Chapter 01: Fundamentals of Inference

Probabilistic Artificial Intelligence

Notes by Kumar Anurag

Why Probability in AI?

Imagine this:

A robot assistant looks outside — it's cloudy.
The forecast says 60% chance of rain.

Should it:

- Carry an umbrella for you?
- Reschedule your outdoor workout?
- Ignore the forecast?

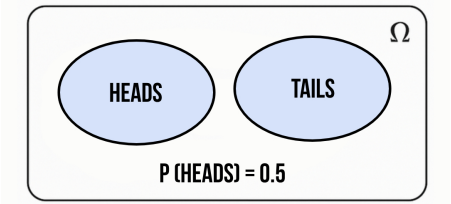
The robot doesn't know for sure—it must *reason under uncertainty*.

Why Probabilistic AI?
Because AI needs a mathematical framework to model *beliefs*, not just facts.

What is a Probability Space?

A probability space is a mathematical world for modeling uncertainty.

$$(\Omega, \mathcal{A}, P)$$



- Ω – All possible outcomes.
  Example: $\Omega = \{\text{Heads}, \text{Tails}\}$
- $\mathcal{A}$ – Set of events (measurable subsets of Ω).
  Example: $\{\emptyset, \{\text{Heads}\}, \{\text{Tails}\}, \Omega\}$
- P – A function that assigns probabilities to events in $\mathcal{A}$, satisfying:
  $P(\Omega) = 1, \quad P(E) \geq 0, \quad$ countable additivity

Probability is a mathematical framework to measure uncertainty.

Sigma-Algebra

σ-algebra (F): A collection of subsets of the sample space, including the sample space itself and the empty set, that is closed under complements and countable unions.

Rules:

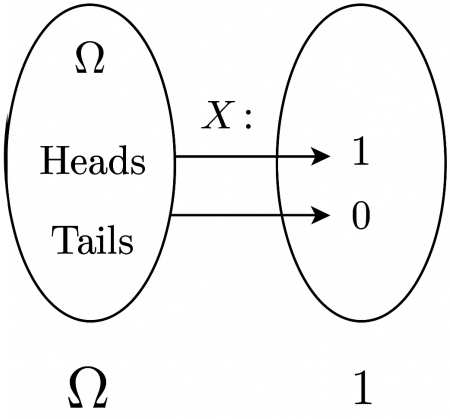
1. **The whole sample space is included**: Ex. for a die, {1,2,3,4,5,6}, this set is always a part of σ-algebra.
2. **Closed under complements**: Ex. if 'rolling a 6' is in the σ-algebra, then 'not rolling a 6' must also be in it.
3. **Closed under unions**: Ex. if we combine events ('rolling a 2' or 'rolling a 4'), the resulting event must also be in the σ-algebra.

What is a Random Variable?

A Random Variable (RV) assigns a number to each outcome in a sample space.

- Technically: A measurable function $X : \Omega \to \mathbb{R}$
- X is the variable; x is the value it takes (realization)
- *Example:* Tossing a coin
  $\Omega = \{\text{Heads}, \text{Tails}\}$
  Define $X(\text{Heads}) = 1$, $X(\text{Tails}) = 0$

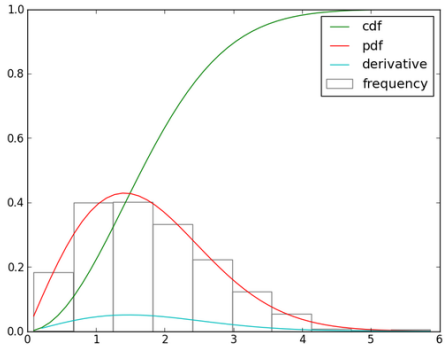
RV turns uncertain outcomes into usable numbers.



What is a Probability Distribution?

A probability distribution tells us how likely each outcome is.

