Probabilities and Random Variables

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

Contents

Helping Material

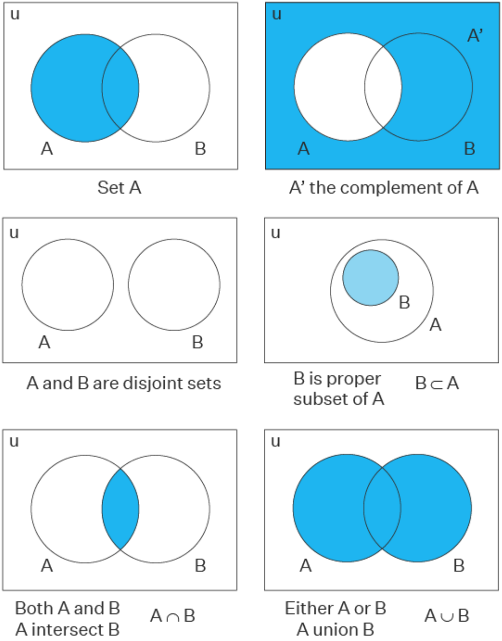
- **Probabilistic Modeling and Reasoning (PMR) Course**
  https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/
- **Multivariate Statistics**
  https://11annah-s-teachings.github.io/
- **Primer on Probabilistic Modeling**
  https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/assets/notes/probabilistic-modelling-primer.pdf

Program

1. Introduction to Probability Theory
2. Properties of a Probability Distribution
   - PDF and PMF
   - Variance
   - Sampling
3. Important Distributions
   - Bernoulli distribution
   - Poisson distribution
   - Normal distribution
   - The Central Limit Theorem
4. Recap & What's Next

We will start at the very beginning:
The realm of probability theory!

Quick set theory reminder:



What is Probability?

- Probability is a number between 0 and 1.
- It tells us how likely an event is to happen.

Event	Probability
Heads in a coin flip	0.5
Rolling a 3 on a die	1/6
Sun rises tomorrow	1
Finding a unicorn	0

Interpretation

Probability helps us reason about uncertainty.

What is a Probability Space?

A probability space is a triple $(\Omega, \mathcal{F}, P)$:

1. **Sample space** Ω: the set of all possible outcomes.
2. **Events** $\mathcal{F}$: a collection of subsets of Ω (events).
3. **Probability function** P: a function $P : \mathcal{F} \to [0, 1]$ assigning probabilities to events.

Example: Rolling a Die

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - they are outcomes $\Rightarrow$ we could write $\Omega = \{a, b, c, d, e, f\}$
- Event A: any even number $\Rightarrow A = \{2, 4, 6\}$
  - $A \subseteq \Omega$, e.g. if $\Omega = \{a, b, c, d, e, f\}$ then $A = \{b, d, f\}$
- $P(\{2\}) = P(\{4\}) = P(\{6\}) = \frac{1}{6} \Rightarrow P(A) = \frac{3}{6}$

What is a Random Variable?

- A **random variable** X:
  - maps outcomes to numbers, i.e., is a **function** $X : \Omega \to \mathbb{R}$.
  - gives *a numerical view of the sample space*
  - we can say "P that X is even" instead of directly referring to Ω.

Example: Rolling a Die

$\Omega = \{a, b, c, d, e, f\}$ are outcomes; X maps them to numbers:

$$X(a) = 1,\ X(b) = 2,\ X(c) = 3,\ X(d) = 4,\ X(e) = 5,\ X(f) = 6$$

So the event $\{b, d, f\} \subseteq \Omega$ becomes X is an even number.

A random variable is like a lens: it translates raw outcomes into numbers.

What is a Distribution?

- A **distribution** tells us how likely each value of a random variable is.
- It is a function: maps values of the random variable to probabilities.

Example: Die Roll

Let X be the result of rolling a fair 6-sided die:

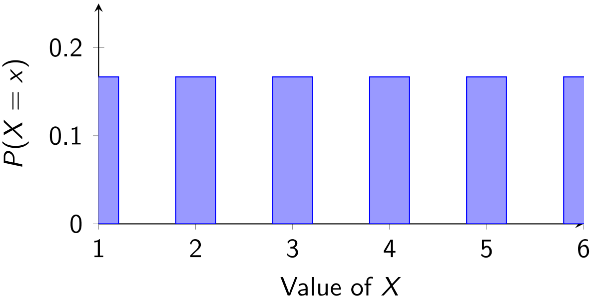
$$P(X = k) = \frac{1}{6}, \quad \text{for } k = 1, 2, 3, 4, 5, 6$$

Uniform distribution over $\{1, 2, 3, 4, 5, 6\}$

The distribution describes the behavior of the random variable.

What is a Distribution?

- A **distribution** tells us how likely each value of a random variable is.
- It is a function: maps values of the random variable to probabilities.



From Outcomes to Distributions

The Full Chain

$$(\Omega, \mathcal{F}, P) \xrightarrow{\text{Random Variable}} X : \Omega \to \mathbb{R} \xrightarrow{\text{Distribution}} P(X = x)$$

Do we need the full chain?

