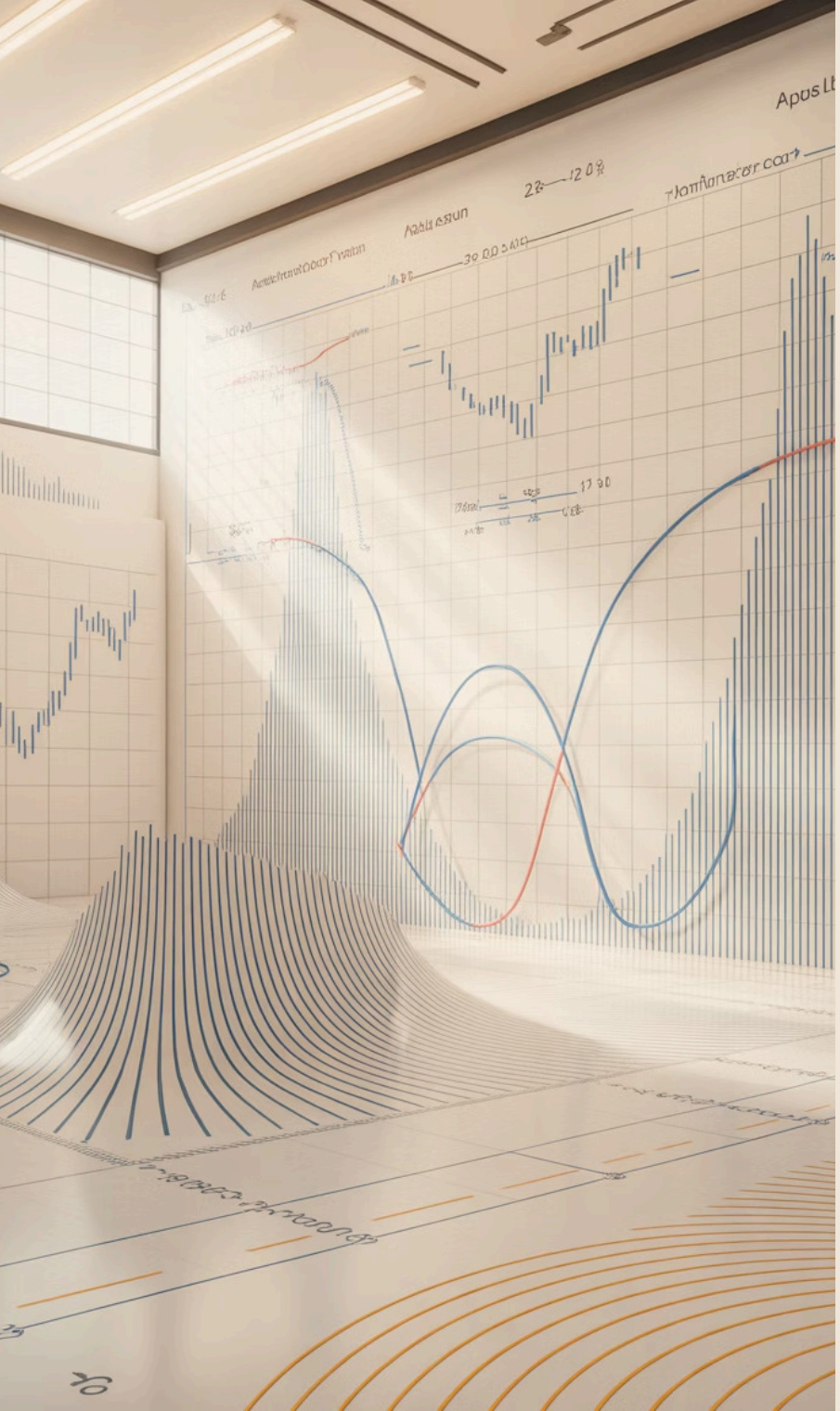
Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)
Corso di Studi in Informatica
Dip. di Matematica e Informatica
Università di Catania

Data Distributions

Understanding how to assign probability values to all possible outcomes of a random variable is fundamental to data analysis and machine learning. This presentation explores probability distributions, from discrete to continuous variables, and their practical applications.



What Are Probability Distributions?

Core Concept

A probability distribution is **a function that assigns probability values to each possible value of a random variable**. It characterizes the variable and defines which outcomes are more likely to observe.

> When a random variable X follows a probability distribution P(X), we write: $X \sim P$ and say that "X follows P".

Discrete Variables

For countable outcomes (like dice rolls or coin tosses), we use **Probability Mass Functions (PMF)**.

Example: Head or tails in a coin flip.

Continuous Variables

For measurable quantities (like height or temperature), we use **Probability Density Functions (PDF)**.

Example: Human body temperature in degrees

Probability Mass Functions (PMF)

For discrete random variable X, the PMF $P(X)$ maps each possible value to its probability:

$$P : \Omega \to [0, 1]$$

A valid PMF must satisfy this crucial normalisation property:

$$\sum_{x \in \Omega} P(x) = 1$$

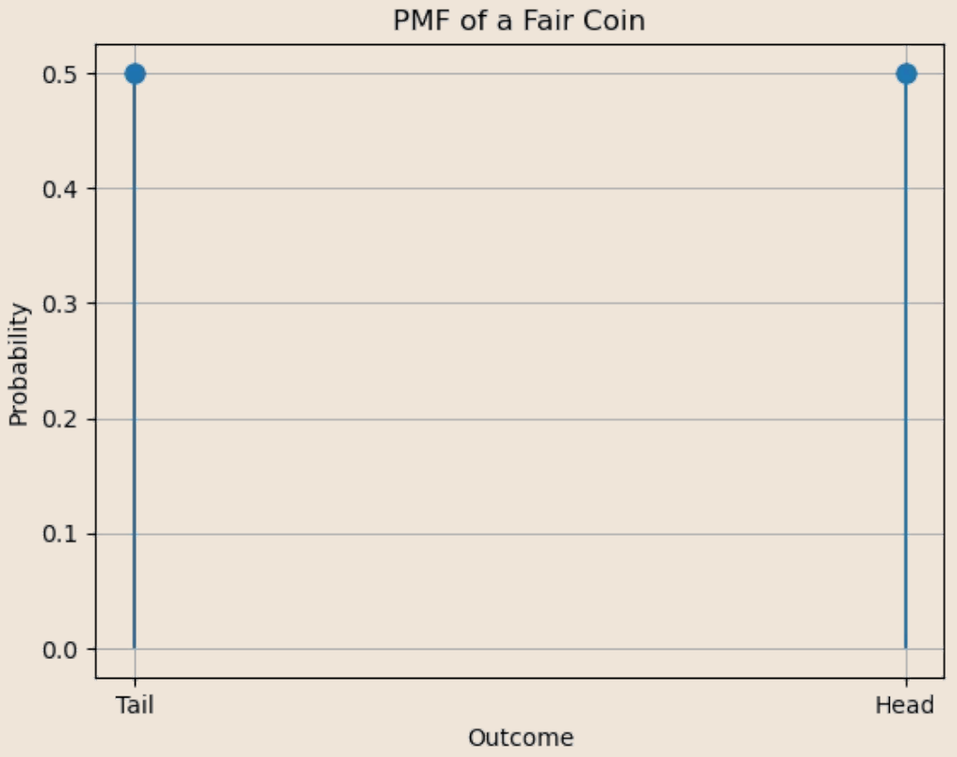
This ensures that probabilities sum to 1, as one outcome must always occur.

Example: Fair vs Biased Coin

Fair Coin

A fair coin follows a discrete uniform distribution where both outcomes have equal probability:

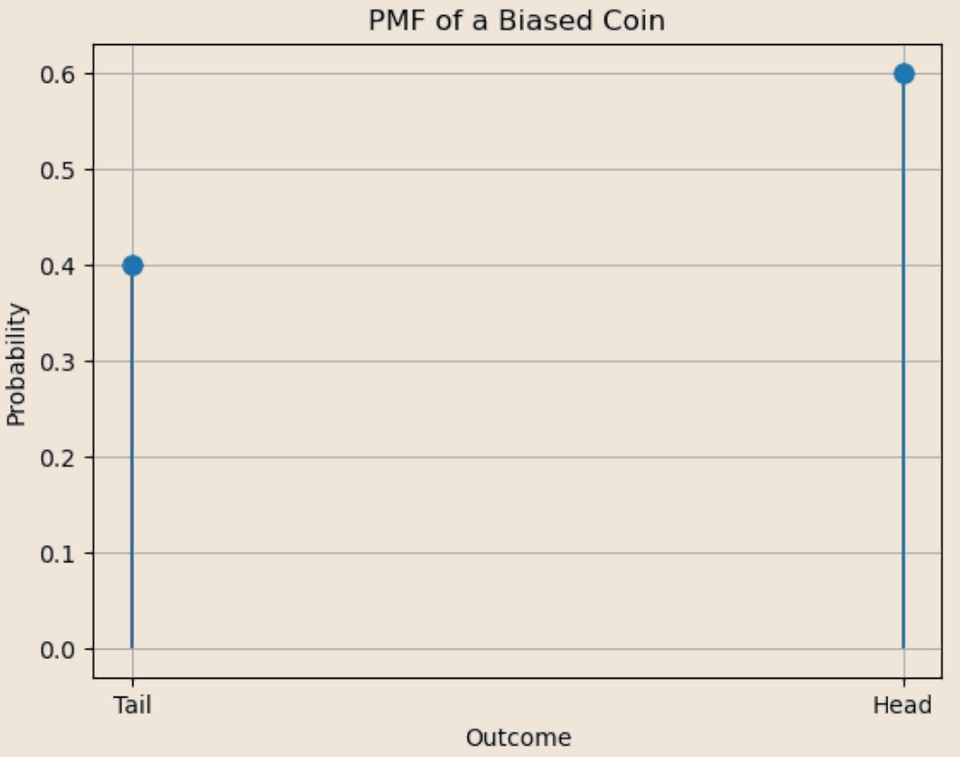
- P(head) = 0.5
- P(tail) = 0.5
- P(head) + P(tail) = 1



Biased Coin

After 10,000 tosses, we observe 6,000 heads and 4,000 tails. Using a frequentist approach:

- P(head) = 6000/10000 = 0.6
- P(tail) = 4000/10000 = 0.4
- Still satisfies: P(head) + P(tail) = 1

