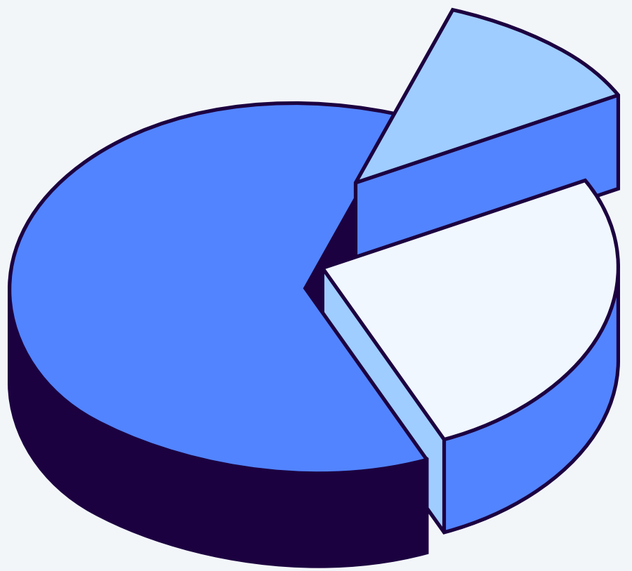
Probability & Statistics Review

لاتنسونا من دعواتكم

الله يوفقنا ويوفقكم وبإذن الله نتقابل في الصيفية 🤲🏻

لو عندكم اي سؤال او اقتراح
هذي حسابتنا في X

Muath Sara Ice



From Uncertainty to Insight

A Practical Journey Through Probability & Statistics for Machine Learning

This guide demystifies the core principles that allow us to quantify randomness, learn from data, and make confident decisions in the face of uncertainty.

What are the Fundamental Rules of Chance?

The Complement Rule

$$P(A^c) = 1 - P(A)$$

The probability that an event *does not* happen is 1 minus the probability that it *does*.

If the probability of a server failing is 2%, the probability of it not failing is 98%.

The Sum Rule (for events that can't happen together)

$$P(A \cup B) = P(A) + P(B)$$

The probability of either event A *or* event B occurring is the sum of their individual probabilities.

The probability of rolling a 1 or a 6 on a single die is `1/6 + 1/6 = 2/6`.

The General Sum Rule (Inclusion-Exclusion)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

To find the probability of A or B, we add their probabilities and subtract the overlap to avoid double-counting.

The probability of drawing a King or a Heart from a deck.

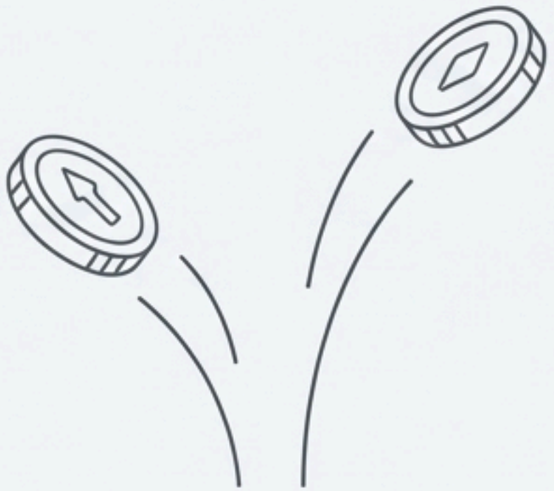
How Do Events Influence Each Other?

Independence

Two events are independent if the outcome of one has no effect on the outcome of the other.

Rule: `P(A ∩ B) = P(A) * P(B)`

Example: The probability of flipping two heads in a row is `0.5 * 0.5 = 0.25`. The first flip doesn't change the second.



Conditional Probability

The probability of event A happening, *given that* event B has already occurred.

Rule: `P(A|B) = P(A ∩ B) / P(B)`

Example: What is the probability of drawing a King, given you've already drawn a face card? There are 12 face cards, 4 of which are Kings. The probabiiity is 4/12.

How Can We Update Our Beliefs with New Evidence?

Bayes' Theorem provides a mathematical framework for updating our beliefs.

