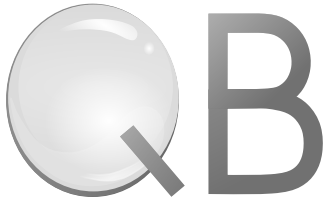
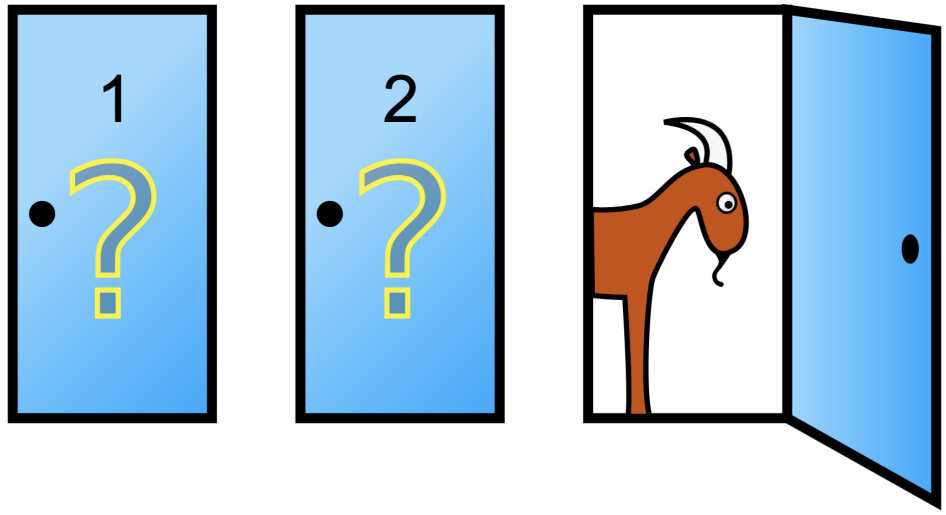


Fundamentals of Quantitative Biology

STATS 01: Basics of Probabilities Theory and Distributions

summer semester 2025

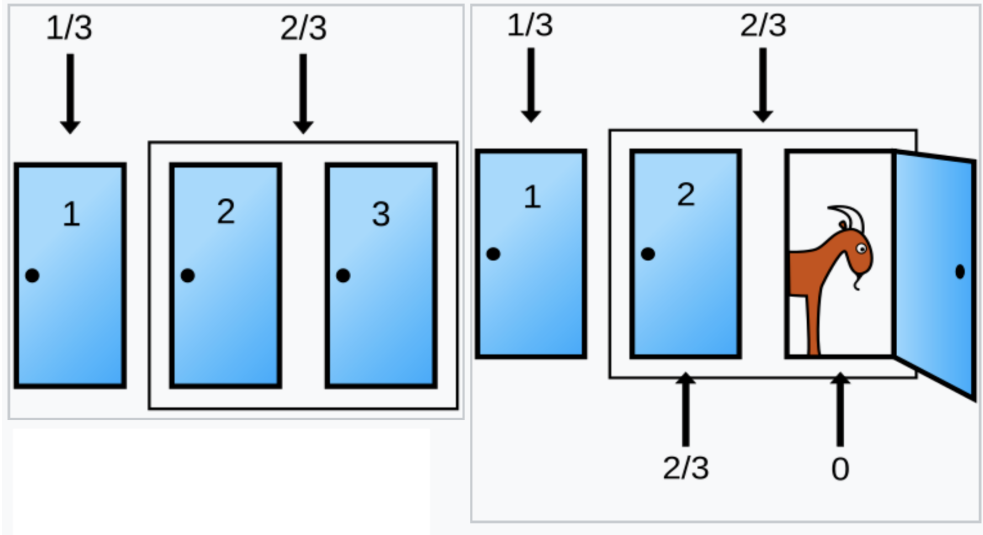
The Monty Hall problem



You pick one of three doors, behind one is a car and the others have goats. After your choice, the host, who knows what's behind the doors, opens another door showing a goat. You then choose to stick or switch doors.
Would you switch or stick to your initial choice?

image from `wikipedia.org/wiki/Monty_Hall_problem`

The Monty Hall problem



Indeed, switching is the better option!

image from wikipedia.org/wiki/Monty_Hall_problem

Random events

Random phenomenon

In a **random phenomenon** the individual outcomes are unpredictable, but over many independent repetitions a regular pattern (or distribution) emerges.

Examples:

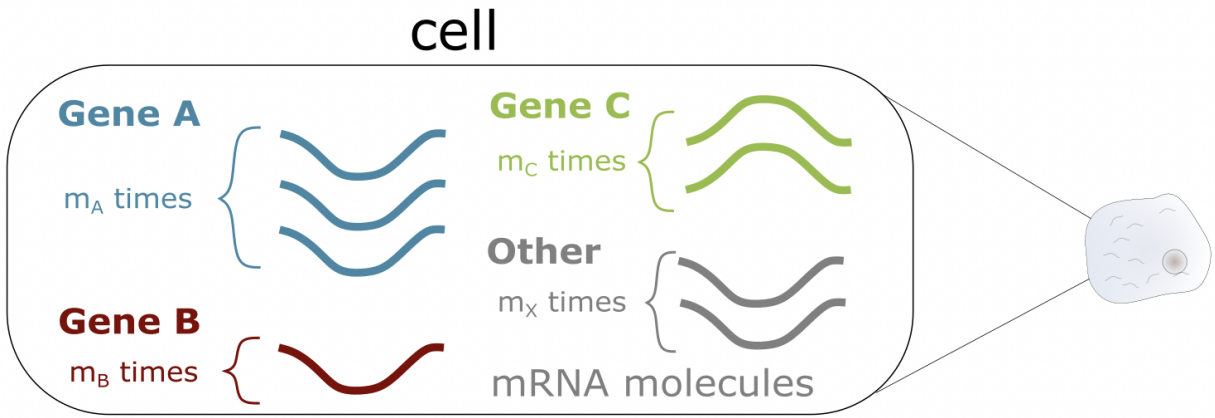
- color of a passing car (random)
- time of sunset (deterministic)
- bus arrival (in between)

Random events in biology

A cell contains $m = 10^3 - 10^6$ mRNA molecules.

- m_A are from gene A
- m_B are from gene B
- m_C are from gene C
- m_X others (such that $m_A + m_B + m_C + m_X = m$)

Draw one random mRNA molecule from a cell. Which gene does it transcribe?

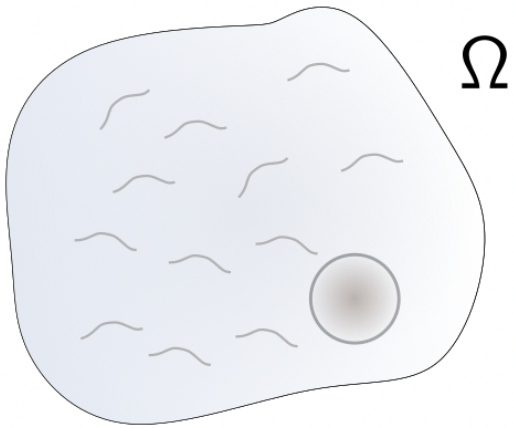


Sample space

The **sample space**, denoted as Ω, of a random phenomenon is the set of all possible outcomes.

Example:

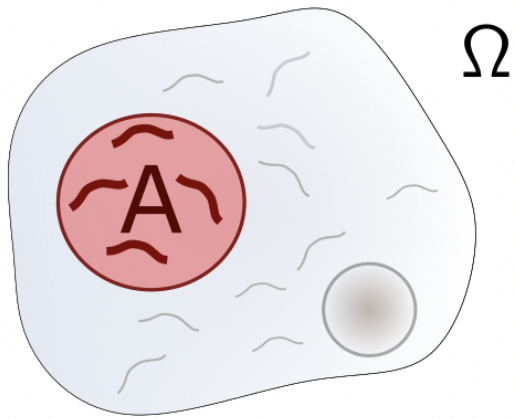
- Ω = set of all mRNA molecules in the cell



Event

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space. Events are usually denoted by upper-case letters: $A, B, \ldots$

Example Let A be the event, that we draw a transcript of gene A.



Classical Definition of Probability (Laplace)

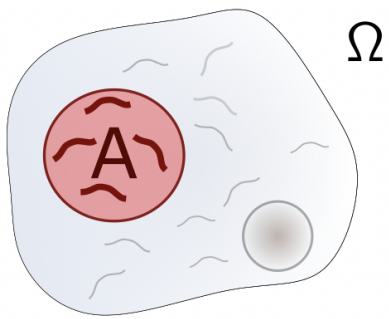
The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

For event $A = \{\text{mRNA from gene A}\} \subseteq \Omega$

$$\mathbb{P}(A) = \frac{\text{size of } A}{\text{size of } \Omega} = \frac{m_A}{m}$$

and $B = \{\text{mRNA from gene B}\} \subseteq \Omega$

$$\mathbb{P}(A) = \frac{\text{size of } B}{\text{size of } \Omega} = \frac{m_B}{m}.$$



Statistical Definition of Probability

Let n be the number of mRNA molecules (sample size) that are randomly sampled from the cell.