- Only in formal probability theory.
- In applications and modeling, we start directly from a **distribution**.
  - Uniform: $P(X = k) = \frac{1}{n}$ for $k \in \{1, \dots, n\}$
  - Bernoulli: $P(X = 1) = p, \quad P(X = 0) = 1 - p$
  - The underlying $(\Omega, \mathcal{F}, P)$ is abstract or implicit.

In probabilistic modeling we often start directly from a distribution, without explicitly defining the sample space or events.

Program

Distribution Properties

1. **Probability Density Function (PDF)** or **Probability Mass Function (PMF)**.
   - $f(x) = P(X = x)$ for discrete variables (PMF)
   - $f(x) = P(a \leq X \leq b)$ for continuous variables (PDF)
2. **Cumulative distribution function (CDF)**
   - $F(x) = P(X \leq x)$
3. **Important Summary Statistics**
   - **Expected value (mean)**: $\mathbb{E}[X]$
   - **Variance (spread)**: $\mathrm{Var}(X)$
   - **Standard deviation**: $\sigma = \sqrt{\mathrm{Var}(X)}$
   - Also: skewness, kurtosis, entropy, etc.
4. **Sampling**
   - Generating random samples from a distribution
   - Allows us to approximate properties of the distribution

PMF and PDF

Probability Mass Function (PMF) – Discrete Random Variables:

$$\boxed{p(x)} = P(X = x)$$

- The probability that a random variable X equals x
- Sum to 1: $\sum_x p(x) = 1$

Probability Density Function (PDF) – Cont. Random Variables:

$$P(a \leq X \leq b) = \int_a^b \boxed{p(x)}\, dx$$

- The probability that a random variable X falls within $[a, b]$
- PDF is the *density*; probabilities result from integrating over intervals
- The value at a single point is not a probability ($p(x)$ can be > 1)
- Sum to 1 over the entire space: $\int_{-\infty}^{\infty} p(x)\, dx = 1$

Cumulative Distribution Function (CDF):

$$F(x) = P(X \leq x)$$

- The probability X is less than or equal to x.
- For discrete variables: $F(x) = \sum_{a \leq x} p(a)$
  - Step function; increases at each outcome
- For continuous variables: $F(x) = \int_{-\infty}^{x} p(t)\, dt$
  - Smooth function; area under the PDF curve up to x
- $F(x)$ is non-decreasing: if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$
- $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$
- For continuous variables: $\boxed{p(x) = \frac{d}{dx} F(x)}$
- For discrete variables: $p(x) = F(x) - F(\alpha)$, where α is the largest value less than x.

Expected Value

The **expected value**, $\mu = \mathbb{E}[X]$, is the average value of a random variable X over its distribution — the mean of the distribution.

Definition: For a function f of an outcome x,

Discrete:

$$\mathbb{E}[x] = \sum_{i=1}^{I} p_i a_i$$

Continuous:

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

Examples:

- **Six-sided die:**
  - $\mathbb{E}[x] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$
  - Note: not an actual outcome, but a statistical average
- **Bernoulli trial (1 with probability p):**
  - $\mathbb{E}[x] = p \cdot 1 + (1 - p) \cdot 0 = p$

Properties of Expectations

1. **Linearity:**

$$\mathbb{E}[f(x)+g(x)] = \mathbb{E}[f(x)] + \mathbb{E}[g(x)], \quad \mathbb{E}[cf(x)] = c\,\mathbb{E}[f(x)]$$

2. **Constant Rule:**

$$\mathbb{E}[c] = c\sum_{i=1}^{I} p_i = c$$

3. **Independence Rule:**

$$\mathbb{E}[f(x)g(y)] = \mathbb{E}[f(x)]\,\mathbb{E}[g(y)]$$

If x and y are independent.

Exercise: Prove the independence rule.

The Variance

The **Variance** measures the spread of a random variable x around its mean μ.

$$\mathrm{Var}[x] = \sigma^2 = \mathbb{E}\left[(x-\mu)^2\right] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

where $\mu = \mathbb{E}[x]$ is the mean.

Properties:

- $\mathrm{var}[cx] = c^2\,\mathrm{var}[x]$
- If x and y are independent:
  $\mathrm{var}[x+y] = \mathrm{var}[x] + \mathrm{var}[y]$

Standard deviation: $\sigma = \sqrt{\mathrm{var}[x]}$

- **Standard deviation** has the same units as x.
- *Variance has different units from x, so not easily interpretable.*

Sampling from a Distribution

Sampling means generating random values $\{x_1, x_2, \ldots, x_N\}$ from a distribution p:

$$x_i \sim p(x), \quad i = 1, 2, \ldots, N$$

Key Points:

- Sampling allows to *easily* approximate important properties of the distribution:
  - Mean: $\mathbb{E}[x] \approx \frac{1}{N} \sum_{i=1}^{N} x_i$ for N samples.
  - Variance: $\text{var}[x] \approx \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2$.
- Some distributions are easy to sample from (e.g., Bernoulli, Gaussian)
- others require advanced methods.

