ELEC3544 · Fall 2026

Data Science with Foundation Models

Yang Liu

The University of Hong Kong



Lecture 3

Probability and Statistics for Data Science

What Do You Need to Know?

You can ask an AI to compute any statistics, run any simulation, generate any visualization. It will do these things very efficiently. So what do YOU need to know?

Three things AI can't do for you:

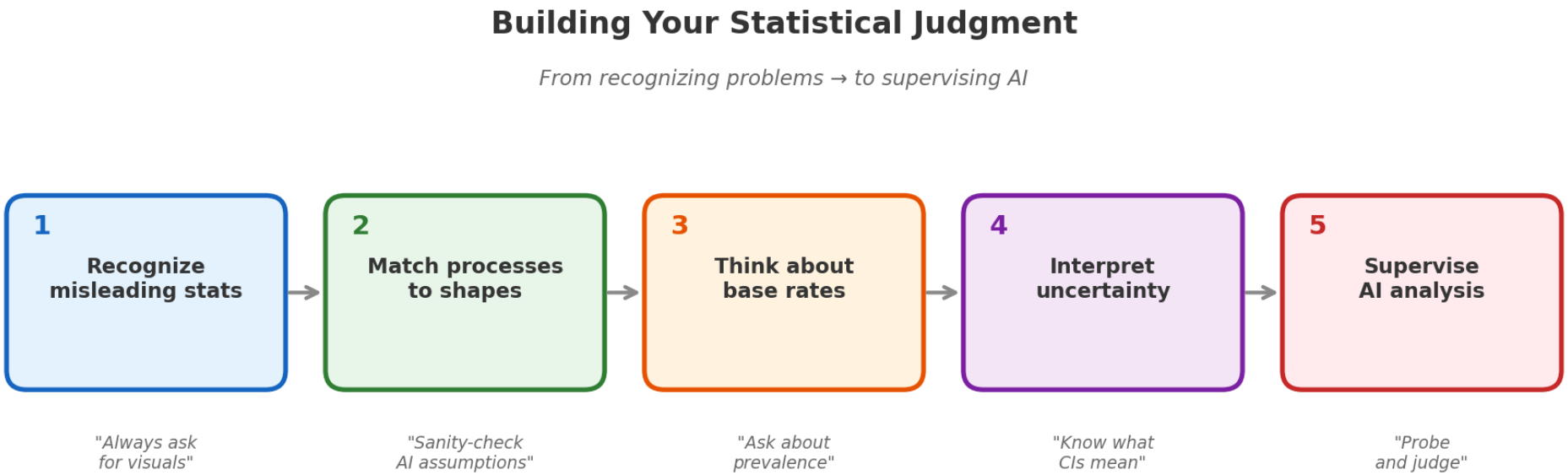
Know your context — AI doesn't understand your domain, your stakeholders, or your constraints.

Ask the right questions — AI answers what you ask. If you ask the wrong question, you get the right answer to the wrong question.

Judge what matters — AI can tell you an effect is statistically significant. It can't tell you whether that effect is worth caring about.

> Today's lecture builds the judgment you need to supervise AI statistical analysis — to catch when it's misleading, to ask probing follow-ups, and to translate statistical results into decisions.

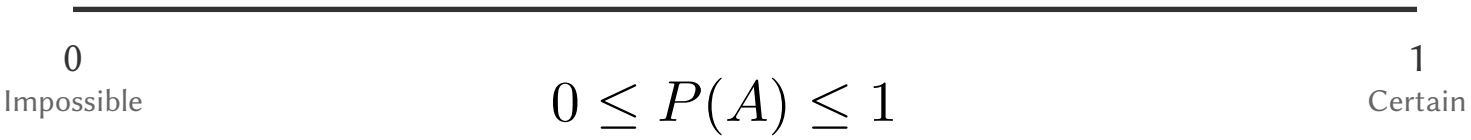
Today's Journey



Each skill annotated: "Always ask for visuals" → "Sanity-check AI assumptions" → "Ask about prevalence" → "Know what CIs mean" → "Probe and judge"

What is Probability?

Probability measures the likelihood of an event occurring, on a scale from 0 to 1.



Two key interpretations:

Frequentist Interpretation

If we repeated the experiment many times, what fraction of trials would show this outcome?

Subjective Interpretation

Our degree of belief or confidence in an outcome. (Connects to Bayesian thinking later.)

Example: Flip a fair coin 1000 times. About 500 will be heads. So P(heads) ≈ 0.5.
The probability reflects the long-run frequency of the outcome.

Random Variables — The Bridge from Events to Numbers

A random variable assigns a numerical value to each possible outcome of a random process. This is how we move from talking about "events" to doing mathematics with "distributions."

Two types:

Discrete

Countable outcomes: dice rolls, counts, categories.
Examples: Number of defects, coin flip results

Continuous

Measurements on a continuum: height, time, temperature.
Examples: Reaction time, stock price

Example: Let X = number of heads in 3 coin flips.
X can take values: 0, 1, 2, or 3. Each value has a probability:
P(X=0) = 1/8, P(X=1) = 3/8, P(X=2) = 3/8, P(X=3) = 1/8

$$X = \text{number of heads in 3 coin flips}$$
$$X \in \{0, 1, 2, 3\}$$
$$P(X = 0) = \frac{1}{8}, \quad P(X = 1) = \frac{3}{8}$$
$$P(X = 2) = \frac{3}{8}, \quad P(X = 3) = \frac{1}{8}$$
$$\sum_x P(X = x) = 1$$

This framework lets us describe uncertainty mathematically — the foundation of statistical inference.

Population vs. Sample — The Fundamental Distinction

Population

The complete set of all individuals or observations we care about.
Often very large or even infinite.
Has true parameters: μ (mean), σ (std dev)

Sample

A subset of the population that we actually observe. What we must work with in practice.

Gives us statistics: $\bar{x}$ (sample mean), s (sample std dev)

Key Insight: We almost never see the whole population.
We use samples to make inferences about populations. This is the core challenge of statistics.

Parameters vs. Statistics — The Naming Convention

Quantity	Population (Greek)	Sample (Roman)
Mean	μ (population mean)	$\bar{x}$ (sample mean)
Standard Deviation	σ (population standard deviation)	s (sample standard deviation)

Notation We'll Use Today

Symbol	What It Represents	Example
n	Sample size (number of observations)	n = 100 students
x_i	The i-th observation in our sample	x_3 = 85 (third score)
$\bar{x}$	Sample mean (our estimate of μ)	$\bar{x}$ = 78.5
s	Sample standard deviation (estimate of σ)	s = 12.3
$P(A)$	Probability of event A	P(rain) = 0.3
$P(A \mid B)$	Probability of A given B has occurred	P(rain | cloudy) = 0.6

📌 Bookmark this slide — we'll use this notation throughout the course.