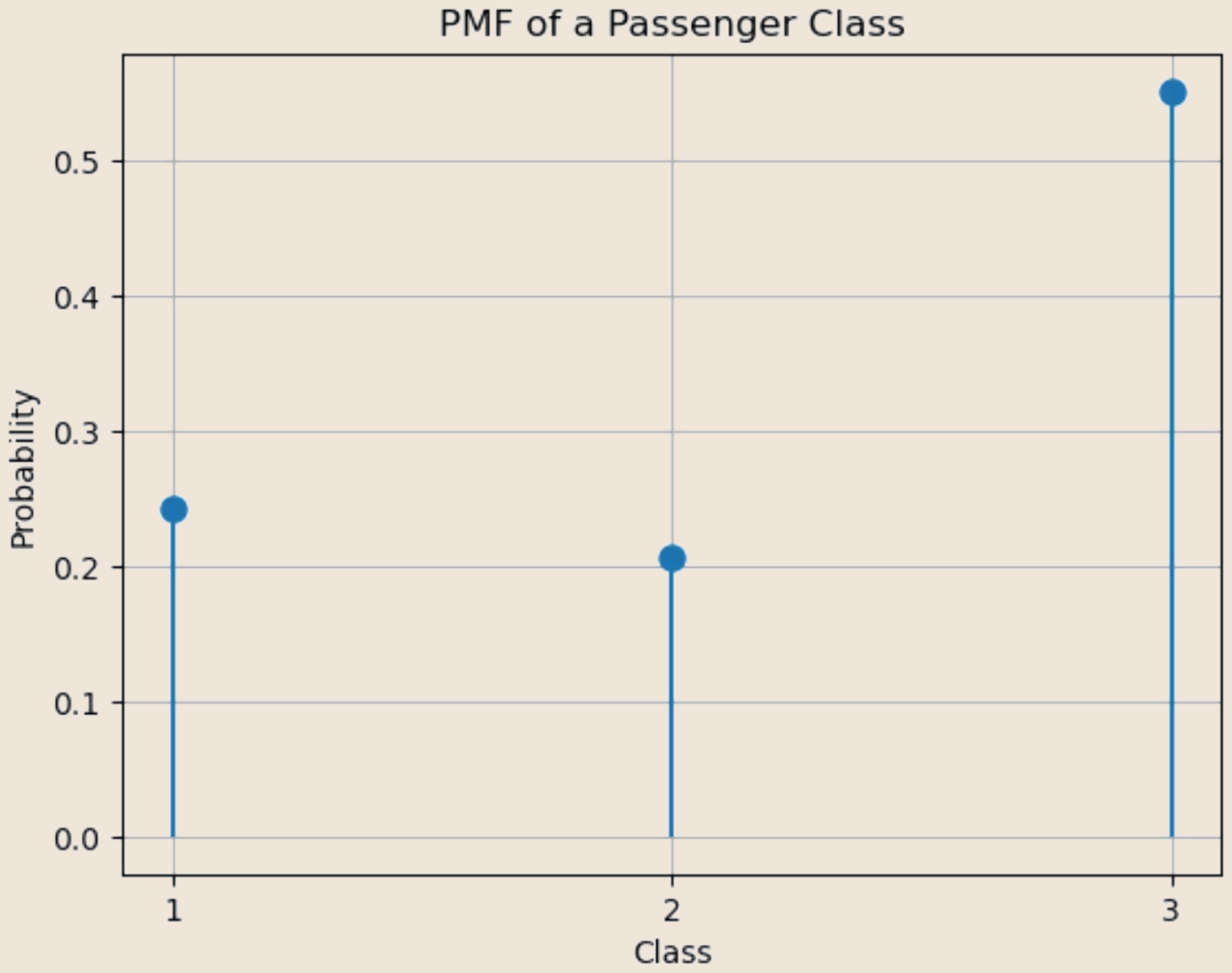


Computing PMFs from Data in Python

```
import pandas as pd
titanic = pd.read_csv('https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.csv'
                      index_col='PassengerId')
pclass = titanic['Pclass'].value_counts(normalize=True).sort_index()

plt.vlines(pclass.index.values, ymin=0, ymax=pclass.values,  linestyles='solid')
plt.plot(pclass.index.values, pclass.values, 'o', markersize=8)
plt.xlabel('Class')
plt.ylabel('Probability')
plt.title('PMF of a Passenger Class')
plt.xticks(pclass.index.values)
plt.grid(True)

# Show the plot
plt.show()
```



Cumulative Distribution Functions (CDF)

For ordered discrete variables, the **CDF** records cumulative probability up to a given value:

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(y)$$

For instance, if H measures height in centimetres, $F(180)$ indicates the probability of selecting someone at most 180 cm tall.

Fair Die: PMF and CDF

Probability Mass Function

For a fair six-sided die:

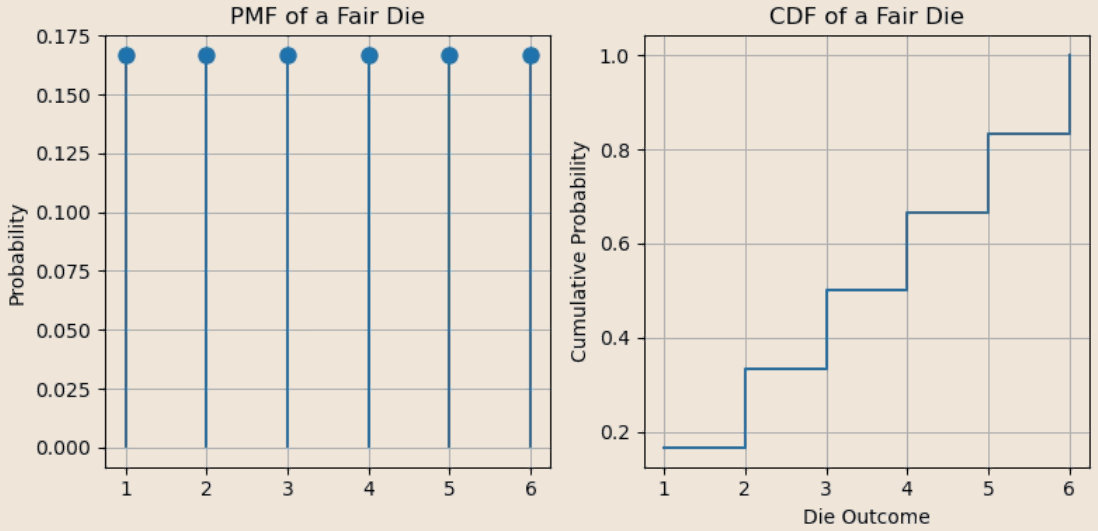
$$f(x) = \frac{1}{6} \text{ for } x \in \{1, 2, 3, 4, 5, 6\}$$

Each outcome has equal probability of occurring.

Cumulative Distribution

The CDF increases in steps:

- $F(1) = \frac{1}{6}$
- $F(2) = \frac{2}{6}$
- $F(3) = \frac{3}{6}$, and so on...



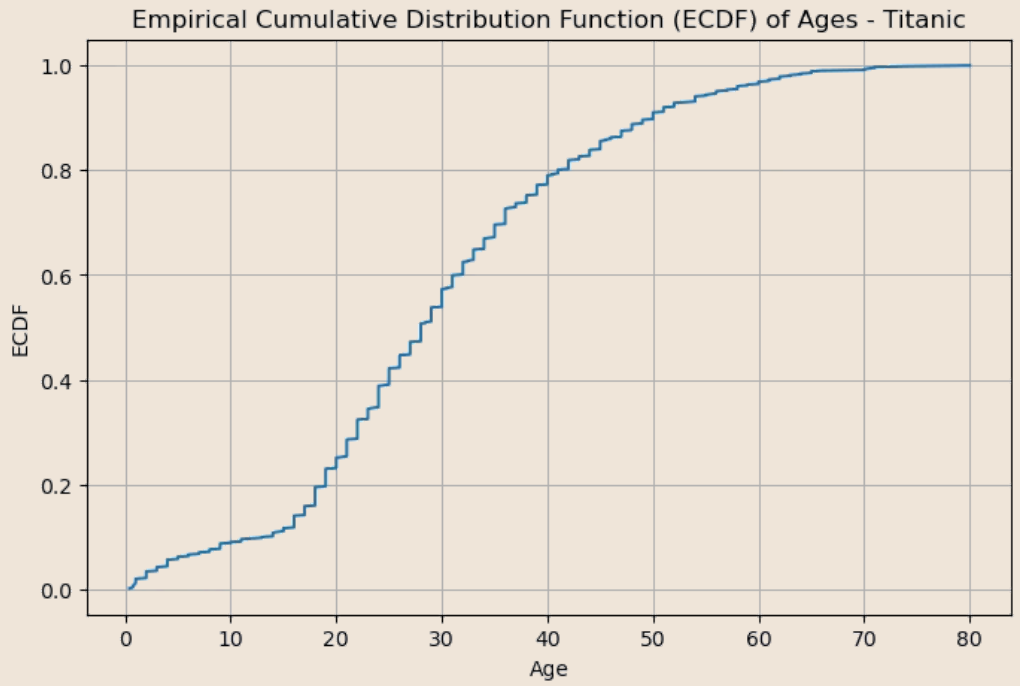
Computing CDFs from Data in Python

In Python, we compute CDFs from data as seen in the case of ECDFs:

```
ages = titanic['Age'].dropna()
ages_sorted = np.sort(ages)

ecdf = np.arange(1, len(ages_sorted)+1) /
len(ages_sorted)

# Plot the ECDF
plt.figure(figsize=(8, 5))
plt.plot(ages_sorted, ecdf)
plt.xlabel('Age')
plt.ylabel('ECDF')
plt.grid(True)
plt.show()
```



Probability Density Functions (PDF)

For continuous random variables, we cannot assign probability to exact values (as $P(X = x) = 0$). Instead, we define a **density function**

$$f : \Omega \to [0, 1]$$

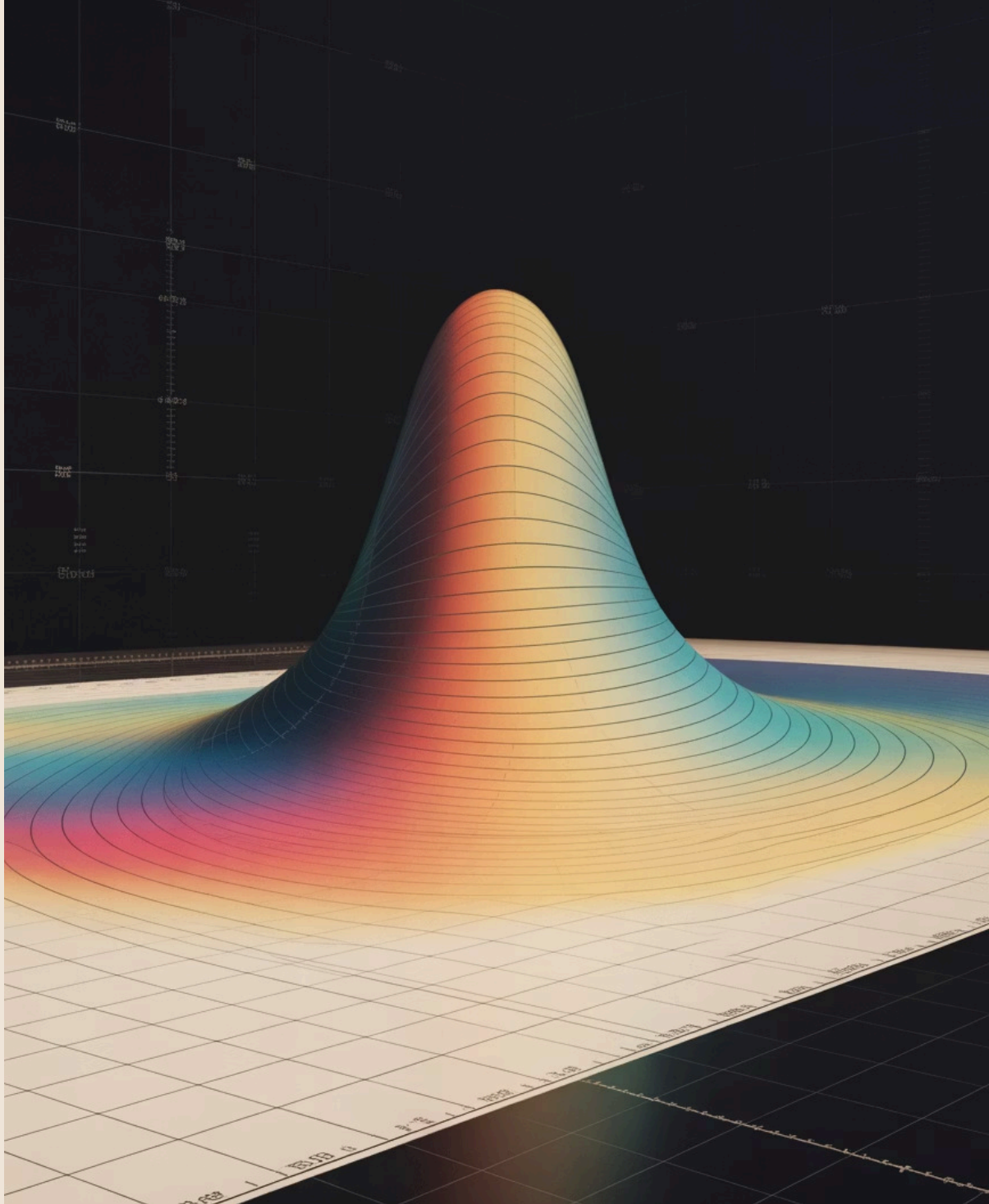
Such that:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