- For discrete random variables -> we use a **PMF (Probability Mass Function)**. *Example:* Rolling a die
  $P(X = 3) = \frac{1}{6}$
- For continuous random variables -> use a **PDF (Probability Density Function)**. *Example:* Gaussian distribution over real numbers
  $P(X = x) = 0$, but we can compute $P(a < X < b)$



- **CDF (Cumulative Distribution Function)** gives probability up to a value:
  $F(x) = P(X \leq x)$

Distributions summarize how random variables behave.

What is a Continuous Distribution?

In a continuous world, outcomes can't be counted — they fill an entire range.

Discrete: When we roll a die -> only 6 outcomes.
Continuous: When we measure temperature -> infinite possibilities!

Key idea: We talk about probability *over intervals*, not single points.

$$P(X = x) = 0 \quad \text{but} \quad P(a < X < b) > 0$$

Example: Height of a person: $P(X = 170\text{cm}) = 0$,
but $P(169.5 < X < 170.5) > 0$

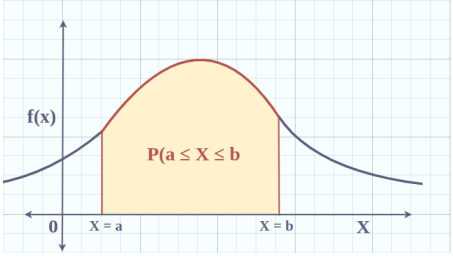
Probability Density Function (PDF)

A PDF describes how "dense" the probability is at each point.

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

Where $f(x)$ is the PDF (e.g., bell curve)

- The PDF can be greater than 1! It's not probability itself
- The **area under the curve** = probability
- Total area under the PDF = 1



Think of it like mass spread over a continuous line.

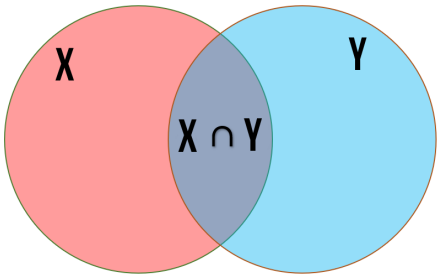
Joint Probability

Joint probability describes the chance of two things happening together.

Example: Weather (X) and whether I carry an umbrella (Y)

$$P(X = \text{Rain},\ Y = \text{Yes}) = 0.3$$

Notation: $P(X, Y)$ or $P(X = x,\ Y = y)$



Joint distribution is like a table of co-occurring probabilities.

We'll use this table on the next slide to compute marginal probabilities.

Marginalization (Sum Rule)

Marginal probability is the probability of one variable, *ignoring* the other.

$$P(X = \text{Rain}) = \sum_{y} P(X = \text{Rain}, Y = y)$$
$$= P(X = \text{Rain}, Y = \text{Yes}) + P(X = \text{Rain}, Y = \text{No})$$

Example: Add across all umbrella choices.

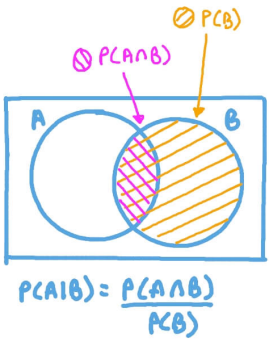
This process is called *marginalization* — you're "summing out" a variable.

Visual: We'll use a probability table and highlight the row/column.

Conditional Probability

Conditional probability is the probability of an event, given that another has happened.

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



Example: What is the chance it's raining *given* we see someone with an umbrella?

Intuition (by my mathematics teacher): We are zooming in on the part of the world where B is true, and asking how often A also happens.

Bayes' Rule: Flipping the Condition

The product rule:

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

Bayes' Rule:

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

Why it matters: Sometimes it's easier to compute $P(B \mid A)$ and $P(A)$ than $P(A \mid B)$ directly.

Used in:

- Medical diagnosis
- Spam filters
- Weather prediction

Understanding Independence

Two events are independent if knowing one tells us nothing about the other.

$$A \perp B \quad \Leftrightarrow \quad P(A \cap B) = P(A) \cdot P(B)$$

Example: Tossing two fair coins:

A: First coin is heads $P(A) = 0.5$

B: Second coin is heads $P(B) = 0.5$

$P(A \cap B) = 0.25 = 0.5 \times 0.5$

Knowing the result of one coin doesn't affect the other. That's independence!

