Introduction to Probability for Machine Learning

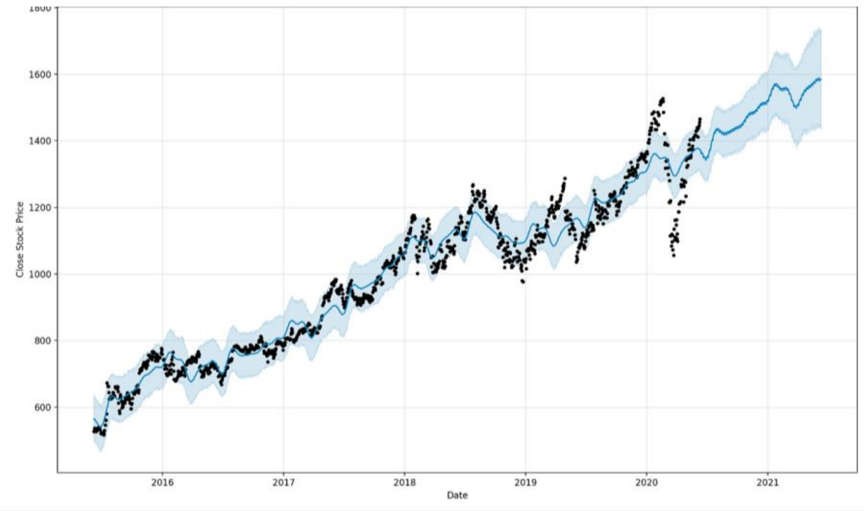
Aditya Arun

IIIT Hyderabad

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

Probability in Real Life



Why Probability?

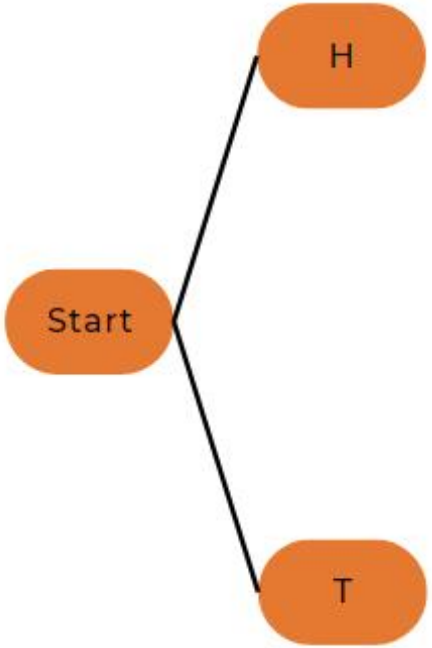
Uncertainty arises through

- Noisy measurements
- Finite size of datasets
- Ambiguity
- Limited Model Complexity
- …

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty

Data as Distribution

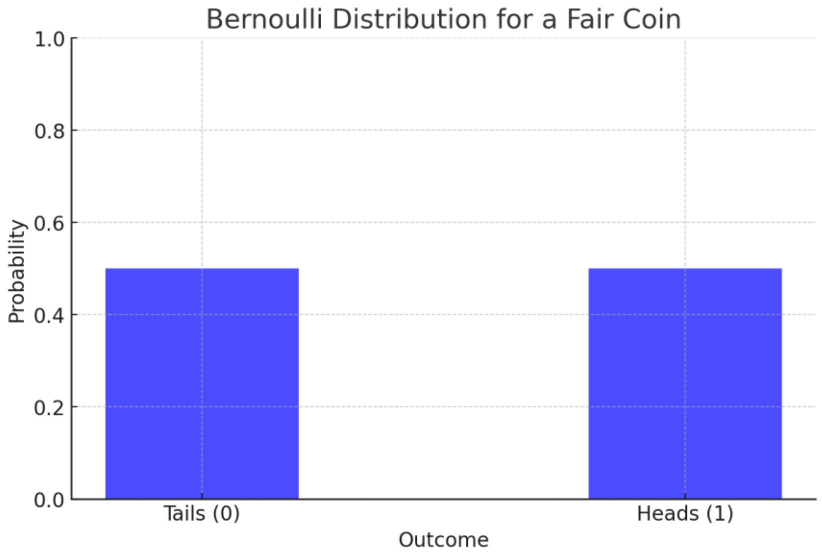
Coin Toss



Heads	Tails
1	0
0	1

- Sample Space $\Omega = \{H, T\}$
- For a fair coin

  $P(X = H) = P(X = T) = 0.5$

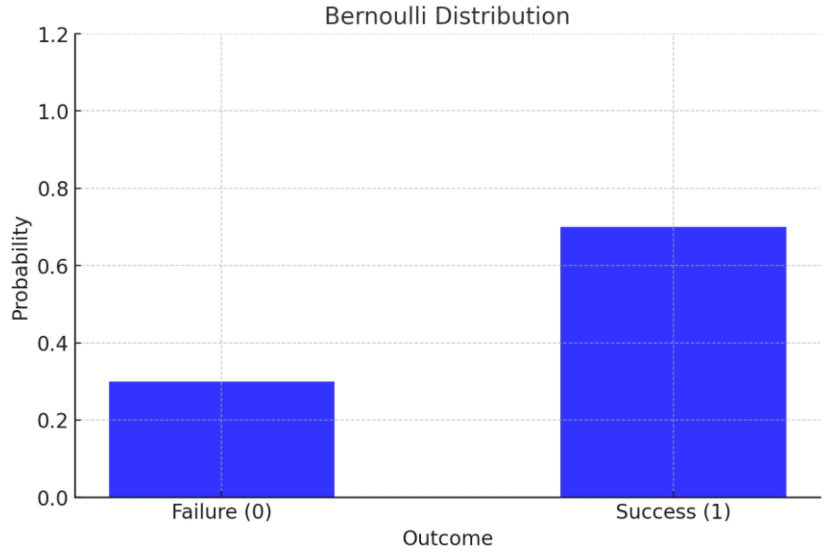


Bernoulli Distribution

- A discrete probability distribution representing a single trial with:
  - Two possible outcomes: Success (1) or Failure (0).
  - Probability of success: p
  - Probability of failure: $1 - p$

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

$$P(X = x) = p^x (1 - p)^{1-x}, \quad x \in \{0, 1\}$$

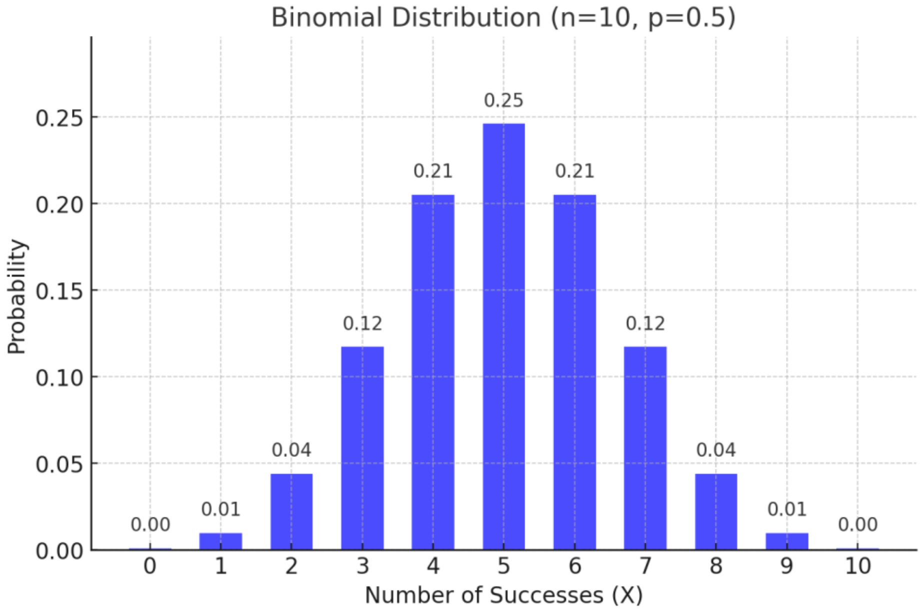


Binomial Distribution

- Extends the Bernoulli distribution to multiple independent trials.
- Focuses on the number of successes (k) in n trials.
  - Tossing a fair coin 10 times and counting the number of heads.