We define probability from density with integration:

$$P(a \leq x \leq b) = \int_{a}^{b} f(x)dx$$

Think of density as "concentration of probability"—higher density means values are more frequently observed in that region.



Uniform Distribution

A random number generator producing values between a and b follows a **uniform distribution**:

$$P(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

01

Zero outside range

$P(x) = 0$ for $x < a$ or $x > b$

02

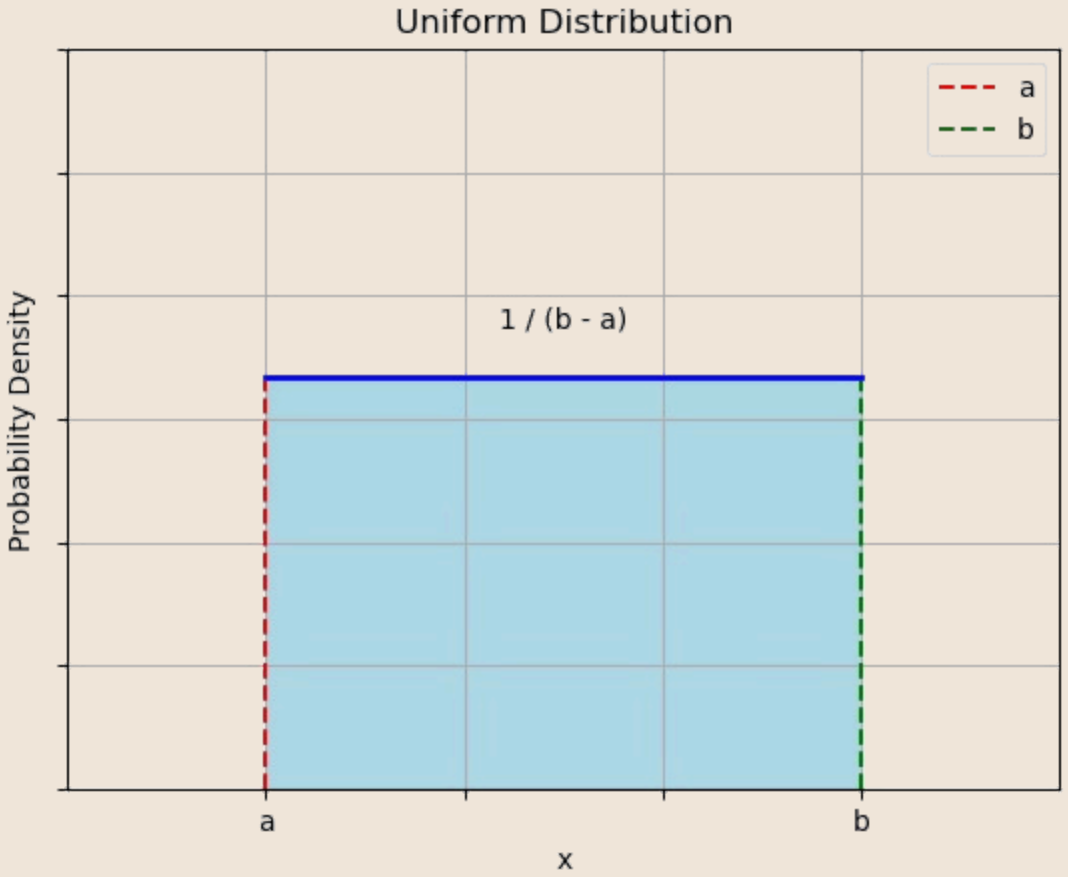
Constant within range

Probability density is uniform across the interval

03

Normalised total

Total area under the curve equals exactly 1 (can be shown)



Approximating PDFs from Data

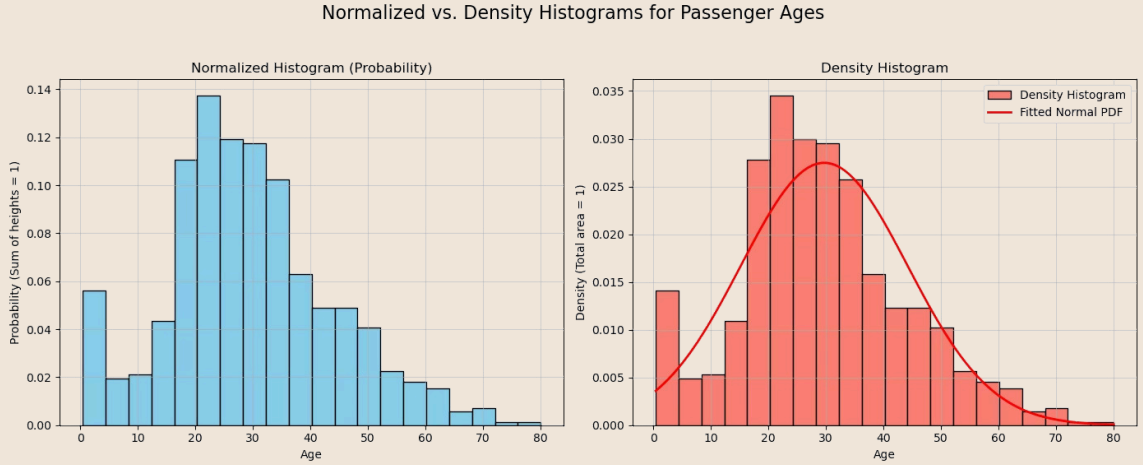
With real data (like Titanic passenger ages), we don't know the true PDF. We approximate it using **density histograms** where the total area of bars equals 1.

In a **density histogram**, we define the height of bin b_j as follows:

$$h_j = \frac{c_j}{w_j \cdot n}$$

This ensures that the area under the histogram H is equal to 1:

$$\int_{-\infty}^{\infty} H(x)dx = \sum_j A_j = \sum_j w_j \frac{c_j}{w_j \cdot n} = \frac{1}{n} \sum_j \frac{w_j}{w_j} \cdot c_j = 1$$




```
ax2.hist(ages, bins=20, density=True)
```

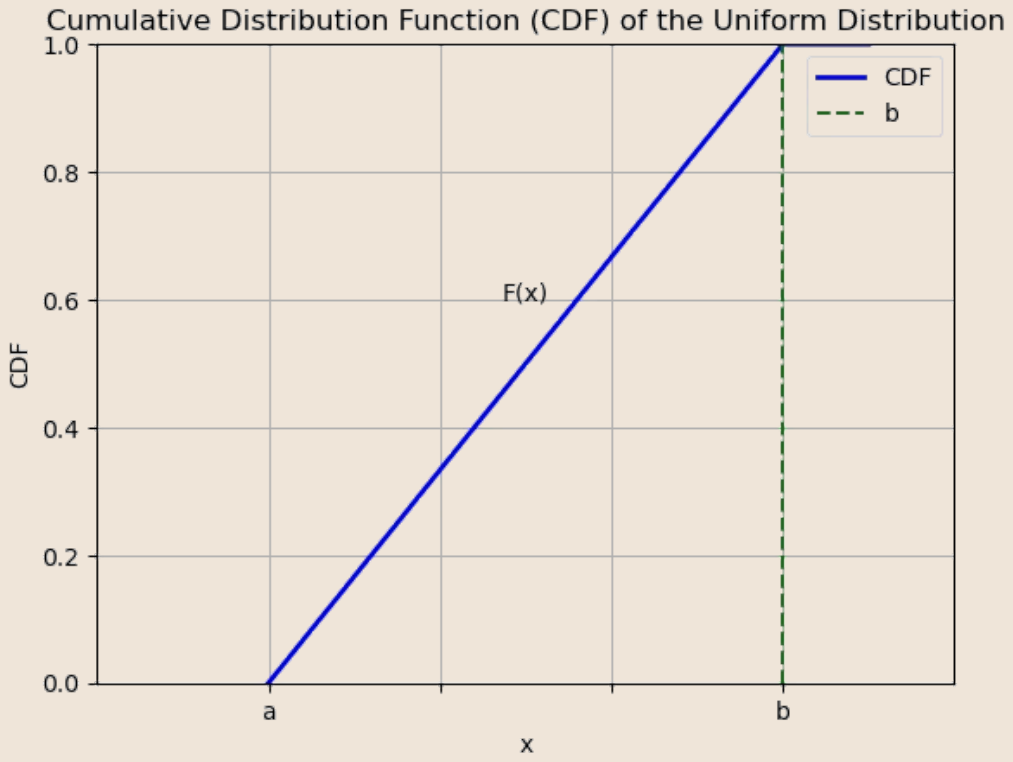
Cumulative Distribution Functions (CDF)

We can define Cumulative Distribution Functions (CDFs) for continuous random variables using their density functions:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

Example: The CDF of the uniform distribution will be given by:

- $F(x) = 0, x < a$
- $F(x) = \frac{x-a}{b-a}, a \leq x \leq b$
- $F(x) = 1, x > b$



Common Probability Distributions

There are several common probability distributions which can be used to describe random events. **These distributions have an analytical formulation which depends generally on one or more parameters.**

When we have **enough evidence that a given random variable is well described by one of these distributions**, we can simply “fit” the distribution to the data (i.e., choose the correct parameters for the distribution) and use the analytical formulation to deal with the random variable.

It is hence useful to know the **most common probability distributions** so that we can recognise the cases in which they can be used.