From Random Variables to Events

How do random variables and events connect?

Key Insight: An **event** is a statement about what values a random variable takes. Events (A, B) are defined by conditions on random variables (X).

Translating between notations:

Random Variable Condition	Event Notation	In Words
$X = 2$	$A = \{X = 2\}$	"X equals 2"
$X > 3$	$B = \{X > 3\}$	"X is greater than 3"
$X \leq 1$	$C = \{X \leq 1\}$	"X is at most 1"

Example: Let X = number of heads in 3 coin flips.
Define event $A = \{X \geq 2\}$ = "at least 2 heads"
Then: $P(A) = P(X \geq 2) = P(X=2) + P(X=3) = 3/8 + 1/8 = \mathbf{1/2}$

When we write $P(A)$ or $P(A|B)$, the events A and B are typically defined by conditions on random variables.

Three Fundamental Probability Rules

Complement Rule $P(\text{not } A) = 1 - P(A)$

Intuition: The probability of something happening plus the probability of it not happening equals 1 (certainty).
Example: If P(rain) = 0.3, then P(no rain) = 0.7

Addition Rule $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

For mutually exclusive events (can't both happen): P(A or B) = P(A) + P(B)
Intuition: Add individual probabilities, subtract overlap to avoid double-counting.
Example: P(roll 1 or 2) = 1/6 + 1/6 = 2/6 (no overlap)

Multiplication Rule $P(A \text{ and } B) = P(A) \times P(B \mid A)$

For independent events (A doesn't affect B): P(A and B) = P(A) × P(B)
Intuition: If A doesn't affect B, just multiply. If it does, use the conditional probability.
Example: P(two heads in a row) = 0.5 × 0.5 = 0.25

These rules form the foundation for all probability calculations — including Bayes' rule.

The Counting Perspective

Many probability problems are counting problems in disguise.

$$P(A) = \frac{\text{number of ways } A \text{ can happen}}{\text{total number of possible outcomes}}$$

This works when all outcomes are equally likely.

Example: Drawing a red card from a standard deck.
26 red cards (hearts + diamonds), 52 total cards.
P(red) = 26/52 = 0.5

Why this matters for Bayes: When we ask "of all positive tests, how many are true positives?" — that's a counting question.
The counting perspective will make Bayes' rule intuitive rather than mysterious.

Summary Statistics — Measuring Location

Where is the center of the data?

Mean ($\bar{x}$) — The Arithmetic Average

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Intuition: The "balance point" of the data — sum all values and divide by count.

⚠ Limitation: Sensitive to outliers — one extreme value can pull the mean dramatically.

Median — The Middle Value

When data is sorted, the median is the middle value. 50% of values fall below, 50% above.
Advantage: Resistant to outliers — extreme values don't affect it.

When to use which: Use median for skewed distributions or when outliers are present (e.g., income, housing prices).
Use mean when the distribution is roughly symmetric and outliers are not a concern.

Summary Statistics — Measuring Spread

How much variation is there in the data?

Standard Deviation *(s)* — Typical Distance from the Mean

Variance is the average of squared deviations from the mean.
Standard deviation is the square root of variance — bringing us back to the original units.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Intuition: "How far, on average, are values from the center?"

Note: We divide by n−1 (not n). This is Bessel's correction — it is based on the degrees of freedom and accounts for the fact that a sample tends to underestimate population variability.

Interquartile Range (IQR) — Range of the Middle 50%

IQR = Q_3 − Q_1 (75th percentile minus 25th percentile). Advantage: Resistant to outliers, like the median.

Pattern: Mean + Std Dev go together (both sensitive to outliers).
Median + IQR go together (both resistant). Choose the pair that fits your data.

Summary Statistics — Measuring Relationships

How do two variables move together?

Covariance — How Variables Vary Together

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

⚠ sample (co-)variance 1/(*n*-1) instead of 1/*n*

- **Positive covariance:** when *X* is above its mean, *Y* tends to be above its mean
- **Negative covariance:** when *X* is above its mean, *Y* tends to be below its mean
- **Problem:** magnitude depends on units — hard to interpret

Pearson Correlation Coefficient, or Correlation (*r*) — Standardized Covariance

$$r = \mathrm{Cov}(X,Y) \,/\, (s_x \cdot s_y)$$

Dividing by the standard deviations removes unit dependence, giving a dimensionless measure.

Symbol Definitions:

Symbol	Meaning
$\mathrm{Cov}(X,Y)$	Covariance between *X* and *Y*
s_x , s_y	Sample standard deviations of *X* and *Y*
r	Pearson correlation coefficient (dimensionless)

Correlation — Properties and Limitations

Range and Interpretation

$r \in [-1, +1]$

- $r = +1$: Perfect positive linear relationship
- $r = 0$: No linear relationship
- $r = -1$: Perfect negative linear relationship

Full Formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

⚠ Critical Limitations

1. **Correlation only measures LINEAR association.** A perfect curved relationship can have $r \approx 0$.
2. **Correlation does not imply causation.** Two variables can be correlated due to a third factor.
3. **Identical correlations can hide completely different data structures** — as we're about to see.

The Tools — Visualizations

Three essential plot types:

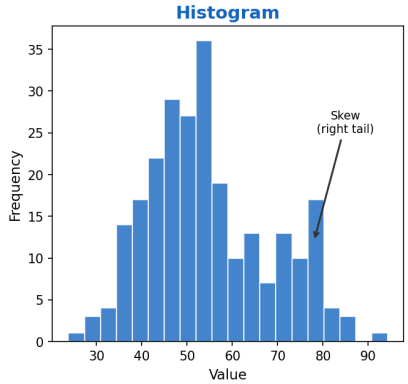
Outlier if: $x < Q_1 - 1.5 \times \text{IQR}$ or $x > Q_3 + 1.5 \times \text{IQR}$

Histogram

Distribution of a single variable.

Reveals: symmetry, skewness, modality (unimodal, bimodal), outliers.

Use when: "What does this variable look like?"