    Q. What is the sample space?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



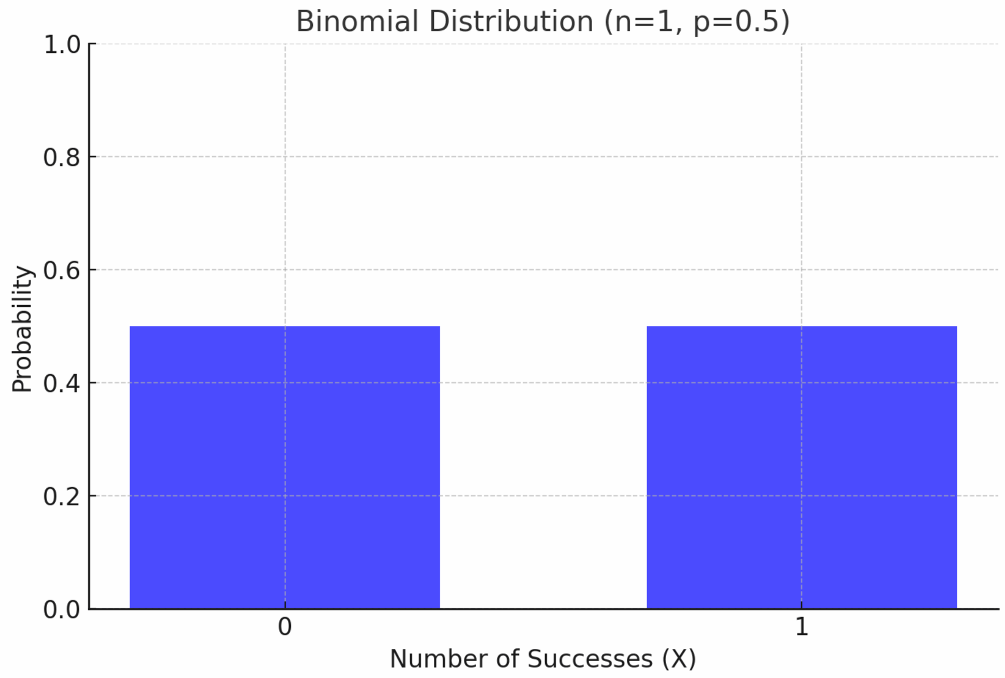
Binomial Distribution

```python
from scipy.stats import binom

# Parameters for the binomial distribution
n = 10  # Number of trials
p = 0.5  # Probability of success
k_values = range(0, n + 1)  # Possible values of k (number of successes)

# Compute the binomial probabilities for each k
pmf_values = [binom.pmf(k, n, p) for k in k_values]

# Display the probabilities
for k, pmf in zip(k_values, pmf_values):
    print(f"P(X = {k}) = {pmf:.4f}")
```



Problem 1: Binomial Distribution

- What is the probability of getting exactly 6 heads in 10 coin tosses if p = 0.5?

$$P(X = 6) = \binom{10}{6}(0.5)^6(1 - 0.5)^4$$

Step 1: Compute $\binom{10}{6}$:

$$\binom{10}{6} = \frac{10!}{6! \cdot (10 - 6)!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

Step 2: Compute p^k and $(1 - p)^{n-k}$:

$$(0.5)^6 = 0.015625, \quad (1 - 0.5)^4 = (0.5)^4 = 0.0625$$

Step 3: Multiply all components:

$$P(X = 6) = 210 \cdot 0.015625 \cdot 0.0625 = 0.205078125$$

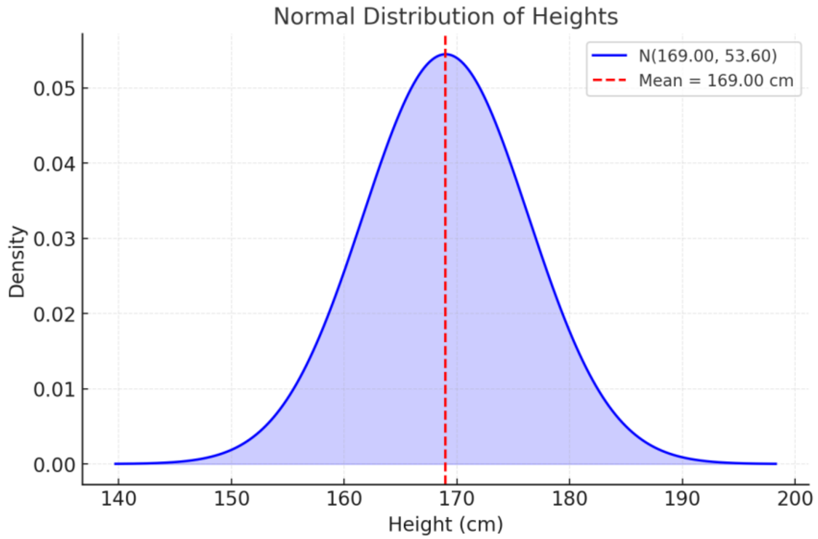
Answer:

$$P(X = 6) = 0.2051 \text{ (rounded to 4 decimal places)}.$$

Heights of individuals in a population

ID	Name	Height (cm)
1	Alice	165
2	Bob	172
3	Charlie	158
4	David	180
5	Eva	170

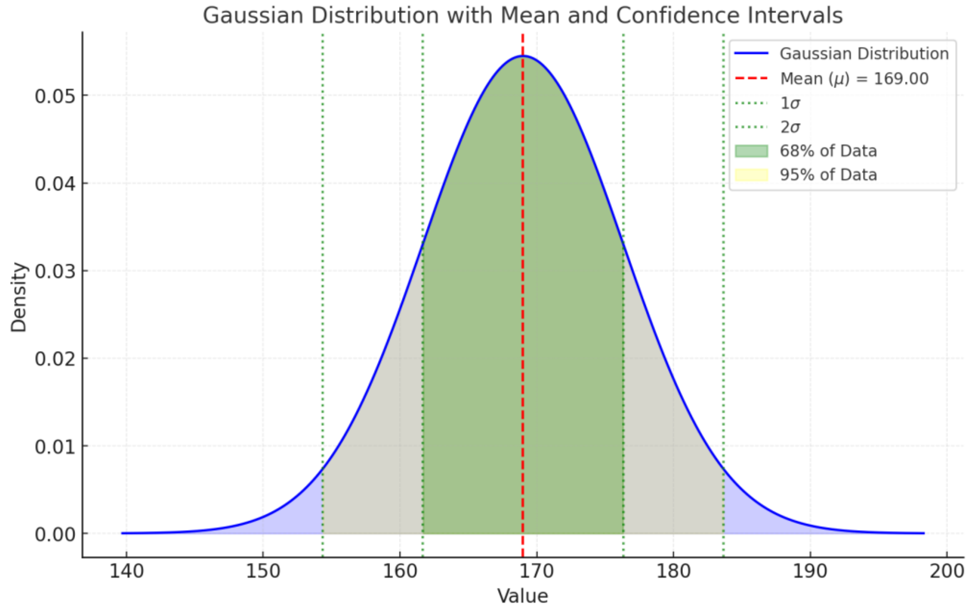
- Height is a continuous data that follows a Normal (Gaussian) Distribution



Gaussian Distribution

- The Gaussian distribution models continuous data symmetrically clustered around a mean(μ).
  - IID data is often modelled using a Gaussian distribution

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Key Characteristics:

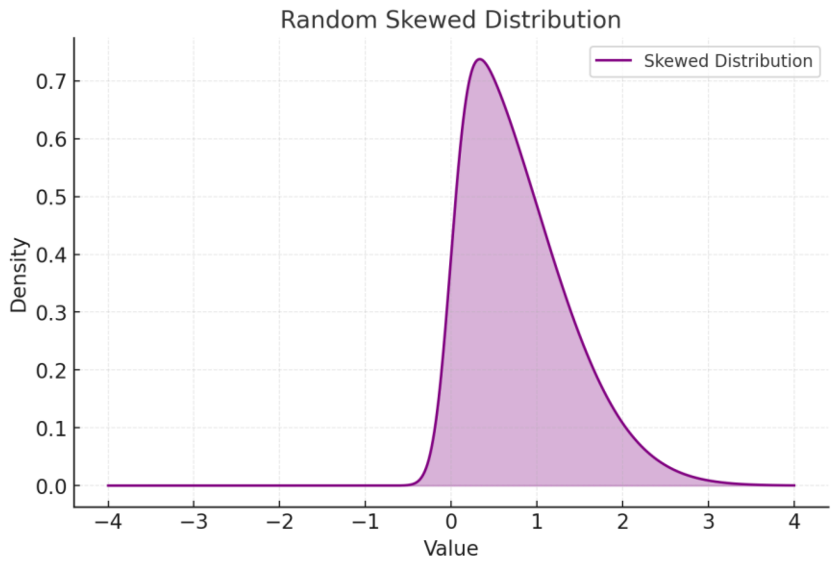
- Symmetrical bell-shaped curve.
- Mean = Median = Mode.
- 68% of data lies within 1 standard deviation (σ).
- 95% of data lies within 2 standard deviations.