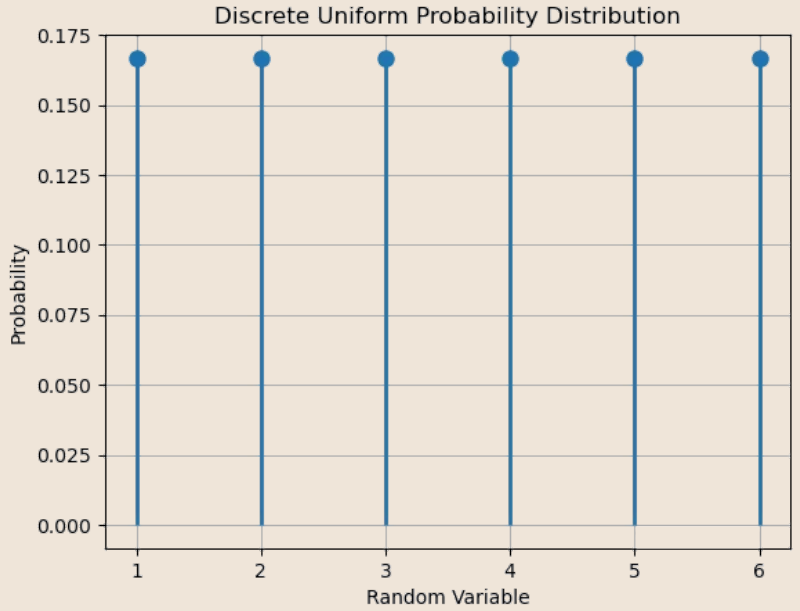
Discrete Uniform Probability Distribution

The discrete uniform distribution is controlled by a parameter $k \in \mathbb{N}$ (number of outcomes) and assumes that all outcomes have the same probability of occurring:

$$P(X = a_i) = \frac{1}{k}$$

where $\Omega = \{a_1, \ldots, a_k\}$

Example: probability of the outcomes of a fair die. Here $k = 6$ and $P(a_i) = \frac{1}{6}$



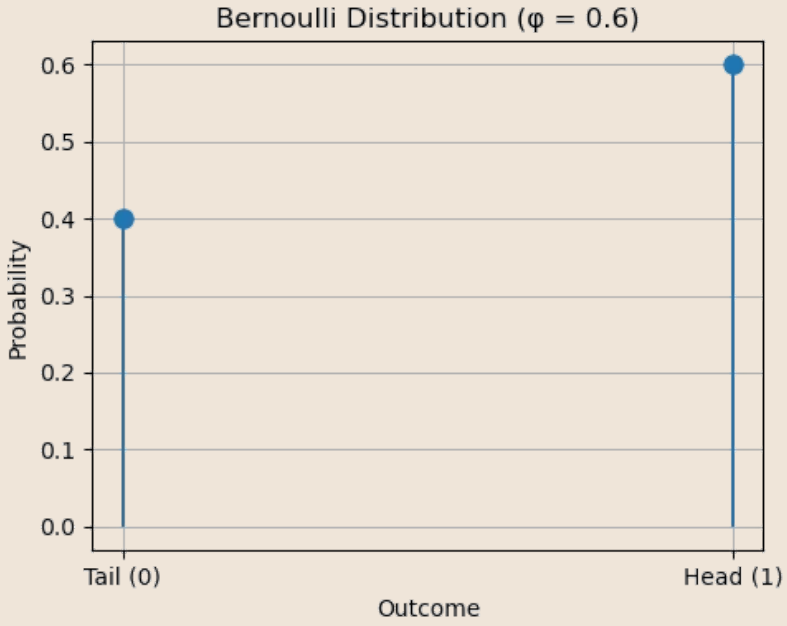
Bernoulli Distribution

The Bernoulli distribution is a discrete distribution over a **single binary random variable**, i.e., the variable X can take only two values: $\{0, 1\}$

The distribution is controlled by a single parameter $\phi \in [0, 1]$ which denotes the probability of the variable to be equal to 1 (the outcome to be positive).

The analytical formulation of the Bernoulli distribution is very simple:

$$P(X = 1) = \phi$$
$$P(X = 0) = 1 - \phi$$



Example: A skewed coin lands on "head" 60% of the times. If we define $X = 1$ when the outcome is "head" and $X = 0$ when the outcome is tail, then the variable follows a Bernoulli distribution with $\phi = 0.6$.

Binomial Distribution

The binomial distribution answers the question:

> What's the probability of exactly k successes in n independent trials?

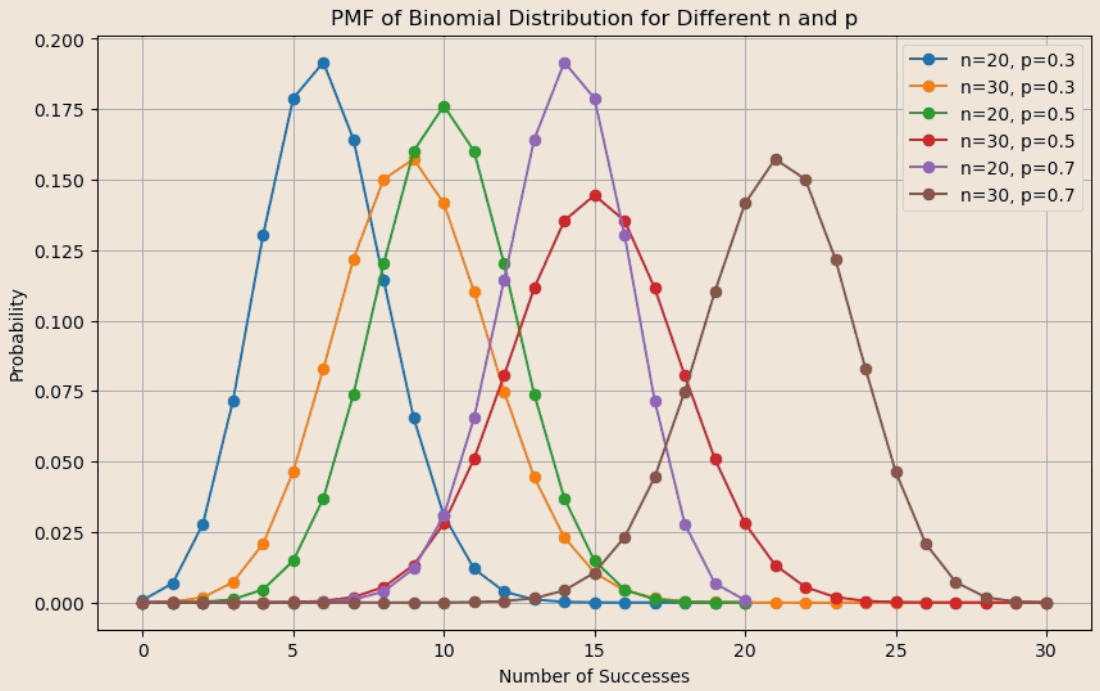
It is defined as:

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

Where:

- k is the number of successes
- n is the number of independent trials
- p is the probability of a success in a single trial

Example: Tossing a fair coin three times for three heads: $P(3) = \binom{3}{3}(0.5)^3 = 0.125$



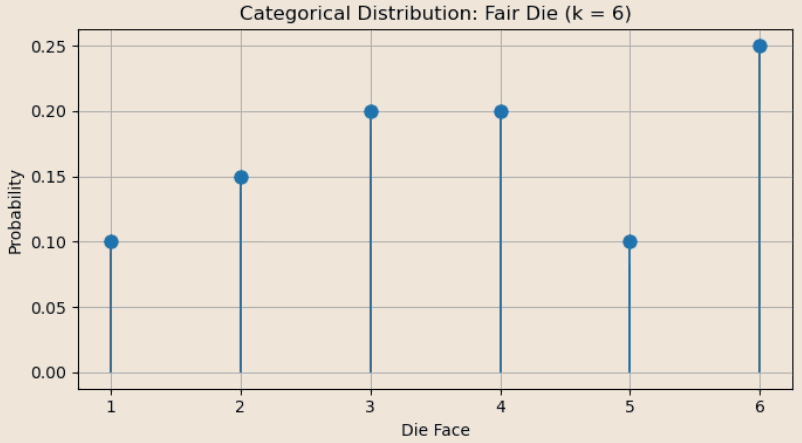
Categorical Distribution

The Categorical distribution is a generalisation of the Bernoulli distribution for a single random variable with more than two possible discrete outcomes. Instead of just two outcomes (0 or 1), a categorical random variable can take on one of k possible outcomes, where each outcome has its own probability.

It is controlled by a set of parameters $p_1, p_2, \ldots, p_k$, where p_i is the probability of outcome i, and $\sum_{i=1}^{k} p_i = 1$.

$$P(X = i) = p_i \quad \text{for } i \in \{1, 2, \ldots, k\}$$

Example: Rolling a single biased six-sided die. If the die is biased such that the probabilities of rolling a 1, 2, 3, 4, 5, or 6 are not all equal, say $P(X = 1) = 0.1$, $P(X = 2) = 0.15$, $P(X = 3) = 0.2$, $P(X = 4) = 0.2$, $P(X = 5) = 0.1$, $P(X = 6) = 0.25$, then the outcome follows a Categorical distribution with $k = 6$ and the specified probabilities.



Multinomial Distribution

The Multinomial distribution generalise the Binomial distribution to the case of more than two possible outcomes per trial (categorical trials). It describes the probability of obtaining a specific combination of counts for each category after conducting a fixed number of independent trials.

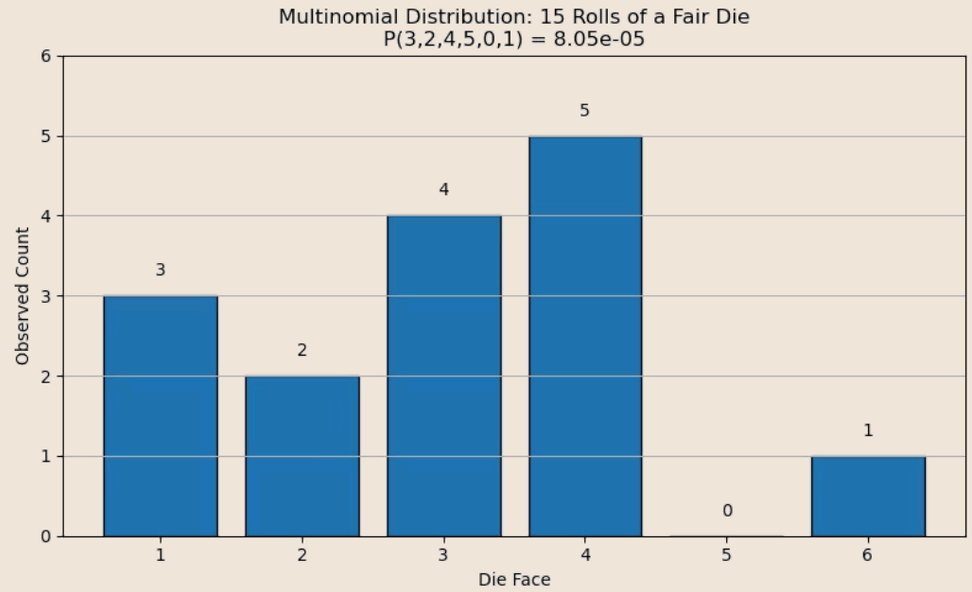
It is controlled by n (total number of trials) and a set of probabilities $p_1, p_2, \ldots, p_k$ for each of the k possible outcomes, where $\sum_{i=1}^{k} p_i = 1$.

$$P(n_1, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Where n_i is the number of times outcome i occurs, such that $\sum_{i=1}^{k} n_i = n$.

