Università degli Studi di Milano Bicocca

Department of Informatics, systems and communication

Foundations of Probability and Statistics

Matteo Suardi

A.Y.: 2024-2025

November 2024

Abstract

This course offers a comprehensive introduction to descriptive statistics, probability, and statistical inference, blending theoretical principles with practical applications using the R programming language.

The probability component covers foundational concepts, event probabilities, Bayes' theorem, random variables, probability distributions, and theoretical underpinnings such as the Law of Large Numbers and the Central Limit Theorem.

The statistical inference segment introduces probabilistic sampling, estimators, and their properties, with a focus on point estimation and maximum likelihood estimators. Interval estimation techniques, including confidence intervals for means and variances, are also covered. Hypothesis testing is presented through the framework of test statistics, significance, and power, encompassing tests on means, variances, proportions, differences between means, and independence.

Contents

1 Descriptive statistics

1.1 Terminology

Population: a set of elements which are the subject of a certain statistical study. The elements of the population are generically called *statistical units*.
A population is called **real** when it is actually existing and visible (example: the buses that circulated in Milan during the last month), while it is called **virtual** if it is abstract or referred to the future (example: patients with a certain disease).

Sample: any subset of the reference population. The number of statistical units present in the sample is called the *sample size*.

Variable: a characteristic of each statistical unit that is detected.

Level/category (of a variable): the distinct values assumed by a variable.

1.2 Types of variables

In statistical science, variables can be broadly classified into two types: **quantitative** and **qualitative** (also called *categorical*).

1.2.1 Quantitative variables

A variable is called **quantitative** when its measurement scale has numerical values that represent different magnitudes. Examples include age, weight, annual income, college GPA, and the number of good friends. Quantitative variables are further divided into two types:

- **Discrete**: a quantitative variable is discrete if it takes values from a countable set, either finite or countably infinite. Examples include the number of children in a family, the number of employees in a company, or the population of a country.

- **Continuous**: a quantitative variable is continuous if it assumes any value within a real interval. This includes variables like GDP of a nation, temperature, height, or weight. Continuous variables can take infinitely many values within a given range.

1.2.2 Qualitative (categorical) variables

A variable is called **categorical** when its measurement scale consists of a set of categories rather than numerical magnitudes. Examples of categorical variables include marital status (e.g., single, married, divorced), primary mode of transportation (e.g., automobile, bicycle, bus), and favorite type of music (e.g., classical, rock, jazz).
Categorical variables can be of two main types:

- **Nominal**: The categories have no inherent order. Examples include religion, preferred shopping destination, and favorite music genre.

- **Ordinal**: The categories have a natural ordering. Examples include perceived happiness (not too happy, pretty happy, very happy), headache severity (none, slight, moderate, severe), and political philosophy (very liberal, moderate, slightly conservative).

1.3 Experimental and observational data

When a sample is taken from a population, we have one of two scenarios:

- **Experimental study**: data collected in replicable and controlled situations (e.g., laboratory experiments).
- **Observational study**: data gathered from existing information (e.g., hotel presence in a season, share prices).

Observational studies may not control for all factors influencing the phenomenon.

1.4 Categories/levels of a variable

Let $x_1, \ldots, x_n$ be the observations of X, the values taken by a variable X for all n statistical units. For each $i = 1, \ldots, n$, x_i is the value taken by X at the i-th statistical unit.
Let $c_1, \ldots, c_k$ be:

- the k distinct levels or categories of X (if X is qualitative or discrete), or
- the k sub-intervals into which we divided the numerical values $x_1, \ldots, x_n$ (if X is categorized as a quantitative variable).

1.5 Absolute frequencies

For each $j = 1, \ldots, k$, let n_j be the number of times that the distinct value c_j appears in the data $x_1, \ldots, x_n$. The values $n_1, \ldots, n_k$ are called **absolute frequencies**.
Given a set A, let $I_A : \{0, 1\}$ be the **indicator function** of A, defined as:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

If X is a qualitative or discrete variable, the absolute frequency n_j for each $j = 1, \ldots, k$ is:

$$n_j = \sum_{i=1}^{n} I_{\{x_i = c_j\}}, \quad j = 1, \ldots, k$$

If X is a categorized variable, and c_j is the sub-interval $(a_{j-1}, a_j]$, for $j = 1, \ldots, k$, then the absolute frequency n_j is:

$$n_j = \sum_{i=1}^{n} I_{\{a_{j-1} < x_i \leq a_j\}}, \quad j = 1, \ldots, k$$

1.6 Relative frequencies

For fair comparison between populations of different sizes, **relative frequencies** are used:

$$\text{relative frequency} = \frac{\text{absolute frequency}}{\text{number of observations}}$$

Let $n_1, \ldots, n_k$ be the absolute frequencies. Then the relative frequencies $f_1, \ldots, f_k$ are:

$$f_j = \frac{n_j}{n}, \quad j = 1, \ldots, k$$

The relative frequencies sum to 1:

$$f_1 + f_2 + \cdots + f_k = \sum_{j=1}^{k} f_j = 1$$

Here's an example of frequency distribution (curriculum of students enrolled in a Master in Economics):

Category	n_j	f_j	p_j
Finance	160	0.40	40%
Marketing	140	0.35	30%
Accounting	100	0.25	25%
	400	1.00	100%

Table 1: Enrolled students in a MSC in Economics

1.7 Absolute cumulative frequencies

Cumulative frequencies are defined if X is a:

- Qualitative ordinal variable,
- Discrete quantitative variable,
- Categorized variable.

Let $n_1, \ldots, n_k$ be the absolute frequencies. The absolute cumulative frequencies $N_1, \ldots, N_k$ are given by:

$$N_j = \sum_{g=1}^{j} n_g, \quad j = 1, \ldots, k$$

1.8 Relative cumulative frequencies

Let $f_1, \ldots, f_k$ be the relative frequencies. The relative cumulative frequencies $F_1, \ldots, F_k$ are given by:

$$F_j = \sum_{g=1}^{j} f_g, \quad j = 1, \ldots, k$$

1.9 Empirical cumulative distribution function (ECDF)

For any $x \in R$, the empirical cumulative distribution function $ecdf(x)$ is defined as:

$$ecdf(x) = \frac{\text{number of observations less than or equal to } x}{\text{total number of observations}}$$

Let $x_1, \ldots, x_n$ be a collection of data points. Then for any $x \in R$:

$$F(x) = \frac{\#\{x_i \leq x\}}{n}$$

where $\#\{x_i \leq x\}$ is the number of observations less than or equal to x, and n is the total number of observations.
If an observation is sampled at random, $F(x)$ represents the probability of obtaining an observation less than or equal to x.
In `R`, the empirical cumulative distribution function can be obtained using the function `ecdf()`.
To obtain the function $ecdF(x)$ from the data $x_1, \ldots, x_n$, consider the ordered data $x_{(1)}, \ldots, x_{(n)}$ from the minimum value $x_{(1)}$ to the maximum value $x_{(n)}$:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

For a given value of x, the value of $F(x)$ is the fraction of observations less than or equal to x, i.e., the fraction of data points to the left of x in the ordered sequence.