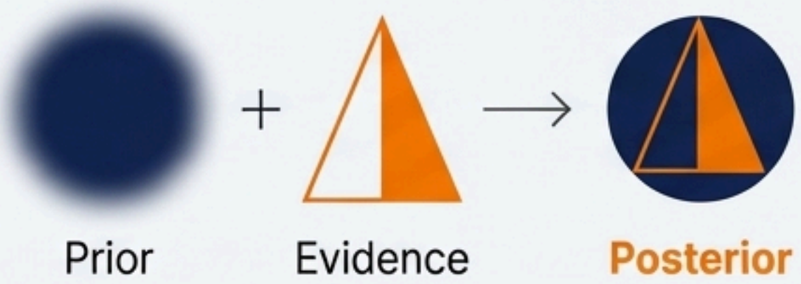
Posterior: What we want to know. Our updated belief about A after seeing evidence B.

Likelihood: How likely is the evidence B if our initial belief A is true?

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Prior: Our initial belief about A before seeing any evidence.

Evidence: The probability of observing the evidence.



Example: Medical Diagnosis

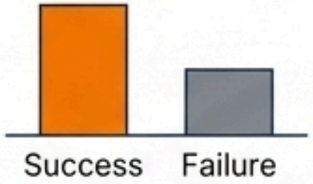
Imagine a test for a disease. We know the test's accuracy (Likelihood) and the disease's prevalence in the population (Prior). Bayes' Theorem lets us calculate the actual probability a person has the disease given a positive test result (Posterior).

What Shape Does Our Randomness Take?

A **Random Variable** is a variable whose value is the outcome of a random event.
We describe its behavior with a **Probability Distribution**.

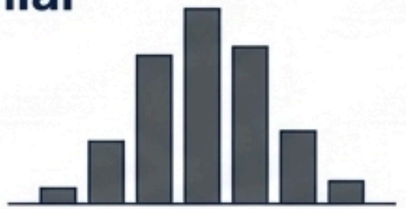
For Discrete Events (countable outcomes)

Bernoulli



A single success/failure trial. (e.g., one coin flip, one click on an ad).

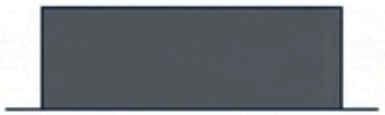
Binomial



The number of successes in a fixed n trials. (e.g., number of heads in 10 flips).

For Continuous Events (a range of outcomes)

Uniform



All outcomes in a range are equally likely. (e.g., a random number generator between 0 and 1).

Normal (Gaussian)

The classic “bell curve,” defined by its mean and standard deviation. (e.g., measurement errors, human height).

How Can We Summarize a Distribution's 'Center'?

Primary Measure: The Expected Value (μ)

The theoretical, long-run average of a random variable. It's a weighted average of all possible outcomes, weighted by their probabilities.