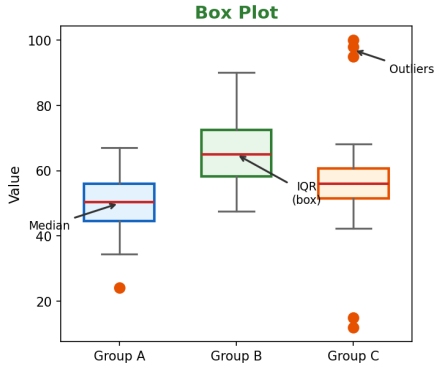
Use: "What does this variable look like?"

Box Plot

Five-number summary: min, Q_1, median, Q_3, max.

Outliers shown as individual points (beyond 1.5×IQR).

Use when: comparing distributions across groups.



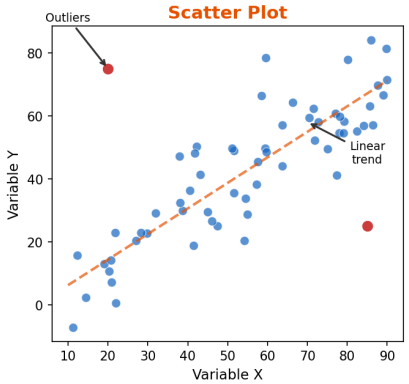
Use: Comparing distributions across groups

Scatter Plot

Relationship between two variables.

Reveals: linear vs. nonlinear, clusters, outliers, strength.

Use when: examining association between X and Y.



Use: Examining association between X and Y

Statistics summarize. Visualizations reveal structure. You need both.

A Test of Our Tools

I have four datasets. Let's analyze them with our tools and see what we learn.

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

These look like different numbers. Let's see what the statistics say.

CTA Demo A: When Summary Statistics Lie

🤖 CTA Demo: When Summary Statistics Lie

What We'll Explore: Can identical summary statistics hide completely different data structures?
Pattern: Verification Through Computation

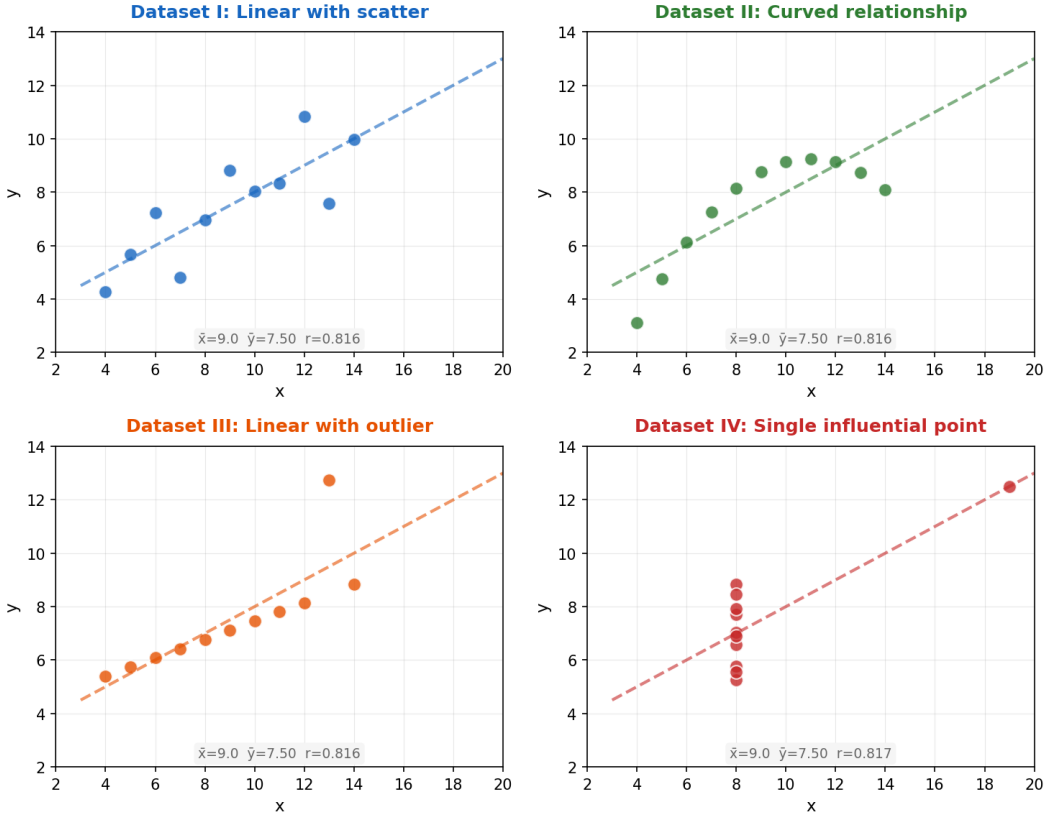
Demo Staging:

Turn 1: "Compute mean of *x*, mean of *y*, standard deviation of *x*, standard deviation of *y*, and correlation for each dataset. Show just the numbers, no plots yet."

⏸️ **PAUSE for prediction:** "Based on these statistics — which are nearly identical — what do you think these datasets look like?"

Turn 2: "Now create a 2×2 grid of scatter plots, one for each dataset."

The Reveal



All four datasets have nearly identical statistics:

$$\bar{x} \approx 9.0 \qquad \bar{y} \approx 7.5 \qquad s_x \approx 3.3 \qquad s_y \approx 2.0 \qquad r \approx 0.816$$

Why This Matters Now

Key Insight: **Ask for the right thing!** The AI agent computed these statistics correctly. Every number is right. But if you had only asked for summary statistics — which is what most people do — you would have concluded these datasets are essentially the same.

What each dataset reveals:

Dataset I: Clean linear relationship — statistics are appropriate

Dataset II: Curved (quadratic) — correlation misses nonlinear structure

Dataset III: One outlier drives correlation away from $r = 1$

Dataset IV: No relationship — one extreme point creates illusion

Anscombe created these in 1973 to show statisticians they need to look at their data.

The lesson is MORE relevant in 2026: AI computes what you ask for, not what you need to see.

Your First Prompting Habit

Always request both summary statistics AND appropriate visualizations.

Example prompt template:

> *"Please analyze this dataset. Provide:*
>
> *(1) Summary statistics including mean, median, standard deviation, and correlation where applicable.*
>
> *(2) Appropriate visualizations — histogram for distributions, scatter plots for relationships.*
>
> *(3) Any patterns or anomalies you notice."*

This prompt explicitly requests visualizations. Don't rely on the AI to know what you need to see.