1.9.1 Properties of the ecdf

- $0 \leq F(x) \leq 1, \quad \forall x \in R$
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to +\infty} F(x) = 1$
- $F(x)$ is non-decreasing
- $F(x)$ is right-continuous

Let X be a discrete variable taking values $c_1, \ldots, c_k$ and assume that the relative cumulative frequencies $F_1, \ldots, F_k$ of X are known.
What is the empirical cumulative distribution function?

$$ecdf(x) = \begin{cases} 0 & \text{if } x < c_1 \\ F_j & \text{if } c_j \leq x \leq c_j + 1 \\ 1 & \text{if} x \geq c_k \end{cases}$$

Given that $c_1 < c_2 < \cdots < c_k$, the ecdf can be constructed as:

$$F(x) = \frac{\text{number of observations } \leq x}{\text{total number of observations}}.$$

1.9.2 Relative cumulative frequencies and ecdf

Let X be a discrete random variable taking values $c_1, c_2, \ldots, c_k$ where $c_1 < c_2 < \cdots < c_k$. Assume the **relative cumulative frequencies** $F_1, F_2, \ldots, F_k$ are known, where:

$$F_j = \sum_{i=1}^{j} f_i, \quad j = 1, \ldots, k,$$

and f_i is the relative frequency of $X = c_i$.

Example

Suppose X takes values $c_1 = 1, c_2 = 2, c_3 = 3$ with relative frequencies $f_1 = 0.2, f_2 = 0.5, f_3 = 0.3$. The relative cumulative frequencies are:

$$F_1 = 0.2, \quad F_2 = 0.7, \quad F_3 = 1.0.$$

The **empirical cumulative distribution function (ECDF)** of X is:

$$F(x) = \begin{cases} 0, & \text{if } x < c_1, \\ F_j, & \text{if } c_j \le x < c_{j+1},\ j = 1, \dots, k-1, \\ 1, & \text{if } x \ge c_k. \end{cases}$$

1.10 An epidemiologic problem

Research question: Is the amount of DDE greater among women who gave birth prematurely?

Frequency distribution: The DDE interval $(0, 180]$ is divided into 10 sub-intervals of length 18. The absolute frequencies are:

DDE Interval (mg/L)	Normal Birth (n=1951)	Premature Birth (n=361)
$(0, 18]$	573	68
$(18, 36]$	906	164
$(36, 54]$	308	65
$(54, 72]$	91	34
$(72, 90]$	40	14
$(90, 108]$	19	10
$(108, 126]$	6	3
$(126, 144]$	5	1
$(144, 162]$	2	1
$(162, 180]$	1	1
Total	1951	361

Relative cumulative frequencies
The relative cumulative frequencies are:

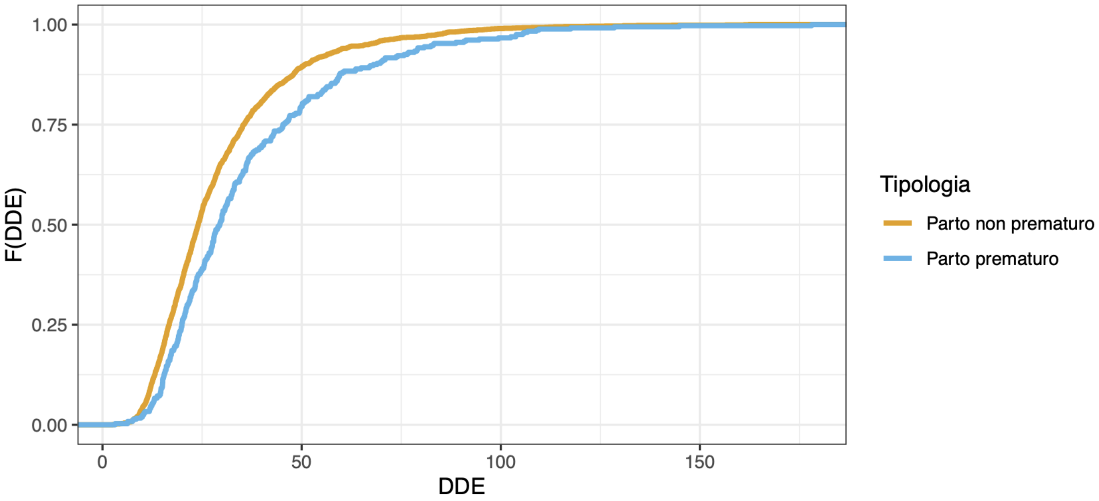
DDE Interval (mg/L)	Normal Birth (Cumulative)	Premature Birth (Cumulative)
$(0, 18]$	0.294	0.188
$(18, 36]$	0.758	0.643
$(36, 54]$	0.916	0.823
$(54, 72]$	0.963	0.917
$(72, 90]$	0.983	0.956
$(90, 108]$	0.993	0.983
$(108, 126]$	0.996	0.992
$(126, 144]$	0.998	0.994
$(144, 162]$	0.999	0.997
$(162, 180]$	1.000	1.000

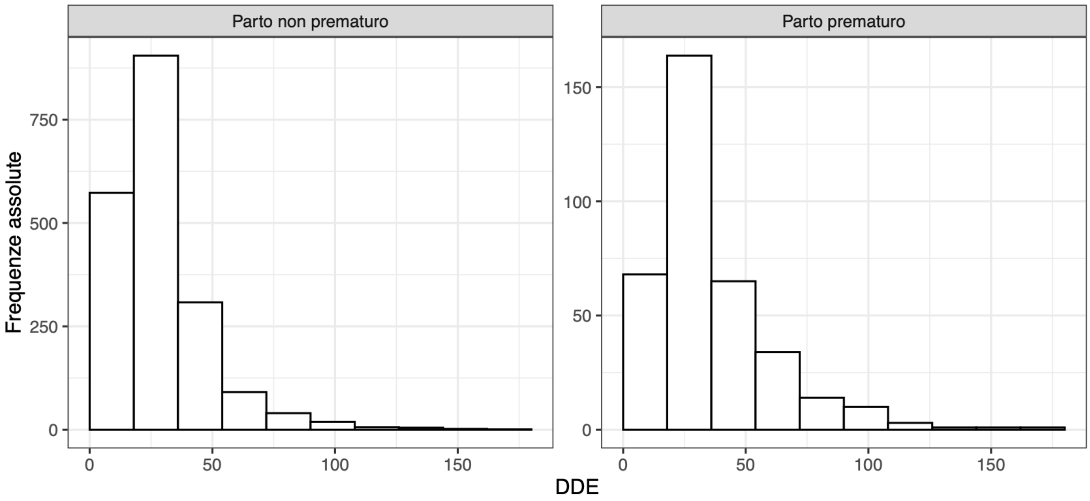
Empirical Cumulative Distribution Function (ECDF)
The ECDF is defined as:

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ \text{Cumulative frequency}, & \text{if } x \in \text{Interval}, \\ 1, & \text{if } x > 180. \end{cases}$$

Analysis and interpretation

- For lower DDE levels ($x \in (0, 36]$), cumulative proportions for premature births are smaller than for normal births.
- For higher DDE levels ($x > 36$), cumulative proportions for premature births are consistently higher.





