Probability and Inference II

- The world is messy — every dataset is full of randomness and error.
- Our goal as data scientists:
  - Predict outcomes
  - Understand relationships between variables
- But we never see the “true” world directly — only noisy samples.
- Probability & inference give us tools to separate real signal from random noise.

Housekeeping

- Please provide code snippets for your work when submitting HW
- Midterm is October 20th
  - In class, written
  - Learning opportunity – I am not here to weed out anyone!

Why Inference Matters in Data Science

- Inference: “a conclusion reached on the basis of evidence and reasoning”
- EDA → Inference → Modeling → Communication.
- Inference = bridge between exploration and modeling.

Two Roles of Inference

- As an end in itself: answering scientific/business questions
  - Don’t be afraid to answer a question with an appropriate statistical method rather than going to something complicated
- As a filter: screening features before modeling.
- Feasibility, can the question even be answered?

Why Inference?

- Data are samples → we never see the full population.
- Patterns could be real or just random noise.
- Inference provides:
  - A language of uncertainty (confidence intervals).
  - A logic of evidence (permutation tests, p-values).

Two Roles of Inference

- End in itself → when the goal is knowledge:
  - "Did this drug lower blood pressure?"
  - "Do users in Region A churn more than in Region B?"
- Filter for features → when the goal is modeling:
  - Screen out noisy or irrelevant predictors.
  - Prioritize features with real signal.
- Both roles matter in DS workflows.

Contrast with Machine Learning

- Inference
  - Answers: How sure are we?
  - Focus: uncertainty, causal/associational claims.
  - Example: “Is screen time associated with lower sleep quality?”
- Machine Learning
  - Answers: How well can we predict?
  - Focus: accuracy, optimization.
  - Example: “Can we predict tomorrow’s churn probability?”
- DS needs both perspectives.

Real-World Examples

- Drug trial (inference end goal):
  - Estimate treatment effect + CI.
  - Report uncertainty as evidence.
- Customer churn (inference + ML):
  - Test which features matter → inference filter.
  - Train predictive model → ML optimization.

Tukey's Lens: Explore vs. Confirm

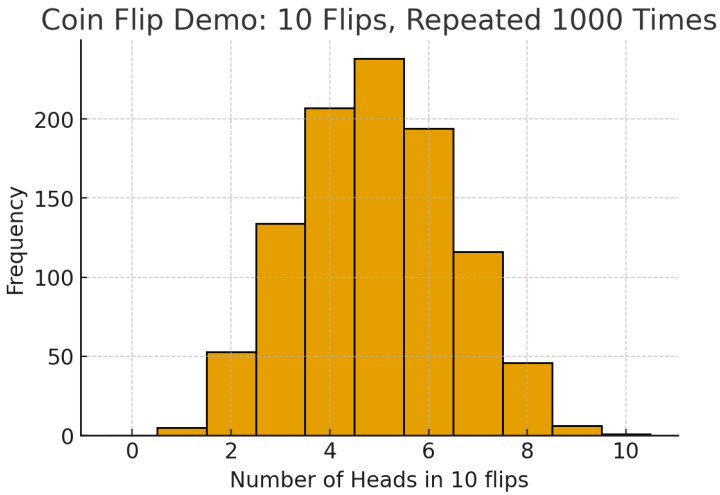
- Exploratory Data Analysis (EDA):
  - Pattern-finding, visualization, hypothesis generation.
- Confirmatory / Inference:
  - Quantify uncertainty, test whether patterns are real.
- Data Science = intersection
  - Explore → Test → Model

Why Samples Differ

- Data scientists rarely see the whole population.
- Two different samples from the same population can give different answers.
- Example:
  - Poll A: 52% candidate support.
  - Poll B: 48% candidate support.
- Who's right?