Example: Given a fair die with 6 possible outcomes, the probability of getting 3 times 1, 2 times 2, 4 time 3, 5 times 4, 0 times 5, and 1 time 6, rolling the dice for 15 times is given by:

$$P(3,2,4,5,0,1) = \frac{15!}{3!2!4!5!0!1!} \cdot 6^{-3} \cdot 6^{-2} \cdot 6^{-4} \cdot 6^{-5} \cdot 6^{0} \cdot 6^{-1} = 8.04 \cdot 10^{-5}$$



Discrete Probability Distributions - Recap

Discrete Uniform

Models a finite set of equally likely outcomes.

$$P(a_i) = \frac{1}{k} \qquad \Omega = \{a_1, \ldots, a_k\}$$

> **Example**: Probability of obtaining a given face when rolling a fair six-sided die

Bernoulli Distribution

Models a single binary trial with parameter ϕ.

$$P(X = 1) = \phi \qquad P(X = 0) = 1 - \phi$$

> **Example:** Probability of obtaining hard or tail in a single coin toss (biased or unbiased)

Binomial Distribution

Models k successes in n independent Bernoulli trials:

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

> **Example**: Probability of obtaining a given number of heads in multiple coin tosses (biased or unbiased)

Categorical Distribution

Generalises Bernoulli to k states with probabilities $p_1, \ldots, p_k$ where $\sum p_i = 1$.

> **Example**: Probability of obtaining a given face in a single die roll (biased or unbiased)

Multinomial Distribution

Models outcomes of n independent categorical trials:

$$P(n_1, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

> **Example**: Number of occurrences of each face of a die when rolled multiple times (biased or unbiased)

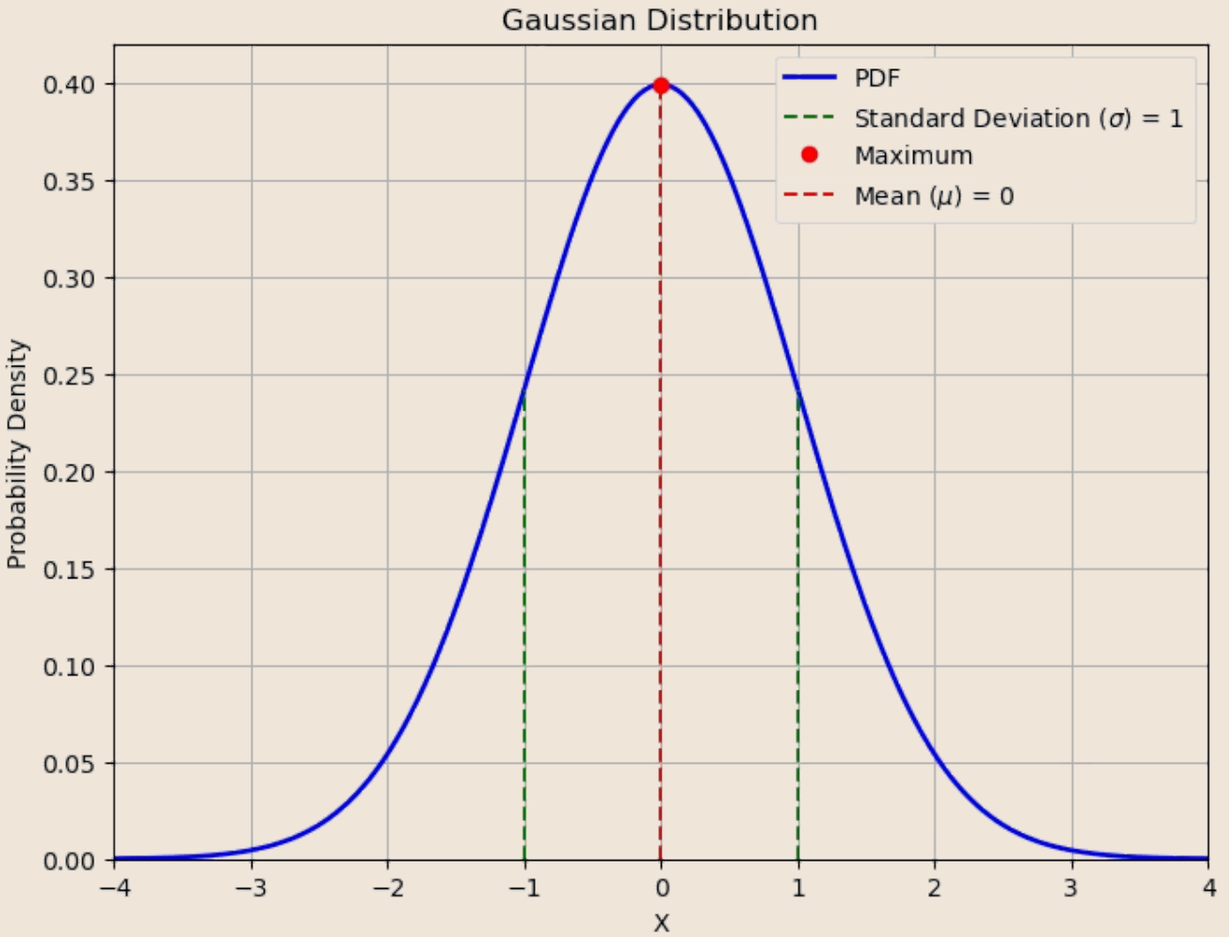
The Gaussian (Normal) Distribution

The most important continuous distribution, characterised by mean μ and standard deviation σ:

$$N(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

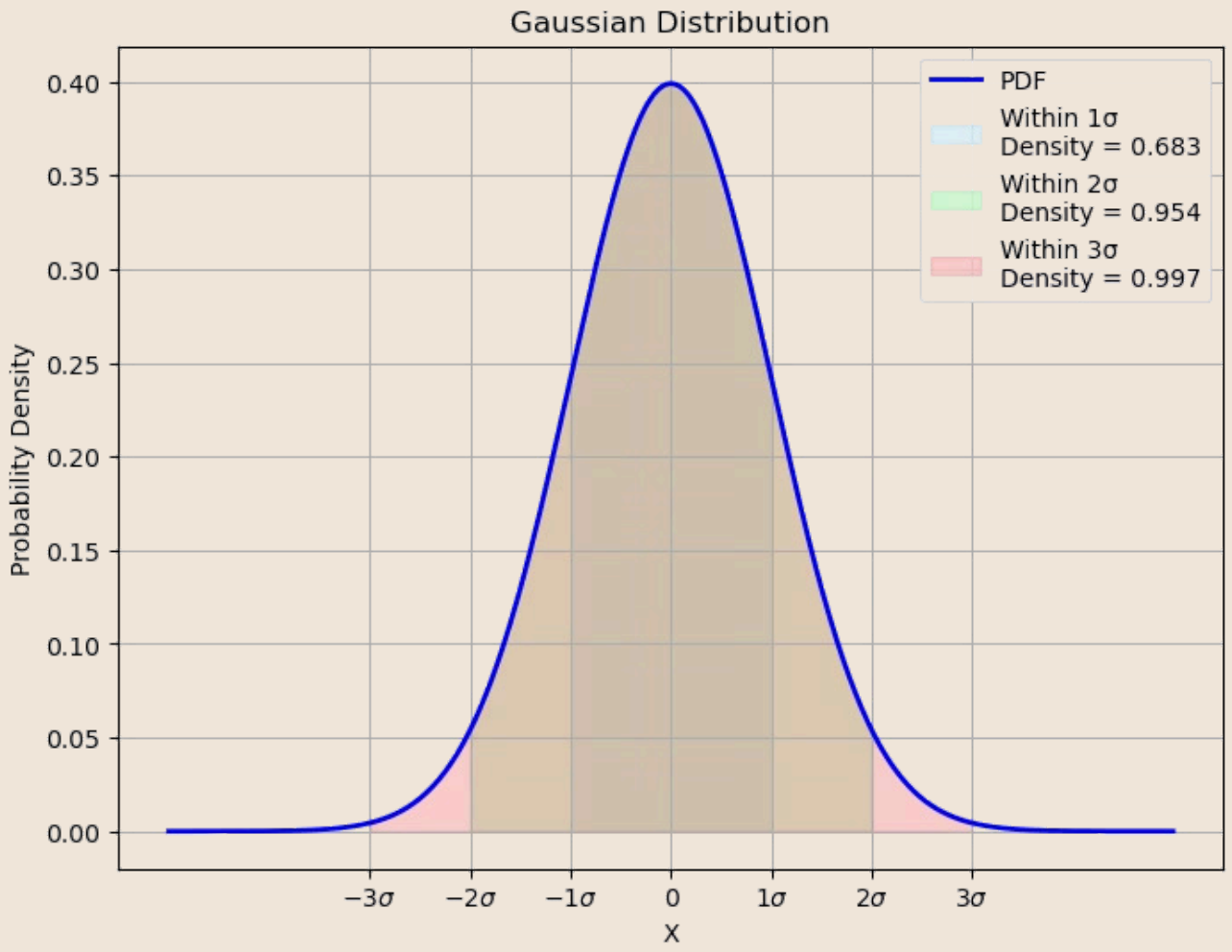
Describes many natural phenomena. Here:

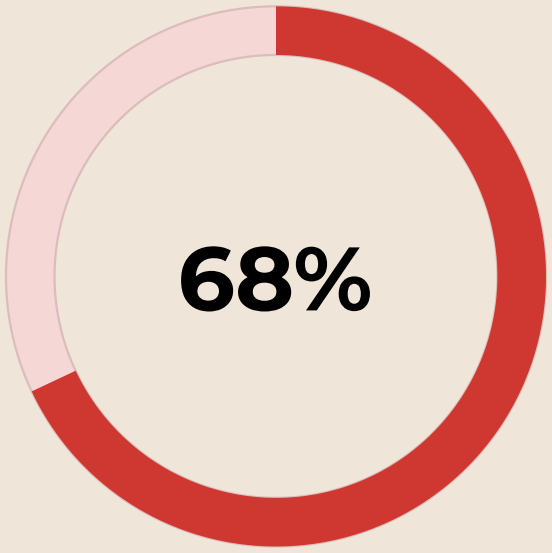
- μ is the most probably element of the distribution - the peak or mode of the distribution
- σ^2 represents how much values concentrate around the mean



Example: In a large population, adult human heights roughly follow a Gaussian (Normal) distribution. Most people cluster around the average height, with fewer individuals being extremely short or extremely tall.