Conditional Independence

Conditional independence: Two events are independent, given a third.

$$A \perp B \mid C \quad \Leftrightarrow \quad P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

Example: Sprinkler (S) and Rain (R) both influence Wet Grass (W).
The events S and R are independent. S doesn't depend on whether it is raining. Similarly, Raining doesn't depend on whether it is Sprinkling.

$P(S, R) = P(S) \cdot P(R)$

Conditional independence is key in graphical models.

Directed Graphical Models

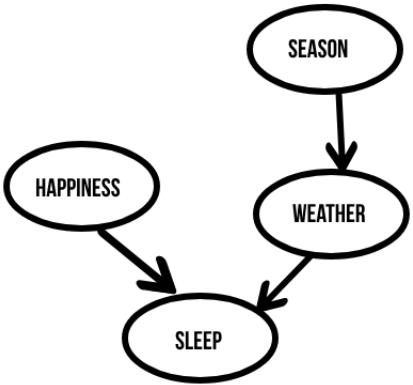
Also called Bayesian Networks.

Idea: Use a directed graph to represent random variables and their dependencies.

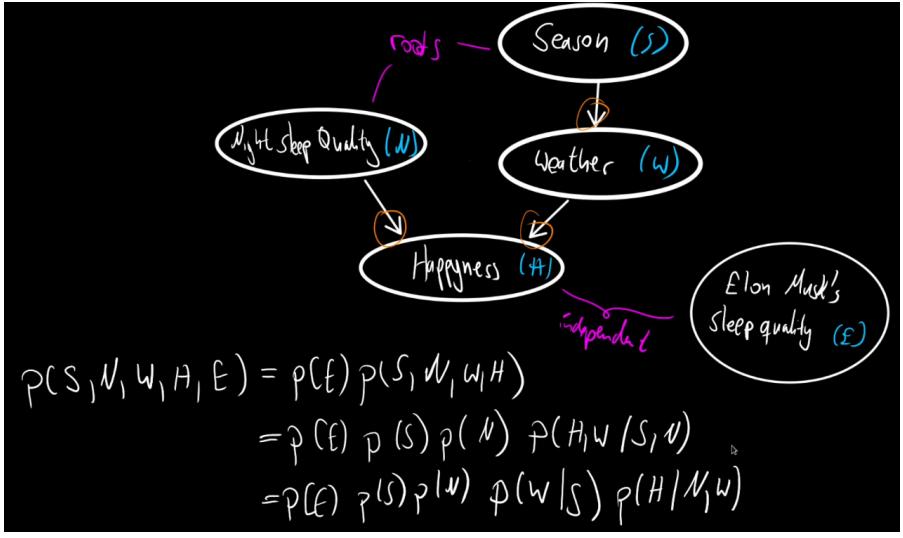
- Nodes = Random Variables
- Arrows = Direct influence / causal connection

Why useful?

- Makes complex distributions manageable
- Helps with reasoning and computation



Factorizing a Joint Distribution



Expectation (Expected Value)

The expected value is the average outcome you'd expect in the long run.

For discrete variables:

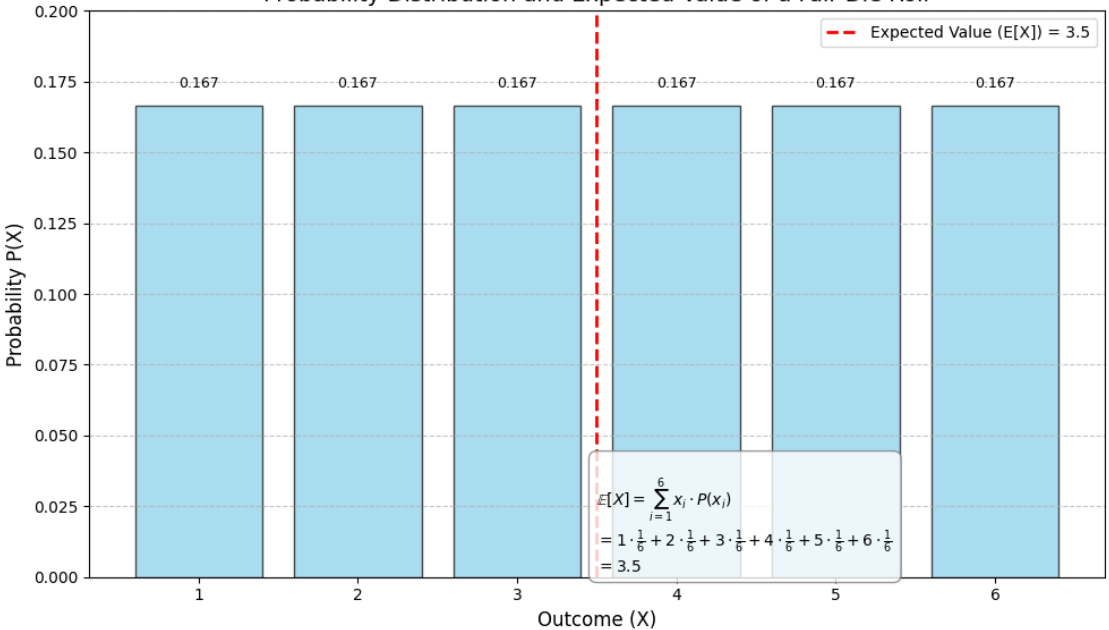
$$\mathbb{E}[X] = \sum_{x} x \cdot P(X = x)$$

For continuous variables:

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

Example: Roll a die $\rightarrow \mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$

Interpretation: Think of it as the "center of gravity" of the distribution.


Probability Distribution and Expected Value of a Fair Die Roll
Expected Value (E[X]) = 3.5
0.167
0.167
0.167
0.167
0.167
0.167
$\mathbb{E}[X] = \sum_{i=1}^{6} x_i \cdot P(x_i)$
$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$
$= 3.5$
Probability P(X)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
1
2
3
4
5
6
Outcome (X)

Variance – How Spread Out is the Distribution?

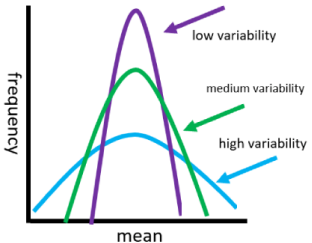
Variance measures how much a random variable deviates from its mean.

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Standard deviation = $\sqrt{\text{Var}(X)}$

Intuition:

- High variance $\rightarrow$ outcomes are spread out
- Low variance $\rightarrow$ outcomes are close to the mean



Imagine measuring the heights of basketball players vs. chess players!

Covariance – Relationship Between Two Variables

Covariance measures how two variables vary together.

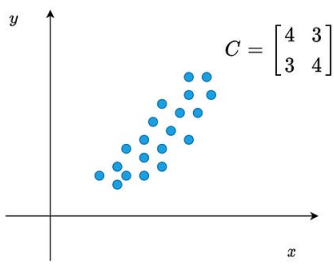
$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Interpretation:

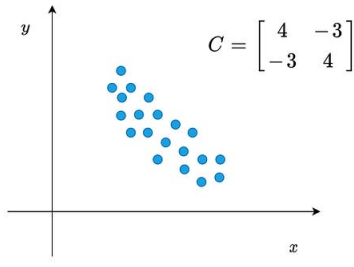
- $\text{Cov}(X, Y) > 0 \rightarrow$ increase together
- $\text{Cov}(X, Y) < 0 \rightarrow$ one increases, the other decreases
- $\text{Cov}(X, Y) = 0 \rightarrow$ uncorrelated (but not always independent!)

Bonus: Correlation is normalized covariance.