✓ VERIFY: Validate

Cross-check quantitative outputs against visual evidence. If the numbers and pictures tell different stories, investigate why.

Why Distributions Matter

A distribution describes the possible values a random variable can take and how likely each value is.

Formal Definitions:

Discrete random variable: Probability Mass Function (PMF) gives $P(X = x)$ for each possible value x.
Continuous random variable: Probability Density Function (PDF) describes relative likelihood;

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Key Principle: If you know the data-generating process, you can predict the distribution shape. This helps your sanity check of AI outputs — if the shape doesn't match the story, something's wrong.

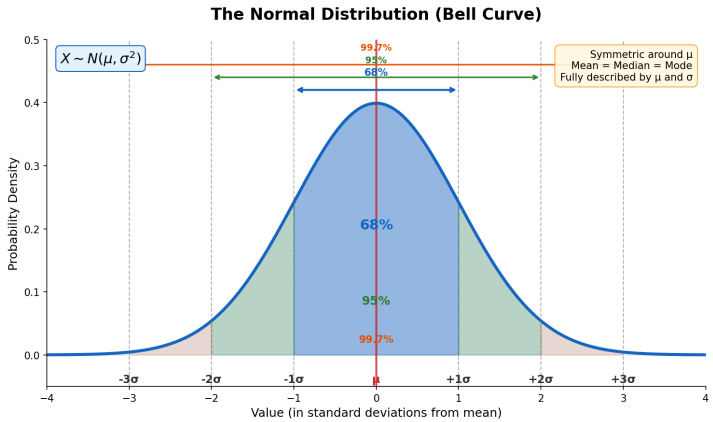
Example: AI fits a normal distribution to your data. But your data is 'time between customer arrivals.' That should be exponential (right-skewed), not normal (symmetric). The AI assumed that doesn't match reality.

The next four slides introduce distribution families by their generative stories.

You don't need to memorize formulas — AI provides those. You need to match the story to the shape.

Normal Distribution — The Bell Curve

The Story: Arises when many small, independent effects add together. The Central Limit Theorem explains why.



$X \sim N(\mu, \sigma^2)$

Symbol	Meaning	Role
μ (mu)	Mean	Center of the distribution
σ^2 (sigma^2)	Variance	Spread (σ is standard deviation)

Probability Density Function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

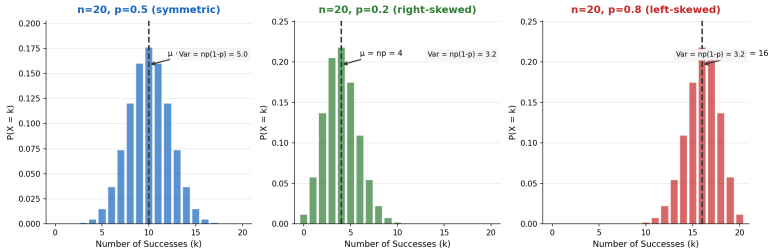
Key Properties:

- Symmetric around μ
- 68-95-99.7 rule: 68% within 1σ, 95% within 2σ, 99.7% within 3σ
- Fully determined by just two parameters (μ and σ)

Examples: Heights, measurement errors, test scores, stock returns (approximately)

Binomial Distribution — Counting Successes

The Story: Arises when counting "successes" in a fixed number of independent trials, each with the same probability of success.



$X \sim \text{Binomial}(n, p)$ where X = number of successes

Symbol	Meaning	Constraints
n	Number of trials	Positive integer
p	Probability of success on each trial	$0 \le p \le 1$

Probability Mass Function:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ , where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

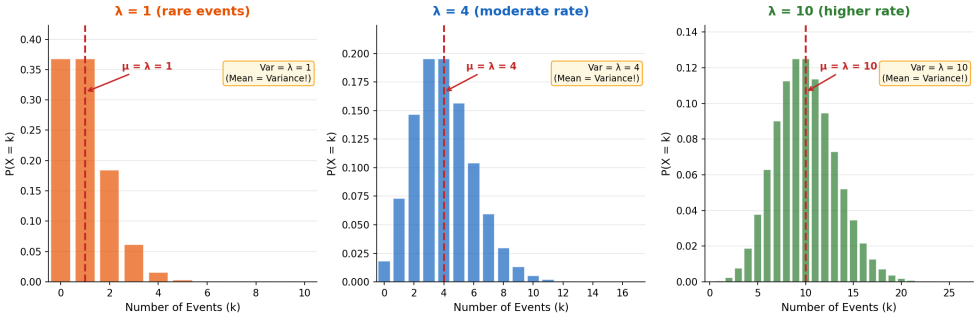
Key Properties:

- Mean: $E[X] = np$
- Variance: $\text{Var}(X) = np(1-p)$
- Shape: symmetric when $p = 0.5$, skewed otherwise

Examples: Heads in 50 coin flips, defective items in a batch of 100, patients responding to treatment

Poisson Distribution — Counting Rare Events

The Story: Arises when counting events in a fixed interval (time, space) when events occur independently at a constant average rate.



$X \sim \text{Poisson}(\lambda)$, X = number of events

Symbol	Meaning	Constraints
λ (lambda)	Average rate of events	$\lambda > 0$

Probability Mass Function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Key Properties:

- Mean: $E[X] = \lambda$
- Variance: $\text{Var}(X) = \lambda$ *(mean equals variance — a distinctive feature!)*
- Shape: Right-skewed for small λ, approaches normal for large λ

Examples: Typos per page, customers per hour, accidents per month, emails per day, mutations per genome region

Exponential Distribution — Waiting Times

The Story: Arises when measuring time between consecutive events in a Poisson process. If events occur at rate λ, waiting time is exponential.