- Then the **absolute frequency** $f(A)$ of event A is the number of mRNA molecules transcribing gene A within the sample. It holds that

$$0 \leq f(A) \leq n.$$

- The **relative frequency** $h(A)$ is given by

$$h(A) = \frac{f(A)}{n} \quad \text{with } 0 \leq h(A) \leq 1.$$

Example: Which Genes are transcribed?

Consider the absolute frequencies $f(A)$ and relative frequencies $h(A)$ of sequenced mRNA molecules transcribing gene A with four different sample sizes $n_1, \ldots, n_4$

n	6	100	1 000	10 000
$f(A)$	3	63	692	6 520
$h(A)$	0.500	0.630	0.692	0.652

- The relative frequencies in the table suggest that $h(A)$ converges to a stable value near 0.65 as the sample size increases.
- The observed relative frequency is therefore used to estimate the unknown probability.

Statistical Definition of Probability

Let $p = \mathbb{P}(E)$ be the probability of the occurrence of the event E, then the probability $\mathbb{P}(E)$ is estimated using the corresponding relative frequency $h(E)$ of a sample with sample size n:

$$\hat{p} = \hat{\mathbb{P}}(E) = h(E) = \frac{f(E)}{n}.$$

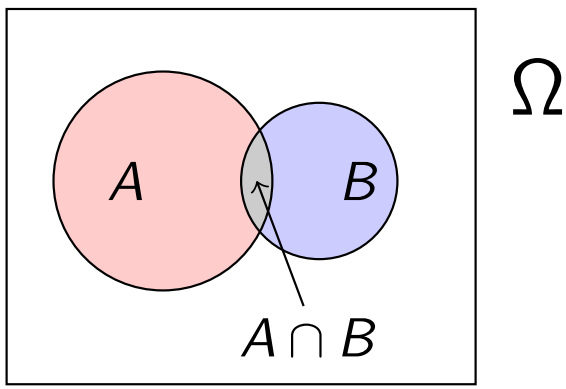
- The sign ˆ("hat") distinguishes the estimated value $\hat{p}$ from the unknown ("true") value p.

Union

The union of two event A and B, denoted $A \cup B$, is the set of outcomes that are contained in either A or B or both.

$\longrightarrow$ It is the event that either event A or B or both occur.

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

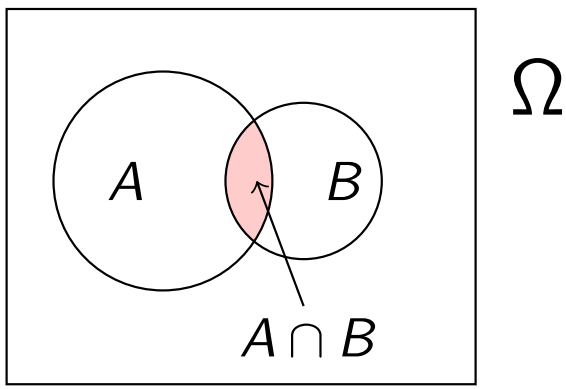


Intersection

The intersection of two event A and B, denoted $A \cap B$, is the set of outcomes that are contained in both A and B.

$\longrightarrow$ It is the event that both event A and B occur.

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$$



Probability Axioms

Let A be an event. The probability of A, denoted $\mathbb{P}(A)$, is assumed to satisfy the following axioms.

Axiom 1 For an event A, $0 \leq \mathbb{P}(A) \leq 1$.

Axiom 2 If Ω denotes the sample space, then $\mathbb{P}(\Omega) = 1$.

Axiom 3 If two event A and B are disjoint, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Axiom 3 more generally:

- If $A_1, A_2, A_3, \ldots$ is a sequence of events with $A_i \cap A_j = \varnothing$ for all $i \neq j$, then

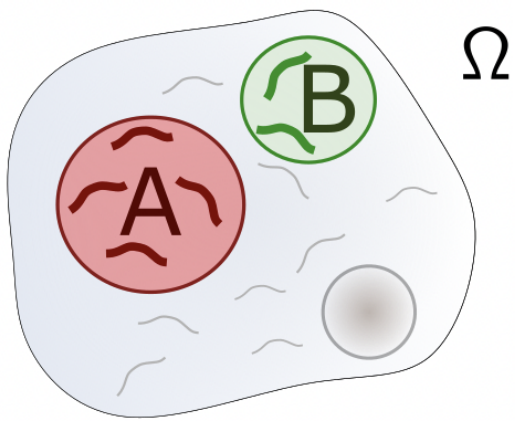
$$\mathbb{P}(A_1 \cup A_2 \cup A_3 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) + \ldots$$

Disjoint event

Two events A and B are said to be **disjoint** if no outcome is contained in both A and B.

- This is equivalent to $A \cap B$ being the empty set, i.e.,

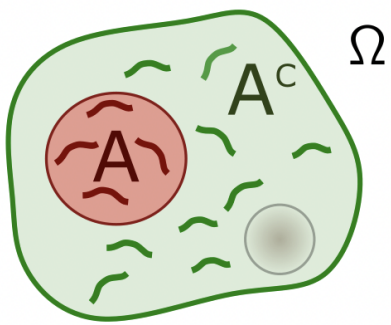
$$A \cap B = \varnothing.$$



Complementary Event

If A is an event, then the complementary event, denoted A^c, is the set of all outcomes in Ω that are **not** contained in A.

- $\Omega = A \cup A^c$ and $A^c \cap A = \varnothing$
- $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$



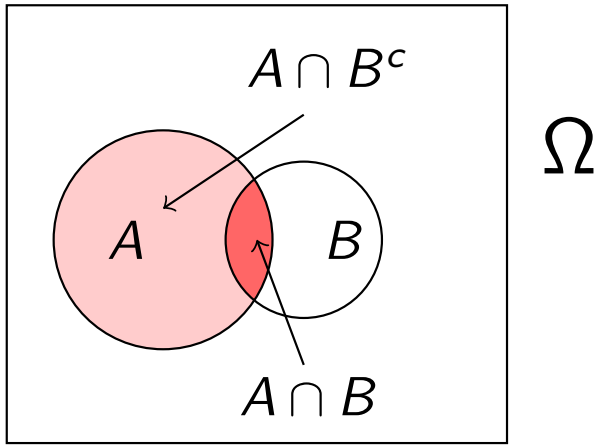
Complementary Event

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

Law of total probability

Consider two events A and B, then

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$$



Sets

If you want to know more, you can watch this video or others on set theory or basic Laplacian probability theory.

Coin Toss

- Toss a fair coin twice. Then the sample space contains:

$$\Omega = \{HH, HT, TH, TT\}$$

- Define event A: first toss is heads: {HH,HT}
- Define event B: second toss is tails: {HT,TT}

First we can see that the overlap or intersection of A and B is
$\{HH, HT\} \cap \{HT, TT\} = \{HT\}$
Now, we can calculate the probabilities of individual events: $\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{2}$
and for the intersection we get: $\mathbb{P}(A \cap B) = \frac{1}{4}$

$$\Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Multiplication rule

Two events A and B are said to be independent, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

- Knowing about the probability of the outcome A does not affect the probability for outcome B. The events are independent!

Random variables

Random variables

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

- We use capital letters to denote a random variable, e.g. X, Y. It is what we *might* get if we do the experiment.
- When we have collected data, and get *actual values*, we denote these by the corresponding lower case letters, e.g. x, y.

Discrete random variables

A **discrete random variable** X can take only a countable number of possible different values $x_1, x_2, \ldots$

Example

- Let X be the number of total mRNA molecules in a cell.
- Let Y be the number of transcripts of gene A in a cell.

Probability mass function

For each possible outcome for a discrete random variable we can assign a probability:

$$\mathbb{P}(X = x_i) = p_i,$$

This reads as the probability that we observe the outcome that the random variable X is x_i is p_i.

Example:

- n random mRNAs from our cell
- X = number of mRNAs from gene A

$$X \in \{0, 1, \ldots, n\} \Rightarrow \mathbb{P}(X = i) = p_i$$

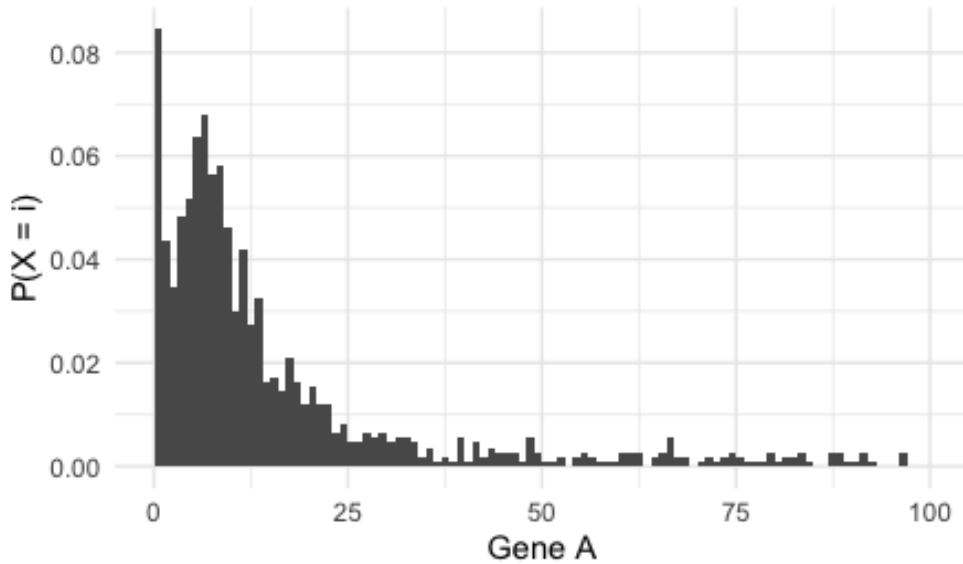
Distribution

- A **probability distribution** is a mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

Probability mass function

The probabilities for each outcome of a discrete random variable can be presented as a table or a formula. The **probability mass function (pmf)** lists the values of the random variable and their probability.