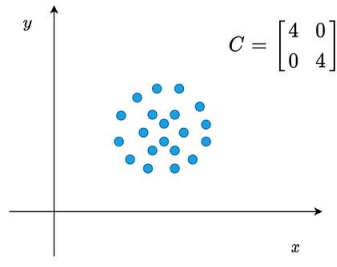
Positive Covariance



Negative Covariance

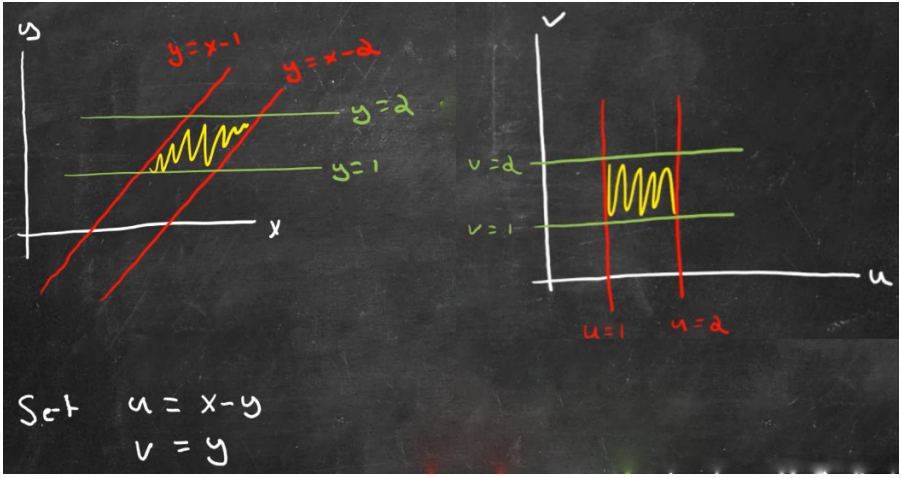


Zero Covariance

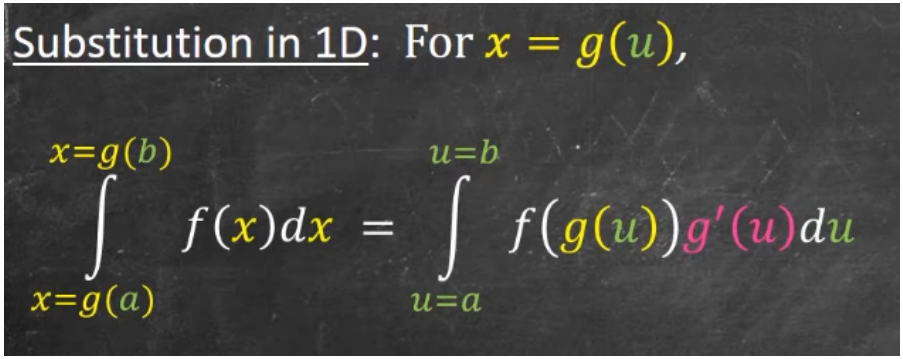


Change of Variables

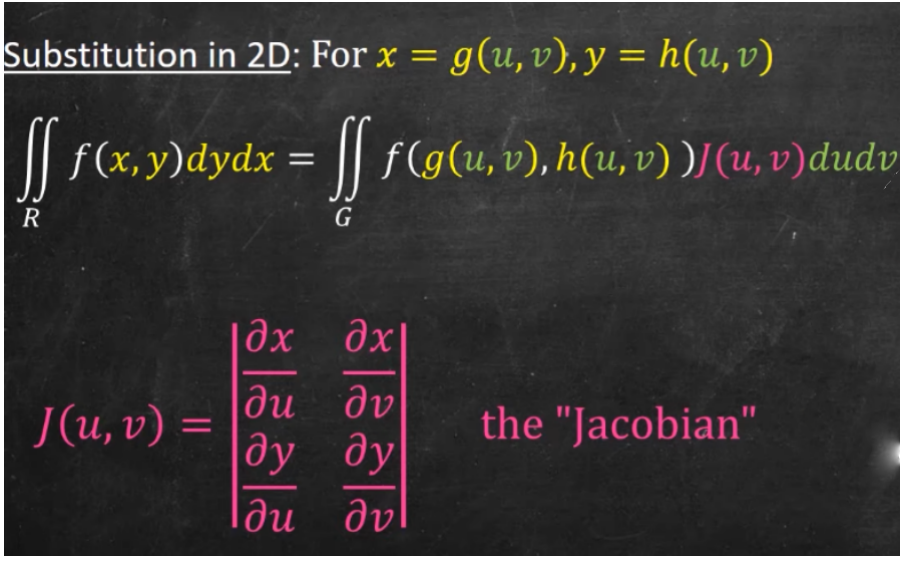
What if we transform a random variable into a new one?