Probability Basics

How to characterize a Probability Distribution?

All probability distributions can be characterized by their **moments**

Moment	Name	Formula (Central)	Interpretation
μ_1	Mean	$\mathbb{E}[X]$	The "center" or average of the distribution
μ_2'	Variance	$\mathbb{E}[(X-\mu_1)^2]$	Spread or dispersion around the mean
μ_3'	Skewness	$\mathbb{E}[(X-\mu_1)^3]$	Asymmetry or "lopsidedness" of the distribution
μ_4'	Kurtosis	$\mathbb{E}[(X-\mu_1)^4]$	Tailedness or "peakedness" of the distribution

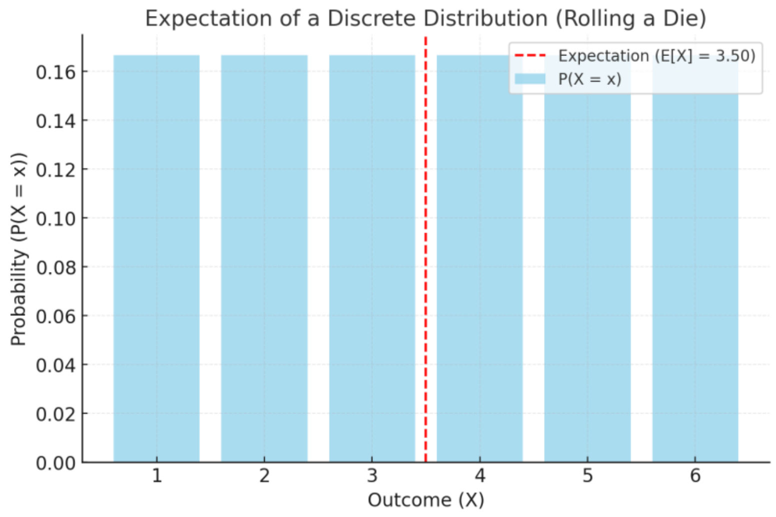


Expectation or Mean - μ

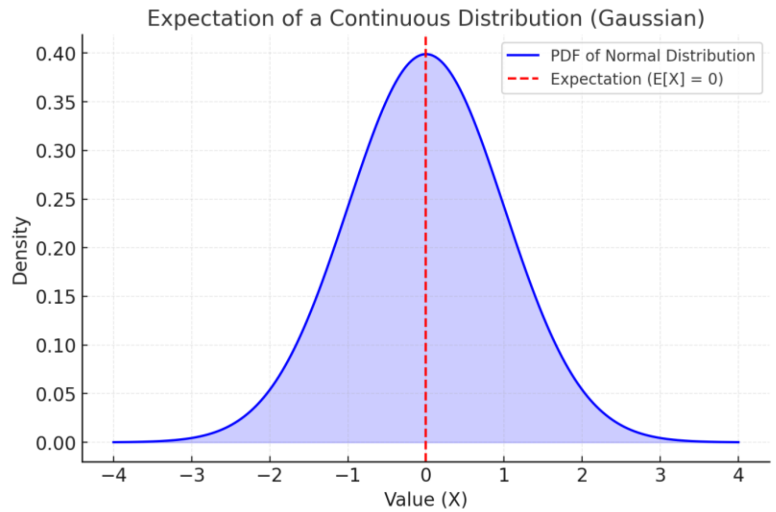
Definition: Expectation measures the “average” or expected value of a random variable.

- For **Discrete Distribution**: $$\mathbb{E}[X] = \sum_x x \cdot P(X = x)$$
  - Expected value of rolling a six-sided die

$$\mathbb{E}[X] = \sum_{x=1}^{6} x \cdot P(X = x) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

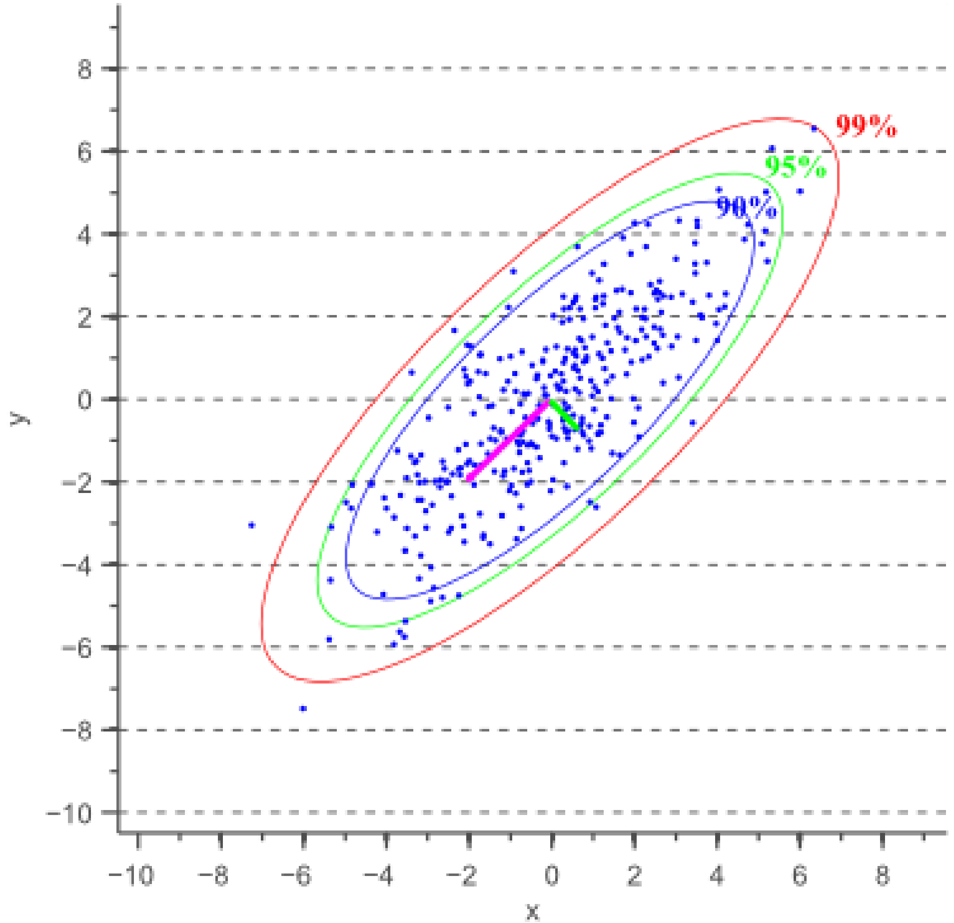


- For **Continuous Distribution**: $$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$
  - For Gaussian, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Variance - VAR(X) or σ^2

Definition: Variance measures the spread of a random variable around its mean (μ).