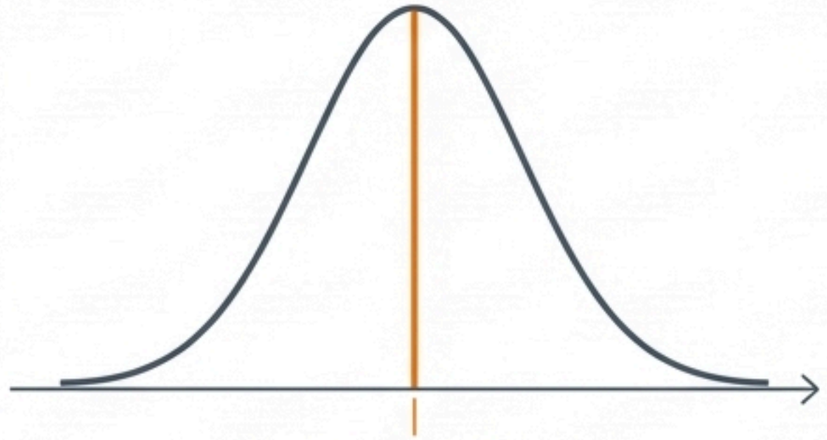
$$E[X] = \sum [x_i * P(X = x_i)]$$

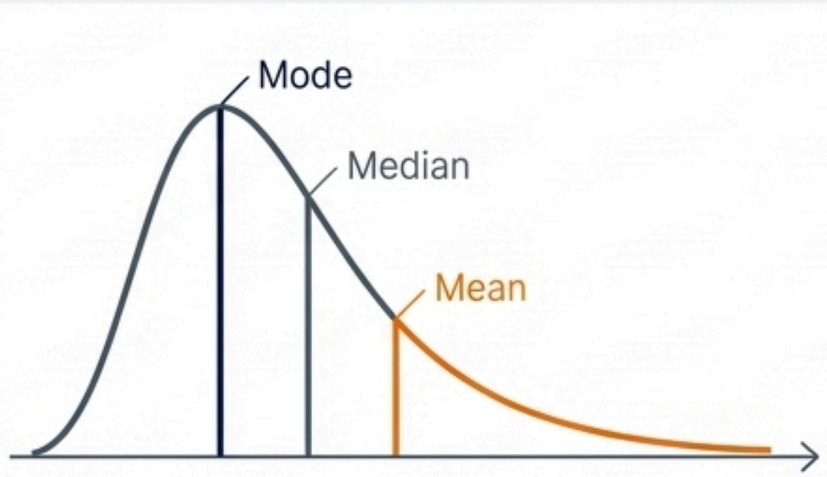
For a single roll of a fair six-sided die, the expected value is: (1*1/6) + (2*1/6) + (3*1/6) + (4*1/6) + (5*1/6) + (6*1/6) = 3.5.

Other Measures of Central Tendency

Median: The middle value (50th percentile) that splits the data in half. Less sensitive to outliers.

Mode: The most frequent value; the "peak" of the distribution.





In a skewed distribution, the mean is pulled towards the tail.

How ‘Spread Out’ Is Our Data?

Primary Measures: Variance and Standard Deviation measure the dispersion of data around the mean.

Rule 1: Variance (σ^2)