For our sequencing experiment (n mRNAs) carried out many times, we may observe the following distribution:



Probability mass function

If X is a discrete random variable, then for a probability mass function to be valid the following must hold

1. $0 \leq \mathbb{P}(X = x) \leq 1$ for all x,
2. $\sum_{\text{all } x} \mathbb{P}(X = x) = 1$.

Continuous random variable

A **continuous random variable** is a random variable whose set of outcomes is an interval on the real line.

For example

- Let X be the height of a random person.
- Let Y be the waiting time at the coffee shop.
- Let Z be the total time your computer takes to restart (including updates).

Probability density function

To describe a continuous random variable, a **probability density function (pdf)** is given. To calculate probabilities that the random variable lies in an interval, the area under the probability density function is calculated.

Probability density function

Let Y be a random variable with pdf $f(y)$, then

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b f(y)\, dy.$$

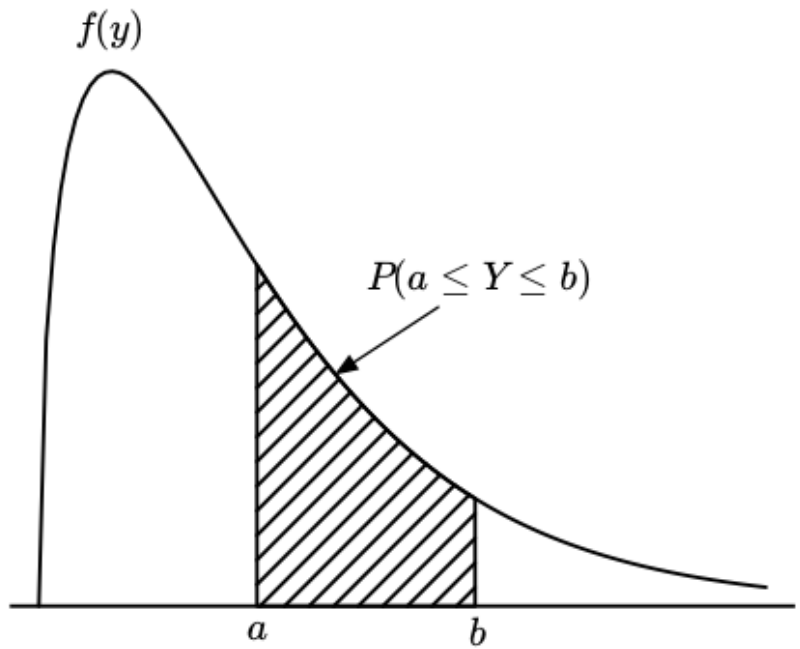
Probabilities are defined by intervals!
A single density value $f(y_i)$ has technically zero probability.

3Blue1Brown video about probability density

Probability density function

For a probability density function, $f(y)$, to be valid, the following must hold:

1. $f(y) \geq 0$ for all $y \in [a, b] \subseteq \mathbb{R} \cup \{+\infty, -\infty\}$ (no negative probabilities),
2. $\int_{-\infty}^{\infty} f(y)\, dy = 1$ (All possibilities add up to certainty).



Cumulative distribution function

The **cumulative distribution function**, denoted $F(x)$, for a random variable X is defined as

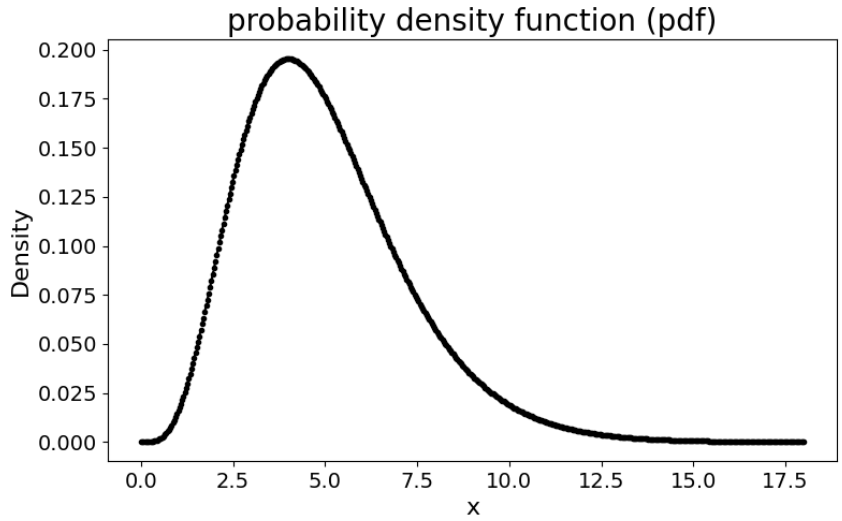
$$F(x) = \mathbb{P}(X \leq x)$$

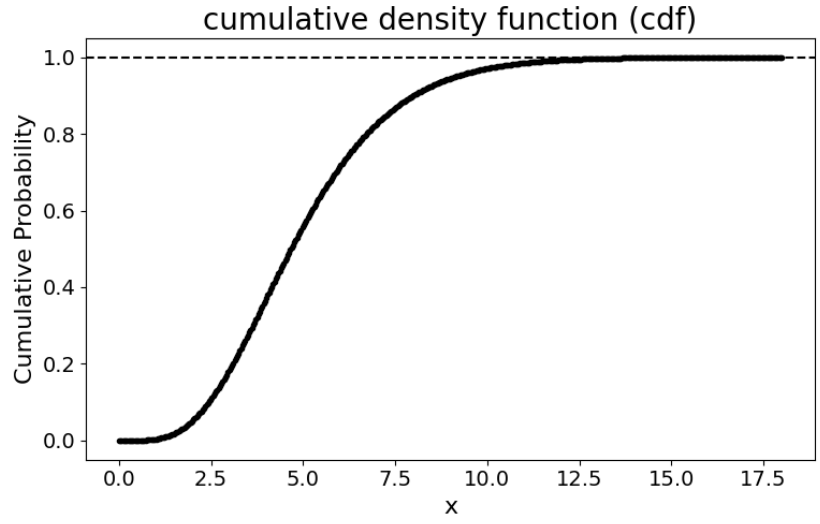
$$= \begin{cases} \sum_{y \leq x} \mathbb{P}(X = y) & \text{discrete} \\ \\ \int_{-\infty}^{x} f(y)\, dy & \text{continuous} \end{cases}$$

Cumulative distribution function

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

$$= \begin{cases} \sum_{y \leq x} \mathbb{P}(X = y) & \text{discrete} \\ \int_{-\infty}^{x} f(y)\, dy & \text{continuous} \end{cases}$$

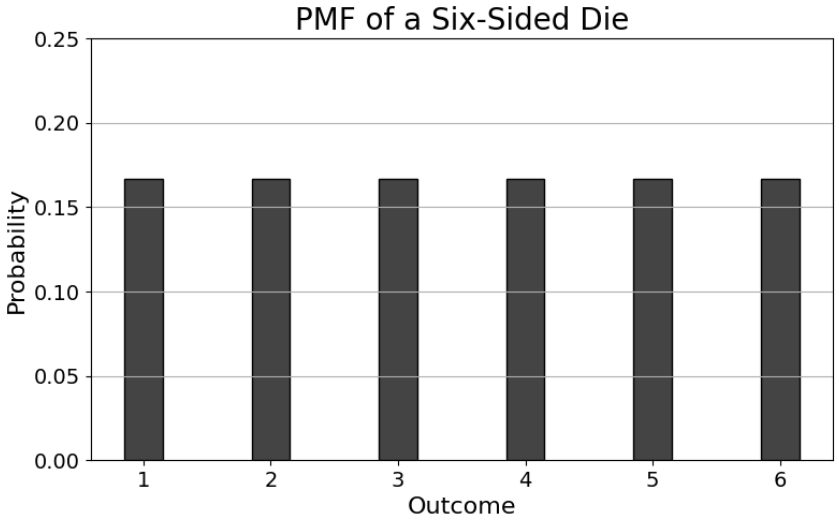


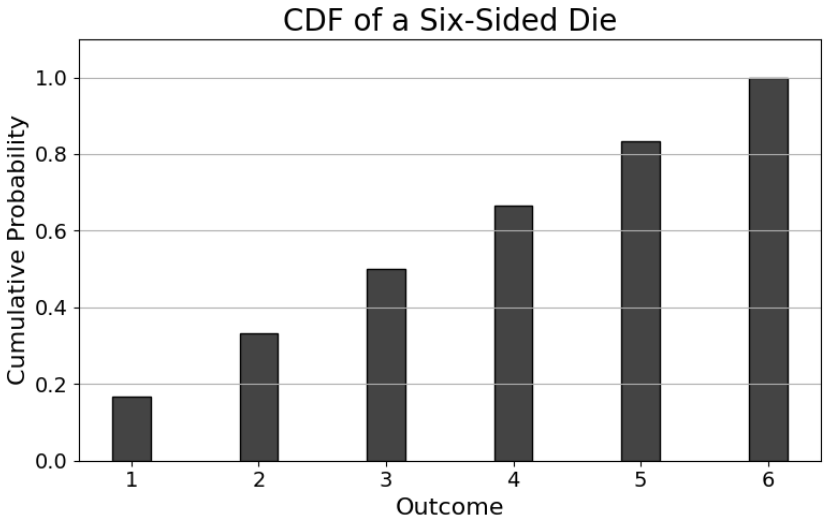


CDF - discrete example

The six-sided die.
The probability mass function reads:

$$f(y) = \begin{cases} \frac{1}{6}, & y \in \{1,2,3,4,5,6\} \\ 0, & \text{otherwise} \end{cases}$$