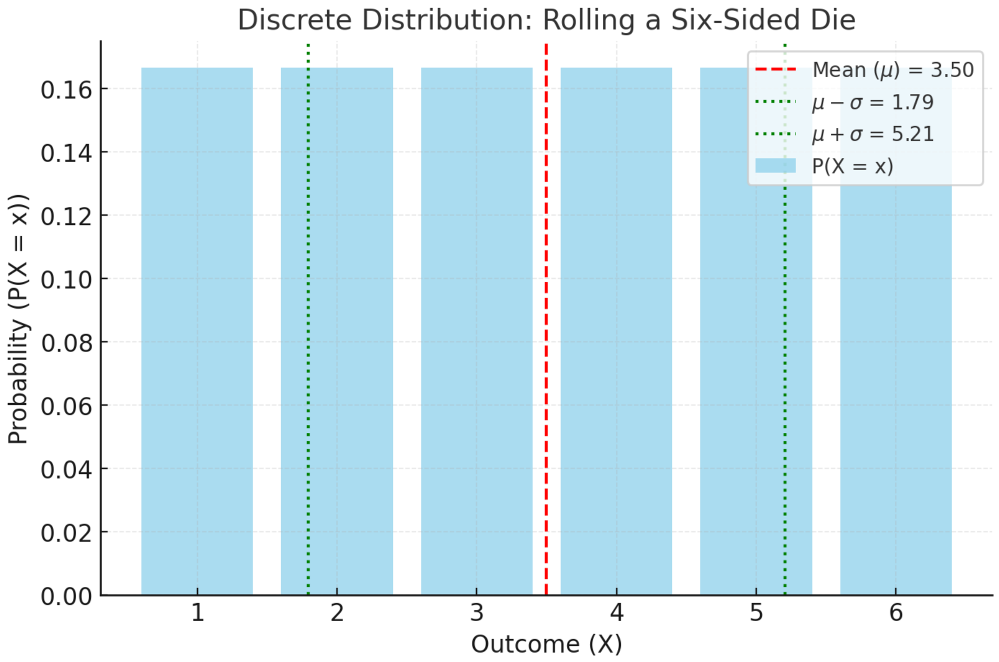
Variance – Discrete Case

Discrete Case:

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 \cdot P(X=x)$$

- **Example:** Variance of rolling a six-sided die, with mean (μ) = 3.5:

$$\mathrm{Var}(X) = \frac{1}{6}[(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2] = \frac{1}{6} \cdot 17.5 = 2.9167$$

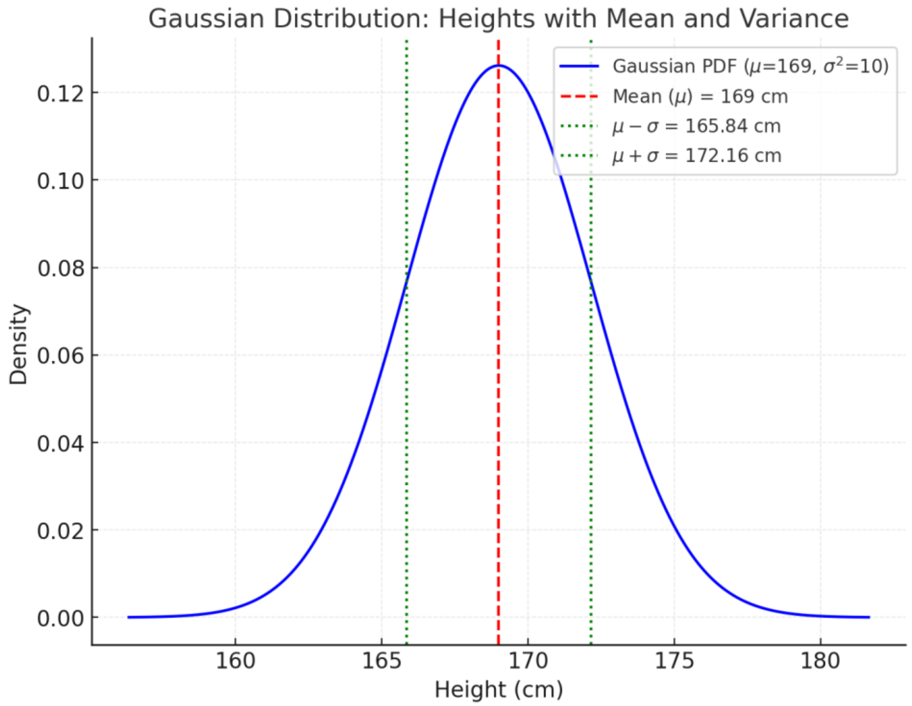


Variance – Continuous Case

Continuous Case:

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx$$

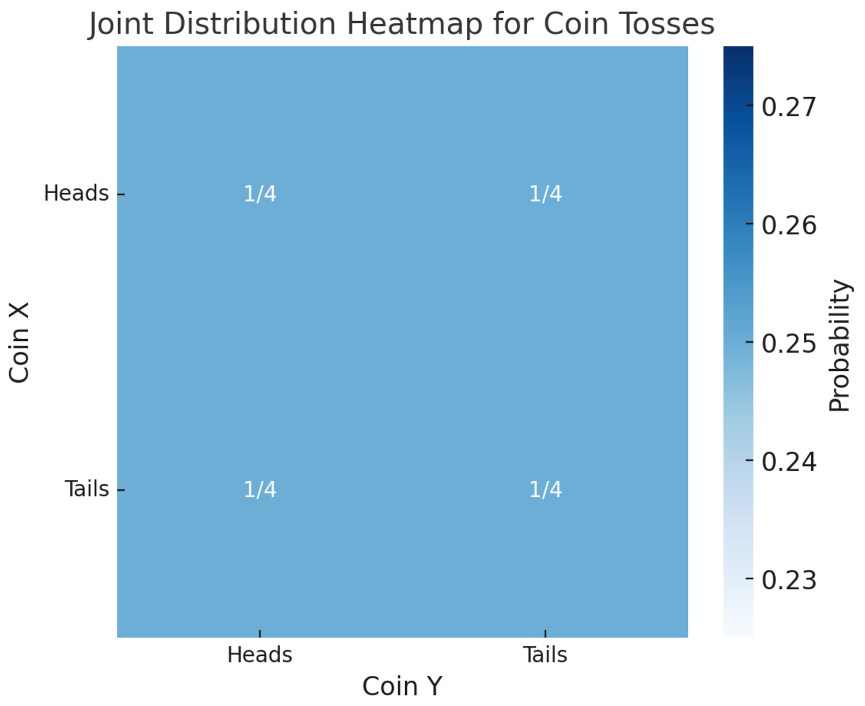
- **Example:** Heights modelled by a Gaussian distribution
  - Mean (μ) = 169 cm.
  - Variance (σ^2) = indicates the spread around 169 = 10

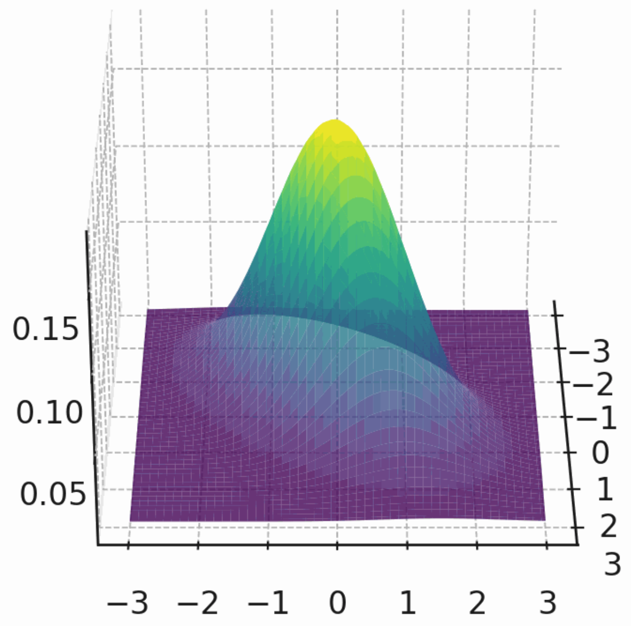


Question: What is standard deviation? How does it compare with variance?

Joint Distribution - $P(X, Y)$

Definition: A **joint distribution** models the probability of two or more random variables occurring together.





Why is joint distribution important?