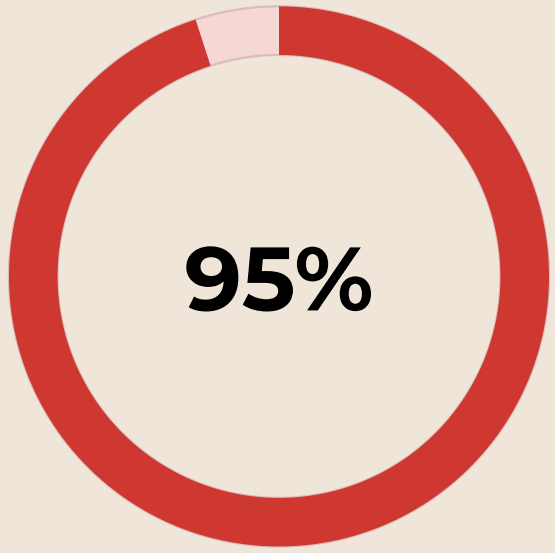
Density of the Gaussian Distribution





Within ±1σ

Approximately 68% of data falls within one standard deviation of the mean



Within ±2σ

Approximately 95% of data falls within two standard deviations

Within ±3σ

Nearly all data (99.7%) falls within three standard deviations

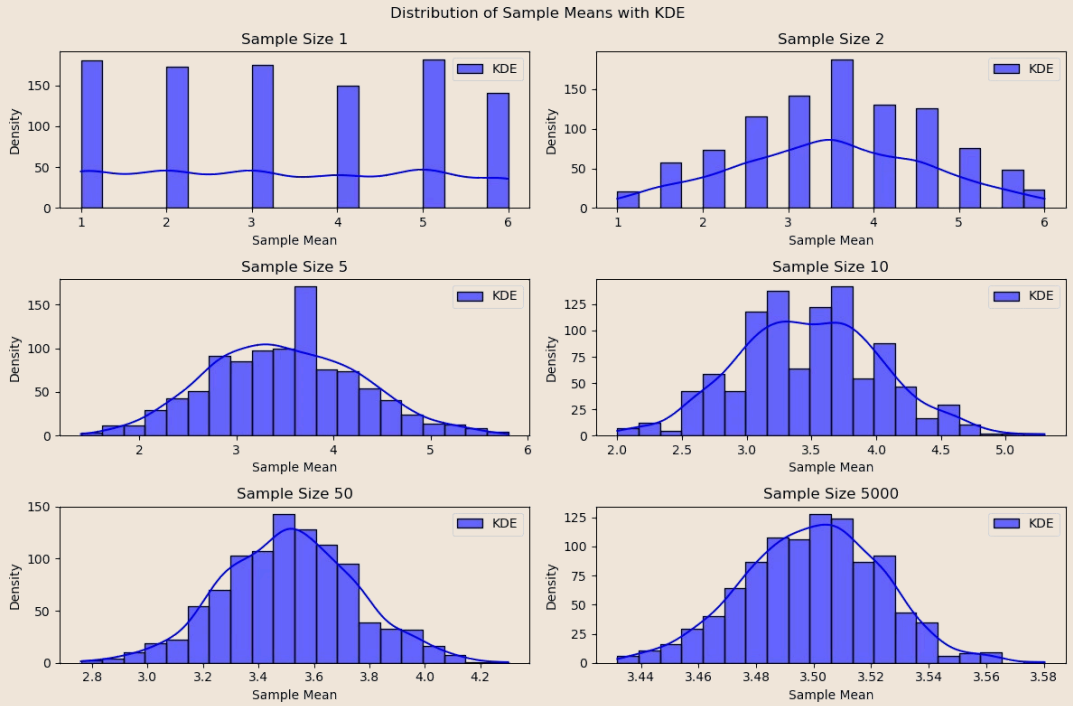
Why the Gaussian Matters: Central Limit Theorem

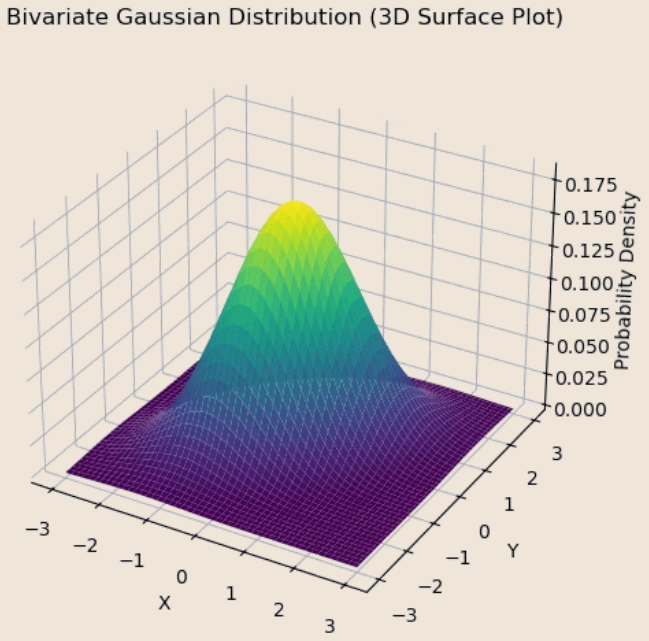
The **Central Limit Theorem (CLT)** explains why Gaussian distributions appear everywhere in nature and data analysis.

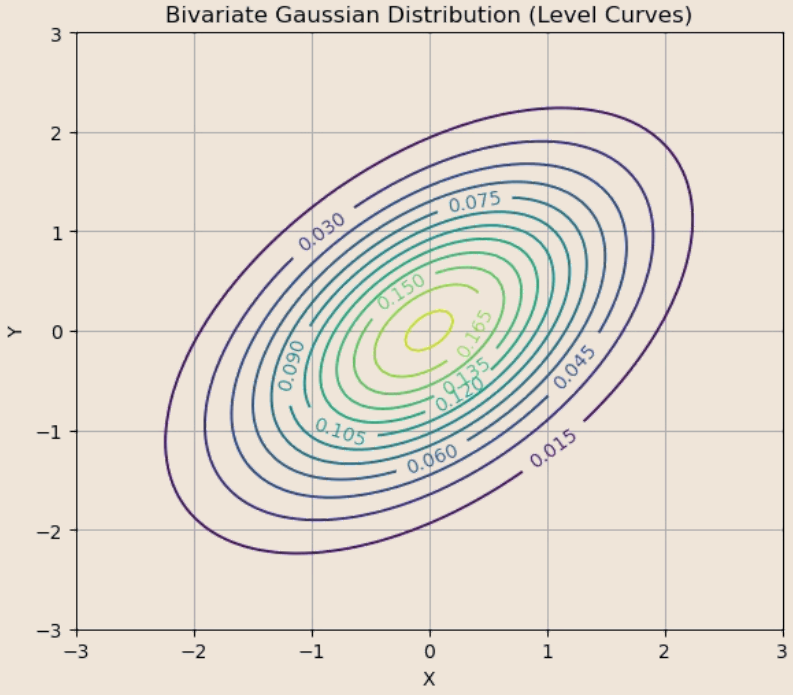
Key insight: The sum (or average) of many **independent random variables** tends towards a normal distribution, regardless of their original distribution.

As the number of variables increases, convergence to the Gaussian becomes more accurate.

Example: Rolling increasing numbers of dice and averaging the results produces increasingly Gaussian-like distributions, even though individual die rolls are uniform.







Multivariate Gaussian Distributions

Gaussian distributions extend to multiple dimensions (e.g., d), parametrised by mean vector μ and covariance matrix Σ:

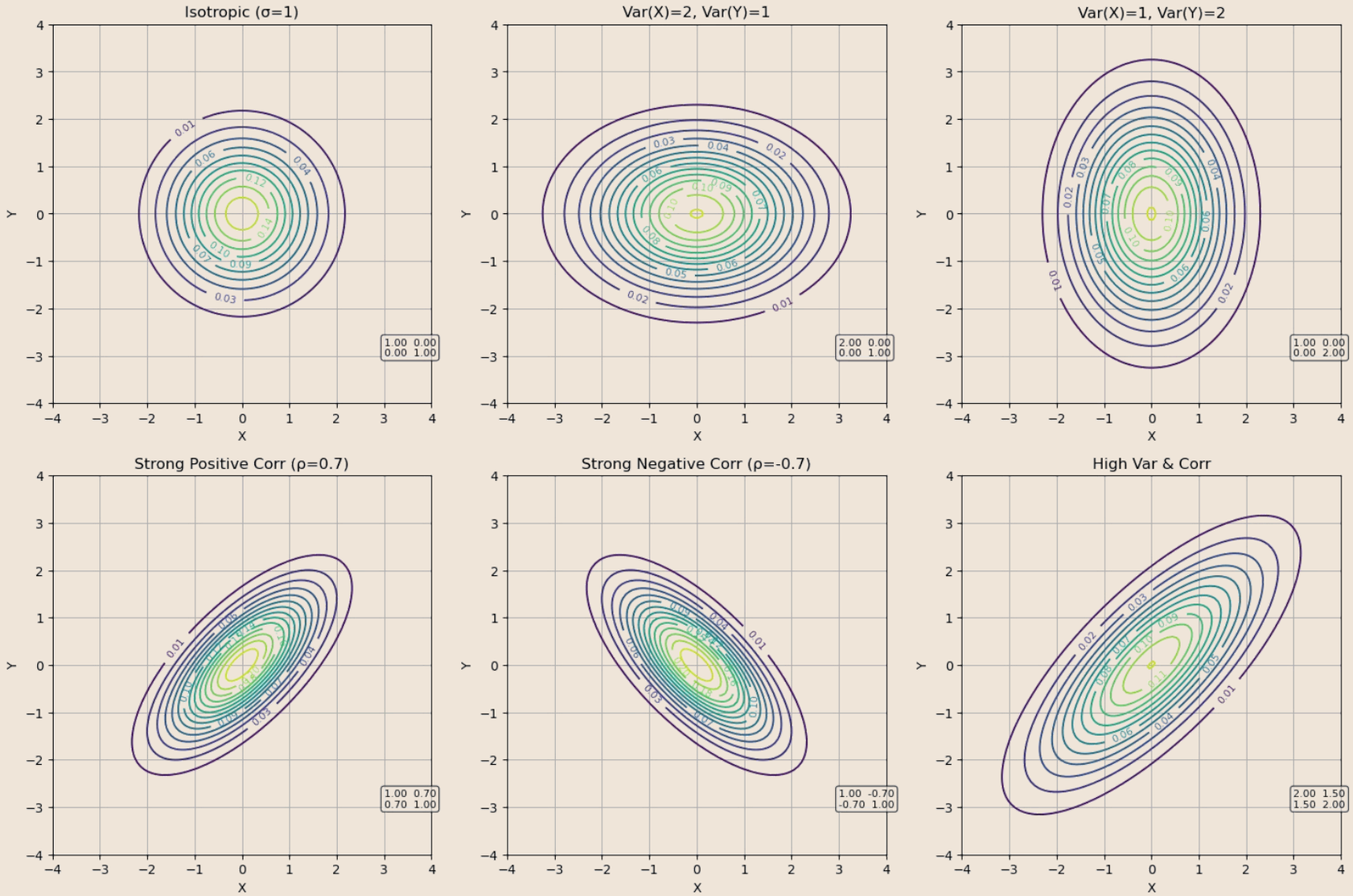
$$N(x;\mu,\Sigma) = \sqrt{\frac{1}{(2\pi)^d \det(\Sigma)}} e^{(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))}$$

When $d = 2$, we can visualize the Gaussian distribution as a 3D structure, where the X-Y plane contains the input values, and the Z axis represents the density values associated with each element. The 3D structured is still bell-shaped.