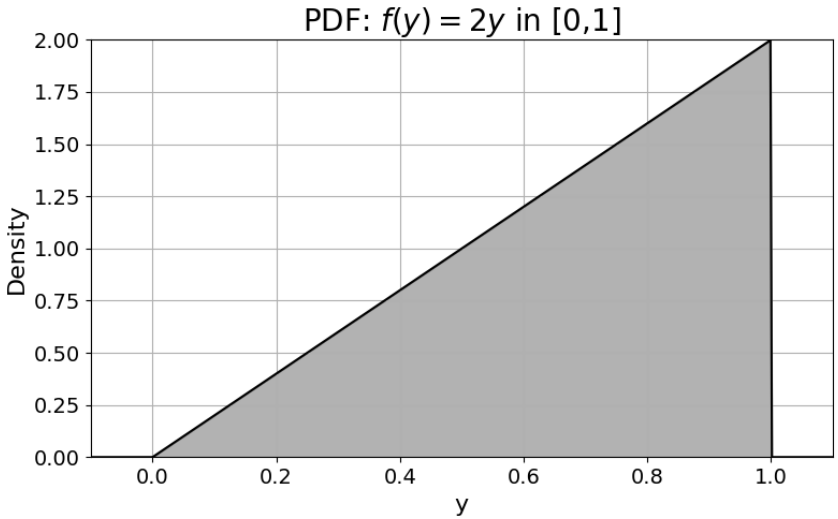
CDF - continuous example

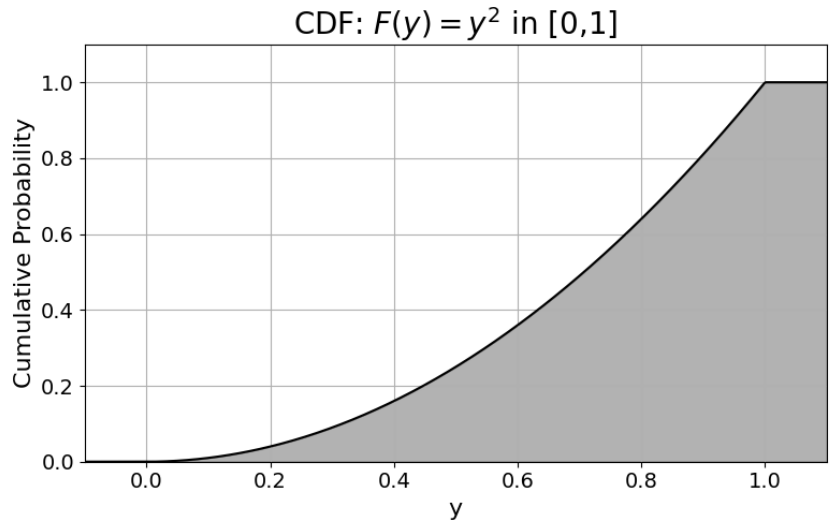
Random variable Y with probability density function

$$f(y) = \begin{cases} 2y, & 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

The CDF reads:

$$F(y) = \begin{cases} 0, & y < 0 \\ y^2, & 0 \leq y \leq 1 \\ 1, & y > 1 \end{cases}$$

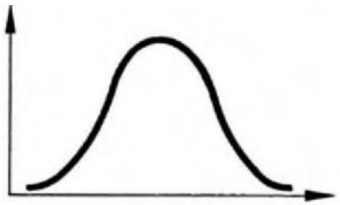




Characteristic features of univariate distributions

Figures of some distribution types

bell-shaped



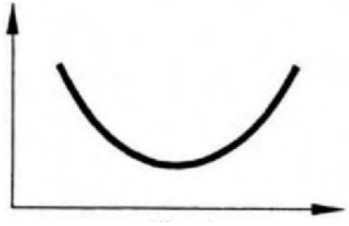
right-skewed

left-skewed

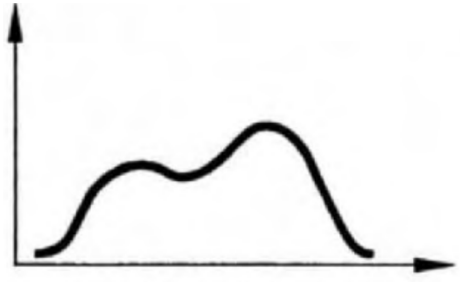
j-shaped

i-shaped

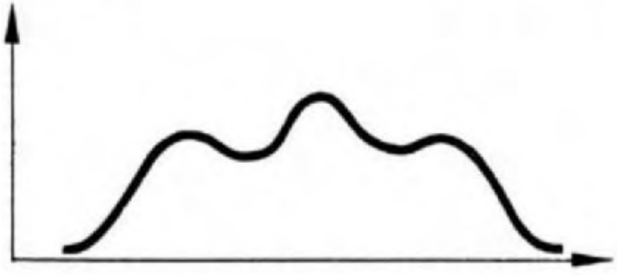
u-shaped



bimodal



multimodal

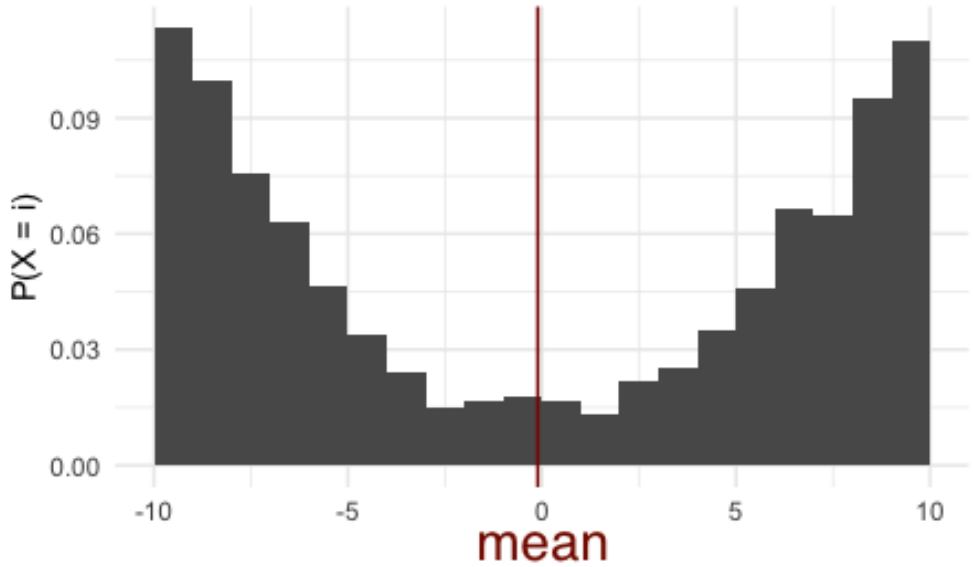


The mean

For measurements $x_1, \ldots, x_n$ the (arithmetic) mean is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- The mean can be interpreted as the centre of mass of the probability distribution, where a seesaw is balanced:



Expected value of a random variable (Math wording)

Consider a discrete random variable X with $\mathbb{P}(X = x) = p(x)$,

- then the *expected value* or *mean*, of X is defined as

$$\mathbb{E}[X] = \sum_{\text{all } x} x \cdot p(x).$$

Consider a continuous random variable X with density $f(x)$.

- The mean of X is defined by

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

The notation μ is also used for the mean.

Computations with means

Let X be a random variable (discrete or continuous), and let a and b be constants, then

1. If $Y = X + b$, then

$$\mathbb{E}[Y] = \mathbb{E}[X] + b$$

2. If $Y = aX$, then

$$\mathbb{E}[Y] = a\,\mathbb{E}[X]$$

3. If $Y = aX + b$, then

$$\mathbb{E}[Y] = a\,\mathbb{E}[X] + b$$

The median

Consider a random variable Y.

- Recall that the cumulative distribution function $F(y)$ is $\mathbb{P}(Y \leq y)$.
- The real number m such that

$$F(m) = \mathbb{P}(Y \leq m) = 0.5$$

is called the median.