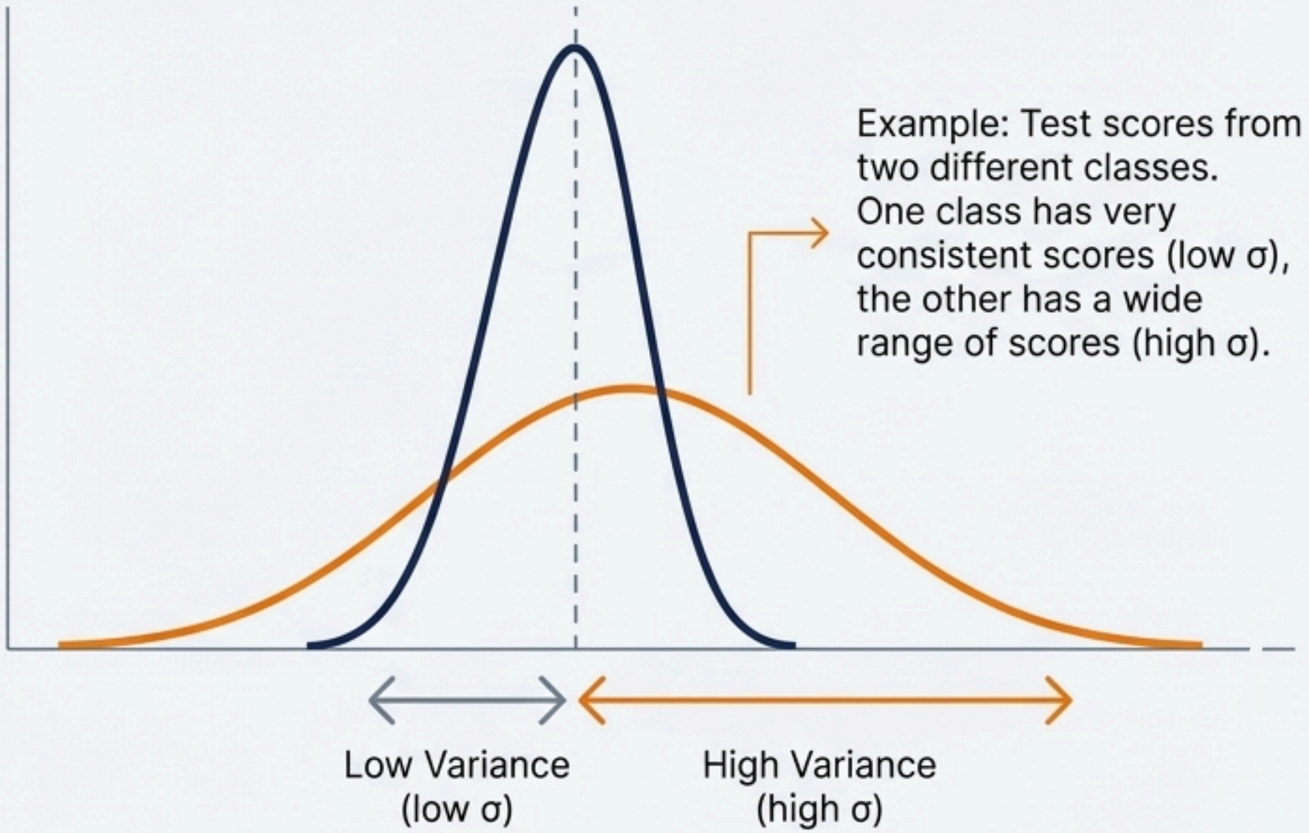
$$\mathrm{Var}(X) = E[(X - \mu)^2]$$

The expected (or average) squared difference of each point from the mean. A larger variance means the data is more spread out.

Rule 2: Standard Deviation (σ)

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

The square root of the variance. It’s more interpretable because it’s in the same units as the original data. It represents a “typical” deviation from the mean.



How Do Variables Move Together?

We often need to understand the relationship between two variables, not just one in isolation.

Measure 1: Covariance

Measures the *direction* of the linear relationship between two variables.

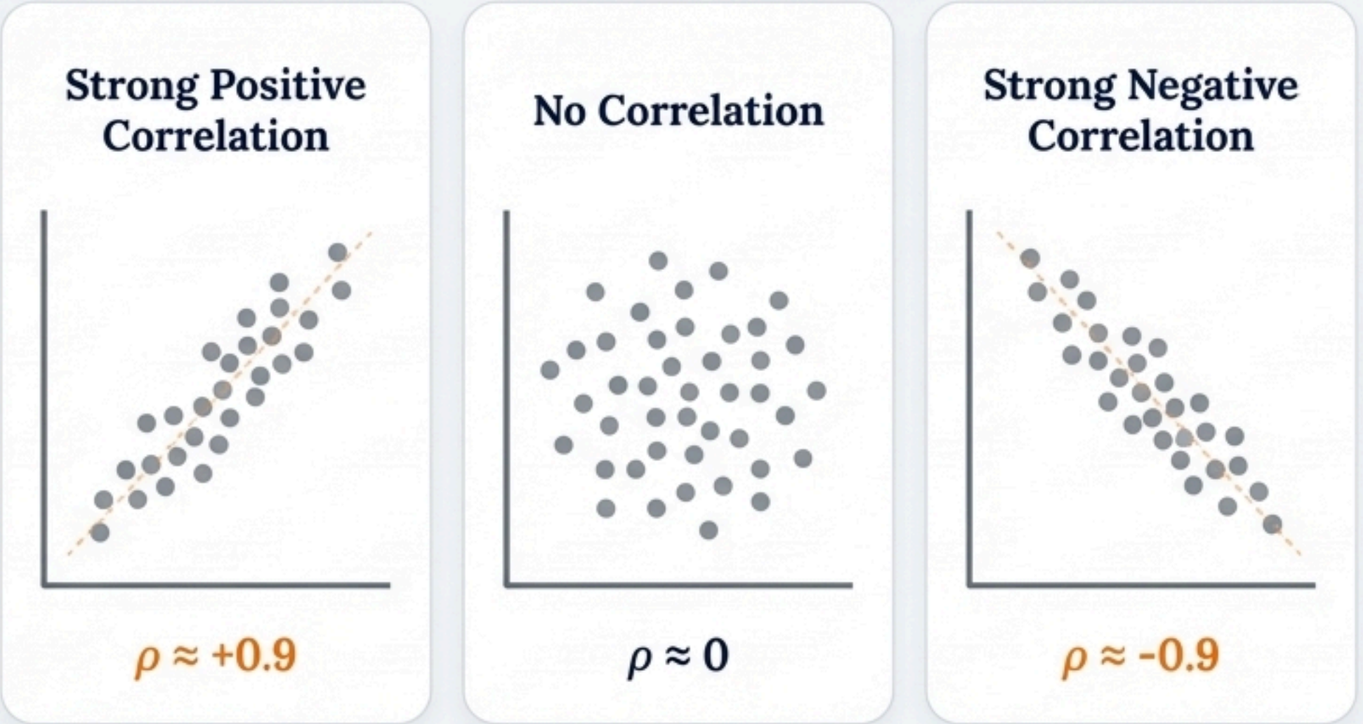
$Cov > 0$: As one variable increases, the other tends to increase.
$Cov < 0$: As one variable increases, the other tends to decrease.