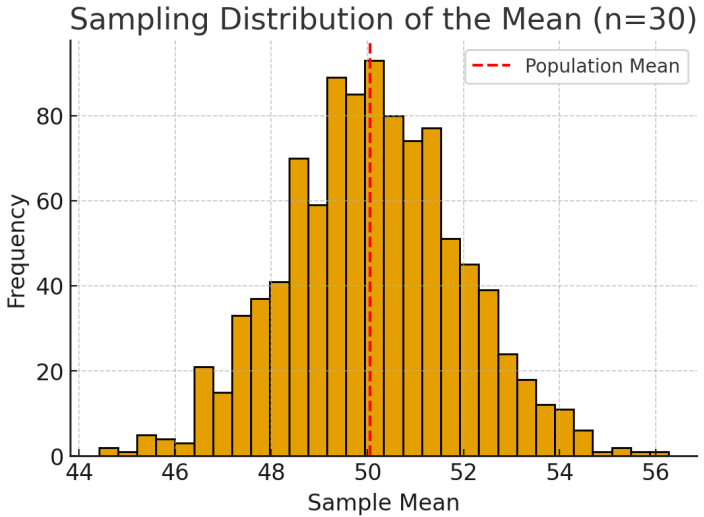
Coin Flip Demo

- Flip a fair coin 10 times: you may get 7 heads.
- Another 10 flips: maybe 4 heads.
- Even though the true p = 0.5, samples vary.



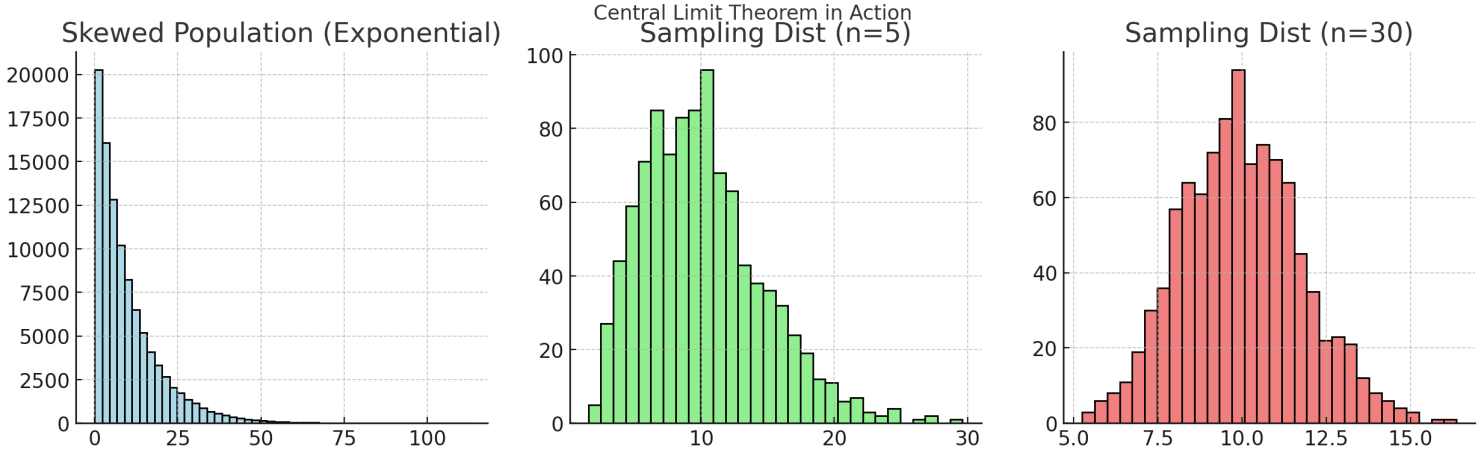
Sampling Distribution Concept

- Imagine repeating sampling many times.
- Plot all the sample means → distribution of estimates.
- Centered at the true population mean.