Sum Rule (Marginalization)

Definition: Marginalization sums or integrates a joint distribution over one variable to find marginal distribution of another

For two random variables X and Y:

Discrete Case:

- Marginal Probability of X: $P(X = x) = \sum_{y} P(X = x, Y = y)$
- Marginal Probability of Y: $P(Y = y) = \sum_{x} P(X = x, Y = y)$

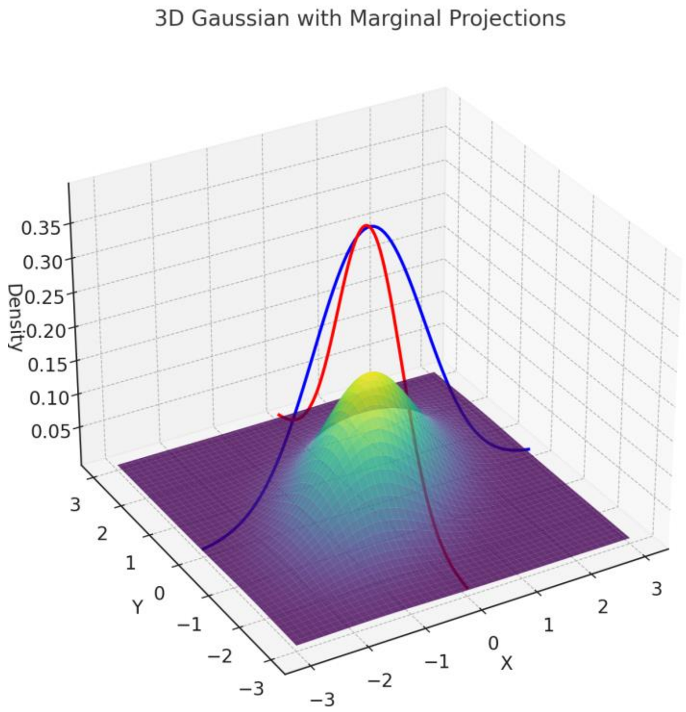
	Heads	Tails	Marginal X
Heads	0.25	0.25	0.50
Tails	0.25	0.25	0.50
Marginal Y	0.50	0.50	nan

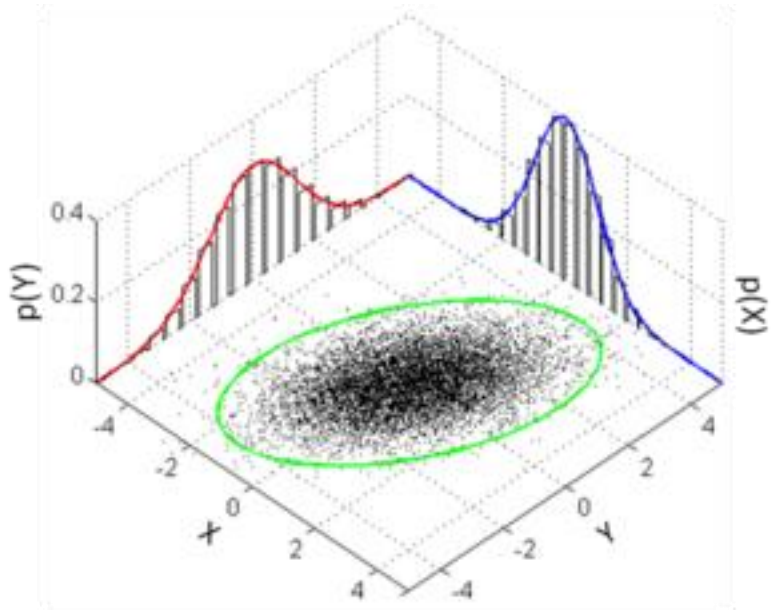
Sum Rule (Marginalization)

For two random variables X and Y:

Continuous Case:

- Marginal probability density of X: $$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$
- Marginal probability density of Y: $$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$





Conditional Probability

Definition: A conditional distribution represents the probability of one random variable given that another variable is fixed at a certain value.

Discrete Case:

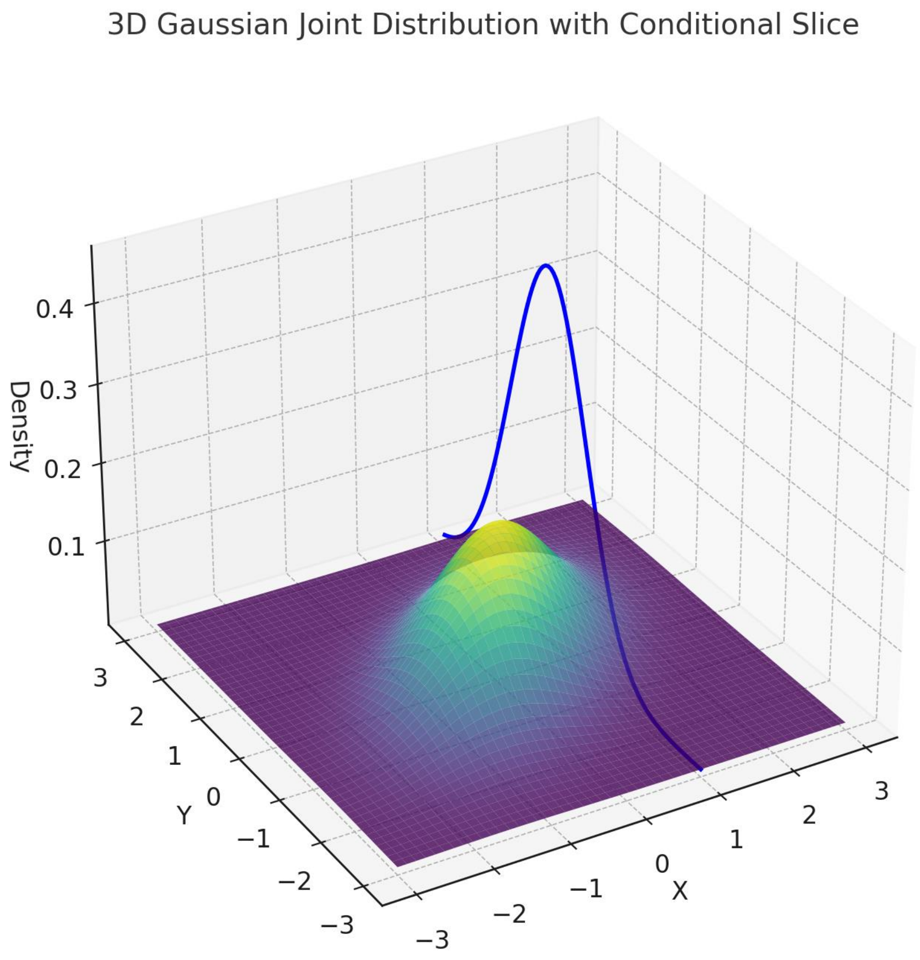
- Conditional Probability: $P(Y = y \mid X = x) = \dfrac{P(X = x, Y = y)}{P(X = x)}$
- Joint probability is divided by marginal probability
- Example: For two dice:

$$P(\text{Die } 2 = 4 \mid \text{Die } 1 = 5) = \frac{P(\text{Die } 1 = 5, \text{ Die } 2 = 4)}{P(\text{Die } 1 = 5)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

Conditional Probability

Continuous Case:

- Conditional probability density: $$f(y \mid x) = \frac{f(x,y)}{f_X(x)}$$
- Where,
  - $f(x, y)$ is the joint PDF of X and Y.
  - $f_X(x)$ is the marginal PDF of X



Product Rule

Definition: The product rule relates the joint probability of two events to their conditional and marginal probabilities.

Discrete Case:

$$P(X, Y) = P(X \mid Y) \cdot P(Y)$$

$P(X, Y)$: Joint probability of X and Y.

$P(X \mid Y)$: Conditional probability of X given Y.

$P(Y)$: Marginal probability of Y.

Continuous case:

$$f(x, y) = f(x \mid y) \cdot f_Y(y)$$

$f(x, y)$: Joint PDF of X and Y.

$f(x \mid y)$: Conditional PDF of X given Y.

$f_Y(y)$: Marginal PDF of Y.

Law of Total Probability

Definition: The Law of Total Probability provides a way to calculate the probability of an event by considering all possible ways it can occur.