Note that the points still lie in a 2D plane. It is also common to represent the Gaussian distribution with a contour plot, where a line connects points in the 2D space which have the same density.

The covariance matrix Σ is $d \times d$ controls the distribution's shape, orientation, and spread. Diagonal elements represent variances along each axis, whilst off-diagonal elements capture correlations between variables.

Effect of Σ in a Multivariate Gaussian



- When the matrix is diagonal, the distribution is "isotropic" (simmetric along both axes);
- High variances deform the shape by controlling individual axes;
- Off-diagonal values introduce correlations between the axes and add tilt.

Estimation of the Parameters of a Gaussian

Given some data (remember, data is values assumed by random variables!), we can obtain the parameters of the Gaussian distribution related to the data with a **maximum likelihood** estimation. In the univariate case, we will compute:

Mean

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

In the multivariate case:

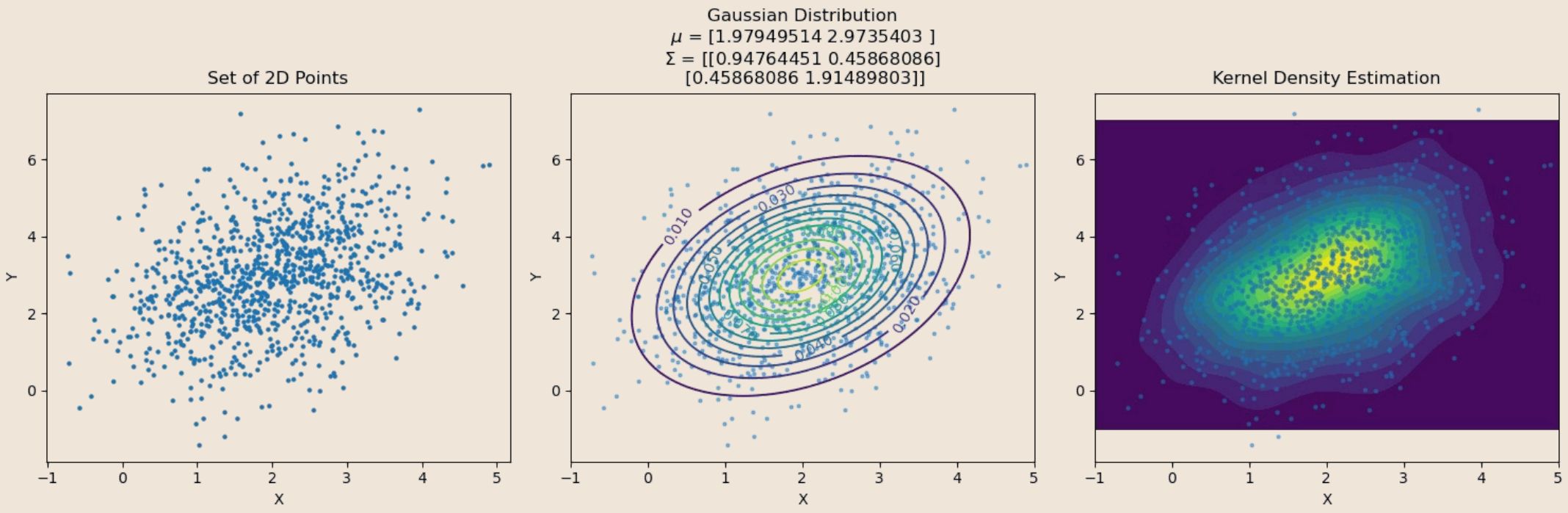
Mean

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

Covariance Matrix

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T$$

The diagram below shows an example in which we fit a Gaussian to a set of data and compare it with a 2D KDE of the data.



Describing Distributions: Key Statistics

Expectation (Mean)

Variance

Covariance

Entropy

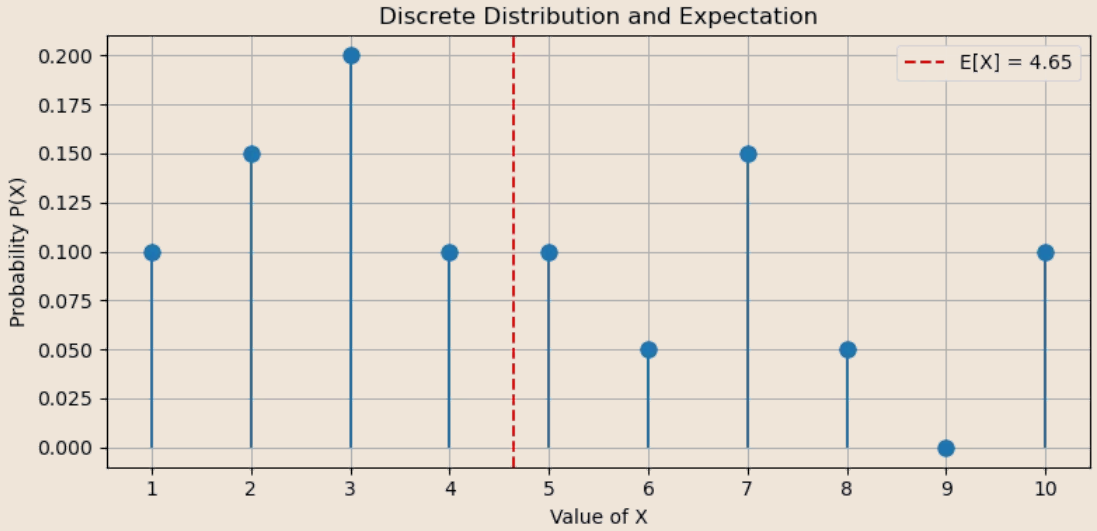
Expectation

The expectation of a random variable is similar to the concept of mean for a sample. Differently from the mean, however, here we weigh more outcomes that are more likely to be observed. In the discrete case:

$$E_{X \sim P}[X] = \sum_{x \in \Omega} xP(x)$$

In the continuous case, the sum becomes an integral:

$$E_{X \sim P}[X] = \int_{x \in \Omega} xf(x)dx$$



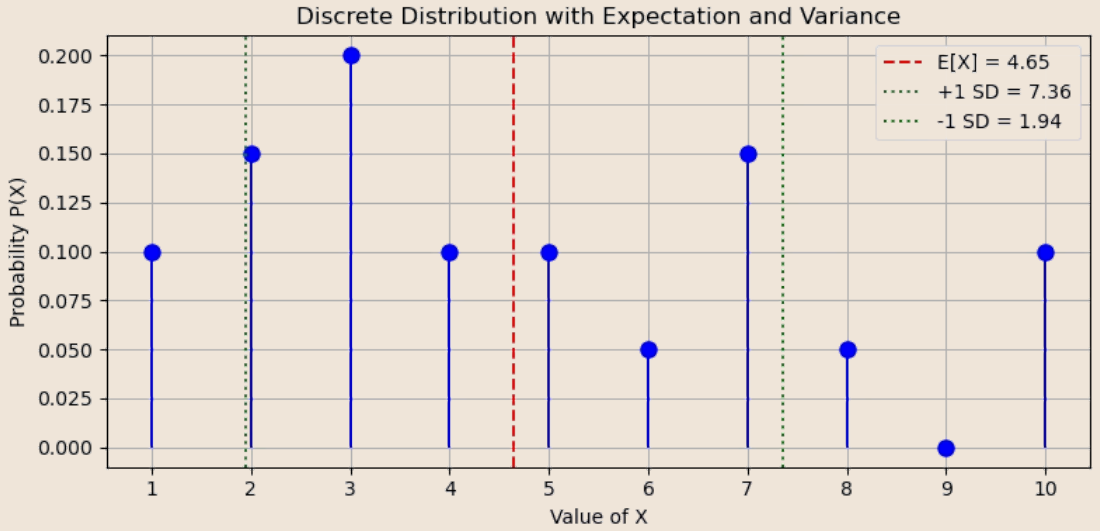
Variance and Standard Deviation

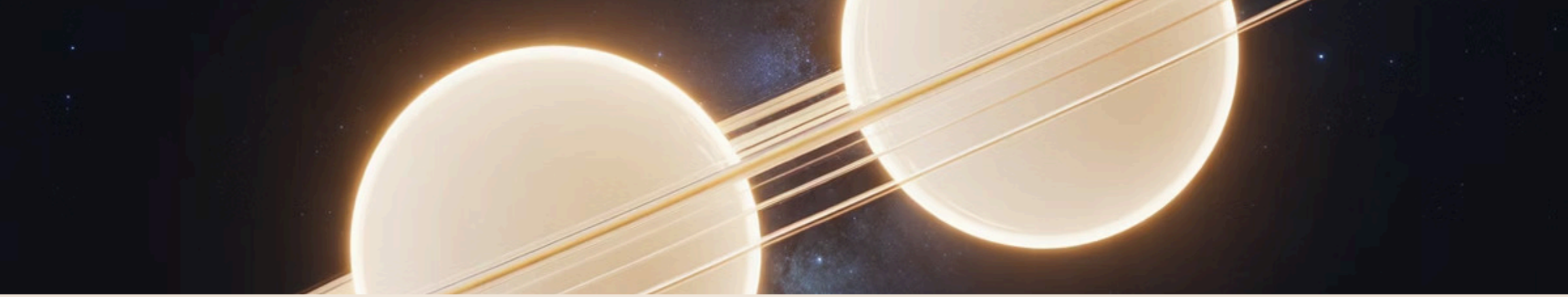
Similarly to the concept of variance of a sample, we describe the variance of a distribution as follows:

$$Var_{X\sim P}[X] = E_{X\sim P}[(X - E_{X\sim P}[X])^2]$$

We define the standard deviation as:

$$Std_{X\sim P}[X] = \sqrt{Var_{X\sim P}[X]}$$





Covariance

Given two random variables $X \sim P_X$ and $Y \sim P_Y$, covariance measures how they are **linearly correlated:**

$$Cov_{X \sim P_X, Y \sim P_Y}(X, Y) = E_{X \sim P_X, Y \sim P_Y}\left[\left(X - E_{X \sim P_X}[X]\right)\left(Y - E_{Y \sim P_Y}[Y]\right)\right]$$

If X is a multi-dimensional variable $X = [X_1, X_2, \ldots, X_n]$, we can compute all the possible covariances between variable pairs: $Cov[X_i, X_j]$. This allows to create a matrix, which is generally referred to as **the covariance matrix**. The general term of the covariance matrix $Cov(X)$ is given by:

$$Cov(X)_{i,j} = \Sigma_{ij} = Cov(X_i, X_j)$$

Self-Information

Self-information aims to quantify the level of information carried by a simple event, based on two observations:

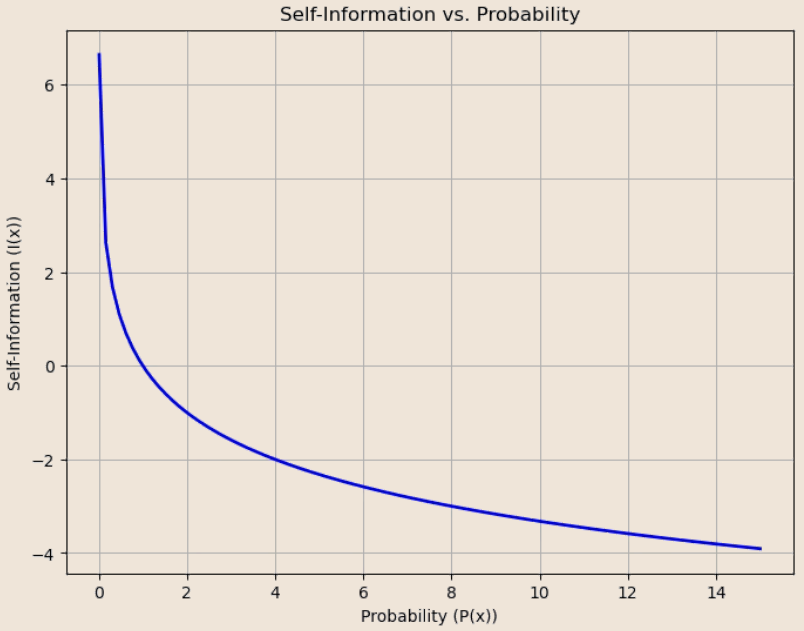
1. Rare events carry more information (e.g., raining in the desert vs raining in London);
2. Self-information of independent events should be additive;

This is defined as follows:

$$I(x) = -\log_2 P(x)$$

The monotonicity of the logarithm ensures observation 1) is valid, while properties on the produce of logarithms confirms observation 2).