$X \sim$ Exponential(λ), X = waiting time until next event

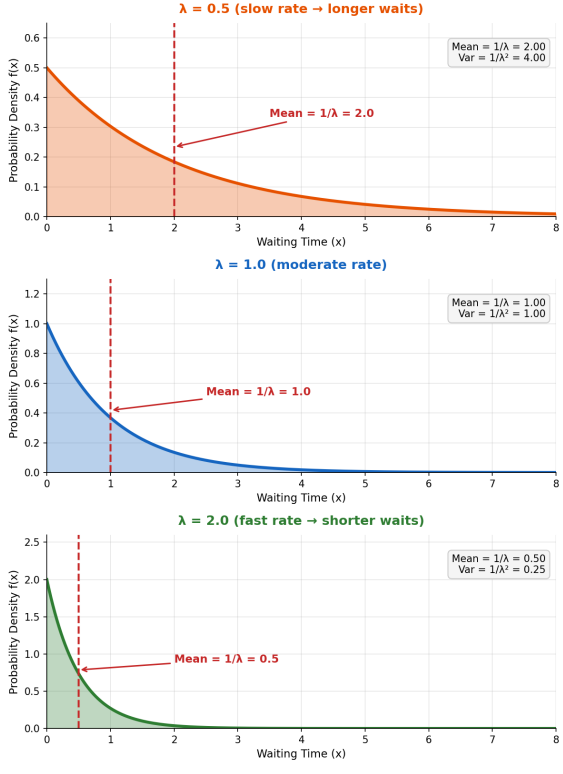
Symbol	Meaning	Constraints
λ (lambda)	Rate parameter	$\lambda > 0$

Probability Density Function:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

Key Properties:

- Mean: $E[X] = 1/\lambda$ (average waiting time)
- Variance: $\text{Var}(X) = 1/\lambda^2$
- **Memoryless property:** $P(X > s+t \mid X > s) = P(X > t)$ — "doesn't remember" how long you've waited
- Shape: Always right-skewed, starts high at x=0 and decays



Connection: Poisson counts events; Exponential measures time between them.
Examples: Time between customers, time until failure

Reasoning Under Uncertainty

How do we update our beliefs when we get new information?

Think conditionally

Distinguish $P(A|B)$ from $P(B|A)$

Remember base rates

The prior probability matters enormously

Interpret intervals

What confidence intervals actually mean

These skills matter because AI can compute probabilities correctly but can't judge whether the right probability is being computed.

Conditional Probability — The Foundation

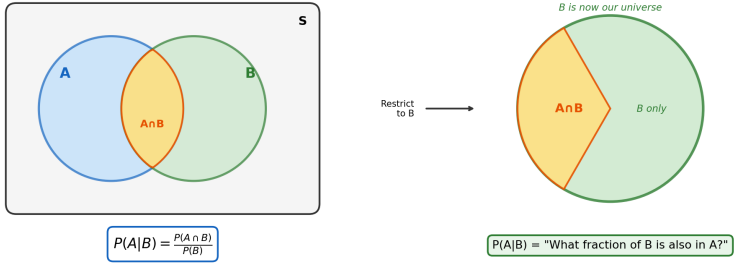
Definition: P($A|B$) = "the probability of A given that B has occurred"

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided } P(B) > 0$$

Intuition: We restrict our universe to only the cases where B happened, then ask how often A also happens within that restricted universe.

Symbol Definitions:

Symbol	Meaning
$P(A\|B)$	Probability of A given B (read: "A given B")
$P(A \cap B)$	Probability of both A and B occurring (the intersection)
$P(B)$	Probability of B (the conditioning event)



Example: In a deck of 52 cards, P(King | Face card) = 4/12 = 1/3. We restrict to face cards (12 total: 4 kings, 4 queens, 4 jacks), then count kings (4).

The Conditional Probability Trap

⚠ **$P(A|B) \neq P(B|A)$ — These are fundamentally different questions!**

Example that illustrates the trap:

P(wet ground | rain) is HIGH

If it rained, the ground is almost certainly wet.

P(rain | wet ground) is LOWER

Wet ground could be from sprinklers, dew, spills...

Why this matters for AI supervision:

- AI might compute the right conditional probability perfectly.
- But if you asked the wrong question, you get a correct answer to the wrong problem.
- Medical tests are the classic case:

P(disease | positive test) ≠ P(positive test | disease)

This asymmetry is the source of enormous confusion in real-world decision making.

The Law of Total Probability

Setup: If events B_1, B_2, ..., B_n partition the sample space (mutually exclusive and exhaustive):

$$P(A) = \sum_i P(A|B_i) \cdot P(B_i)$$

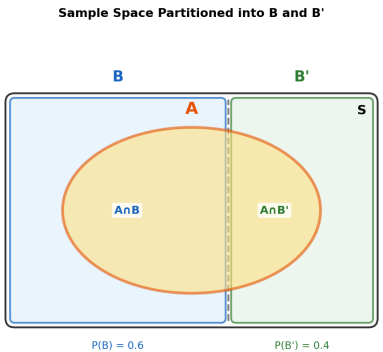
Intuition: To find the total probability of A, sum up the probability of A in each "bucket" weighted by how likely each bucket is.

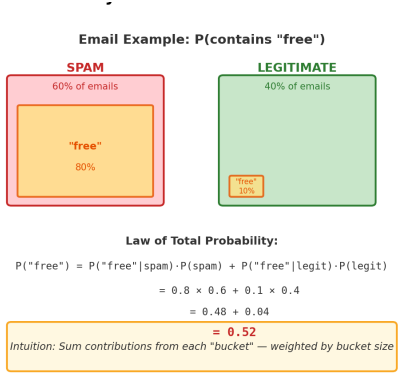
Simple case (two buckets):

$$P(A) = P(A|B) \cdot P(B) + P(A|B') \cdot P(B')$$

Example: Email classification

- 60% of emails are spam, 40% are legitimate
- P(contains "free" | spam) = 0.8,
P(contains "free" | legitimate) = 0.1
- P(contains "free") = 0.8 × 0.6 + 0.1 × 0.4 = 0.48 + 0.04 = **0.52**





This becomes essential for Bayes' rule — the denominator P(E) is computed this way.

Bayes' Rule — Updating Beliefs with Evidence

The question Bayes answers: Given that we observed evidence E, how should we update our belief about hypothesis H?

$$P(H|E) = P(E|H) \cdot P(H) / P(E)$$

Each component named:

Term	Name	Meaning
$P(H\|E)$	**Posterior**	Updated belief after seeing evidence
$P(E\|H)$	**Likelihood**	How likely is this evidence if H is true?
$P(H)$	**Prior**	Initial belief before seeing evidence
$P(E)$	**Evidence**	Total probability of seeing this evidence

> **Key Insight:** The prior $P(H)$ — the base rate — has enormous influence on the posterior. A rare hypothesis stays unlikely even with seemingly strong evidence.

Bayes' Rule — The Counting Perspective

Remember our counting approach to probability? Bayes becomes intuitive when you count.

Setup: Imagine 1000 people.
Some have a disease (D), some don't (D'). A test gives positive (T^+) or negative (T^-) results.
The question: Of all positive tests, how many are true positives?