Quartiles

Similarly we can introduce **quartiles** Q_1, Q_2, Q_3:

For a random variable Y the
first quartile is the point Q_1 such that

$$\mathbb{P}(Y \leq Q_1) = 0.25.$$

second quartile (median) is the point Q_2 such that

$$\mathbb{P}(Y \leq Q_2) = 0.5,$$

i.e., it is the median.
third quartile is the point Q_3 such that

$$\mathbb{P}(Y \leq Q_3) = 0.75.$$

Percentiles

In full generality one introduces quantiles, where the qth quantile is the constant b such that

$$\mathbb{P}(Y \leq b) = q$$

for a random variable Y.

Sometimes people use **percentiles**.

Simply multiply q with 100 to obtain %

Computing the sample median

of discrete measurements \random variables

1. Order the n observations from smallest to the largest (ascending order).
2. If n is odd, then the median M is the centre observation, i.e., observation number $(n+1)/2$ in the list.
3. If n is even, then the median M is the mean of the two middle numbers, i.e., observations number $n/2$ and $(n+1)/2$ in the list.

Sample quartiles

1. Arrange the observations from the smallest to the largest and locate the median M.
2. The first quartile Q_1 is the median of the observations who are to the left of the median in the ordered list.
3. The third quartile Q_3 is the median of the observations who are to the right of the median in the ordered list.

Example

Consider the following dataset:

$$3, 7, 8, 12, 13, 14, 18, 21, 23, 27$$

- The median is the middle value of an ordered dataset.
- For an even number of observations, it is the average of the two middle numbers.
- In our dataset:

$$3, 7, 8, 12, 13, 14, 18, 21, 23, 27$$

- Median $= \frac{13+14}{2} = 13.5$

Example

Consider the following dataset:

$$3, 7, 8, 12, 13, 14, 18, 21, 23, 27$$

- First Quartile (Q1): The median of the first half of the data.
- Third Quartile (Q3): The median of the second half of the data.
- In our dataset:

$$3, 7, 8, 12, 13 \quad \text{(first half)}$$

$$14, 18, 21, 23, 27 \quad \text{(second half)}$$

- $Q1 = 8$
- $Q3 = 21$

Percentiles

Consider the following dataset:

$$3, 7, 8, 12, 13, 14, 18, 21, 23, 27$$

Percentiles indicate the value below which a given percentage of observations fall.

- The formula for the position L_k of the k-th percentile (P_k) is:

$$L_k = (n + 1) \cdot \frac{k}{100}$$

  where n is the number of observations, and k is the desired percentile.

- For L_{90} (the position for the 90th percentile):

$$L_{90} = (10 + 1) \cdot \frac{90}{100} = 11 \cdot 0.9 = 9.9$$

- Since 9.9 is not an integer, we interpolate between the 9th and 10th values:

$$P_{90} = 23 + 0.9 \cdot (27 - 23) = 23 + 3.6 = 26.6$$

Mode of a distribution

For a discrete random variable X with probability mass distribution $\mathbb{P}(X = x) = p(x)$, the mode is defined as

$$mode = \underset{x \in \{x_1, \ldots, x_n\}}{\arg\max} \; p(x)$$

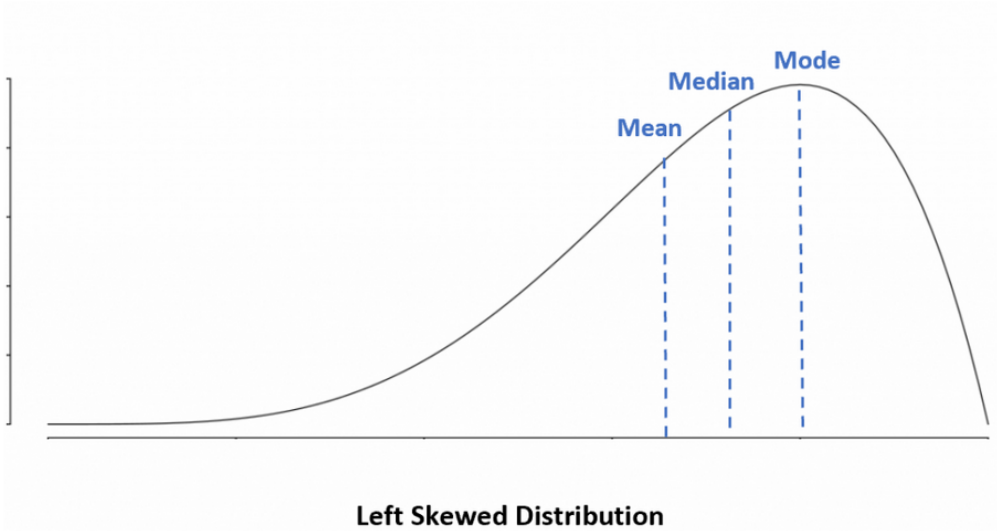
Example

- The mode is the value that appears most frequently in a dataset.
- In the dataset:

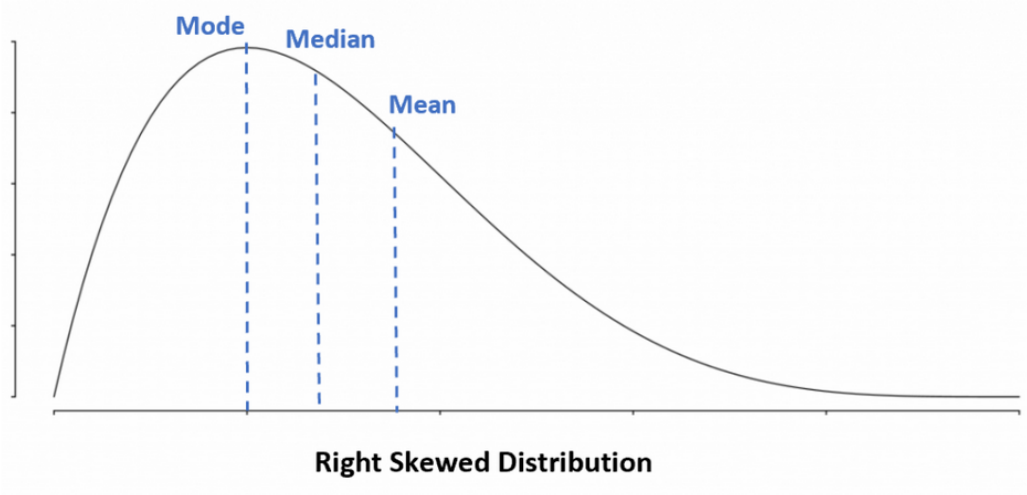
$$3, 7, 8, 8, 13, 14, 14, 14, 18, 21$$

- The number 14 appears most frequently (3 times).
- Mode = 14
- When there are no repeated values the dataset is 'mode-less'.

Skewness



Mode > Median > Mean



Mode < Median < Mean

Characteristic features of univariate distributions

— measures of mean variation —
(variance)

Interpretation of mean variation \variance

- The mean variation of a random variable determines the spread of the distribution.

The variance

For a random variable X with expected value $\mathbb{E}[X]$, the variance, $\mathrm{Var}(X)$ or σ^2, is defined as

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

Variance of a random variable

For a discrete random variable X with mean μ and $\mathbb{P}(X = x) = p(x)$

- the variance of X is

$$\sigma^2 = \sum_{\text{all } x} (x - \mu)^2 p(x).$$

Consider a continuous random variable X with mean μ and density $f(x)$.

- The variance of X is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx.$$

Standard deviation

The standard deviation, σ, is the positive square root of the variance.

$$\sigma = \sqrt{\mathsf{Var}(X)}$$

Linear combinations

Let X be a random variable (discrete or continuous), and let a and b be constants, then

1. If $Y = X + b$, then

$$\mathrm{Var}(Y) = \mathrm{Var}(X)$$

recall $\mathbb{E}(Y) = \mathbb{E}(X) + b$

2. If $Y = aX$, then

$$\mathrm{Var}(Y) = a^2 \mathrm{Var}(X)$$

recall $\mathbb{E}(Y) = a\mathbb{E}(X)$

3. If $Y = aX + b$, then

$$\mathrm{Var}(Y) = a^2 \mathrm{Var}(X)$$

Sample variance

Consider data

$$x_1, x_2, \ldots, x_n.$$

- The sample mean equals

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The sample variance is defined to be

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- $\bar{x}$ and s^2 describe the distribution of data the same way as μ and σ^2 describe probability distributions.

Standard error

The standard error of $\bar{X}$ is the square-root of $\mathrm{Var}(\bar{X})$:

$$\bar{X} = \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation and n is the sample size.

The sample standard error is given by:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}.$$

Variation and interquartile range

Calculation of the (variation) range:

$$V = x_{\text{max}} - x_{\text{min}},$$

where x_{max} is the largest data point and x_{min} is the smallest data point.

Calculation of the interquartile range (IQR):

$$\text{IQR} = Q_3 - Q_1,$$

where Q_3 is the third quartile and Q_1 is the first Quartile.

