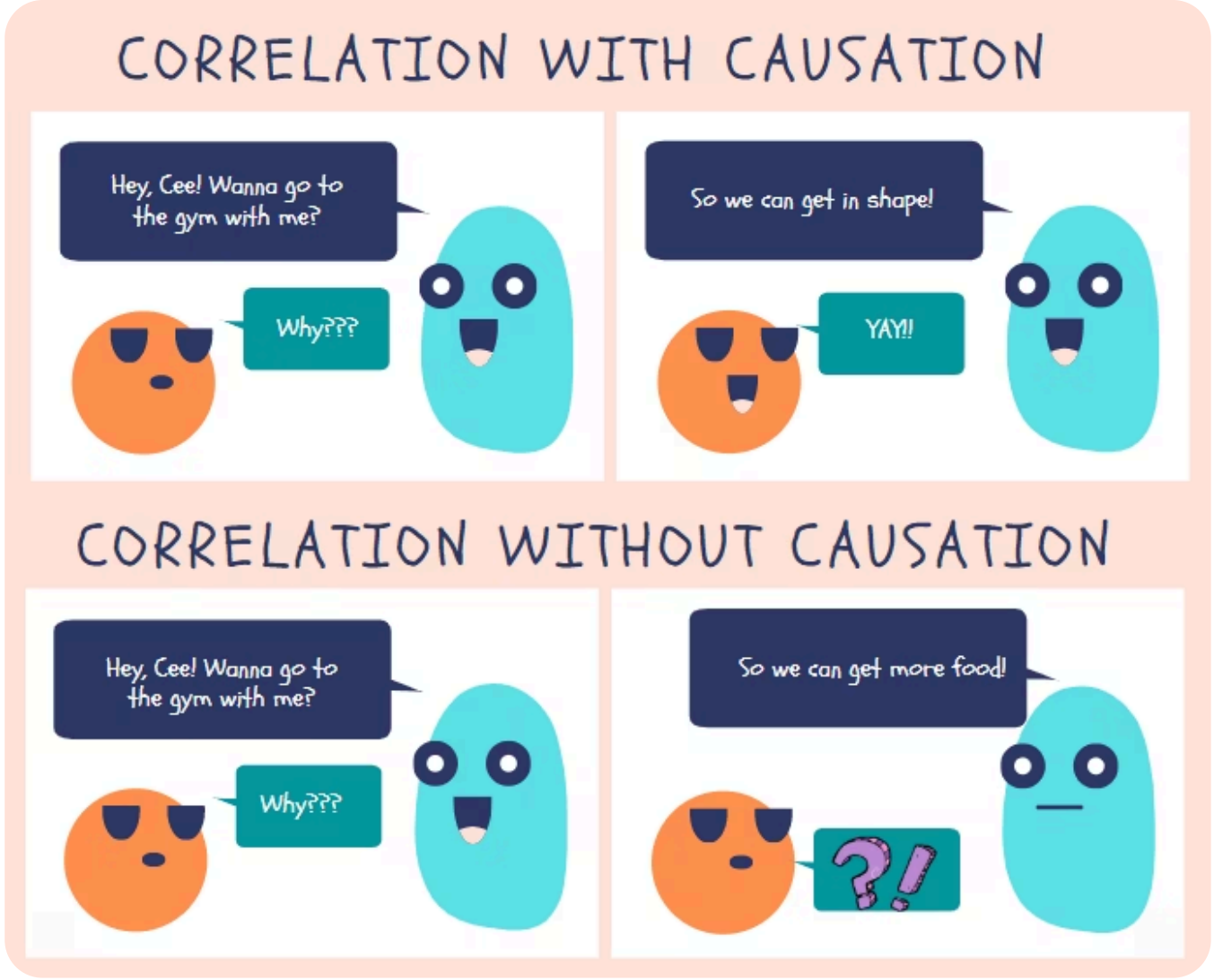
➡️ Causation

What it is: One variable directly influences or produces a change in another.

- Direct Link: A clear cause-and-effect relationship (e.g., turning a light switch on causes the light to illuminate).
- Requires Experimentation: Often needs controlled experiments to prove.