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Measure 2: Correlation Coefficient (ρ)

A standardized version of covariance that is always between -1 and 1. It measures both the *strength* and *direction* of a linear relationship.

$$\rho = \frac{Cov(X, Y)}{\sigma_X * \sigma_Y}$$



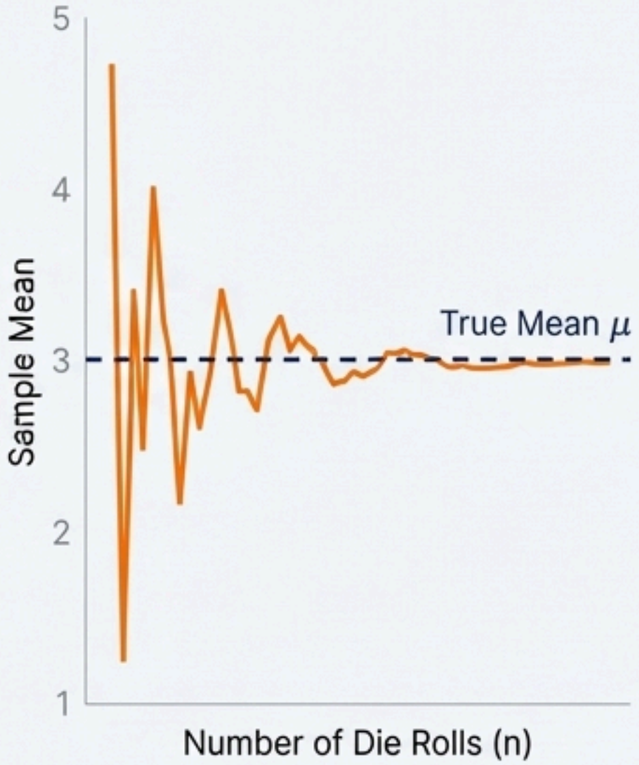
The Two Theorems That Make Inference Possible

Theorem 1: The Law of Large Numbers (LLN)

As you collect more and more data (as the sample size n grows), the sample mean $\bar{x}$ is guaranteed to converge to the true population mean μ.

Why It Matters

This is the principle that justifies data collection. More data leads to more accurate estimates.

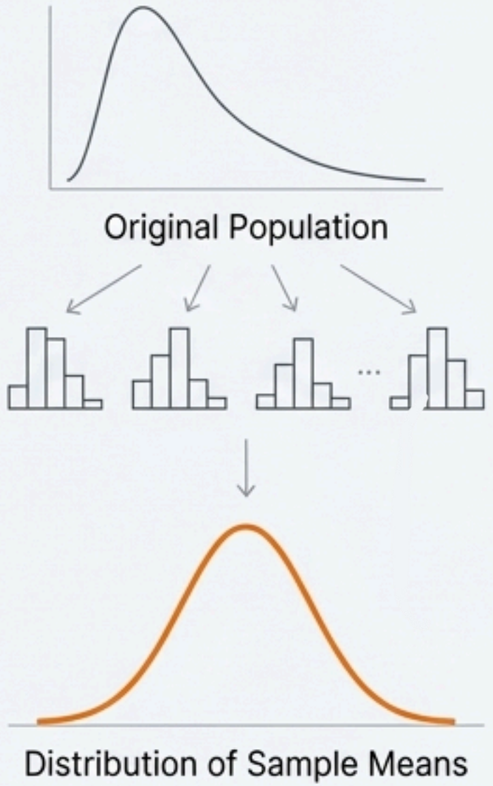


Theorem 2: The Central Limit Limit Theorem (CLT)

No matter the shape of the original population's distribution, the distribution of *sample means* will be approximately Normal (a bell curve) if the sample size is large enough.

Why It Matters

This is why the Normal distribution is so ubiquitous. It allows us to make inferences about the mean without needing to know the population's underlying distribution.



How Do We Estimate What We Can't See?

The Goal: Point Estimation

We use a **sample statistic** (e.g., sample mean $\bar{x}$) as our best guess for an unknown **population parameter** (e.g., population mean μ).