Change of Variables - 1D



Change of Variables - 2D



Plausible vs. Logical Inference

Logical inference: Uses strict rules — only what's definitely true.
Plausible inference: Operates under uncertainty — what's *likely* true.

Example:

Logical Inference

If it rains, the ground is wet.
It is raining.
$\Rightarrow$ *The ground is wet.*

Plausible Inference

The ground is wet.
$\Rightarrow$ *Maybe it rained? Maybe the sprinkler was on?*

Key Insight: Plausible inference lets us reason *backwards* from evidence.
This is the foundation of Bayesian thinking!

Where Do Priors Come From?

Priors represent what we believe before seeing any data.

But where do they come from?

- **Subjective belief:** Our personal knowledge or intuition.
  (e.g., "Coins are fair unless proven otherwise.")
- **Empirical estimates:** Based on past data.
  (e.g., "Spam words appear in 70% of past spam emails.")

Note: Priors are not magic — they reflect assumptions. Make them wisely.

What Are Conjugate Priors?

When you do Bayesian inference, you're updating a prior belief using data to get a posterior belief:

Bayes' Rule:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Sometimes, after multiplying the prior and the likelihood, the result (posterior) belongs to the same family as the prior.

That's called a **conjugate prior**.

Step 1 — Prior Belief

We want to estimate the probability of heads, θ, for a coin (baised).

Before flipping the coin, we assume:

$$\theta \sim \text{Beta}(\alpha = 2,\ \beta = 2)$$

This means:

- We believe the coin is fair (symmetric Beta)
- It's like we've seen 1 head and 1 tail before

This is our prior belief. Now let's collect data...

Step 2 — Observe Data

We flip the coin 10 times and observe:

$$7 \text{ heads}, 3 \text{ tails}$$

This is modeled as:

$$X \sim \text{Binomial}(n = 10, \theta)$$

Now we update our prior using Bayes' Rule...

Step 3 — Posterior Belief

Using Bayes' Rule, our new belief is:

$$\theta \mid \text{data} \sim \text{Beta}(\alpha + \text{heads},\ \beta + \text{tails})$$

In our case:

$$\theta \mid \text{data} \sim \text{Beta}(2 + 7,\ 2 + 3) = \text{Beta}(9, 5)$$

Interpretation:

- We've updated our belief based on observed data
- Posterior is still a Beta distribution — conjugacy!

This makes Bayesian updating fast and intuitive.

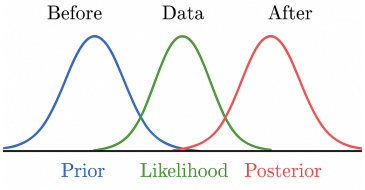
Tractable Gaussian Inference

Gaussian distributions make Bayesian inference easy and exact.

Suppose we have:

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2) \quad \text{(Prior)}$$

$$x \mid \theta \sim \mathcal{N}(\theta, \sigma^2) \quad \text{(Likelihood)}$$



Then the posterior is also Gaussian:

$$\theta \mid x \sim \mathcal{N}(\mu_{\text{post}}, \sigma_{\text{post}}^2)$$

Key Benefits:

- Conjugate: Gaussian prior + Gaussian likelihood = Gaussian posterior
- Easy to compute mean and variance updates
- Widely used in Kalman filters, linear regression, etc.

Thank you!

THE UNIVERSITY OF NEW MEXICO

OPTIMIZATION AND ESTIMATION LAB

ONE.UNM.EDU