The counting approach:

Step 1: Start with base rate — How many of 1000 have the disease?
Step 2: Apply sensitivity — Of those with disease, how many test positive?
Step 3: Apply false positive rate — Of those without disease, how many test positive?
Step 4: Count — True positives / All positives = $P(D|T^+)$

This natural frequency approach makes Bayes intuitive — no formula memorization needed.
Just count people in each category.

CTA Demo B: The Medical Test Problem

🤖 CTA Demo: The Medical Test Problem

What We'll Explore: How base rates dramatically affect the meaning of a positive test

Pattern: Verification Through Computation

The Setup:

- Disease prevalence: **1%** (1 in 100 people have it)
- Test sensitivity: **99%** (if you have it, 99% chance of testing positive)
- False positive rate: **5%** (if you don't have it, 5% chance of testing positive)
- **Question: If you test positive, what's the probability you have the disease?**

Then: "Ask CTA to solve using the counting approach with 10,000 people."

The Medical Test: Bayes' Rule

E = positive test; H = disease present; H' = disease absent

Disease prevalence: 1% (1 in 100 people have it) $\longrightarrow \mathrm{P}(H) = 1\%,\ \mathrm{P}(H') = 99\%$

Test sensitivity: 99% (if you have it, 99% chance of testing positive) $\longrightarrow \mathrm{P}(E|H) = 99\%$

False positive rate: 5% (if you don't have it, 5% chance of testing positive) $\longrightarrow \mathrm{P}(E|H') = 5\%$

Question: If you test positive, what's the probability you have the disease? $\longrightarrow \mathrm{P}(H|E) = ?$

$$\mathrm{P}(E) = \mathrm{P}(E|H)\mathrm{P}(H) + \mathrm{P}(E|H')\mathrm{P}(H')$$
$$= 0.99 \times 0.01 + 0.05 \times 0.99 = 0.0594 = 5.94\%$$

$$\mathrm{P}(H|E) = \frac{\mathrm{P}(E|H)\mathrm{P}(H)}{\mathrm{P}(E)} = \frac{0.99 \times 0.01}{0.0594} = \frac{1}{6} \approx 16.67\%$$

Among people who test positive, about 1 in 6 has the disease.

The Base Rate Reveal

The calculation (counting with 10,000 people):

- 100 have disease (1% of 10,000)
- 9,900 don't have disease
- Of 100 with disease:

99 test positive

- Of 9,900 without:

495 test positive

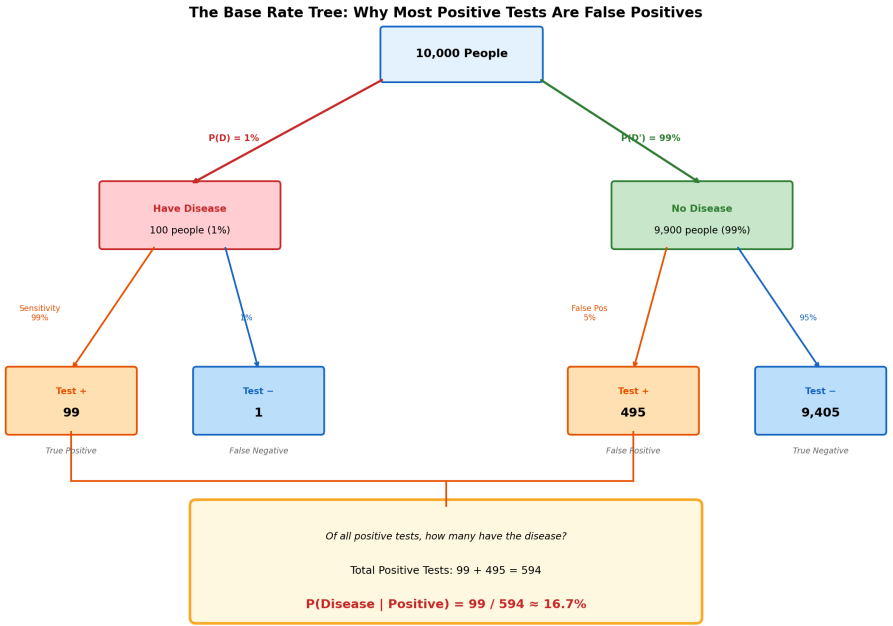
- Total positive tests:

99 + 495 = **594**

P(disease | positive)

= 99 / 594

≈ 16.7%



Key Insight: A 99% accurate test gives only ~17% confidence!
The low base rate (1%) means most positive tests are false positives.
There are simply so many more healthy people to generate false alarms.

Why Base Rates Are Counter-intuitive

The psychological trap:

We focus on the test's accuracy (99%) and ignore how rare the condition is.

The math:

When the base rate is low, even a small false positive rate generates many false positives because you're applying it to a huge population of healthy people.

Real-world implications:

Medical screening	Airport security	Fraud detection	Spam filters

⚠️ **When AI reports "high confidence" in a rare event, always ask: what's the base rate?**

AI computes P(evidence | hypothesis) correctly — but you need P(hypothesis | evidence).
These are not the same thing, and the base rate is what connects them.

From Point Estimates to Intervals

Quantifying Uncertainty

The problem with point estimates:
A sample mean gives us a single "best guess" — but different samples give different estimates. How confident should we be that our estimate is close to the truth?

Point Estimate

$\bar{x} = 170$ cm

"The average height is 170 cm"

But how precise is this?



Confidence Interval

[167 cm, 173 cm]

"We estimate the true mean is between 167 and 173 cm"

Key Insight: Intervals capture uncertainty. A narrow interval means we're fairly precise; a wide interval means there's a lot of uncertainty.
When AI gives you a point estimate, always ask: "What's the uncertainty around this?"

Constructing a Confidence Interval

The formula and its components

$$\text{CI} = \bar{x} \pm z^* \times \frac{s}{\sqrt{n}}$$

$$\text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin of Error}$$

Each component explained:

Symbol	Name	Meaning
$\bar{x}$	**Sample mean**	Center of the interval — our best estimate
z^*	**Critical value**	1.96 for 95% CI, 2.58 for 99% CI — sets confidence level
s	**Sample std dev**	Measures data variability — more spread ⟶ wider CI
n	**Sample size**	Number of observations — larger n ⟶ narrower CI
$s/\sqrt{n}$	**Standard error**	How much $\bar{x}$ varies across samples — the key quantity

Larger n ⟶ narrower CI (more data = more precision) | Higher confidence ⟶ wider CI (casting a wider net)

What "95% Confidence" Really Means

The crucial distinction most people miss

The true parameter μ is FIXED. The interval VARIES.