Program

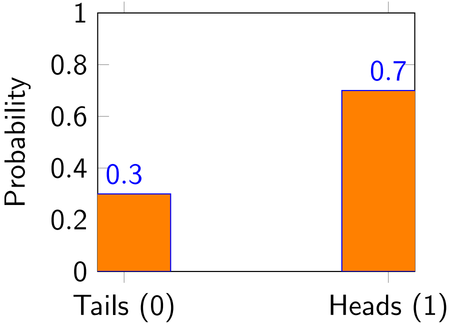
Bernoulli Distribution

Definition: Models procedures with two outcomes: success (1), failure (0)
Example: Toss a biased coin with $p = 0.7$ of landing heads (success):

$$X \sim \text{Bernoulli}(p), \quad P(X = 1) = p, \quad P(X = 0) = 1 - p$$

Properties:

- $\mathbb{E}[X] = p = 0.7$
- $\text{Var}(X) = p(1 - p) = 0.21$



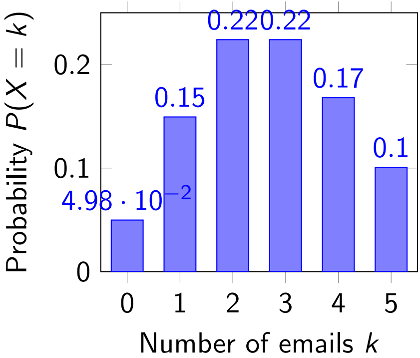
Poisson Distribution

Definition: Models the number of events in a time interval, given the events occur independently and at an average rate λ.
Example: Number of emails received per hour

$$X \sim \text{Poisson}(\lambda), \quad P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

Properties:

- $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda = 3$, expected number of emails per hour.
- Probability of receiving k emails in an hour: $P(X = k) = \frac{3^k e^{-3}}{k!}$



Univariate Gaussian: Definition and Properties

Definition: A univariate Gaussian (Normal) distribution is defined as:

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

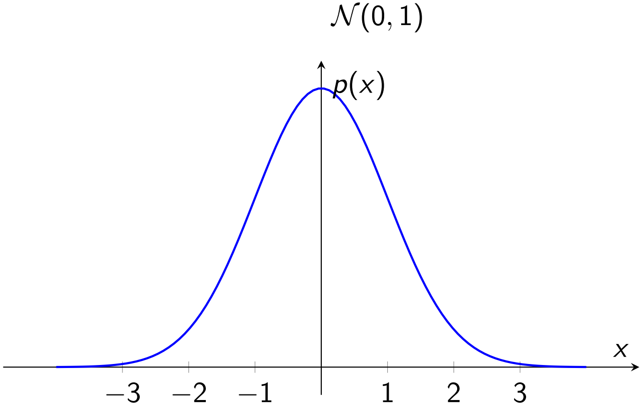
Parameters:

- μ: mean (center of the distribution)
- σ^2: variance (spread of the distribution)

Properties:

- Symmetric around μ
- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\mathrm{var}[x] = \sigma^2$
- A Sum of independent Gaussians is Gaussian

Univariate Gaussian



Multivariate Gaussian: Definition and Properties

Definition: A d-dimensional multivariate Gaussian is defined as:

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Parameters:

- $\boldsymbol{\mu} \in \mathbb{R}^d$: mean vector
- $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$: covariance matrix

Mean and Covariance:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}, \quad \mathrm{Cov}[\mathbf{x}] = \boldsymbol{\Sigma}$$

Why Are We Obsessed with Gaussians?

Gaussians (a.k.a. Normal distributions) are everywhere:

- **Mathematically convenient:** Closed-form expressions for mean, variance, marginalization, conditioning, etc.
- **Defined by just two parameters:** Mean μ and variance (or covariance) σ^2/Σ
- **Stable under linear transformations:** Linear combinations of Gaussians are still Gaussian
- **Pop up in nature:** Measurement errors, heights, weights, noise, and many other phenomena
- **Crucial in ML:** Gaussian assumptions simplify models (e.g., Gaussian Naive Bayes, GPs, Kalman filters)

And there is the Central Limit Theorem (CLT)...

The Central Limit Theorem (CLT)

Why that obsession with Gaussians?

What is the CLT? If you add up many independent random outcomes,

the sum tends to follow a Gaussian distribution.

Why? Random variation averages out. The “bell curve” emerges naturally when:

- Each variable has a **bounded mean and variance**
- The values aren’t too extreme or weird

We will check that in the exercises later:

Program

Recap & What's Next

So far, we've seen:

- What probability distributions are.
- How they define a **PDF or PMF**, a **CDF**, and key properties like **expectation** and **variance**.
- That we can **sample** from them to simulate uncertainty.

What's missing? *(Coming next!)*

- **Probabilistic Modeling:**
  - How to use *parametric* distributions to model real-world phenomena.
- **Probabilistic Inference:**
  - How to fit parameters to data.
  - Via *optimization* or *Bayesian inference*.
  - Using fitted models to make predictions, decisions, and analyses.