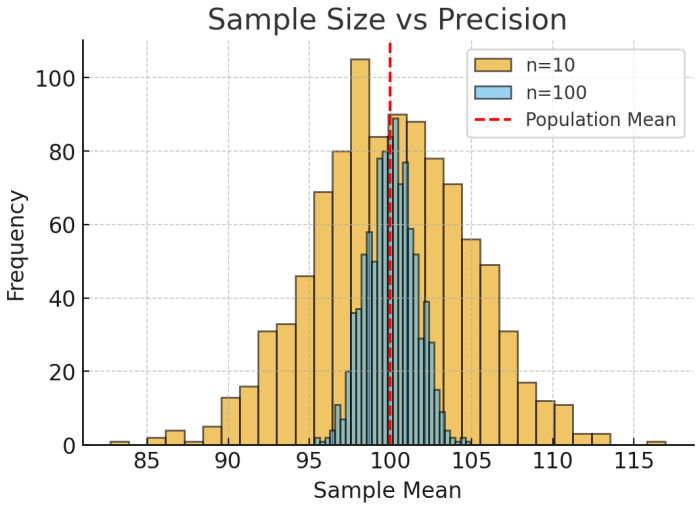
Central Limit Theorem

- As sample size ↑:
  - Sampling distribution gets narrower.
  - Shape approaches normal, even if population not normal.
- The law that makes inference possible.



Bigger Samples = More Precision

- n = 10 → wide variability.
- n = 100 → narrower variability.



Example: Polling Margins of Error

- Media report: “Candidate leads by 2% ± 3% margin of error.”
- That “±3%” comes directly from sampling variability.
- Polls differ because each sample is a different slice of the population.

$$\text{MOE} = z \times SE = 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

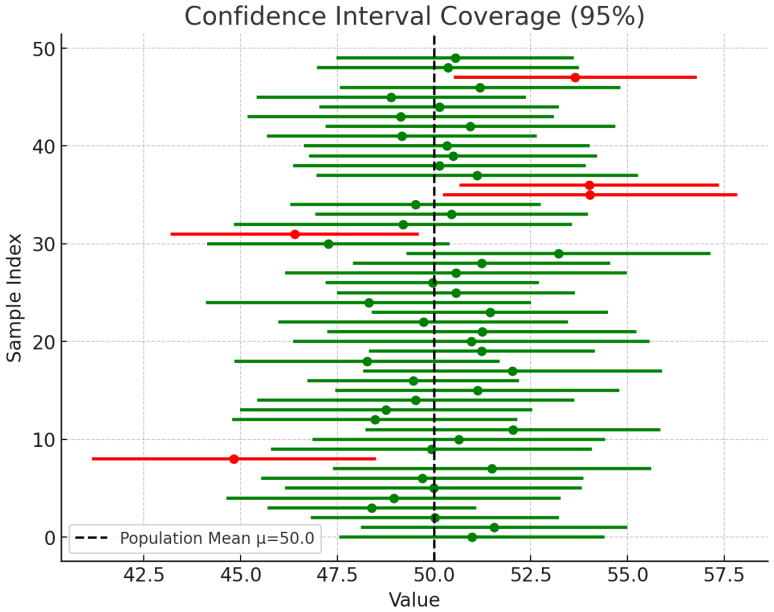
Transition to Confidence Intervals

- We now understand:
  - Samples vary.
  - Larger n = more stability.
- If as a data scientist you do not have enough data do not take the project!
- Next: how do we quantify the uncertainty in one sample estimate?
- Answer: Confidence Intervals.

What is a Confidence Interval?

- A range of plausible values for a population parameter.
- Based on sampling variability.
- Example: “Average sleep = 6.8 hours (95% CI: 6.3 – 7.3).”
- Key message: Don’t trust just a point estimate.

Many Intervals



Misinterpretations (What CI is NOT)

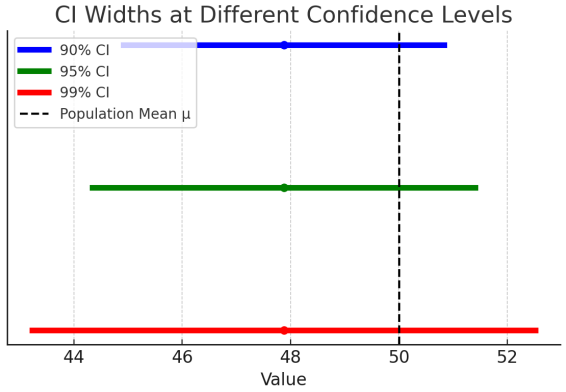
- NOT: 95% chance μ is in this interval.
- NOT: 95% of the data lie in this interval.
- IS: Our method will cover μ in ~95% of repeated samples.