Conclusion: Premature births are associated with higher DDE levels.

1.11 Histograms

A histogram is a graphical representation of the frequency distribution of a variable, useful for visualizing how data is distributed. For categorized variables:

- The **width** of each rectangle corresponds to the interval length.
- The **height** of each rectangle represents the absolute or relative frequency.

1.11.1 Problem with unequal interval lengths

When intervals have different lengths:

- Using absolute or relative frequencies for the height distorts the representation.
- Larger intervals may dominate the histogram visually, even if their frequencies are low.

Solution: density-based histograms
To adjust for unequal interval lengths, the height of each rectangle should be proportional to the **density** of the observations in that interval. The density is:

$$d_j = \frac{\text{relative frequency}}{\text{range of the interval}} = \frac{f_j}{a_j - a_{j-1}}, \quad j = 1, \dots, k,$$

where:

- f_j: Relative frequency of the j-th interval.
- $a_j - a_{j-1}$: Length of the j-th interval.

1.11.2 Key properties of density

- The **area** of each rectangle reflects the relative frequency:

$$\text{Area of a rectangle} = \text{height (density)} \times \text{width (interval length)} = \text{relative frequency}.$$

- The total area of all rectangles equals 1 when using relative frequencies, making the histogram a valid representation of a probability distribution.

1.12 Ogive

The ogive assumes a uniform distribution of data within each subinterval and provides a piecewise linear approximation of the cumulative distribution function (CDF).
Let:

- a_{j-1} and a_j be the bounds of the j-th subinterval, $j = 1, \dots, k$,
- $F_0 = 0$, the cumulative frequency at the start.

The cumulative function $F(x)$ is defined as:

$$F(x) = \begin{cases} 0, & \text{if } x < a_0, \\ F_{j-1} + d_j(x - a_{j-1}), & \text{if } a_{j-1} \leq x < a_j,\ j = 1, \dots, k, \\ 1, & \text{if } x \geq a_k, \end{cases}$$

where:

- $w_j = a_j - a_{j-1}$: The width of the j-th subinterval,
- f_j: The absolute frequency in the j-th subinterval,

- $d_j = \frac{f_j}{w_j}$: The density within the j-th subinterval.

Construction of the ogive

1. Compute the cumulative frequencies F_j up to each subinterval:

$$F_j = F_{j-1} + f_j.$$

2. Interpolate linearly between a_{j-1} and a_j:

$$F(x) = F_{j-1} + d_j(x - a_{j-1}).$$

3. Plot the points (a_j, F_j) for $j = 0, \ldots, k$, and connect them with straight lines.

1.12.1 Ogive in in R

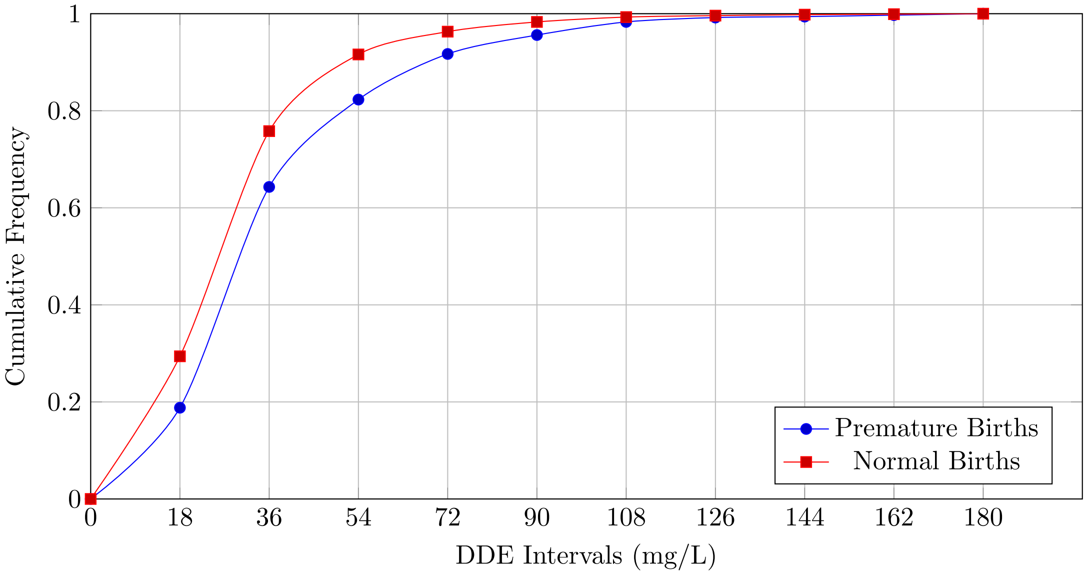
The `agricolae` package in R can be used to construct and plot the ogive.

```
# Example data: DDE levels
data <- c(573, 906, 308, 91, 40, 19, 6, 5, 2, 1) # Absolute frequencies
intervals <- seq(0, 180, by=18) # Subinterval bounds

# Load agricolae package
install.packages("agricolae")
library(agricolae)

# Generate cumulative frequencies
ogive <- graph.freq(data, intervals, ogive = TRUE)

# Plot the ogive
plot(ogive, type="l", main="Ogive of DDE Levels", xlab="DDE (mg/L)", ylab="Cumulative Frequency")
```



1.13 Measures of central tendency

1.13.1 Arithmetic mean (average)

The arithmetic mean of n observations $x_1, x_2, \dots, x_n$ is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Properties of the mean

- **Mean of a constant variable:** If $x_1 = x_2 = \cdots = x_n = a$, then $\bar{x} = a$.
- **Internality:** The mean satisfies:

  $$x_{(1)} \leq \bar{x} \leq x_{(n)}$$

  where $x_{(1)} = \min(x_1, \dots, x_n)$ and $x_{(n)} = \max(x_1, \dots, x_n)$.
- **Linearity:** If $y_i = a + bx_i$, then:

  $$\bar{y} = a + b\bar{x}$$
- **Deviations sum to zero:**

  $$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$
- **Minimizes squared deviations:**

  $$\bar{x} = \arg\min_{a \in R} \sum_{i=1}^{n}(x_i - a)^2$$

1.13.2 Weighted arithmetic mean

If weights $w_1, \dots, w_n$ are assigned to $x_1, \dots, x_n$, the weighted mean is:

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

1.13.3 Median

The median is the value that divides a dataset into two equal parts:

- 50% of observations are less than or equal to the median.
- 50% of observations are greater than or equal to the median.