μ (true population mean)

[Frequentist view] A specific number that exists in reality. It doesn't move around. We just don't know what it is.

The interval $[\bar{x} \pm z^* \cdot s/\sqrt{n}]$

Depends on which sample you happened to draw. Different sample ⟶ different interval.

μ = true population parameter → **FIXED (not random)** $[\bar{x} - z^* \frac{s}{\sqrt{n}}, \bar{x} + z^* \frac{s}{\sqrt{n}}]$ → **RANDOM (depends on sample)**

The correct interpretation:

If we repeated the sampling procedure many times and constructed an interval each time, about 95% of those intervals would contain the true parameter μ.

Analogy: A machine makes nets calibrated to catch a fish 95% of the time. Once you have a specific net, it either catches the fish or it doesn't — the fish doesn't have a "95% probability" of being caught.

The Common Misconception

And why it matters

WRONG: "There's a 95% probability that the true mean is in this interval."

$$\text{Either } \mu \in [167, 173] \text{ or } \mu \notin [167, 173] \text{ — no probability applies}$$

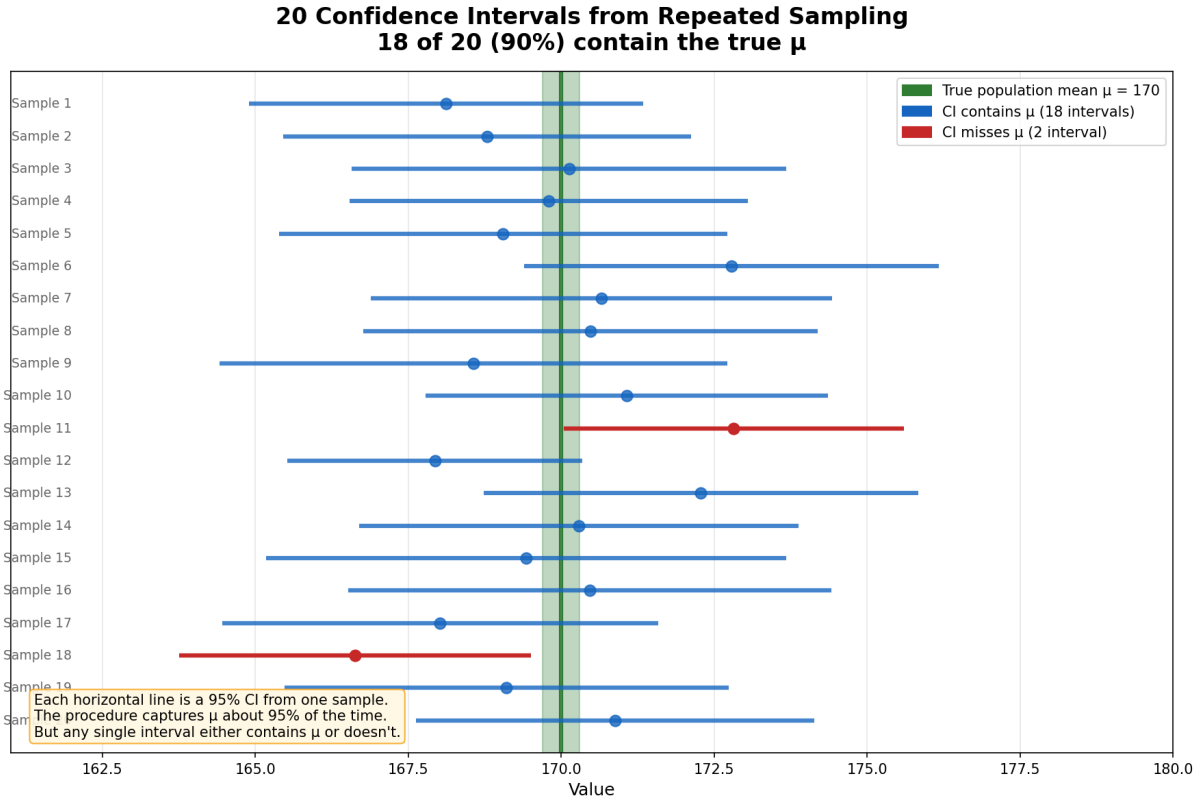
Why it's wrong:

Probability describes uncertainty about random things. Once you've computed a specific interval from your data, there's nothing random left. The true mean μ is either inside your interval or it isn't. You just don't know which.

Why it matters:

The misconception leads to overconfidence. If your sample happened to be unusual, your interval might be one of the 5% that misses entirely — and you'd have no way of knowing. The correct interpretation keeps you appropriately humble.

The Common Misconception



CTA Demo C:
What Confidence Intervals Really Mean

> 🤖 **CTA Demo: What Confidence Intervals Really Mean**
>
> What We'll Explore: The correct interpretation of confidence intervals (it's not what most people think)
>
> Pattern: Iterative Refinement

Common misconception to test:

"There's a 95% probability that the true mean is in this interval." ← WRONG

Demo Prompt:

Turn 1: "Explain what a 95% confidence interval means."

Turn 2: "Is it correct to say there's a 95% probability the true value is in the interval?"

Turn 3: "Demonstrate with a simulation showing many intervals from repeated sampling."

Interpreting AI-Generated Intervals

When AI gives you a confidence interval, ask:

1. **What assumptions were made?** (Normality? Independence? Sufficient sample size?)
2. **Is the interval for a mean, proportion, or something else?** (Different formulas apply)
3. **What's the practical significance of the width?** (A narrow interval isn't useful if the range still spans "good" to "bad")

Example prompt for clarity:

> *"Please provide a 95% confidence interval for the mean. Explain what assumptions you're making and whether they're reasonable for this data."*

✓ VERIFY: Inspect

Examine intermediate results and assumptions. Don't just accept the interval; check that the underlying assumptions are met.

A narrow interval from a biased sample is worse than a wide interval from a representative sample.

Making Decisions with Data

How do we use data to make decisions about claims?

Understand hypothesis testing

The logic of "innocent until proven guilty"

Interpret *p*-values correctly

What they mean and don't mean

These skills matter because hypothesis testing is one of the most misused statistical tools, and AI will compute *p*-values without questioning whether they answer your actual question.