The Pitfall: The Third Variable (Confounding Factor)

"Often, a hidden 'third suspect' is actually responsible for both observed changes, leading us to falsely assume causation."



credit: https://devskrol.com/2020/07/17/correlation-vs-causation/

For example, high ice cream sales and increased swimming incidents are correlated, but neither causes the other. The "third variable" — warm weather — causes both!

The Detective's Equation: Decoding Linear Regression

"Once we've found a connection (correlation), the next step for a savvy detective is to predict! Linear Regression is like your statistical crystal ball, helping you predict one variable based on another."

Predicting the Outcome

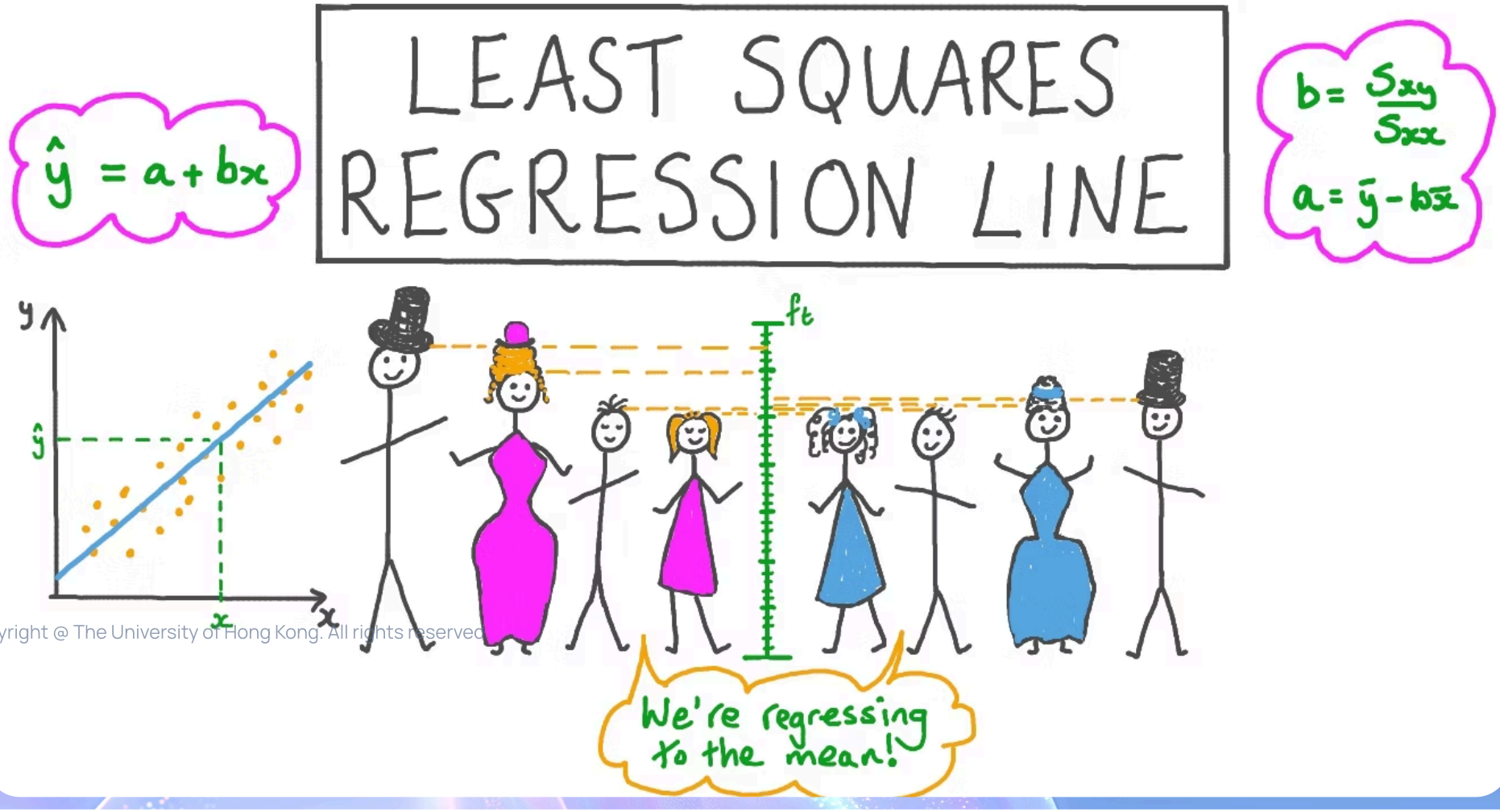
It's about finding a straight line that best describes the relationship between two variables. Think of it as drawing the clearest path through all your clues!