Calculation:
Let $x_1, x_2, \dots, x_n$ be the dataset, and $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the sorted data ($x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$).

- **Odd Number of Observations:**

  $$Me = x_{(\frac{n+1}{2})}$$

- **Even Number of Observations:**

$$Me = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Properties of the median

- **Robustness:** The median is not sensitive to outliers.
- **Range:** The median satisfies:
$$x_{(1)} \leq Me \leq x_{(n)}$$
- **Minimizes Absolute Deviations:**
$$\sum_{i=1}^{n} |x_i - Me| \leq \sum_{i=1}^{n} |x_i - a|, \quad \forall a \in R.$$

Example: odd number of observations
Dataset: $5, 1, 8, 3, 2$

- Sorted: $1, 2, 3, 5, 8$
- Median:
$$Me = x_{(\frac{5+1}{2})} = x_{(3)} = 3$$

Example: even number of observations
Dataset: $4, 7, 1, 2, 6, 5$

- Sorted: $1, 2, 4, 5, 6, 7$
- Median:
$$Me = \frac{x_{(3)} + x_{(4)}}{2} = \frac{4+5}{2} = 4.5$$

Computing median from frequencies
To compute the median using cumulative relative frequencies:

- If $F_j > 0.5$, the median falls in the corresponding bin.
- If $F_j = 0.5$, the median is the midpoint of the bin.

1.13.4 Measures of dispersion: variance, standard deviation, range, IQR

- **Variance:** The variance σ^2 measures the spread of the data:

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

For a sample, the sample variance is:

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

- **Standard deviation:** The standard deviation is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2}.$$

- **Range:** The range is the difference between the maximum and minimum values:

$$\text{Range} = x_{\max} - x_{\min}.$$

- **Interquartile range (IQR):** The IQR is the range of the middle 50

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

where Q_1 and Q_3 are the first and third quartiles, respectively.

1.13.5 Other means

- **Geometric mean:**

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}.$$

- **Harmonic mean:**

$$\text{Harmonic mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

1.13.6 Bivariate statistics

Bivariate statistics analyze the relationship between two variables.

- **Covariance**: The covariance measures the degree to which two variables change together:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

For a sample:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- **Correlation coefficient**: The correlation coefficient r is a normalized measure of the linear relationship between two variables:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively. The correlation coefficient ranges from -1 to 1:

-
  - $r = 1$: Perfect positive correlation.
  - $r = -1$: Perfect negative correlation.
  - $r = 0$: No linear correlation.

- **Chisini mean**: The Chisini mean generalizes means by solving the equation:

$$f(M) = \frac{1}{n} \sum_{i=1}^{n} f(x_i),$$

  for M, where f is a function defined on the data.

- **Other means for bivariate data**
  - **Harmonic mean for two variables:**

$$H(X, Y) = \frac{2 \cdot X \cdot Y}{X + Y}.$$

  - **Weighted mean:** For weights $w_1, \ldots, w_n$:

$$\text{Weighted mean} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

1.14 Summary statistics in R

- `mean()`: Compute the arithmetic mean.
- `median()`: Compute the median.
- `var()`: Compute the variance.
- `sd()`: Compute the standard deviation.
- `cov()`: Compute the covariance.
- `cor()`: Compute the correlation coefficient.

2 Probability

2.1 Random experiment

A **random experiment** is an action or process that leads to one of several possible outcomes. Although all possible outcomes are known, the actual outcome of the experiment is uncertain.

Example 1: tossing a coin
When tossing a coin, the possible outcomes are:

Heads (H) or Tails (T)

However, you cannot predict whether you will get heads or tails on a specific toss.

Example 2: drawing a ball from an urn
Consider an urn with 3 balls: one red, one green, and one blue. The possible outcomes when drawing a ball are:

Red, Green, Blue

But you do not know which ball will be drawn before performing the experiment.

Example 3: stock prices
When observing the price of a stock, the possible outcomes could be:

Price Increase, Price Decrease, Price Stays the Same

Even though the possibilities are known, the exact change in the stock price remains unpredictable.

Example 4: number of people at the emergency room
On a Saturday night, the number of people arriving at an emergency room in one hour is a random outcome. The possible outcomes are non-negative integers:

$$0, 1, 2, 3, \ldots$$

But the exact number of people cannot be known in advance.

Key Points:

- All possible outcomes of a random experiment are known.
- The specific outcome that will occur is uncertain.

2.2 The space of elementary events (sample space) Ω

In probability theory, the **sample space** Ω represents the set of all possible outcomes of a random experiment.

2.2.1 Sample Points ω

Each individual outcome of the experiment is called a **sample point**, denoted by ω. This is just one specific result that can occur when you perform the experiment. For example: if you flip a coin, one possible outcome is **heads** (denoted as H), and another is **tails** (denoted as T). Here, each of these outcomes H and T is a sample point.

2.2.2 The sample space Ω

The entire set of possible outcomes (all the sample points) forms the **sample space** Ω. So, for a coin toss:

$$\Omega = \{H, T\}$$

This means the sample space contains two outcomes: heads and tails.

2.2.3 Finite sample space $\#\Omega$

When the number of possible outcomes is limited or finite, we say that the sample space is *finite.* The symbol $\#\Omega$ represents the **cardinality** (i.e., the number of elements or outcomes) in the set Ω.