For a finite or countable partition $B_1, B_2, \dots, B_n$ of the sample space (where B_i are mutually exclusive and collectively exhaustive events):

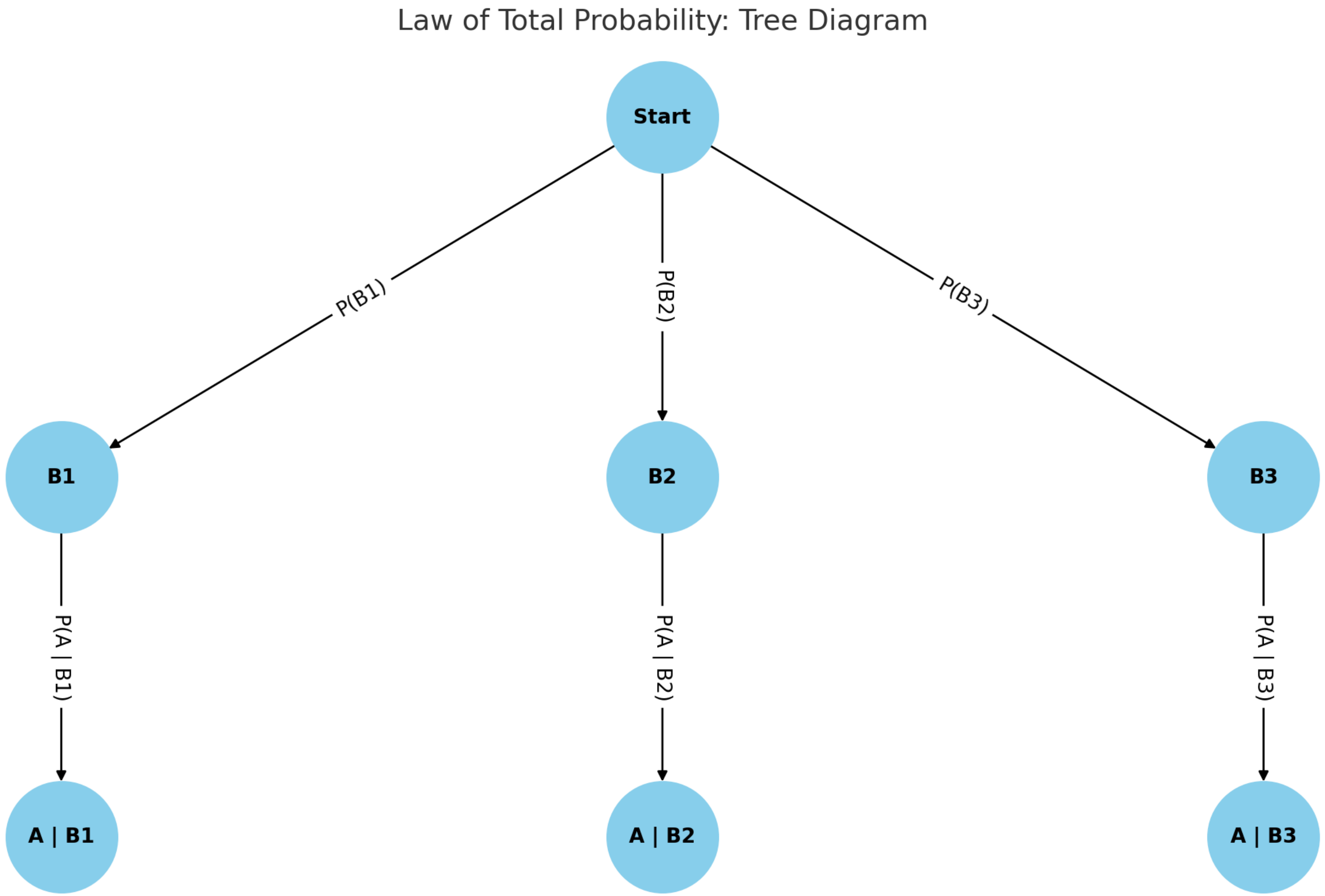
1. Discrete Case:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

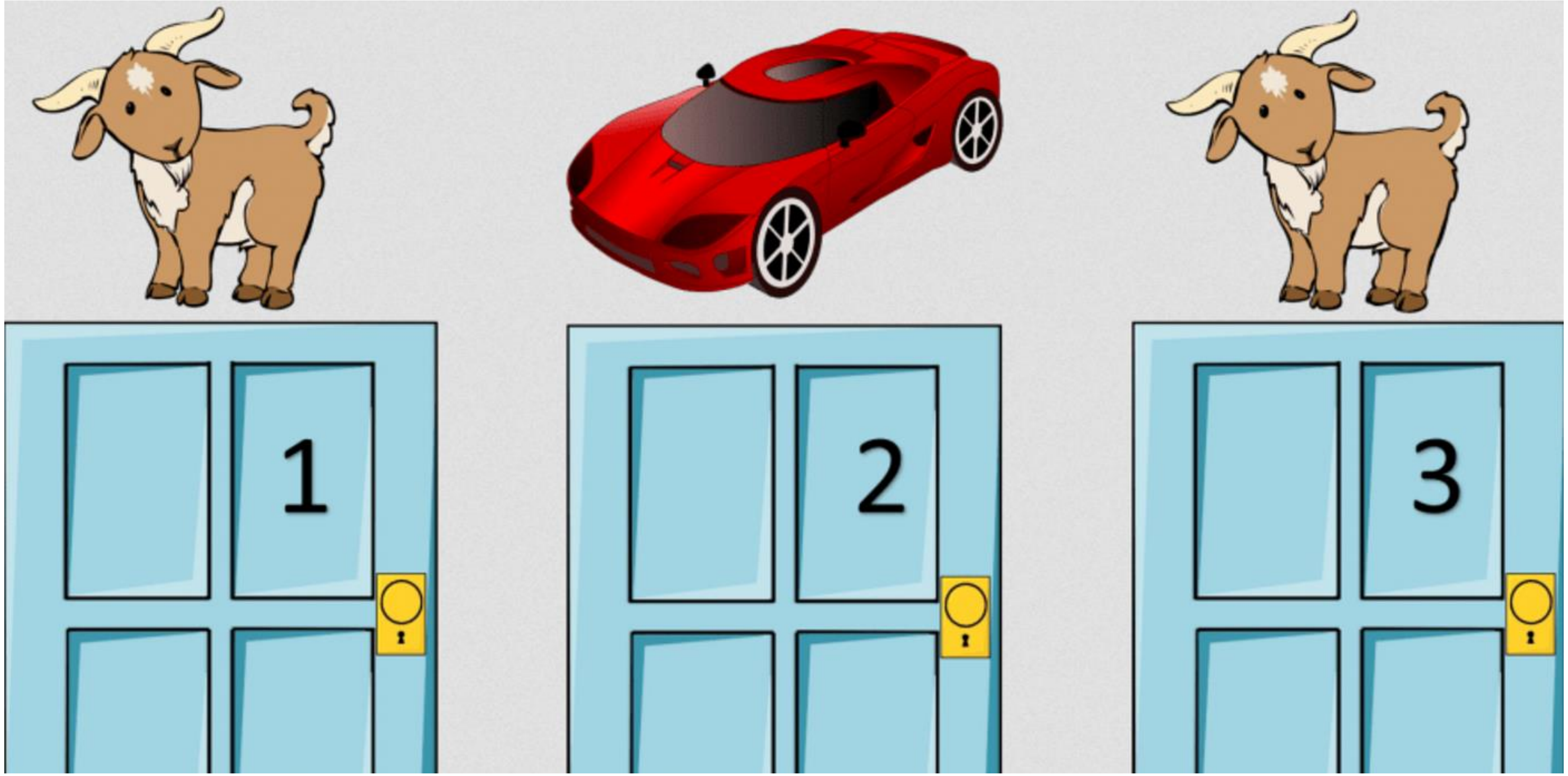
2. Continuous Case:

$$P(A) = \int_{-\infty}^{\infty} P(A \mid B = b) \cdot f_B(b)\, db$$

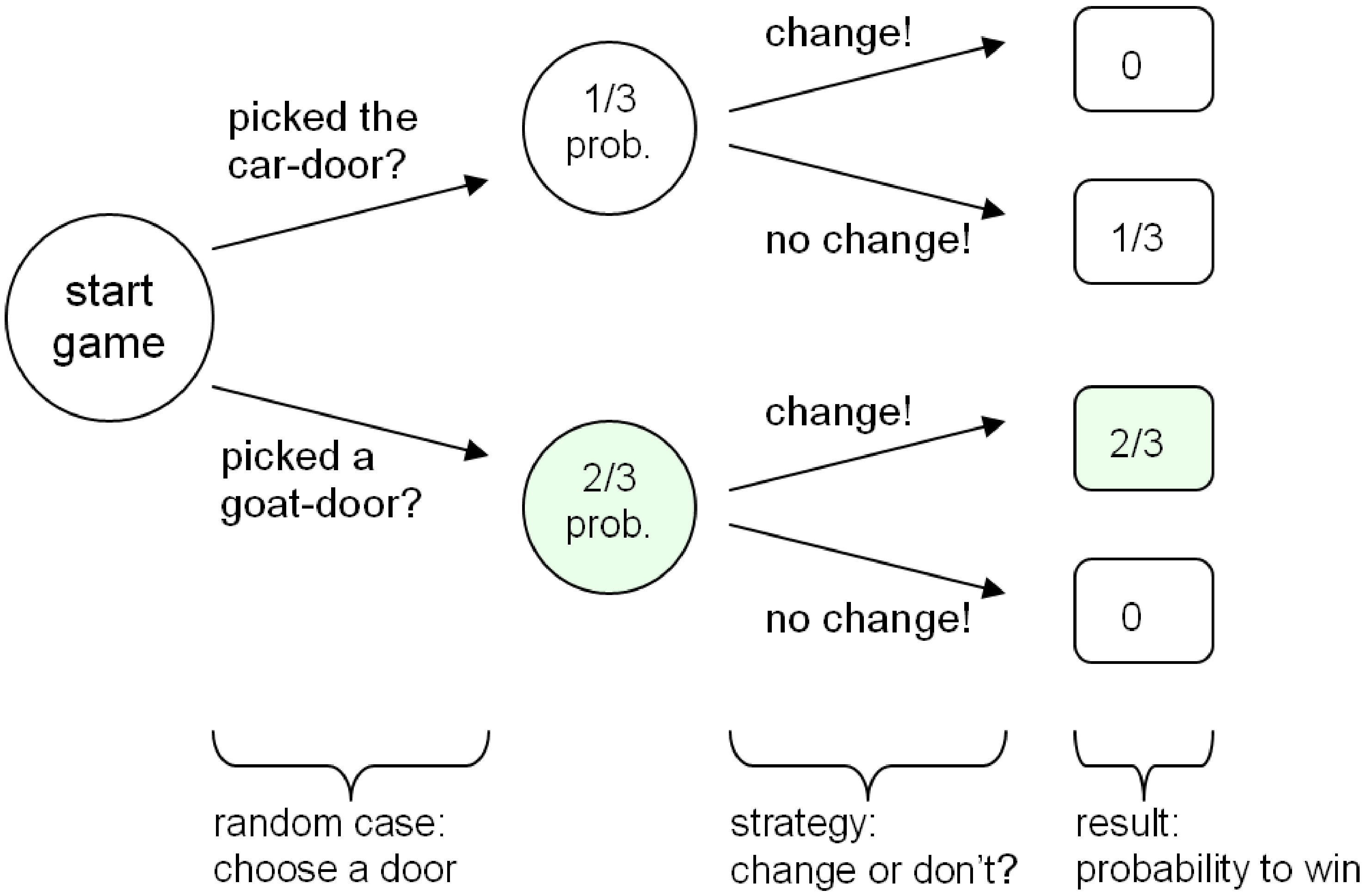
Law of Total Probability



Problem 2: Monty Hall Problem



Problem 2: Solution hint using conditional probability



Bayes Theorem

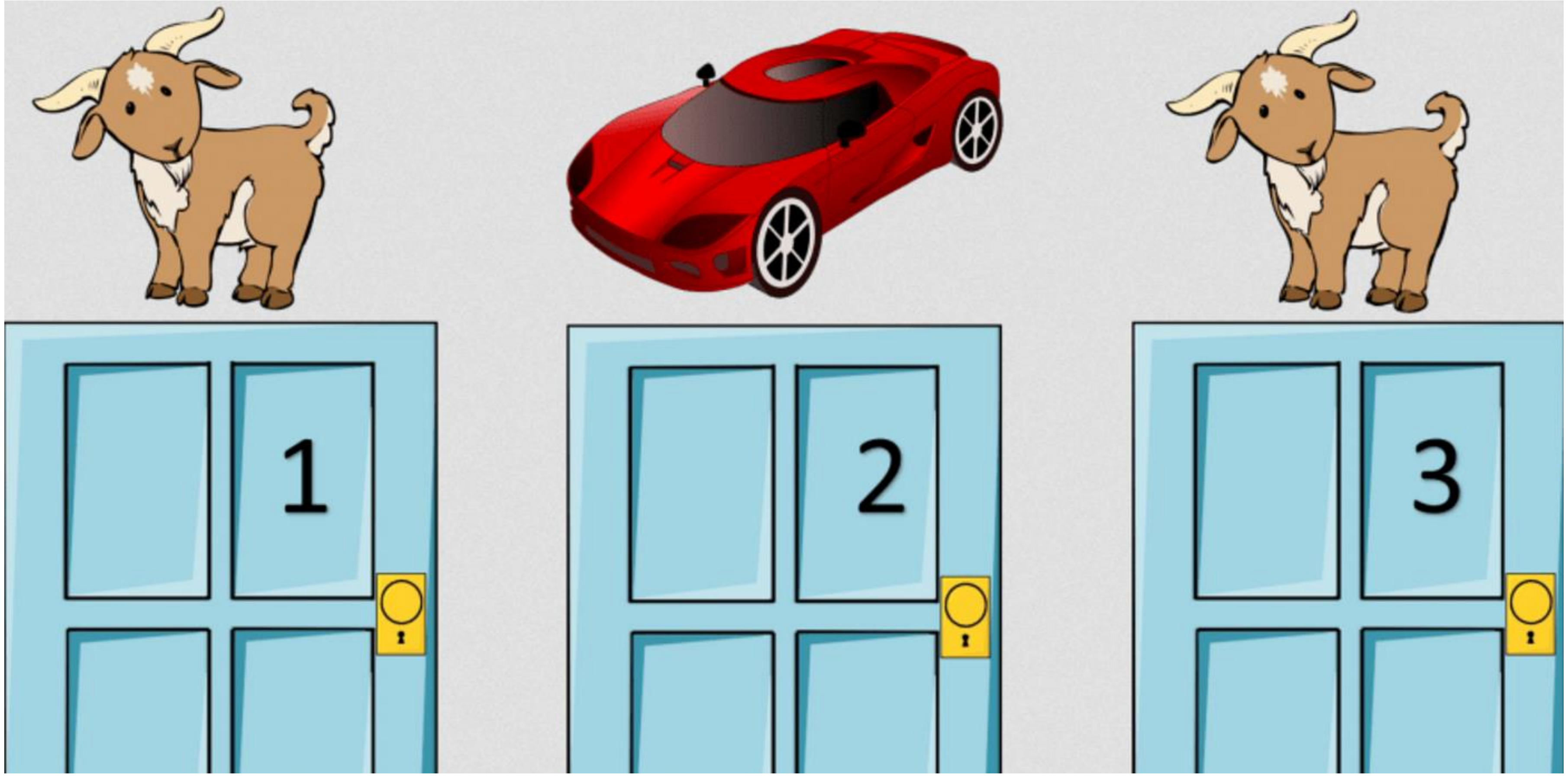
Definition: Bayes Theorem is a fundamental concept in probability that relates conditional probabilities and helps update beliefs in light of new evidence.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Where:

- $P(A \mid B)$: Posterior probability (probability of A given B).
- $P(B \mid A)$: Likelihood (probability of B given A).
- $P(A)$: Prior probability of A.
- $P(B)$: Evidence (total probability of B).

Revisiting Monty Hall’s problem through Bayesian Lens



Monty Hall's Problem: Solution from Bayesian Lens

Problem setting:

- You chose one door (Door A)
- The host Monty, who knows what's behind each door, opens another door (Door B), that has goat
- You are given a choice to either stick with your original door or switch (to Door C).