(Univariate) Distributions

Discrete Distributions

Bernoulli distribution

A **Bernoulli distribution** is a discrete probability distribution for a Bernoulli trial – a random experiment that has only two outcomes (usually called a "Success" or a "Failure").
Consider

$$X = \begin{cases} 1, & \text{if success,} \\ 0, & \text{if failure.} \end{cases}$$

Probability mass function:

$$\mathbb{P}(X = 1) = p \quad \text{and} \quad \mathbb{P}(X = 0) = 1 - p$$

and

- $\mathbb{E}[X] = p$
- $\mathrm{Var}(X) = p(1 - p)$.

Binomial distribution

Consider tossing a fair independent coin 10 times.

- What is the probability of 2 heads?

$\Rightarrow$ Nothing else than 10 independent Bernoulli experiments

Binomial probability mass function

If $X \sim B(n, p)$, then

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

with

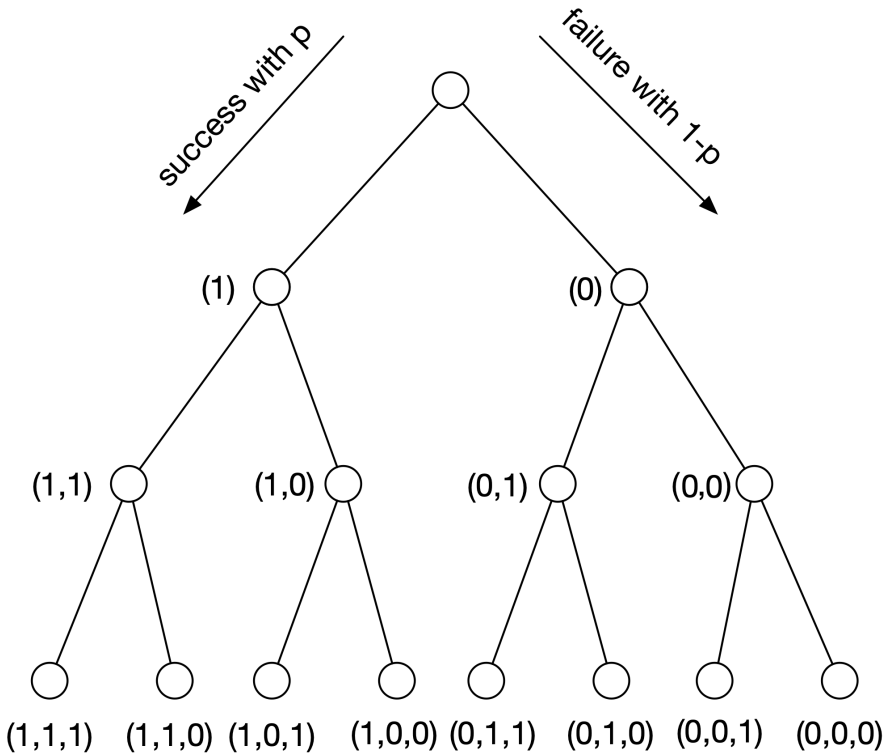
$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

and

$$0! = 1.$$

Binomial distribution

The binomial coefficient



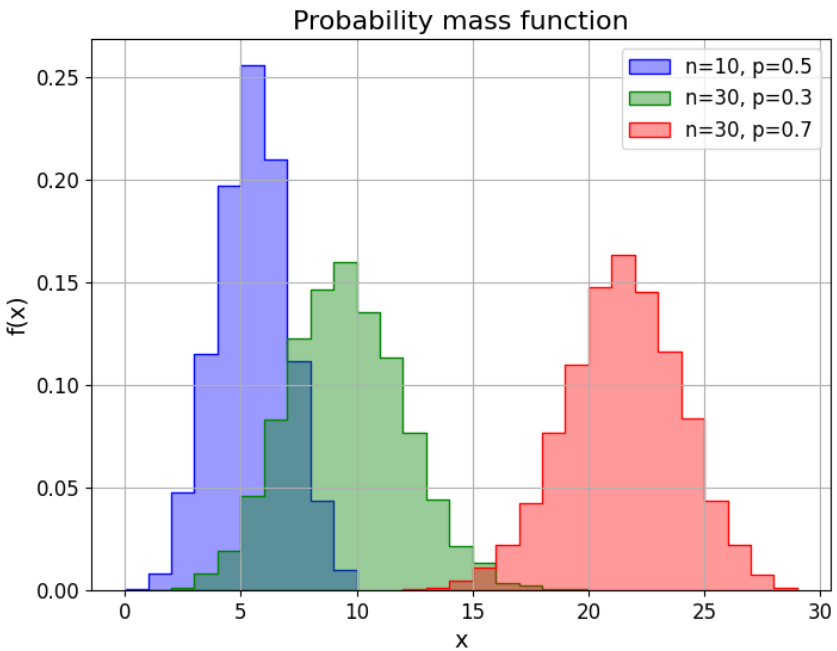
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

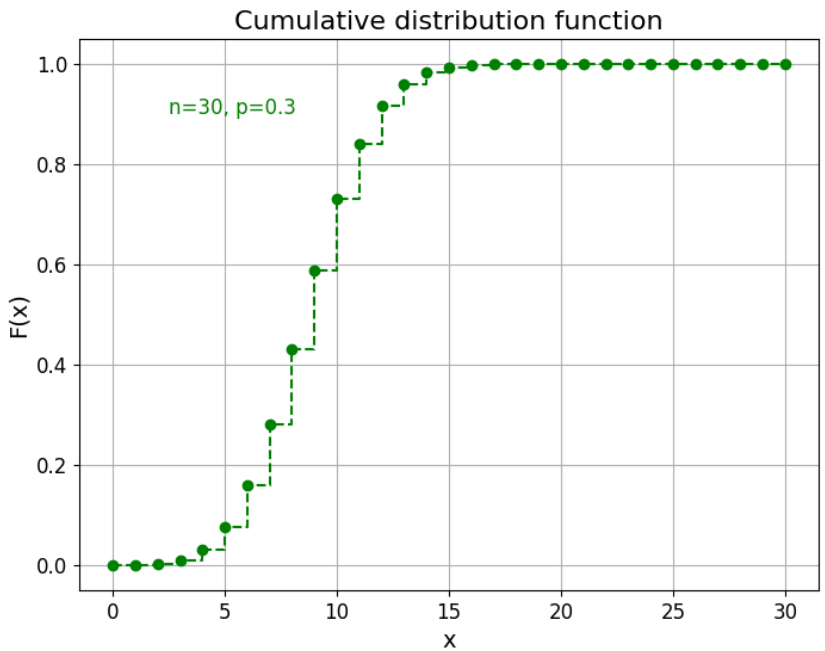
n: number of trials
k: number of successes

n \ x	0	1	2	3	4	...
1	1	1				
2	1	2	1			
3	1	3	3	1		
4	1	4	6	4	1	
...						

Binomial distribution

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





Mean and variance of a binomial experiment

If a random experiment can be approximated by a binomial process $X \sim B(n, p)$, then the average and the variance are given by

- $\mathbb{E}[X] = np$
- $\mathrm{Var}(X) = np(1-p)$,

with p being the individual success probability and n is the number of trials.

Binomial Distribution: PCR Test Example

Consider a PCR test that correctly identifies a virus 90%of the time. Let X be the number of correct identifications out of n tests.

- $p = 0.9$ (probability of correct identification)

A person gets a positive test but wants to be certain and makes another one. There are 2 possibe outcomes:

$$[1, 1] : p \cdot p = 0.9 \cdot 0.9 = 0.81$$
$$[1, 0] : p \cdot (1 - p) = 0.9 \cdot (1 - 0.9) = 0.09$$

There is still an almost 10% chance to get a negative false result.

Binomial Distribution: PCR Test Example

What if we test three times in a row?
Possible outcomes for $n = 3$:

$$[1, 1, 1] : p \cdot p \cdot p = 0.9 \cdot 0.9 \cdot 0.9 = 0.729$$
$$[1, 1, 0] : p \cdot p \cdot (1 - p) = 0.081$$
$$[1, 0, 1] : p \cdot (1 - p) \cdot p = 0.081$$
$$[0, 1, 1] : (1 - p) \cdot p \cdot p = 0.081$$
$$[1, 0, 0] : p \cdot (1 - p) \cdot (1 - p) = 0.009$$
$$[0, 1, 0] : (1 - p) \cdot p \cdot (1 - p) = 0.009$$
$$[0, 0, 1] : (1 - p) \cdot (1 - p) \cdot p = 0.009$$
$$[0, 0, 0] : (1 - p) \cdot (1 - p) \cdot (1 - p) = 0.001$$

Number of possibilities: 2^n.

Binomial Distribution: PCR Test Example

What if we test three times in a row?
Possible outcomes for $n = 3$:

$$
\begin{aligned}
[1,1,1] &: p \cdot p \cdot p = 0.9 \cdot 0.9 \cdot 0.9 = 0.729 \\
[1,1,0] &: p \cdot p \cdot (1-p) = 0.081 \\
[1,0,1] &: p \cdot (1-p) \cdot p = 0.081 \\
[0,1,1] &: (1-p) \cdot p \cdot p = 0.081 \\
[1,0,0] &: p \cdot (1-p) \cdot (1-p) = 0.009 \\
[0,1,0] &: (1-p) \cdot p \cdot (1-p) = 0.009 \\
[0,0,1] &: (1-p) \cdot (1-p) \cdot p = 0.009 \\
[0,0,0] &: (1-p) \cdot (1-p) \cdot (1-p) = 0.001
\end{aligned}
$$

Number of possibilities: 2^n.


