Probabilities and Random Variables

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

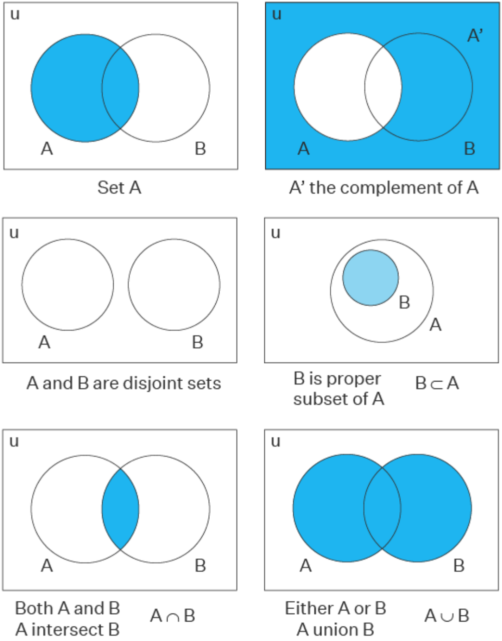
Contents

Helping Material

- **Primer on Probabilistic Modeling**
  https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/assets/notes/probabilistic-modelling-primer.pdf

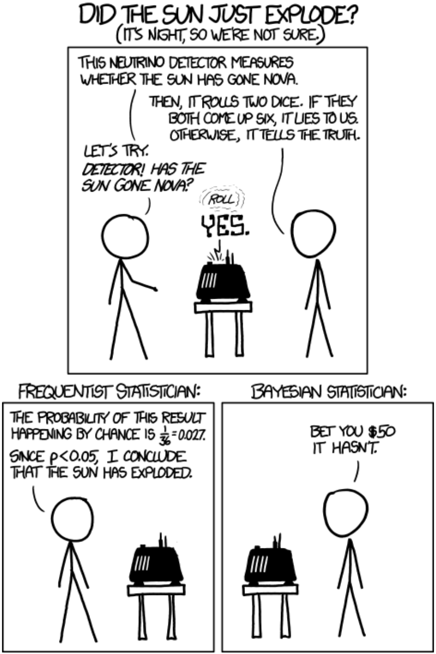
**We will start at the very beginning:
The realm of probability theory!**

Quick set theory reminder:



QUESTION:

What is your understanding of the term "probability"?


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

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:

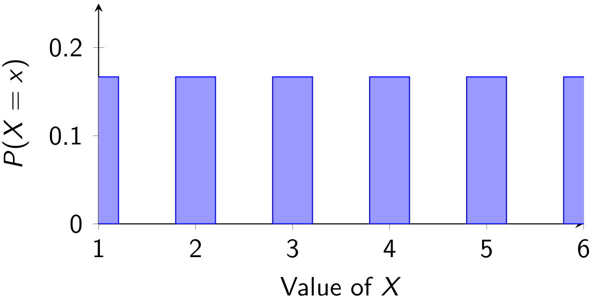
$$P(X = k) = \frac{1}{6}, \quad \text{for } k = 1, 2, 3, 4, 5, 6$$

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.

What Can We Do With a Distribution?

A probability distribution allows us to:

1. **Define a probability density function (PDF)** or a **probability mass function (PMF)**.
   - Gives the relative likelihood of each outcome.
2. **Define a cumulative distribution function (CDF).**
   - $F(x) = P(X \leq x)$, accumulates probability up to x.
3. **Summarize properties of the distribution.**
   - Most important: **expected value (mean)** and **variance (spread)**.
   - Also: skewness, kurtosis, entropy, etc.
4. **Sample from the distribution.**
   - Generate artificial data consistent with the modeled uncertainty.

PMF and PDF

Probability Mass Function (PMF) – Discrete Random Variables:

$$P(X = x) = p(x)$$

- Gives the probability that the random variable equals a specific value.
- Probabilities add up over all possible values and sum to 1.

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

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

- Gives the *density* of the random variable around a point; probabilities come from areas under the curve.
- The value at a single point is not a probability (can be > 1).

Cumulative Distribution Function (CDF)

Cumulative Distribution Function (CDF):

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

- Gives the probability that the random variable takes a value less than or equal to x.
- For discrete variables: a step function; for continuous variables: a smooth, increasing curve.

Properties:

- $F(x)$ is non-decreasing and satisfies: $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$.
- For continuous variables: $p(x) = \frac{d}{dx} F(x)$.

Expectation

Expectation is a property of a probability distribution, representing a probability-weighted average.

Definition: For a function f of an outcome x,

Discrete:

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

Continuous:

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

- Subscript $P(x)$ often dropped when context is clear.
- Notation variants: $\mathbb{E}[f]$, $\mathcal{E}[f]$, $\langle f \rangle$.
- If $f(x) = x$, then $\mathbb{E}[x]$ is the **mean**.

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:

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

Because probabilities sum to one.