The "Line of Best Fit"

This special line minimizes the distance to all the data points. It's the most accurate trend you can draw to represent the correlation you found.

Making Informed Guesses

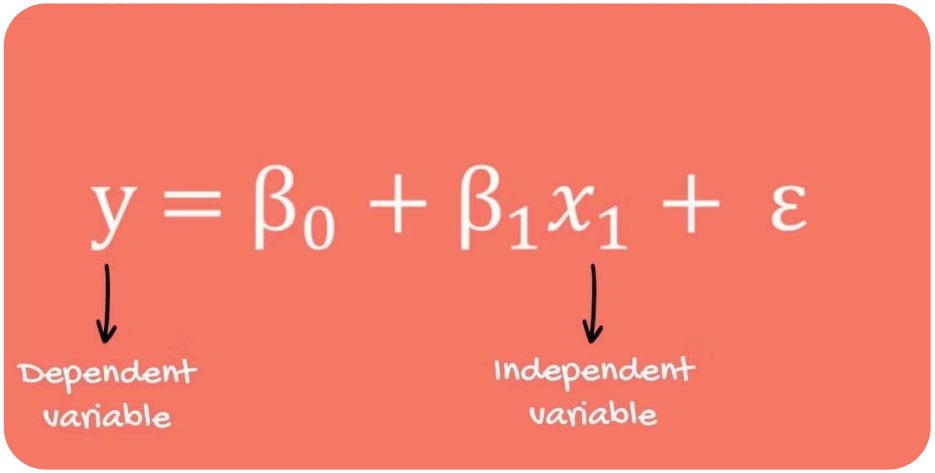
Once you have this line, you can plug in a new value for one variable and predict what the other variable might be. It helps us forecast unknown scenarios.



HKU Med

The Detective's Equation: Decoding Linear Regression

Our "crystal ball" formula uses linear regression to predict one 'mystery' (like Life Expectancy) based on various 'clues'. Think of it as a precise way to map out how our evidence connects to the outcome.



credit: https://medium.com/analytics-vidhya/machine-learning-simple-linear-regression-using-python-7d13e8ac8300

1

Y: The Mystery

This is what we're trying to predict or explain (e.g., Life Expectancy).

2

β_0 (Beta Nought): The Baseline

The expected value of Y when all our 'clues' (X values) are zero. Our starting point!

3

β_1 (Beta One): Clue's Impact

These coefficients show how much Y changes for each unit increase in a specific 'clue' (X), holding other clues constant.

4

X_1: The Clues

These are our independent variables or factors we believe influence the mystery (e.g., Healthcare Spending, Education Years).

5

ε (Epsilon): Unexplained Factors

The part of the mystery our clues don't quite solve – the random variation or factors we haven't accounted for.

A bigger Beta means a stronger link!

Your Assessment: The Detective's Next Case

Having sharpened our detective skills, it's time to apply them to a real-world investigation: global health challenges. This is our project roadmap, guiding us from identifying a mystery to proposing innovative solutions, all powered by data.

Identify the Health Mystery & Data

Our first task is to uncover a global health problem that can be explored with readily available datasets (e.g. https://data.gov.hk/en/ or https://www.kaggle.com/datasets). This could involve life expectancy, disease prevalence, or factors like diet and mental health.

Review the Case Files

Before any new analysis, we'll conduct a literature review. What have other 'detectives' already learned about this health problem? This background research helps us formulate hypotheses and understand the context.

Analyze the Evidence

Using our statistical tools—correlation, regression, and t-tests—we'll sift through the data. We'll identify significant factors associated with our chosen health problem, looking for patterns and relationships.

Propose Health Policies

Based on our data-driven insights, we'll propose practical health policies designed to prevent or mitigate the identified problem. This is about translating evidence into actionable strategies.

Innovate with Health Technology

Finally, we'll explore how emerging health technologies can contribute to solving the problem. Could AI diagnostics, wearable sensors, or digital therapeutics offer new ways to address the challenge?