3Blue1Brown video about Binomial distributions

Poisson distribution

If repetitive events are known to occur independently at a given rate λ, then the number of such events occuring in a given time or space X, has the **Poisson distribution**.

$$X \in \{0, 1, 2, \ldots, \infty\}.$$

The probability mass function is given by:

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}$$

which gives the probability of observing exactly x events in a fixed interval of time or space.

Poisson distribution

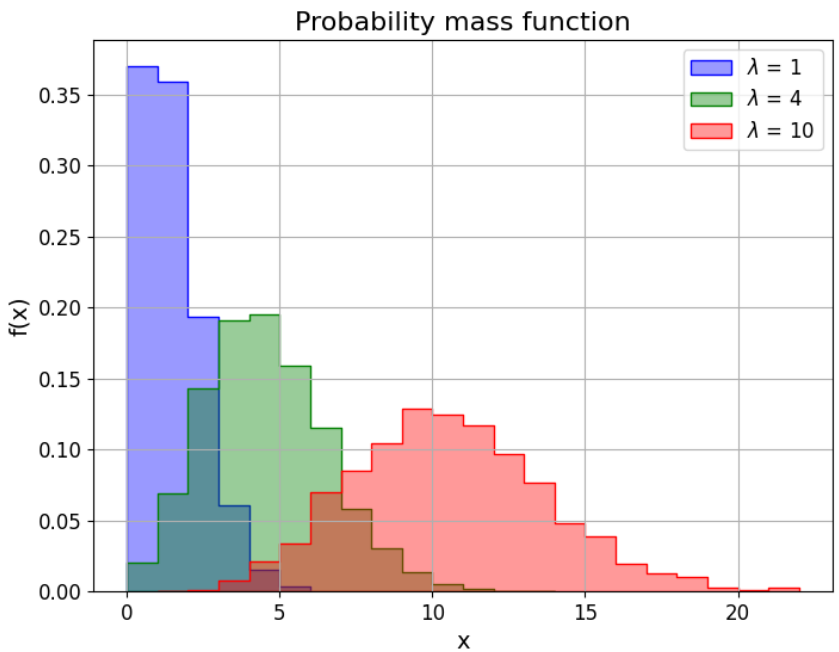
The mean and variance of the Poisson distribution are equal:

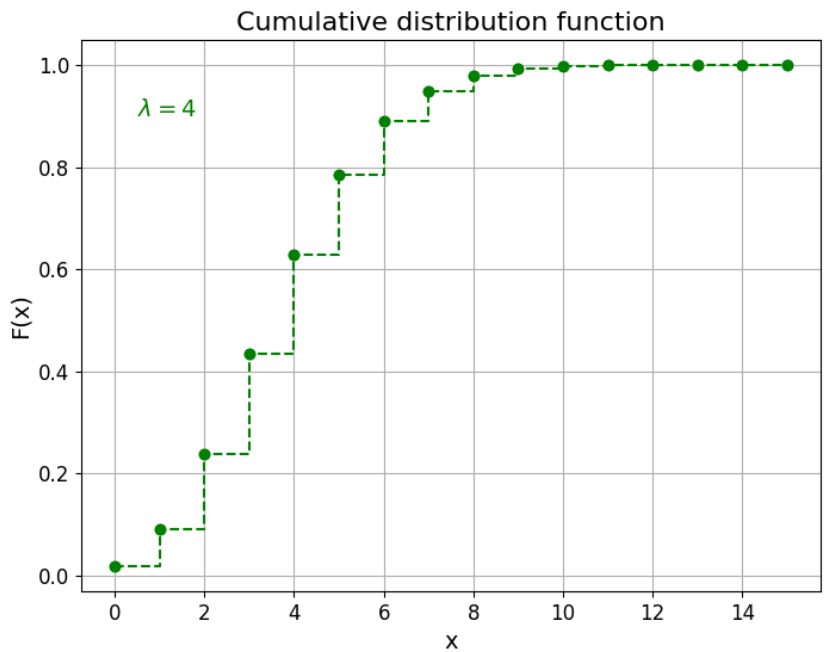
- $\mathbb{E}[X] = \lambda$
- $\mathrm{Var}(X) = \lambda$

Example - DNA mutations:
The number of mutations occurring in a given length of DNA within a fixed time period can follow a Poisson distribution. Human DNA has a mutation rate of approx. 10^{-5} per base per 20 year generation. That means in 100 years we have $5 \cdot 10^{-5} \cdot 3 \cdot 10^{9} = 1.5 \cdot 10^{5}$ expected mutations in the entire human genome. Given that approx. 1% are protein coding regions we can estimate 1500 mutations there.

Poisson distribution

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}$$





Continuous Distributions

Continuous distributions

For a continuous random variable, X, we define $f_X(x)$ to be the probability density function (pdf) of X. In order to be a valid pdf, $f_X(x)$ must be strictly non-negative and satisfy:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

All functions that satisfy the above constraints are pdfs for a continuous random variable.

Exponential distribution - Waiting time distribution

The exponential distribution describes the waiting time X until the next event occurs, where the time depends on the parameter $\lambda > 0$:

The pdf of X is:

$$f_\lambda(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, x \in \mathbb{R}^+.$$

The cdf of X is:

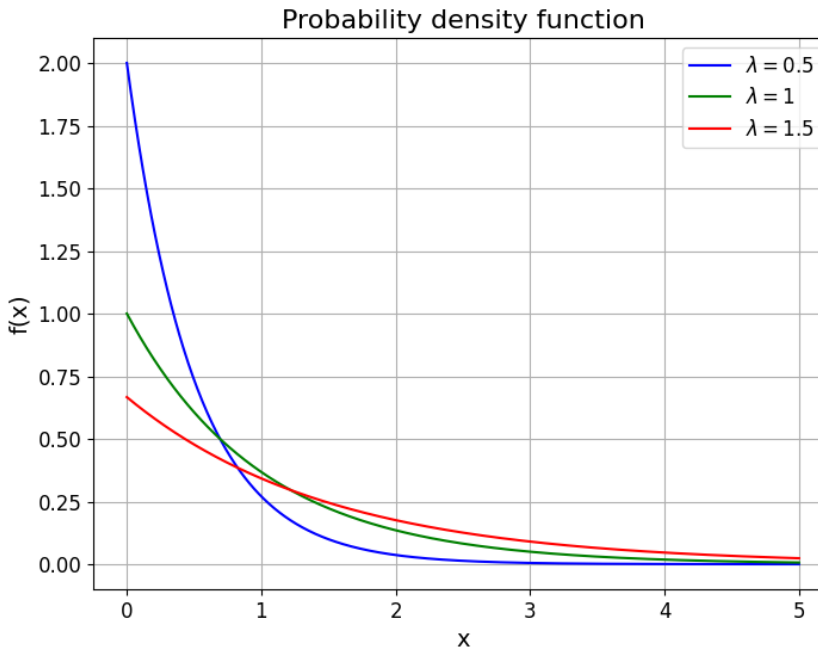
$$F_\lambda(x) = 1 - e^{-\lambda x}.$$

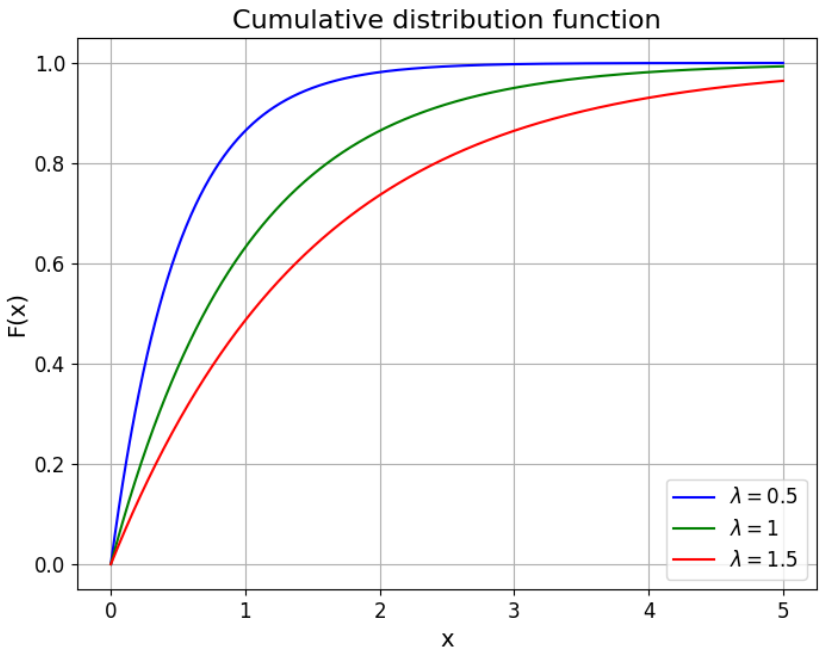
The moments of X are:

$$\mathbb{E}[X] = \frac{1}{\lambda}$$
$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Exponential distribution

$$f_\lambda(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, x \in \mathbb{R}^+.$$





Exponential distribution

Examples

- Radioactive Decay: The time between successive emissions in a radioactive substance follows an exponential distribution.
- MTTF (mean time to failure): The time until failure of certain components in a machine or device.
- Biomedical: Time until the occurrence of certain biological events, such as cell division, or the spread of diseases

The system needs to be memoryless - an event occurring in the future is independent of past events.