Anatomy of a CI

Estimate $\pm$ (Critical Value $\times$ Standard Error)

$$\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Confidence	z*
90%	1.645
95%	1.960
99%	2.576



- **Sample size (n):** bigger n → narrower CI.
- **Variability (σ):** more variability → wider CI.
- **Confidence level (C):** higher confidence (99%) → wider interval.

CIs as a Data Readiness Test

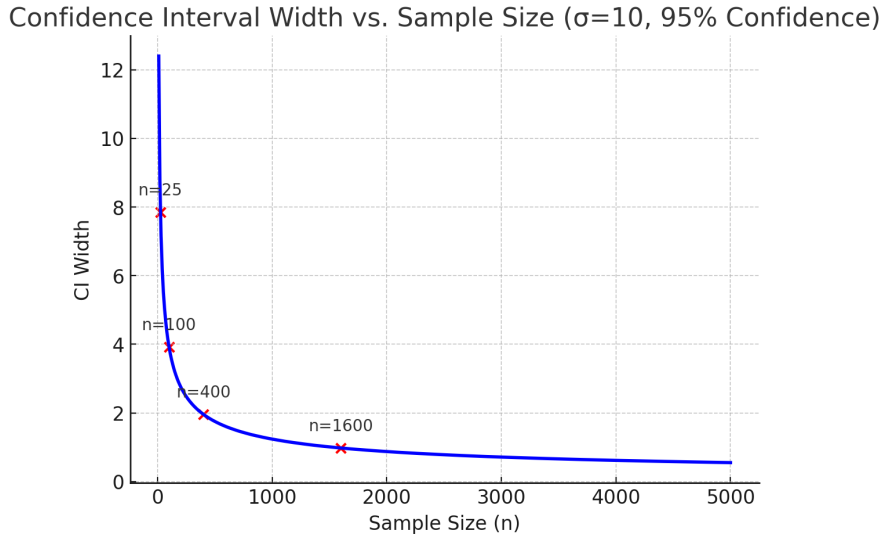
- We've seen: CIs measure uncertainty around estimates.
- Now: How can we use them to decide whether it's even worth building a machine learning model?
- Key idea:
  - If your estimates are already precise and stable, ML may not add much.
  - If they're noisy and uncertain, that's a signal that more modeling—or more data—might help.

What CI Width Tells You

- Narrow CI → Stable process
  - You already understand the pattern; not much left to “learn.”
  - ML model won’t improve much.
- Wide CI → Noisy or variable process
  - There’s still signal to uncover.
  - ML could help find complex structure or interactions.
- CI that includes 0 → No clear effect
  - Probably no predictive signal yet — collect more data or rethink the question.

CI Width for Data Sufficiency

- Formula: CI width ≈ $2z^* \frac{\sigma}{\sqrt{n}}$
- Inverse relationship with n:
  - Wider CI → too little data.
  - Narrow CI → enough data for stable estimation.
- Example logic:
  - CI width goal = ±2%.
  - If current CI = ±8%, need 16× more data fo that precision (inverse of sqrt(n)).
- So: CIs are a quantitative way to check data readiness.



Use CI for Predictive Signal Before ML

- Instead of jumping straight to a model:
  - Estimate simple relationships (correlations, mean differences).
  - Look at confidence intervals around those estimates.
  - Ask: Do any exclude 0? Are they wide or narrow?
- If most CIs include 0 → little stable signal → ML will overfit.
- If some are significant and tight → reliable signal → proceed to modeling.

Example: Go / No-Go Logic for ML

Example scenario:

You're evaluating whether to build a churn prediction model.

Feature	Estimated Effect	95% CI	Interpretation
Calls to support	+0.25	(0.10, 0.40)	Stable signal → predictive
Region	+0.03	(–0.15, +0.21)	Unclear → weak
Age	–0.01	(–0.05, +0.03)	No signal
Random noise feature	0.00	(–0.10, +0.10)	Overfitting risk

Takeaway: Focus only on variables with **tight, directional CIs**.