The Method: Maximum Likelihood Estimation (MLE)

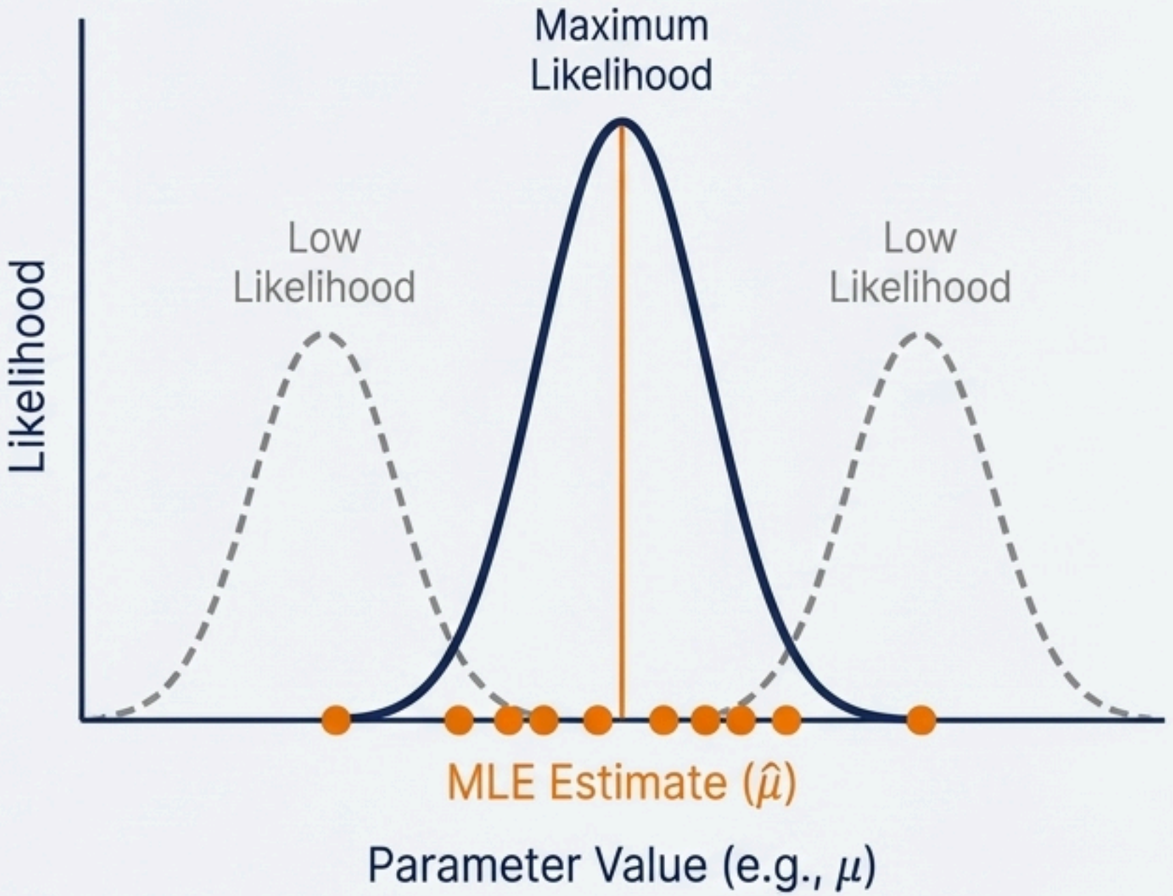
"Given the data we observed, what are the parameter values that would make this data the **most likely** to have occurred?"

How it Works: We construct a "Likelihood Function" and find the parameter values that maximize it.

MLE in Action

Bernoulli Example: For a series of coin flips with k heads in n trials, the MLE for the probability of heads (p) is simply the sample proportion: $\hat{p} = k/n$.

Linear Regression: Minimizing the Mean Squared Error (MSE) is equivalent to finding the Maximum Likelihood Estimate for a model's weights, assuming the errors are normally distributed.



The Link Between Bayesian Priors and ML Regularization

A Different Approach: Bayesian Inference

Rule:

`Posterior Belief ∝ Likelihood * Prior Belief`

From MLE to **MAP** (Maximum A Posteriori)

MLE Formula: `max( P(Data|θ) )`

MAP Formula: `max( P(Data|θ) * P(θ) )`

The Connection to Machine Learning

Regularization is a technique to prevent overfitting by adding a penalty to the cost function for large model weights.