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 Mean (Expected Value)

The mean of a distribution is the *expected value* of numerical outcomes:

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

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):**

$$\mathbb{E}[x] = p \cdot 1 + (1 - p) \cdot 0 = p$$

Change of units: If x is in metres and we want cm:

$$\mathbb{E}[100x] = 100\,\mathbb{E}[x]$$

The Variance

Variance measures the average squared distance from the mean:

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

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: $\sigma = \sqrt{\mathrm{var}[x]}$ Same units as x, often used as a measure of spread.

Variance: Change of Units and Normalization

Change of Units:

- If x is in metres, then x^2 is in m^2.
- Variance changes with units:
  $\mathrm{var}[100x] = 100^2\,\mathrm{var}[x]$

Normalization:

- Given mean μ and variance σ^2, to normalize x:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \quad \Rightarrow \quad \mathbb{E}[x_{\text{norm}}] = 0, \quad \mathrm{var}[x_{\text{norm}}] = 1$$

Note: Variance has different units from x, so it's not always directly interpretable.

Sampling from a Distribution

Sampling means generating random values that follow a given probability distribution.

Notation: If x is a random variable sampled from distribution P, we write:

$$x \sim P$$

Key Points:

- Some distributions are easy to sample from (e.g., Bernoulli, Gaussian), others require advanced methods.
- 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: $\mathrm{var}[x] \approx \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mathbb{E}[x])^2$.

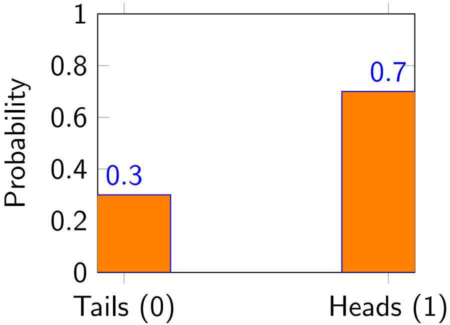
Bernoulli Distribution

Definition: Models a procedure with two outcomes: success (1) an failure (0). **Example: Coin toss**: Toss a biased coin with probability $p = 0.7$ of landing heads (success).

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

Properties:

- Mean: $\mathbb{E}[X] = p = 0.7$
- Variance: $\mathrm{Var}(X) = p(1 - p) = 0.21$



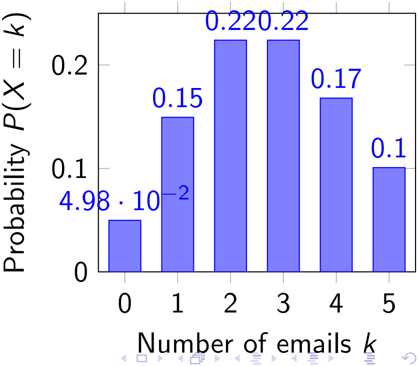
Poisson Distribution

Definition: Models the number of events in a fixed interval of time or space, given the events occur independently and at a constant 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:

- Mean and variance: $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda = 3$, 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: $\mathrm{var}[x] = \sigma^2$

Univariate Gaussian: Example and Additional Properties

Example:

- Let $x \sim \mathcal{N}(3, 4)$
- Then:

$$\mathbb{E}[x] = 3 \qquad \text{var}[x] = 4 \qquad \text{std}[x] = \sqrt{4} = 2$$

Additional Properties:

- Linear transformation: If $y = ax + b$, then $y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Sums of independent Gaussians are 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:

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

Multivariate Gaussian: Example and Key Properties

Example: Let $\mathbf{x} \sim \mathcal{N}\left(\begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2 & 1\\1 & 3\end{bmatrix}\right)$

- $\boldsymbol{\mu} = \begin{bmatrix}1\\2\end{bmatrix}$
- $\boldsymbol{\Sigma} = \begin{bmatrix}2 & 1\\1 & 3\end{bmatrix}$

Properties:

- Marginals are Gaussian
- Affine transformations preserve Gaussianity
- If $\mathbf{x}_1$ and $\mathbf{x}_2$ are jointly Gaussian, then $\mathbf{x}_1 \mid \mathbf{x}_2$ is Gaussian

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:

Recap

What have we learned so far?

- Distributions and Random Variables are fundamental to model phenomena involving uncertainty.
- They are formally defined over a probability space, but we often do not need that formalism in practice; we start directly with a **distribution**.
- The **expected value** (mean) and **variance** are key properties of distributions.
- Sampling from a distribution is key for estimating the mean and variance.
- Some probability distributions are particularly important in probability theory and statistics, e.g., The **Bernoulli**, **Poisson**, and **Gaussian**.

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 distributions to model real-world phenomena.
  - Introducing *parametric* families (e.g., Normal, Poisson).
- **Probabilistic Inference:**
  - How to fit parameters to data.
  - Via *optimization* or *Bayesian inference*.
  - Using fitted models to make predictions, decisions, and analyses.

Next time: From theory to action — modeling, fitting, and applying distributions in machine learning.