CI - Summary

- CIs aren't just for uncertainty in inference — they're a data-science diagnostic tool.
- They help answer:
  - Do we have enough data?
  - Is there stable signal to justify ML?

Permutation Test

- We've talked about uncertainty (CIs).
- But how do we test whether a difference we see could just be random?
- Enter permutation tests — the most intuitive way to test the "what if nothing's really going on?" question.

The Logic of a Permutation Test

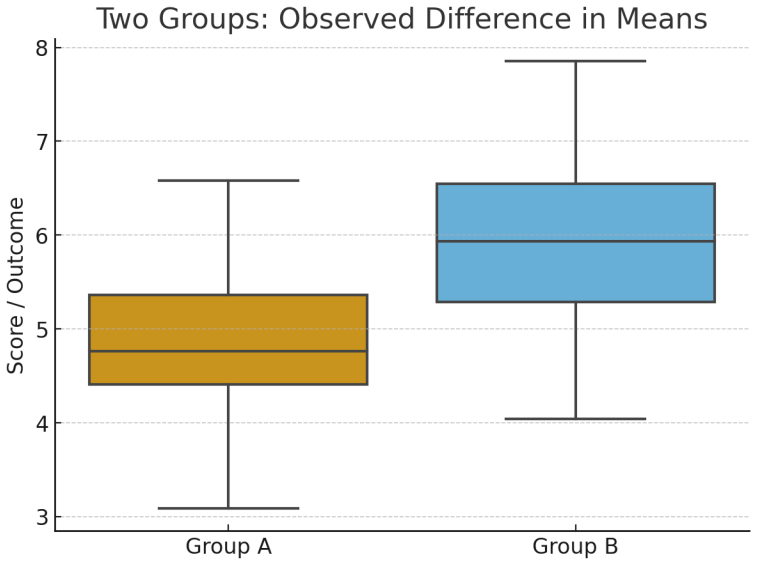
- Imagine two groups, A and B.
- You measure an outcome (e.g., sales, test scores, click rates).
- You see a difference — but could it be by chance?
- Idea:
  - If there's no real difference, group labels shouldn't matter.
  - So, shuffle the labels and see what kinds of differences you get.

Step-by-Step Procedure

- Compute the observed difference (mean_B – mean_A).
- Shuffle the group labels randomly.
- Recompute the difference.
- Repeat thousands of times to build a “null world.”
- See where the real observed difference lies in that null distribution.

Visual Example: Two Groups

- Show Group A and Group B with their means (e.g., 5.3 vs 6.1).
- Observed difference = 0.8.
- Question: “Is that real or random?”

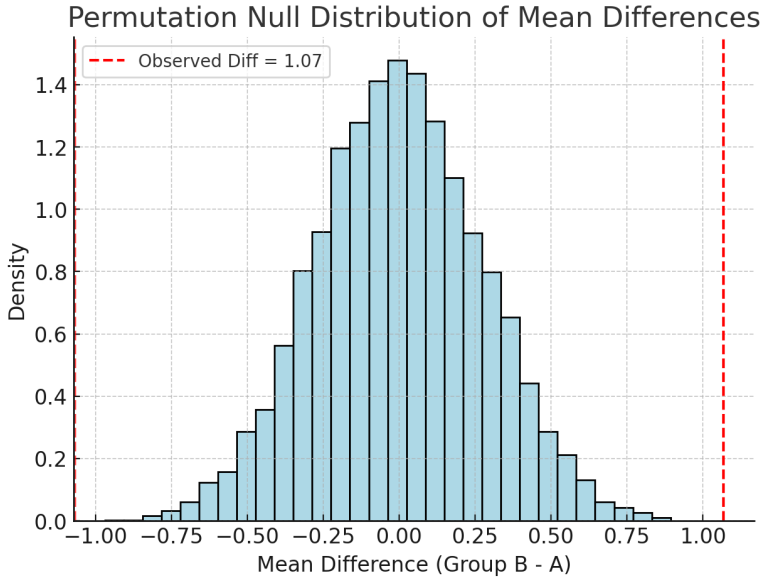


Shuffle Under the Null

- Now randomly reassign half the labels (simulate the world where there's no true difference).
- Compute the new difference.
- Repeat many times.
- Pseudo code: combine vectors A & B, pick a random split, compute mean difference, repeat

Build the Null Distribution

- Collect all shuffled differences → histogram.
- Centered around 0 (since under null, no difference).
- Mark the observed difference with a vertical line.