The Insight: They are Mathematically Equivalent

`Minimize(Cost) = Minimize(-Log-Likelihood - Log-Prior)`

This is the Regularization Penalty

Key Equivalencies

L2 Regularization (Ridge)	is equivalent to MAP estimation with a **Gaussian Prior** (a belief that weights should be small and centered around zero). 
L1 Regularization (Lasso)	is equivalent to MAP estimation with a **Laplace Prior**. 

How Can We Express Our Uncertainty with a Range?

Concept: The Confidence Interval (CI)

Instead of a single "best guess" (a point estimate), a CI provides a range of values that is likely to contain the true population parameter.

The Anatomy of a CI

```
Point Estimate ± Margin of Error
```

The **Margin of Error** depends on:

1. The variability in the data (**standard error**).
2. How confident we want to be (the **critical value**).

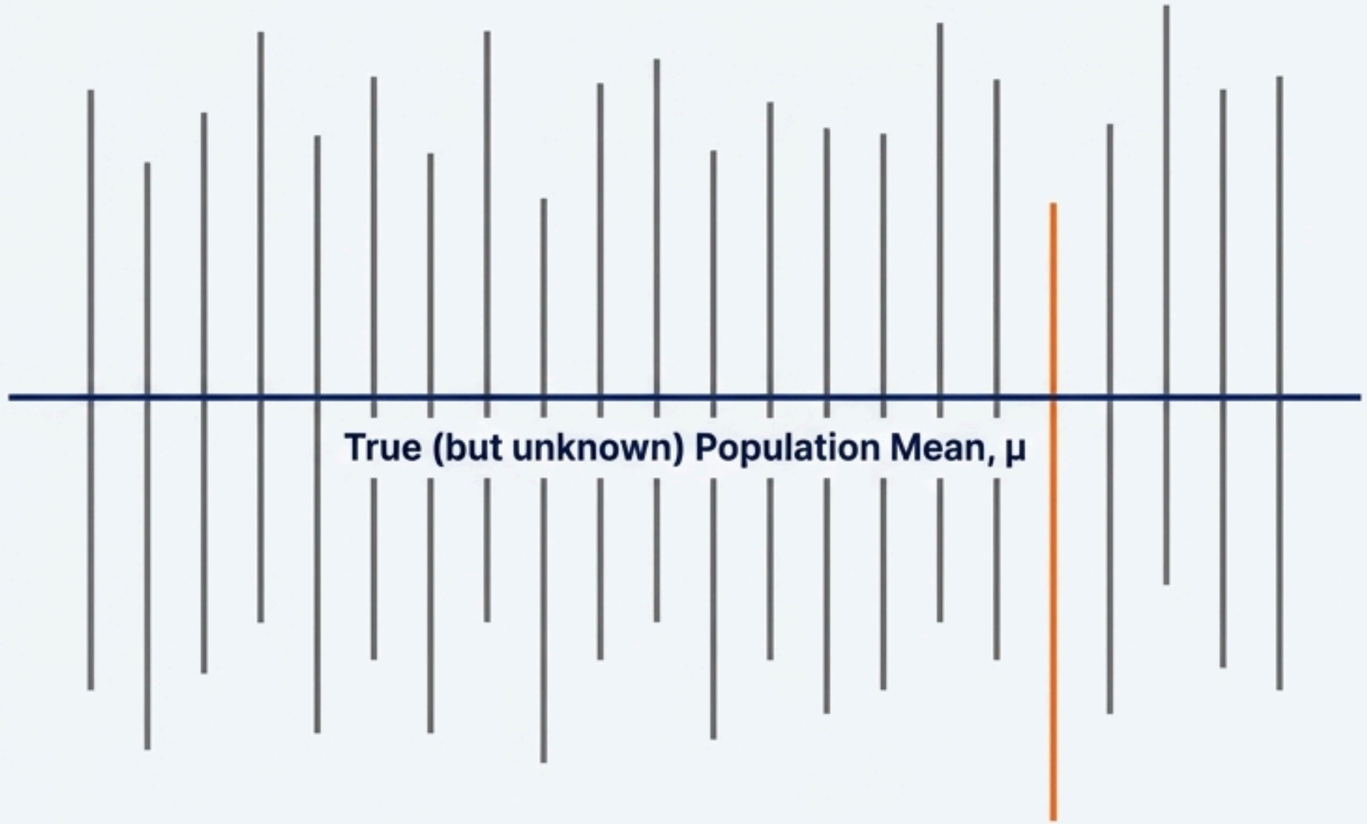
The Critical Interpretation

Incorrect View

A “95% confidence interval” does **NOT** mean there is a 95% probability the true mean is in our specific interval.