Examples
1. Tossing a coin k times

- $k = 1$ (one coin toss):

  $$\Omega = \{H, T\}$$

  This means there are two possible outcomes: heads (H) or tails (T). The **cardinality** (number of elements) of Ω is:

  $$\#\Omega = 2$$

- $k = 2$ (two coin tosses):

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

  There are 4 possible outcomes: both heads (HH), heads first and tails second (HT), tails first and heads second (TH), or both tails (TT). The cardinality is:

  $$\#\Omega = 4$$

- $k = 3$ (three coin tosses):

  $$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

  There are 8 possible outcomes. The cardinality is:

  $$\#\Omega = 8$$

2. Rolling dice k times

- $k = 1$ (rolling one die):
  A standard dice has 6 faces, numbered from 1 to 6. The sample space Ω is:

  $$\Omega = \{1, 2, 3, 4, 5, 6\}$$

  Each number corresponds to a possible outcome when rolling the die. The cardinality is:

  $$\#\Omega = 6$$

- $k = 2$ (rolling two dice):
  When rolling two dice, each die can show a number from 1 to 6, so the sample space is all the possible pairs of outcomes, where each pair represents the result of both dice:
  $$\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$$
  There are $6 \times 6 = 36$ possible outcomes. The cardinality is:
  $$\#\Omega = 36$$
- $k = 3$ (rolling three dice):
  Now, you roll three dice, and the sample space consists of all possible triples of outcomes:
  $$\Omega = \{(1,1,1), (1,1,2), \ldots, (6,6,6)\}$$
  There are $6 \times 6 \times 6 = 216$ possible outcomes. The cardinality is:
  $$\#\Omega = 216$$

2.2.4 Key points

- **Sample space** Ω: Set of all possible outcomes of a random experiment.
- **Sample point** ω: A single outcome within Ω.
- **Cardinality** $\#\Omega$: The number of elements in Ω.

Practical Example
Consider flipping a coin twice. What is the probability of getting two heads in a row? The sample space is:
$$\Omega = \{HH, HT, TH, TT\}$$
The event "getting two heads" is $\{HH\}$.
Since each outcome is equally likely and there are 4 possible outcomes, the probability is:
$$P(\text{two heads}) = \frac{1}{\#\Omega} = \frac{1}{4}$$

2.2.5 Events and sample space

Let Ω be a finite set that contains all possible outcomes of an experiment. This set Ω is called the **sample space**.
An event A is any subset of the sample space Ω. In mathematical notation:
$$A \subseteq \Omega$$

Is Ω an event?
Yes, Ω is an event because it is a subset of itself:
$$\Omega \subseteq \Omega$$

This is called the ***sure event.***

Is the empty set $\emptyset$ an event?

Yes, the empty set $\emptyset$ is also an event. Since $\emptyset$ has no elements, but is still a subset of Ω, it is called the ***impossible event***:

$$\emptyset \subseteq \Omega$$

Vacuously true statement
The statement, *"If $\omega \in \emptyset$, then $\omega \in \Omega$"* is vacuously true because $\emptyset$ has no elements.

Outcome of a random experiment
When we conduct a random experiment, exactly one outcome ω is obtained. We say that event A occurs if $\omega \in A$.

Notation: sets vs. tuples

- Curly brackets {} denote **sets** (order doesn't matter):

$$\{a, b, c\} = \{b, c, a\}$$

- Round brackets () denote **tuples** (order matters):

$$(a, b, c) \neq (b, c, a)$$

Example: three coin flips
The sample space Ω of three coin flips is:

$$\Omega = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HHH}\}$$

Some example events:

- We get tails at least once:

$$A = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}\}$$

- We get heads twice:

$$B = \{\mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}\}$$

- We get tails at most once:

$$C = \{\mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HHH}\}$$

- We get heads three times:

$$D = \{\mathrm{HHH}\}$$

2.3 Inclusion relationship between events

In probability theory, events can have a relationship called **inclusion**, where one event is entirely contained within another. This relationship is expressed in terms of **subsets.**

2.3.1 Subset relationship ($A \subseteq B$)

If $A \subseteq B$, it means that every outcome ω that belongs to event A also belongs to event B. In simpler terms, if event A happens, then event B must also happen because all outcomes in A are included in B.

Example: Let's say you have two events:

- A: "It rains today."
- B: "It rains sometime this week."

If it rains today (A), then it must also rain this week (B), so $A \subseteq B$.

2.3.2 Equality of events ($A = B$)

Two events A and B are equal ($A = B$) if and only if both of the following are true:

- $A \subseteq B$ (every element of A is in B),
- $B \subseteq A$ (every element of B is in A).

In other words, $A = B$ when both events have the exact same outcomes. This means that A occurs if and only if B occurs, meaning the two events are identical.

2.3.3 Proper subset ($A \subset B$)

Event A is a proper subset of B ($A \subset B$) if:

- $A \subseteq B$ (all elements of A are in B),
- $A \neq B$ (event A is not equal to B, meaning B contains at least one outcome that A does not have).

In other words, A is part of B, but B has more outcomes than A.

2.3.4 Important properties

- **Every event A is a subset of Ω:** $A \subseteq \Omega$ for every event A because every event A is formed from outcomes in the sample space Ω.
- **The empty set $\emptyset$ is a subset of every event:** $\emptyset \subseteq A$ for every event A because $\emptyset$ has no elements, so it automatically satisfies the subset condition.

Example
Let's consider an experiment where we flip a coin three times. The sample space is:

$$\Omega = \{\text{TTT}, \text{TTH}, \text{THT}, \text{HTT}, \text{THH}, \text{HHT}, \text{HTH}, \text{HHH}\}$$

Now, let's define two events:

- A: "We get tails on both the second and third tosses" $A = \{\text{TTT}, \text{HTT}\}$
- B: "We get tails on the second toss" $B = \{\text{TTT}, \text{TTH}, \text{HTT}, \text{HTH}\}$

Notice that A is a subset of B ($A \subseteq B$) because every outcome in A is also in B:

- TTT is in both A and B.
- HTT is in both A and B.

However, B has more outcomes than A, such as TTH and HTH, which are not part of A. Therefore, A is a proper subset of B ($A \subset B$).

2.4 Operations among events

When working with events in probability, we can perform various set operations that correspond to logical combinations of events. Let's break down the operations using the events A and B and their relationships.

2.4.1 Union of events ($A \cup B$)

The union of two events A and B, denoted $A \cup B$, is the set of all outcomes that belong to at least one of the two events A or B. The union occurs if either event A or event B (or both) occur.

$$A \cup B = \{\omega \in \Omega : \omega \text{ belongs to } A \text{ or } B\}$$

Example: Let's assume we flip a coin three times. The sample space is:

$$\Omega = \{\text{TTT}, \text{TTH}, \text{THT}, \text{HTT}, \text{THH}, \text{HHT}, \text{HTH}, \text{HHH}\}$$

Define two events:

$$A = \{\text{TTT}, \text{TTH}, \text{THT}, \text{THH}\} \quad \text{(Tails on the first toss)}$$

$$B = \{\text{THH}, \text{HHT}, \text{HTH}\} \quad \text{(One tail and two heads)}$$

The union of A and B is:

$$A \cup B = \{\text{TTT}, \text{TTH}, \text{THT}, \text{THH}, \text{HHT}, \text{HTH}\}$$

2.4.2 Intersection of events ($A \cap B$)

The intersection of two events A and B, denoted $A \cap B$, is the set of outcomes that belong to both events A and B. The intersection occurs if both A and B occur simultaneously.

$$A \cap B = \{\omega \in \Omega : \omega \text{ belongs to both } A \text{ and } B\}$$

Example: Using the same events A and B from the coin toss example, the intersection of A and B is:

$$A \cap B = \{\text{THH}\}$$

This outcome (THH) is the only one where both A (tails on the first toss) and B (one tail and two heads) happen.