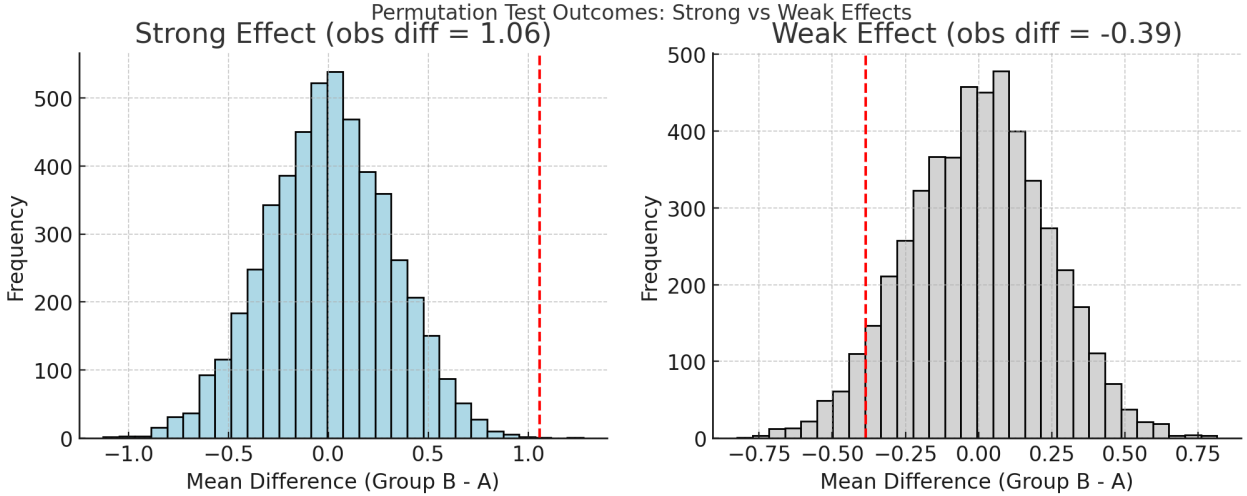


Estimating the p-value

- p-value = fraction of shuffled differences as extreme as the observed one.
- If only a few shuffled values are as extreme → unlikely under null → evidence of real effect.

$$p = \frac{\text{Number of shuffled differences } \geq |\text{observed difference}|}{\text{Total number of shuffles}}$$

Worked Example



Scenario	Observed Difference (mean_B − mean_A)	p-value	Interpretation
Strong effect	0.74	**0.54%**	Very unlikely under the null — strong evidence of a real difference.
Weak effect	0.47	**7.78%**	Could easily arise by chance — weak or inconclusive evidence.