The Logic of Hypothesis Testing

Analogy: "Innocent until proven guilty" — We assume the boring explanation is true, then ask: how surprising is our data?

Key Terms:

Term	Symbol	Meaning
Null hypothesis	H_0	The default assumption (no effect, no difference)
Alternative hypothesis	H_1 or H_a	What we're trying to find evidence for
Test statistic	T	A number from the data measuring departure from H_0
P-value	p	P(data this extreme or more | H_0 is true)
Significance level	α	Threshold for "surprising enough" (typically 0.05)

The Logic: If data is very surprising under H_0 (small *p*-value), we have evidence against H_0. If $p < \alpha$, we reject H_0. Otherwise, we fail to reject (not the same as accepting H_0!).

$$\text{If } p < \alpha, \text{ reject } H_0 \quad \text{(evidence against null)} \qquad \text{If } p \geq \alpha, \text{ fail to reject } H_0 \quad \text{(insufficient evidence)}$$

The Hypothesis Testing Procedure

Step 1: State hypotheses — Define H_0 (null) and H_1 (alternative)

Step 2: Choose significance level α — Typically 0.05 (willing to be wrong 5% of time)

Step 3: Collect data and compute test statistic — Measure departure from H_0

Step 4: Calculate p-value — Probability of data this extreme or more extreme if H_0 is true

Step 5: Decision — If $p < \alpha$, reject H_0; otherwise, fail to reject H_0

Example: Testing whether a coin is fair

- H_0: $p = 0.5$ (coin is fair)
- H_1: $p \neq 0.5$ (coin is biased)
- Data: Flip 100 times, observe 65 heads
- **Question: Is 65/100 surprising enough under "fair coin" to reject H_0?**

What the *p*-Value Actually Means

p-value = P(observing data this extreme or more extreme | H_0 is true)

The p-value is a probability about the DATA, not about the hypothesis.

What it IS:

- A measure of how surprising our data would be if H_0 were true
- Smaller p = more surprising = stronger evidence against H_0

What it is NOT:

- NOT P(H_0 is true)
- NOT P(H_1 is false)
- NOT P(you made an error)

⚠ A p-value of 0.03 does NOT mean "3% chance the null is true." It means "if the null were true, there's a 3% chance of seeing data this extreme."

CTA Demo D: *P*-Value Misconceptions

🤖 CTA Demo: *P*-Value Misconceptions

What We'll Explore: Testing common misconceptions about what *p*-values mean

Pattern: Challenging for Completeness

Demo Prompt:

Turn 1: "If I get a p-value of 0.03, what does that mean?"

Turn 2: "Does p = 0.03 mean there's a 3% chance the null hypothesis is true?"

Turn 3: "If I replicate the study, what's the probability I'll get a significant result again?"

Note: CTA should correctly identify these as misconceptions.
Watch for subtle errors in phrasing — this is where AI can still mislead.

The *P*-Value Misconception Reveal

Misconception 1: "$p = 0.03$ means 3% chance H_0 is true"
Reality: *P*-value tells you nothing about P(H_0 is true). It's P(data | H_0), not P(H_0 | data).
This confusion swaps the conditional!

Misconception 2: "$p < 0.05$ means the effect is important"
Reality: Statistical significance ≠ practical significance.
A tiny, meaningless effect can be statistically significant with large enough *n*.

Misconception 3: "$p = 0.03$ means 97% chance of replication"
Reality: Replication probability depends on effect size, sample size, and power — not the original *p*-value.
Many "significant" findings fail to replicate.

AI will compute *p*-values correctly.
The danger is in misinterpreting what they mean — and AI explanations can still contain subtle errors.

Statistical vs. Practical Significance

Key distinction:

Statistical significance	**Practical significance**
$p < \alpha$ (unlikely under H_0)	Is the effect large enough to matter?

Example: Blood pressure drug study

- Study finds new drug lowers blood pressure by 0.5 mmHg
- With n = 100,000 patients, this is statistically significant ($p < 0.001$)
- **But 0.5 mmHg is clinically meaningless — normal variation is ±10 mmHg!**

Effect Size — Measuring Magnitude:

Cohen's $d = (\bar{x}_1 - \bar{x}_2) / s_{pooled}$

Small: $d \approx 0.2$ Medium: $d \approx 0.5$ Large: $d \approx 0.8$

Key Insight: Always ask for effect sizes alongside p-values. A significant result with a tiny effect size may not be worth acting on.

🤖 **CTA Learning exercise.**

Today's Key Takeaways

Five skills you built today:

1. **Always visualize** — Statistics can hide structure (Anscombe's Quartet)

2. **Know your distributions** — Match the generative story to the shape

3. **Remember base rates** — $P(H|E) \neq P(E|H)$, and priors matter enormously

4. **Interpret intervals correctly** — 95% is about the procedure, not the parameter

5. **Distinguish significance types** — Statistical ≠ practical; always ask for effect sizes

> **Unifying theme:** AI computes correctly; your job is to ask the right questions and interpret wisely.

Prompting Habits Summary

Habits we developed today:

1. *"Provide summary statistics AND appropriate visualizations."*
2. *"Explain what assumptions you're making and whether they're reasonable."*
3. *"Show me the calculation step by step using natural frequencies / counting."*
4. *"What is the effect size, not just the p-value?"*
5. *"Is this result practically significant in the context of [your domain]?"*

Meta-habit: *"Always ask CTA to explain its reasoning, not just give answers."*

Trust but VERIFY. AI is a powerful tool, not an oracle.

Form your course project group

3 people/group

Deadline Sep 23 at 4 pm

If your grouping info is not received by then,

the TAs will assign you to a group.

Homework #1 will be released later through emails.

TA contacts

Homework #1

Tuesday 4-5PM

Yushi Du:
u3012725@connect.hku.hk
https://us04web.zoom.us/j/78561597481

Zhengyang Hu:
u3010250@connect.hku.hk

Thursday 3-4PM

Jun Dai:
jundai332@gmail.com
https://rochester.zoom.us/j/97202452759

Lulu Cai:
u303011209@connect.hku.hk

TA office hours on Zoom! Two slots (1 hour each) on Tuesday and Thursday

Questions?

CONTACT INFORMATION



yangliu.mit.edu

Instructor: Prof. Yang Liu

Email: yliu12@hku.hk

Office Hours: 5:00 - 5:50 PM, Thursdays

Location: CB-714, Chow Yei Ching Building

Let me know if you want remote!

🤖 **CTA is your 24/7 interactive and responsive learning partner!**