Define events:

Let,

- A_1: The car is behind the door you initially chose (Door A).
- A_2: The car is behind the door Monty does not open (Door C).
- A_3: The car is behind the door Monty opens (Door B)

Monty Hall's Problem: Solution from Bayesian Lens

- Step 1: Assign Prior Probabilities

$$P(A_1) = P(A_2) = P(A_3) = {}^{1}\!/_{3}$$

- Step 2: Define Evidence (Monty opens the door B and reveals the goat)

$$B = \text{“Monty opens Door B, and reveals a goat”}$$

- Step 3: Compute Likelihoods: $P(B|A)$
  - If A_1 (car behind door A): Monty has two doors to choose from (B or C), each with goat. He randomly opens one.

$$P(B|A_1) = {}^{1}\!/_{2}$$

  - If A_2 (car behind door C): Monty must open door, as door C has car.

$$P(B|A_2) = 1$$

  - If A_3 (car behind door B): Monty cannot open door B (it has car), making this scenario impossible.

$$P(B|A_3) = 0$$

Monty Hall's Problem: Solution from Bayesian Lens

- Step 4: Compute marginal probability $P(B)$

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$$

upon substitution: $P(B) = \left(\frac{1}{2}.\frac{1}{3}\right) + \left(1\,.\frac{1}{3}\right) + \left(0.\frac{1}{3}\right) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.$

- Step 5: Compute posterior probabilities
  - $P(A_1|B) = \frac{P(B|A_1)\,P(A_1)}{P(B)} = \frac{1/_2 . 1/_3}{1/_2} = \frac{1}{3}.$
  - $P(A_2|B) = \frac{P(B|A_1)P(A_2)}{P(B)} = \frac{1.1/_3}{1/_2} = \frac{2}{3}.$
  - $P(A_3|B) = \frac{P(B|A_3)P(A_3)}{P(B)} = \frac{0.1/_3}{1/_2} = 0.$

Conclusion:

- If you stick with your initial choice (Door A): Probability of winning $P(A_1|B) = {}^1/_3$,
- If you switch to the remaining door (Door C): Probability of winning $P(A_2|B) = {}^2/_3$.

Optimal strategy: Always switch, as it doubles your chances of winning the car!

Problem 3: Medical Diagnosis

Problem: A certain disease affects 1 in 1,000 people ($P(Disease) = 0.001$). A test for the disease has:

- **Sensitivity** (true positives): 99% ($P(Positive\ Test \mid Disease) = 0.99$).
- **Specificity**: 95% ($P(Negative\ Test \mid No\ Disease) = 0.95$)
  (We can equivalently say that the false positive rate ($P(Positive\ Test \mid No\ Disease$) is 5%)

You take the test, and the result is positive. What is the probability you actually have the disease ($P(Disease \mid Positive\ Test)$)?

Hint: Use Bayes Rule: $P(\ Disease \mid Positive\ Test\ ) = \frac{P(positive\ test|disease)P(disease)}{P(positive\ test)}$

Ans: **1.94%**

Solution hint

$$P(\text{Positive Test}) = P(\text{Positive Test} \mid \text{Disease}) \cdot P(\text{Disease}) + P(\text{Positive Test} \mid \text{No Disease}) \cdot P(\text{No Disease})$$

Where:

- $P(\text{No Disease}) = 1 - P(\text{Disease}) = 0.999,$
- $P(\text{Positive Test} \mid \text{No Disease}) = 1 - \text{Specificity} = 0.05.$

1. Compute $P(\text{Positive Test})$:

$$P(\text{Positive Test}) = (0.99 \cdot 0.001) + (0.05 \cdot 0.999)$$

$$P(\text{Positive Test}) = 0.00099 + 0.04995 = 0.05094$$

2. Compute $P(\text{Disease} \mid \text{Positive Test})$:

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{0.99 \cdot 0.001}{0.05094}$$

$$P(\text{Disease} \mid \text{Positive Test}) = \frac{0.00099}{0.05094} \approx 0.0194$$

Probability in Machine Learning

Formulating probabilistic objective in ML problems

- We have data $\mathcal{D}$ and we assume it is sampled from some distribution
- How do we figure out the **parameters that best "fit" that distribution**?

Revisiting Bayes Theorem

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) \cdot P(\theta)}{P(\mathcal{D})}$$

- θ: model parameters
- $\mathcal{D}$: observed data
- $P(\theta|\mathcal{D})$: Posterior (probability of the parameters given the data).
- $P(\mathcal{D}|\theta)$: Likelihood (probability of data given the parameters)
- $P(\theta)$: Prior belief about parameters

Maximum Likelihood Estimate (MLE)

- Objective: find θ that maximizes the likelihood:

$$\theta_{\mathrm{MLE}} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Assumes no prior knowledge about θ ($P(\theta)$ is uniform)

- Likelihood translates into the data-fit term in ML objective functions (recall least squares objective).

Maximum a Posteriori Estimate (MAP)

- Objective: find θ that maximizes the likelihood:

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

- Incorporates prior knowledge about $P(\theta)$:

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(\mathcal{D} \mid \theta) \cdot P(\theta)$$

- The prior acts as a regularization term in the objective function

MLE vs MAP

Aspect	MLE	MAP
Objective	Maximizes the likelihood $P(\mathcal{D} \mid \theta)$	Maximizes the posterior $P(\theta \mid \mathcal{D})$
Incorporates Prior?	No	Yes
Formula	$\theta_{\text{MLE}} = \arg\max_\theta P(\mathcal{D} \mid \theta)$	$\theta_{\text{MAP}} = \arg\max_\theta P(\mathcal{D} \mid \theta) \cdot P(\theta)$
Interpretation	Only considers the fit to the observed data.	Considers both data fit and prior knowledge about θ.
Objective in ML	Corresponds to minimizing only the loss term (data-fit).	Corresponds to minimizing loss + regularization.
Prior Assumption	Assumes a uniform prior (or no prior).	Allows for specific priors (e.g., Gaussian, Laplace).
Example in ML	Logistic regression without regularization.	Ridge regression (Gaussian prior), Lasso (Laplace prior).
When to Use?	When no prior knowledge is available or justified.	When prior knowledge or beliefs about θ exist.

Problem 4: MLE vs MAP

Problem Statement:

Suppose you are trying to estimate the probability θ of getting heads when flipping a biased coin. You perform a small experiment by flipping the coin 10 times and observe 7 heads and 3 tails.

Problem 4: Solution Setup

1. Observations:
   - Number of flips: $n = 10$
   - Number of heads: $x = 7$
2. Likelihood Function

   The likelihood of observing x heads in n flips, given θ, follows a binomial distribution:

$$L(\theta) = P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

3. Prior knowledge (for MAP)
   - For MLE: no prior knowledge about θ (uniform prior, $P(\theta) = 1$)
   - For MAP: Assume Beta prior, $B(\alpha, \beta)$ to encode prior belief about θ

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

   For example, for a roughly fair coin, let $\alpha = 2$ and $\beta = 2$

Problem 4: Solution – Computing MLE

To find MLE, maximize the likelihood:

$$\hat{\theta}_{MLE} = argmax_{\theta} L(\theta)$$

Likelihood is proportional to:

$$L(\theta) \propto \theta^{x}(1-\theta)^{n-x}$$

Taking logarithm (log-likelihood):

$$\log L(\theta) = x \log \theta + (n-x)\log(1-\theta)$$

Differentiate w.r.t. θ and set to zero:

$$\frac{d \log L(\theta)}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$$

Solve for θ:

$$\hat{\theta}_{MLE} = \frac{x}{n} = \frac{7}{10} = 0.7$$

Problem 4: Solution – Computing MAP

For MAP, maximize posterior $P(\theta|x)$, which is proportional to $P(x|\theta)P(\theta)$:

$$P(\theta|x) \propto \theta^{x}(1-\theta)^{n-x}.\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

Taking logarithm:

$$\log P(\theta|x) = (x+a-1)\log\theta + (n-x+\beta-1)\log P(1-\theta)$$

Differentiate w.r.t. θ and set to zero:

$$\frac{d\log P(\theta|x)}{d\theta} = \frac{x+\alpha-1}{\theta} - \frac{n-x+\beta-1}{1-\theta} = 0$$

Solve for θ:

$$\hat{\theta}_{MAP} = \frac{x+\alpha-1}{n+\alpha+\beta-2} = \frac{7+2-1}{10+2+2-2} = \frac{8}{12} = 0.6667$$

Problem 4: Solution – Comparing MLE vs MAP

- MLE Estimate: $\hat{\theta}_{MLE} = 0.7$
  - MLE maximizes the likelihood based solely on the observed data.
- MAP Estimate: $\hat{\theta}_{MAP} = 0.6667$
  - MAP incorporates prior knowledge, pulling the estimate slightly closer to the prior belief (coin being fair)

Key Takeaway:

- MLE focuses only on the data and is prone to overfitting with small datasets.
- MAP balances observed data with prior beliefs, making it more robust for small sample sizes or when prior knowledge is available.

Recap

- Data as distributions (Bernoulli, Binomial, Gaussian)
- Basic Probability Concepts (expectation, variance, joint probability, sum rule, conditional probability, product rule, law of total probability, Bayes theorem)
- Probability in Machine Learning (infer parameters using maximum likelihood estimate (MLE) and maximum a posteriori estimate (MAP))