Advantages of Permutation Tests

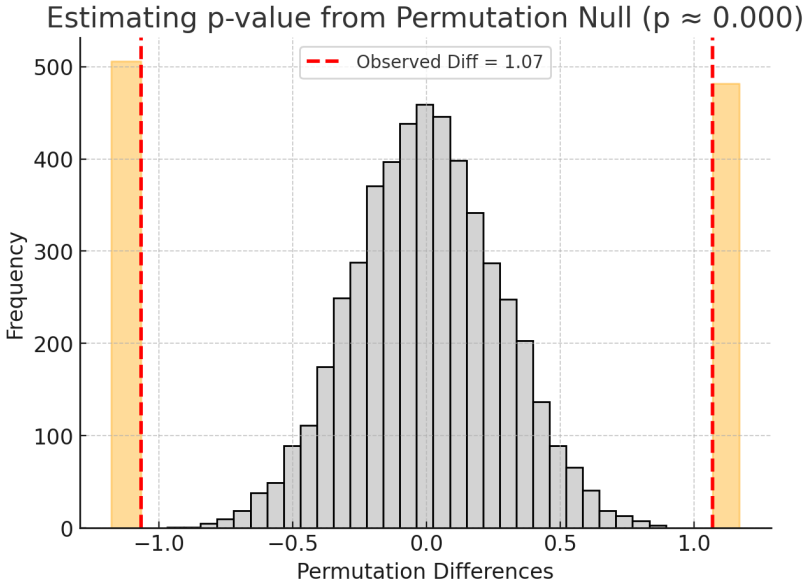
- No assumptions about distributions.
- Works with any test statistic (mean, median, correlation, etc.).
- Intuitive link to randomness.
- Can validate analytic tests (t-test ≈ permutation test under large n).

Limitations

- Computationally expensive for large datasets.
- Must preserve the structure of the data (exchangeable labels).
- Randomization must be valid (no dependencies).
- Doesn’t easily generalize to complex models without adaptation.

Connection to p-values

- Permutation tests build the null distribution.
- p-values summarize it:
  - Fraction of null-world outcomes as extreme as observed.
- So permutation tests explain where p-values come from.



p-values: Surprise Under the Null

- From permutation tests → we already computed the proportion of extreme shuffled results.
- That proportion is the p-value.
- Big idea:
  - The p-value tells us how surprising our data would be if the null hypothesis were true.

p-Value vs Permutation

$$p = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(|d_i| \geq |d_{\text{obs}}|)$$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$p = P(|Z| \geq |z_{\text{obs}}|) = 2(1 - \Phi(|z_{\text{obs}}|))$$

What a p-value Means

- p-value = measure of data–model compatibility.
- Small p → data are less compatible with the null.
- Large p → data are plausible under the null.
- **Not** a probability the null is true.
- **Not** evidence strength on a graded scale.

Link to Confidence Intervals

- CI and p-value are two views of the same logic.
- If $0 \notin$ 95 % CI $\rightarrow$ $p < 0.05$.
- If $0 \in$ 95 % CI $\rightarrow$ $p > 0.05$.
- CI shows range of plausible values; p compresses it to a single number.

Avoiding Threshold Thinking

- $p < 0.05$ is a convention, not a magic boundary.
  - $p = 0.049$ and $p = 0.051$ → essentially the same evidence.
- Always report: effect size + CI + p.
- Context matters:
  - Medicine → stricter (0.01)
  - Marketing → looser (0.10)
  - ML → combine with practical metrics.

Measuring How Much It Matters

- Small p-value → “something’s happening.”
- But we care: how much?
- Effect size = signal strength.
- In data science, this is what separates statistical detection from actionable insight.

What Is an Effect Size?

- Quantifies the magnitude of a relationship or difference.
- Independent of sample size.
- Types:
  - Mean difference
  - Correlation (r, R^2)
  - Standardized difference (Cohen's d)
  - Odds ratio or risk ratio

$$d = \frac{\bar{X}_2 - \bar{X}_1}{s_p}$$

d	Magnitude
0.2	Small
0.5	Medium
0.8	Large

Effect Size vs Sample Size

- p-value depends on n; effect size does not.
- Big n → small p even for trivial effects.
- Example:
  - n = 100,000 → $p < 0.001$ for 0.1% change
  - But $d \approx 0.01$ → meaningless in practice.

Effect Sizes as a Go/No-Go

- Before you build a model, check effect sizes:
  - Are any relationships large and stable?
  - Or are all near zero (no real structure)?
- Wide or near-zero CIs → ML will model noise.
- Stable, strong effects → real predictive opportunity.

Summary

- Classical stats: Is there an effect?
- Data science: Is it big enough to matter, predict, or act on?
- Effect size connects the two.

Concept	Question it answers	DS connection
p-value	"Is there signal beyond chance?"	Detects presence
Effect size	"How strong is the signal?"	Measures predictability
CI	"How stable is that estimate?"	Quantifies reliability