2.4.3 Complementation of events (A^c)

The complement of an event A, denoted A^c (or $\Omega \setminus A$), is the set of outcomes in the sample space Ω that do not belong to A. The complement occurs if A does not occur.

$$A^c = \{\omega \in \Omega : \omega \text{ does not belong to } A\}$$

Example: Using the same event $A = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{THH}\}$, the complement of A is:

$$A^c = \{\mathrm{HTT}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HHH}\}$$

This event corresponds to "heads on the first toss" (since tails on the first toss is excluded).

2.4.4 Difference between events ($A \setminus B$)

The difference between two events A and B, denoted $A \setminus B$, is the set of outcomes that belong to A but not B. It occurs if A happens, but B does not.

$$A \setminus B = \{\omega \in A \text{ and } \omega \notin B\}$$

Example: Using the same events $A = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{THH}\}$ and $B = \{\mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}\}$, the difference $A \setminus B$ is:

$$A \setminus B = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}\}$$

This event represents outcomes where there are tails on the first toss, but there are at most one head.

2.4.5 Example of operations with three coin flips

Let's summarize the operations using three coin flips, where $\Omega = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HHH}\}$:

$$A = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{THH}\} \quad \text{(Tails on the first toss)}$$

$$B = \{\mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}\} \quad \text{(One tail and two heads)}$$

1. **Union** ($A \cup B$):
$$A \cup B = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}, \mathrm{THH}, \mathrm{HHT}, \mathrm{HTH}\}$$
This event includes outcomes where either A or B (or both) occur.

2. **Intersection** ($A \cap B$):
$$A \cap B = \{\mathrm{THH}\}$$
This outcome is the only one that satisfies both events.

3. **Complement** (A^c):
$$A^c = \{\mathrm{HTT}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HHH}\}$$
This event includes outcomes where tails do not appear on the first toss.

4. **Difference** ($A \setminus B$):
$$A \setminus B = \{\mathrm{TTT}, \mathrm{TTH}, \mathrm{THT}\}$$
These are the outcomes where tails appear on the first toss but heads appear at most once.

2.5 Properties of set operations among events

When working with events in probability, several important **properties** govern how events interact with each other through union, intersection, and other operations. These properties are similar to basic set operations in mathematics.

2.5.1 Commutative property

The **commutative property** states that the **order** in which we take the union or intersection of two events does not matter.

- **Union**:

$$A \cup B = B \cup A$$

This means that the union of A and B is the same whether you write A first or B first.

- **Intersection**:

$$A \cap B = B \cap A$$

This means that the intersection of A and B is the same regardless of the order.

Example: Consider the events $A = \{\text{TTT, TTH}\}$ and $B = \{\text{TTH, HTT}\}$.

$$A \cup B = \{\text{TTT, TTH, HTT}\}$$

and

$$B \cup A = \{\text{TTH, HTT, TTT}\}$$

Both are equal, confirming the commutative property for union. Similarly,

$$A \cap B = \{\text{TTH}\}$$

and

$$B \cap A = \{\text{TTH}\}$$

which confirms the commutative property for intersection.

2.5.2 Distributive property

The **distributive property** allows us to distribute union and intersection over each other.

- **Distributive over intersection**:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- **Distributive over union**:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Example: Let's say $A = \{\text{TTT, TTH}\}$, $B = \{\text{TTH, HTT}\}$, and $C = \{\text{HTH, HHH}\}$. We can check:

$$A \cap (B \cup C) = \{\text{TTH}\}$$

and

$$(A \cap B) \cup (A \cap C) = \{\text{TTH}\} \, (\text{since} \, A \cap C = \emptyset).$$

Thus, both sides are equal, confirming the distributive property.

2.5.3 Associative property

The **associative property** allows us to group events without affecting the result.

- **Union**:

$$A \cup (B \cup C) = (A \cup B) \cup C$$

- **Intersection**:

$$A \cap (B \cap C) = (A \cap B) \cap C$$

Thanks to the associative property, we can write:

$$A \cup B \cup C \quad \text{and} \quad A \cap B \cap C$$

without needing parentheses to specify the grouping.
Example: Let's use $A = \{\text{TTT, TTH}\}$, $B = \{\text{TTH, HTT}\}$, and $C = \{\text{THH, HHT}\}$.

$$A \cup (B \cup C) = \{\text{TTT, TTH, HTT, THH, HHT}\}$$

and

$$(A \cup B) \cup C = \{\text{TTT, TTH, HTT, THH, HHT}\}$$

This confirms the associative property.

2.5.4 Disjoint (or incompatible) events

Two events A and B are disjoint or incompatible if they cannot occur together. This means that their intersection is the empty set:

$$A \cap B = \emptyset$$

Example: Consider the events $A = \{\text{TTT}\}$ (all tails) and $B = \{\text{HHH}\}$ (all heads). These events cannot happen simultaneously, so they are disjoint.

$$A \cap B = \emptyset$$

2.5.5 Mutually exclusive events

A set of events $\{H_1, H_2, \ldots, H_k\}$ is mutually exclusive (or pairwise exclusive) if no two events can occur together. In other words, for every pair of events, their intersection is empty:

$$H_i \cap H_j = \emptyset \quad \text{for all} \quad i \neq j$$

Example: Consider three events in a coin flip: - $H_1 = \{\text{TTT}\}$ (all tails), - $H_2 = \{\text{HHH}\}$ (all heads), - $H_3 = \{\text{THH}\}$ (two heads and one tail). These events are mutually exclusive because no two of them can happen at the same time.

2.5.6 Exhaustive events

A set of events $\{H_1, H_2, \ldots, H_k\}$ is exhaustive if together they cover the entire sample space. This means their union is the entire sample space:

$$\Omega = H_1 \cup H_2 \cup \cdots \cup H_k$$

Example: In the case of flipping a coin twice, let's define the events: - $H_1 = \{\text{HH}\}$ (two heads), - $H_2 = \{\text{TT}\}$ (two tails), - $H_3 = \{\text{HT, TH}\}$ (one head, one tail). These events are exhaustive because:

$$\Omega = \{HH, TT, HT, TH\} = H_1 \cup H_2 \cup H_3$$

2.5.7 Partition of the sample space

If a set of events is **mutually exclusive** and **exhaustive**, it forms a **partition** of the sample space. A partition divides the sample space into distinct, non-overlapping events that together cover the entire space. ***Example***: Let's flip a coin twice. The sample space is:

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

Define the events: - H_1 = {HH}, - H_2 = {HT, TH}, - H_3 = {TT}. These events are mutually exclusive (no overlap) and exhaustive (they cover the whole sample space), so they form a partition of Ω.