Self-information is measures in bits if the logarithm is base 2 and nats if the logarithm is base e.



Entropy: Measuring Uncertainty

Entropy is defined as the expected self-information. For discrete variables:

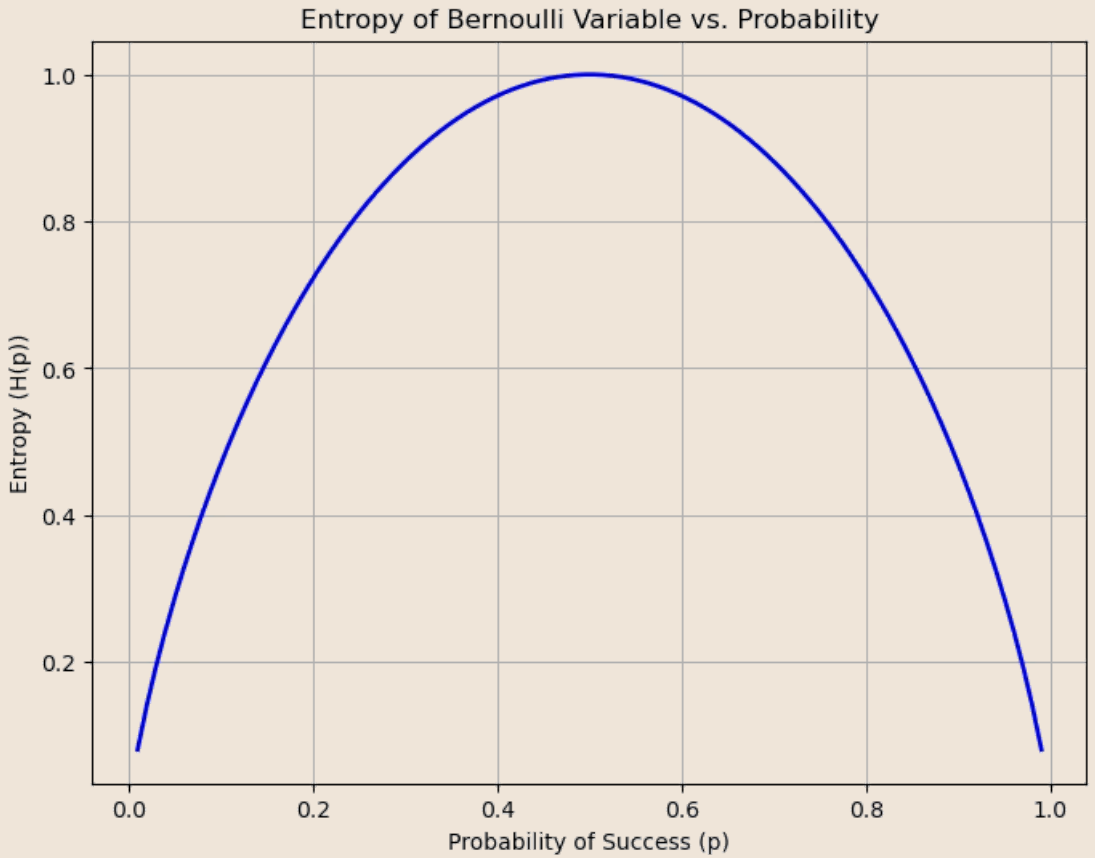
$$H(X) = -\sum_{x \in \Omega} P(x) \log_2 P(x)$$

For continuous ones:

$$h(X) = -\int_{x \in \Omega} f(x) \log_2 f(x) dx$$

Examples:

- Fair coin: $H(X) = 1$ bit (maximum uncertainty)
- Biased coin (99% heads): $H(X) \approx 0.08$ bits (low uncertainty)



Describing Distributions: Key Statistics (Recap)

Expectation (Mean)

The weighted average of all possible values:
$E[X] = \sum xP(x)$ for discrete variables.

Variance

Measures spread around the mean:
$Var[X] = E[(X - E[X])^2]$.
Standard deviation σ is the square root of variance.

Covariance

Measures how two variables change together:
$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$
.

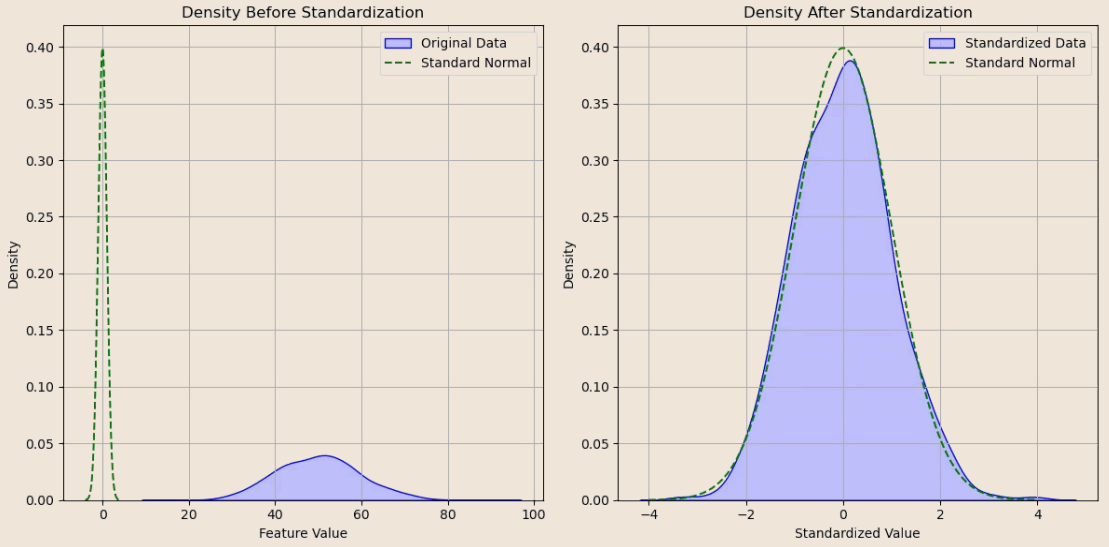
Entropy

Quantifies uncertainty:
$H(X) = -\sum P(x) \log_2 P(x)$.
Higher entropy means more unpredictability.

Standardisation: Creating Comparable Scales

Standardisation transforms any random variable X to have mean 0 and variance 1:

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - E[X]}{\sqrt{Var[X]}}$$

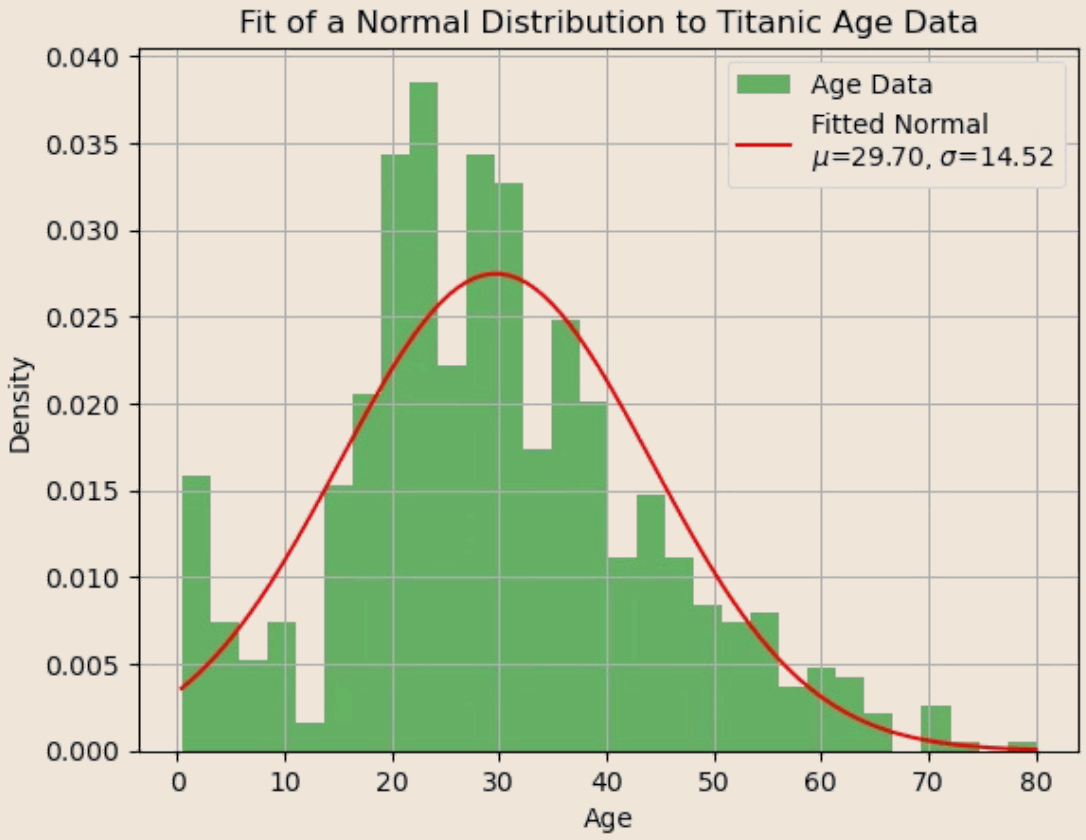


Fitting Distributions to Real Data in Python

In practice, we estimate distribution parameters from observed data using techniques like **maximum likelihood estimation**.

Python's scipy.stats provides automated fitting functions.

Remember: Choose distributions based on data characteristics and domain knowledge. The Gaussian is common but not always appropriate—discrete data may need binomial, categorical, or other distributions.



Conclusions and Next Steps

We Have Explored:

- Data distributions to model uncertainty
- Probability Mass Functions
- Probability Density Functions
- Common probability distributions
- Expected values, variances, covariances, entropy

In next lectures, we will see how probability distributions are used in inferential statistic

References

- Parts of chapter 1 of [1];
- Most of chapter 3 of [2];
- Parts of chapter 8 of [3].

[1] Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[2] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. https://www.deeplearningbook.org/

[3] Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.