The Normal Distribution

Normal distribution

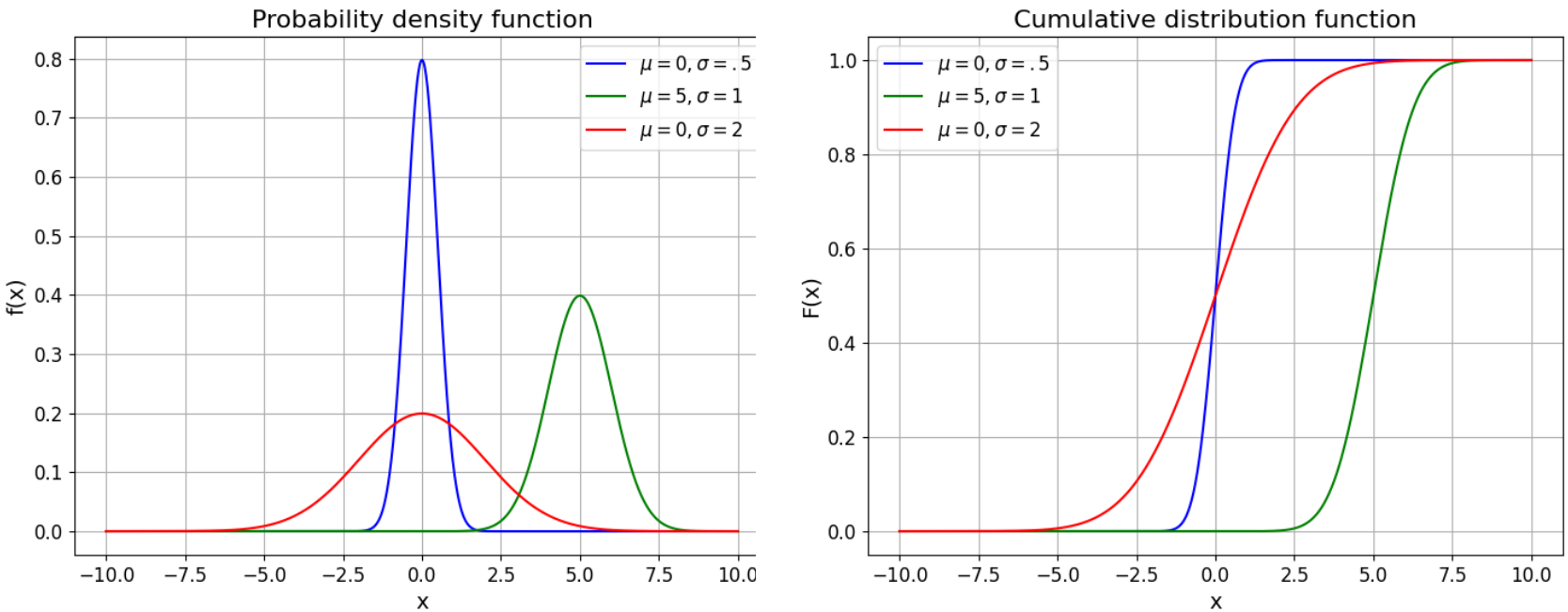
Consider a random variable X which has a **normal distribution** with mean μ and variance σ^2.

- We denoted this
$$X \sim \mathcal{N}(\mu, \sigma).$$
- X is a continuous random variable and is thus described by a probability density function.
- The probability density function of X is
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

The normal distribution is an example where mode = mean = median.

Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$



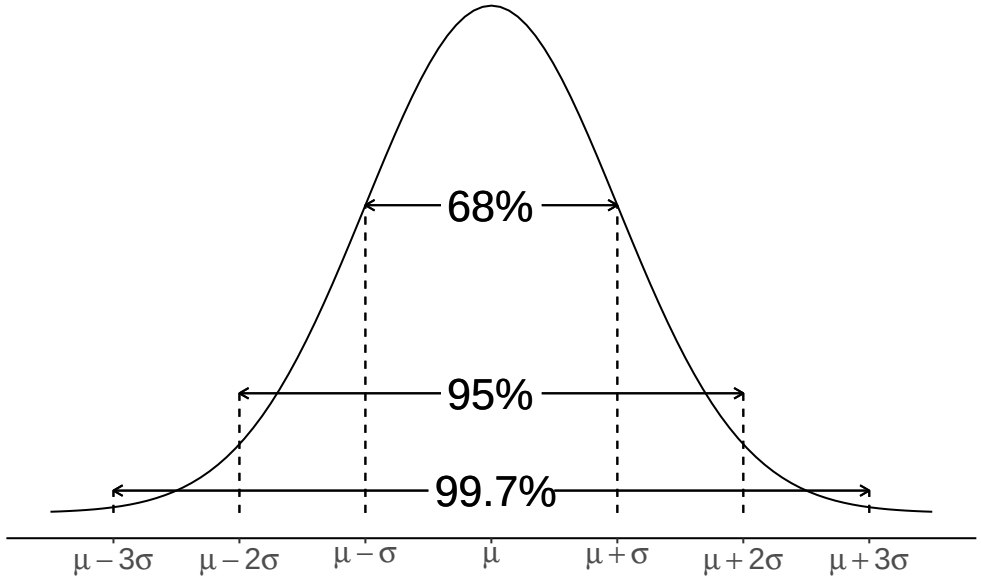
Normal distribution

Examples

- Blood Pressure: Blood pressure measurements in a healthy population typically follow a normal distribution.
- Exam Scores: In large classes, the distribution of exam scores often approximates a normal distribution.
- Measurement Errors: In scientific and engineering experiments, measurement errors often follow a normal distribution due to the Central Limit Theorem.

A random variable is normally distributed if it results from the sum of a large number of independent, identically distributed random variables.

68-95-99.7% Rule



For the area under a Normal curve,

- 68% lies between $\mu - \sigma$ and $\mu + \sigma$,
- 95% lies between $\mu - 2\sigma$ and $\mu + 2\sigma$,
- 99.7% lies between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Standardisation

The **standard normal distribution** is a normal distribution with a mean of 0 and standard deviation of 1. It is often denoted by Z.

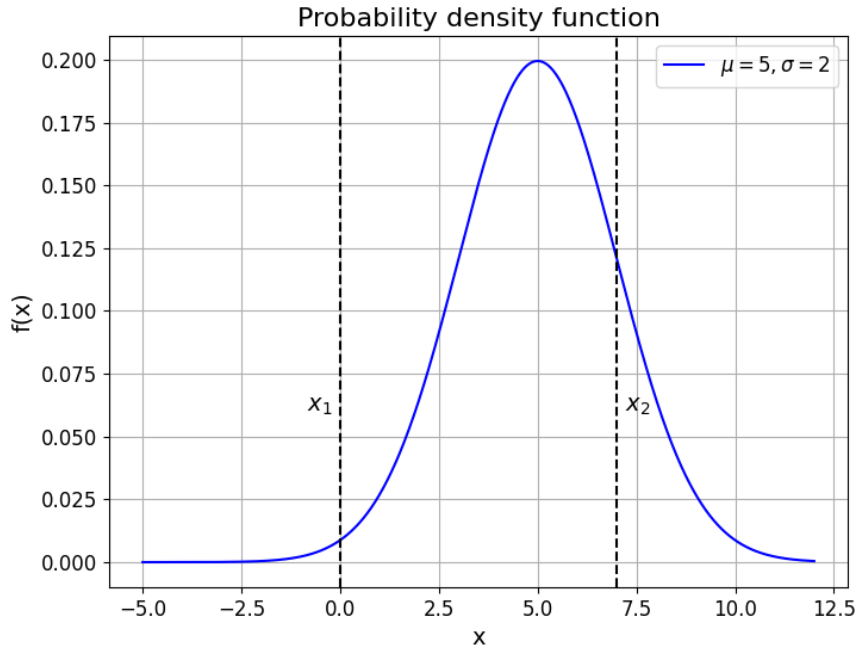
Sometimes it is required to transform a $X \sim \mathcal{N}(\mu, \sigma)$ into a $Z = \mathcal{N}(\mu = 0, \sigma = 1)$ to account for better comparability and to make computations easier.

Let X be a normal random variable with mean μ and standard deviation σ, then the transformation

$$z = \frac{X - \mu}{\sigma}$$

leads to a new normally distributed variable z with $\mu = 0$ and $\sigma = 1$.

The z - score



Let $x_1 = 0$ and $x_2 = 7$ be two random variables that need to be compared to a process described by a normal distribution with $\mu = 5$ and $\sigma = 2$. Then, $z_1 = (x_1 - \mu)/\sigma = -2.5$ (x_1 is $2.5 \cdot \sigma$ below the mean) and $z_2 = (x_2 - \mu)/\sigma = 1$ (x_2 is one σ above the mean).