Summary

- **Commutative property**:

$$A \cup B = B \cup A \quad \text{and} \quad A \cap B = B \cap A$$

- **Distributive property**:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad \text{and} \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- **Associative property**:

$$A \cup (B \cup C) = (A \cup B) \cup C \quad \text{and} \quad A \cap (B \cap C) = (A \cap B) \cap C$$

- **Disjoint events**:

$$A \cap B = \emptyset \quad \text{(A and B cannot happen together).}$$

- **Mutually exclusive events**:

$$H_i \cap H_j = \emptyset \quad \text{for all} \quad i \neq j$$

- **Exhaustive events**:

$$\Omega = H_1 \cup H_2 \cup \cdots \cup H_k \quad \text{(events cover the entire sample space).}$$

- **Partition**: A set of events that is both mutually exclusive and exhaustive.

2.6 De Morgan's Laws

De Morgan's laws describe the relationship between the union and intersection of sets when applying the complement operator.

2.6.1 De Morgan's laws for two sets

1. The complement of the union of two sets A and B is the intersection of their complements:

$$(A \cup B)^c = A^c \cap B^c$$

2. The complement of the intersection of two sets A and B is the union of their complements:

$$(A \cap B)^c = A^c \cup B^c$$

2.6.2 De Morgan's laws for a collection of sets

If we extend De Morgan's laws to any given collection of sets $\{A_i : i \in I\}$, the following rules hold:

1. The complement of the union of a collection of sets is the intersection of their complements:

$$\left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} A_i^c$$

2. The complement of the intersection of a collection of sets is the union of their complements:

$$\left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} A_i^c$$

Example: Consider three sets:
$A = \{1, 2, 3\}$
$B = \{3, 4, 5\}$
$C = \{5, 6, 7\}$
Let's calculate the complement of their union and intersection:

1. **Union and its complement**:

$$A \cup B \cup C = \{1, 2, 3, 4, 5, 6, 7\}$$

If the universal set $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, then:

$$(A \cup B \cup C)^c = U \setminus (A \cup B \cup C) = \{8, 9\}$$

This corresponds to:

$$A^c \cap B^c \cap C^c = \{8, 9\}$$

2. **Intersection and its complement:**

$$A \cap B \cap C = \emptyset$$

Since the three sets have no elements in common, their intersection is empty. The complement of the empty set is the universal set U:

$$(A \cap B \cap C)^c = U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

This corresponds to:

$$A^c \cup B^c \cup C^c = \{1, 2, 4, 6, 7, 8, 9\}$$

2.7 Probability on a finite sample space: classical approach

When the sample space Ω is finite, and we have reason to believe that all outcomes of the random experiment are equally likely, we can assign probabilities to events using the classical approach. For any event $A \subseteq \Omega$, the probability of A is given by the following ratio:

$$P(A) = \frac{\text{number of favorable outcomes for the event } A}{\text{total number of outcomes}} = \frac{\#A}{\#\Omega}$$

where $\#A$ is the number of elements in the set A, and $\#\Omega$ is the total number of elements in the sample space Ω. In this way, we have defined a function P on 2^Ω and the following three properties are satisfied:

1. $P(A) \geq 0$ for every $A \subseteq \Omega$,
2. $P(\Omega) = 1$,
3. if $A \subseteq \Omega$ and $B \subseteq \Omega$ are disjoint (i.e. $A \cap B = \emptyset$), then:

$$P(A \cup B) = P(A) + P(B)$$

Examples of random experiments with equally likely outcomes

1. **Rolling a dice**
The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, and each outcome is equally likely. Therefore:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

2. **Tossing an unbalanced coin**
The sample space is $\Omega = \{T, H\}$ (Tails and Heads), and if the coin is balanced, each outcome is equally likely:

$$P(\{T\}) = P(\{H\}) = \frac{1}{2} = 0.5$$

3. **Drawing a ball from an urn**
Suppose we have an urn containing 90 balls, each numbered from 1 to 90. The probability of drawing any specific ball is:

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{90\}) = \frac{1}{90}$$

Example: flipping a coin three times
Let's consider the random experiment of flipping a coin three times. The sample space is:

$$\Omega = \{TTT, TTH, THT, HTT, THH, HHT, HTH, HHH\}$$

Since there are three flips, the number of outcomes is:

$$\#\Omega = 2^3 = 8$$

1. **Event A: tails at least twice**

The event A consists of the following outcomes:

$$A = \{TTT, TTH, THT, HTT\}$$

Thus, the probability of A is:

$$P(A) = \frac{\#A}{\#\Omega} = \frac{4}{8} = 0.5$$

2. **Event C: heads at least once**
The event C consists of the following outcomes:

$$C = \{TTH, THT, HTT, THH, HHT, HTH, HHH\}$$

Thus, the probability of C is:

$$P(C) = \frac{\#C}{\#\Omega} = \frac{7}{8} = 0.875$$

3. **Event D: Heads in the first two tosses**
   The event D consists of the following outcomes:

$$D = \{HHT, HHH\}$$

   Thus, the probability of D is:

$$P(D) = \frac{\#D}{\#\Omega} = \frac{2}{8} = 0.25$$

2.8 Kolmogorov's three axioms

Probability is a mathematical way to measure the likelihood of an event happening. The foundation of modern probability theory is based on three key rules, known as **Kolmogorov's axioms**. These axioms define the properties that any probability function must satisfy.

2.8.1 Axiom 1: *Non-negativity*

The first axiom states that the probability of any event A must be a non-negative number. In other words, the probability can't be less than zero:

$$P(A) \geq 0$$

This makes sense because probabilities represent real-world situations, and negative probabilities would have no meaning.

Example:

- The probability of flipping a coin and getting "heads" is $P(\text{heads}) = 0.5$.
- The probability of rolling a die and getting a 3 is $P(3) = \frac{1}{6}$.

Both of these probabilities are greater than or equal to zero, as required by the first axiom.

2.8.2 Axiom 2: *Probability of the sure event is 1*

The second axiom states that the probability of the entire sample space (i.e., the event that something happens, no matter what) is 1:

$$P(\Omega) = 1$$

Here, Ω represents the **sample space**, which is the set of all possible outcomes. This axiom reflects the fact that something will happen for sure — there's no uncertainty about that.

Example: If you roll a die, the possible outcomes are $\Omega = \{1, 2, 3, 4, 5, 6\}$. The probability of rolling some number (anything from 1 to 6) is 1:

$$P(\Omega) = 1$$

because no matter what, you're guaranteed to roll one of these numbers.