Correct View

It means: “If we were to repeat this sampling process 100 times, we expect that 95 of the confidence intervals we construct would capture the true, unknown parameter.” It’s a statement about the long-run reliability of the method.



A Formal Framework for Making Data-Driven Decisions

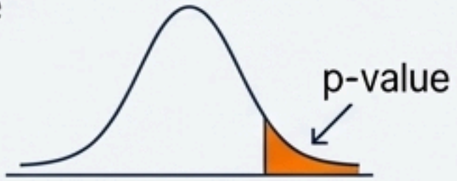
Hypothesis Testing is a procedure for deciding between two competing claims about a population.

Step 1: Define the Hypotheses

Null Hypothesis (H_0): The default assumption of 'no effect' or 'status quo." This is what we seek to find evidence against. (e.g., "The average user engagement time is 10 minutes.")

Alternative Hypothesis (H_a): The claim we want to test. (e.g., "The average user engagement time is *greater than* 10 minutes.")

Step 2: Weigh the Evidence with a p-value



p-value Definition: The probability of observing our sample result (or something more extreme), *assuming the null hypothesis is true*.

Interpretation:

- Small p-value: Our data is very surprising if H_0 is true. Strong evidence *against* H_0.
- Large p-value: Our data is not surprising. We have no reason to doubt H_0.

Step 3: Make a Decision

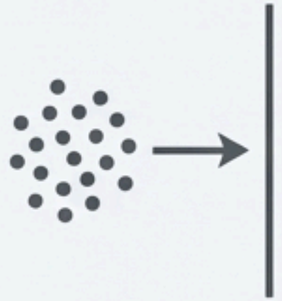
The Rule: We compare the p-value to a pre-set significance level (α, usually 0.05).

Outcome 1: If **p-value < α**, we **Reject the Null Hypothesis**.

Outcome 2: If **p-value ≥ α**, we **Fail to Reject the Null Hypothesis**.

Choosing the Right Statistical Test

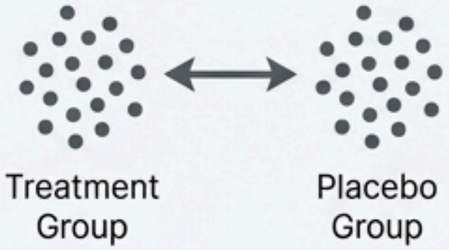
The Workhorse: The t-distribution. We use t-tests whenever the population standard deviation (σ) is unknown and we must estimate it with the sample standard deviation (s). This is nearly always the case in practice.



Comparing a group to a standard.

One-Sample t-Test

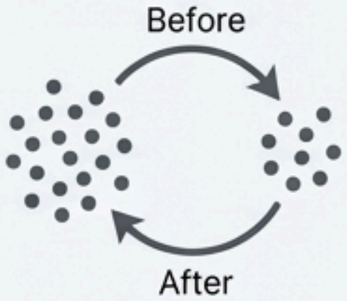
Is the average test score in this class **significantly** different from the national average of 80?



Comparing two independent groups.

Two-Sample t-Test

Does the treatment group have a **significantly** different average recovery time than the placebo group?



Comparing two related measurements.

Paired t-Test

Is there a **significant** change in weight for participants from *before* to *after* a diet program? (We test the average *difference*).

Insight in Action: The A/B Test

The Scenario

A company wants to know if a new website design (Version B) increases the user conversion rate compared to current design (Version A).

The Method

1. Randomly show Version A or B to users.
2. Collect data on conversion rates for each group.
3. Perform a **Two-Sample Test for Proportions**.
4. Calculate the p-value.

The Statistical Translation

We use hypothesis testing to determine if the observed difference is real or just due to random chance.

- **Null Hypothesis (H_0):** The conversion rates are the same. $p_A = p_B$. (The new design has no effect).
- **Alternative Hypothesis (H_a):** The new design is better. $p_B > p_A$.

The Decision & Insight

If the p-value is low (e.g., < 0.05), we reject H_0. We have statistical evidence that Version B is genuinely better.

The Outcome: We confidently launch the new design, knowing it's backed by data, not just intuition. This is the final step from uncertainty to a quantifiable, valuable insight.