2.8.3 Axiom 3: *Countable additivity*

The third axiom is a bit more technical. It says that if you have a collection of **disjoint events** (events that cannot happen at the same time), the probability of one of those events happening is the sum of their individual probabilities:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

This means that if you have events $A_1, A_2, A_3, \ldots$ that don't overlap (they can't all happen together), then the probability of any one of these events happening is simply the sum of the probabilities of each event.

Example: Suppose you have three disjoint events:

- A_1: Rolling a 1 on a dice.
- A_2: Rolling a 2 on a dice.
- A_3: Rolling a 3 on a dice.

Since these events can't happen at the same time (you can't roll a 1 and a 2 in the same roll), their probabilities add up:

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$

With each probability equal to $\frac{1}{6}$, we get:

$$P(A_1 \cup A_2 \cup A_3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = 0.5$$

So, the probability of rolling either a 1, 2, or 3 is 0.5.

Summary of the axioms

1. **Nonnegativity**: probabilities must be non-negative: $P(A) \geq 0$.
2. **Probability of the sure event**: the probability of the whole sample space (something happening for sure) is 1: $P(\Omega) = 1$.
3. **Countable additivity**: for disjoint events, the probability of one of them happening is the sum of their individual probabilities.

These three axioms form the backbone of probability theory and ensure that all probabilities are consistent and logical.

2.8.4 Consequences of the three axioms of Kolmogorov

1. **The probability of the impossible event**
   The first consequence is that the probability of the empty set, $\emptyset$, is zero:

   $$P(\emptyset) = 0$$

   The empty set represents an "impossible" event (an event that cannot occur). However, caution is needed here. This doesn't mean that every event with a probability of 0 is impossible. You can have an event A such that:

   $$P(A) = 0 \quad \text{and} \quad A \neq \emptyset$$

This can happen in cases with infinitely many possible outcomes (such as selecting a specific real number from an interval $[0, 1]$).

2. **Finite additivity (for disjoint events)**
If $A_1, A_2, \ldots, A_n$ are ***pairwise disjoint*** events (meaning no two events can happen at the same time, i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), the probability of their union equals the sum of the probabilities of each event:
$$P\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} P(A_j)$$
This is called ***finite additivity*** and it's a consequence of Kolmogorov's third axiom.

Example:
Imagine you roll a fair die. The possible outcomes are 1, 2, 3, 4, 5, and 6. Let's define the following pairwise disjoint events:

- A_1: Rolling a 1
- A_2: Rolling a 2 or 3
- A_3: Rolling a 4, 5, or 6

Since these events are disjoint (no overlap between them), the probability of rolling either 1, or 2 or 3, or 4, 5, or 6 is:
$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$$
Where:
$$P(A_1) = \frac{1}{6}, \quad P(A_2) = \frac{2}{6}, \quad P(A_3) = \frac{3}{6}$$
So, the total probability is:
$$P(A_1 \cup A_2 \cup A_3) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1$$
This makes sense because these events cover all possible outcomes of rolling the die.

3. **Monotonicity**
$$\textit{If } A \subseteq B, \textit{ then } P(A) \leq P(B).$$
This means that if event A is a subset of event B (i.e., whenever A happens, B also happens), the probability of A cannot be greater than the probability of B. This is logical since B includes everything that A can happen in, and maybe more.
Example:

- Let A be the event that a card drawn from a deck is a red card.
- Let B be the event that a card drawn from a deck is a card (which includes all possibilities). Clearly, $A \subseteq B$ and thus $P(A) \leq P(B)$.

4. **Bounds of probability:**
$$\textit{For every event } A, \ 0 \leq P(A) \leq 1.$$
This is a basic property of probabilities: the probability of any event is always between 0 and 1. A probability of 0 means the event will never happen, and a probability of 1 means the event will always happen.

Example:

- The probability of rolling a number between 1 and 6 on a standard die is 1 (since it's guaranteed).
- The probability of rolling a 7 on a standard die is 0 (since it's impossible).

5. **Complement rule**
For any event A,
$$P(A^c) = 1 - P(A)$$
The probability that event A does not happen is equal to 1 minus the probability that it does happen. This makes sense because the total probability must sum to 1 (either the event happens or it doesn't).

6. **Finite sub-additivity**
For any finite number of events A_1, A_2,..., A_n:
$$P\left(\bigcup_{j=1}^{n} A_j\right) \leq \sum_{j=1}^{n} P(A_j)$$
The probability that at least one of several events happens is less than or equal to the sum of the probabilities of those individual events. This is because when you sum the probabilities, you might be counting overlaps (i.e., situations where more than one event happens) more than once.

Example: If you flip a coin twice, let A_1 be the event that the first flip is heads, and A_2 the event that the second flip is heads. We know:
$$P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$$
In this case, both $P(A_1) = P(A_2) = 0.5$, so the sum would be 1. But $P(A_1 \cup A_2)$, which is the probability of getting heads at least once, is actually less than 1.

7. **Inclusion-exclusion principle**
For any two events A and B,
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
When you calculate $P(A \cup B)$, which is the probability that either A or B happens, you need to subtract $P(A \cap B)$, which is the probability that both happen. This is because when you add $P(A)$ and $P(B)$, you're counting the cases where both A and B happen twice, so you subtract it once to correct for that.

Example:

Let A be the event that you draw a red card from a deck (probability $P(A) = 0.5$).
Let B be the event that you draw a heart (probability $P(B) = 0.25$).
Since the event of drawing a heart is already included in the event of drawing a red card, we have $P(A \cap B) = 0.25$. Thus,
$$P(A \cup B) = 0.5 + 0.25 - 0.25 = 0.5$$

2.9 Where to define probability

2.9.1 Understanding the basic terms

1. **Set function (probability function)**
   The function P is called a ***probability function***. It assigns a probability to subsets of a set Ω (which represents the possible outcomes of some random experiment).
   Mathematically, we can write:
   $$P : \mathcal{F} \rightarrow [0, 1]$$
   This means that the function P takes an element of $\mathcal{F}$, which is a subset of Ω, and returns a value between 0 and 1. This value represents the probability of that event happening.

2. Ω
   This is the ***sample space***, or the set of all possible outcomes of an experiment. For example, if you're flipping a coin, $\Omega = \{\text{Heads}, \text{Tails}\}$.

3. . $\mathcal{F}$
   This is a collection of subsets of Ω, also called a ***sigma-algebra*** (or σ-algebra). It represents the different "events" that can happen based on the outcomes in Ω.

   So, if $\Omega = \{\text{Heads}, \text{Tails}\}$, then some subsets of Ω could be:
   $$\mathcal{F} = \{\emptyset, \{\text{Heads}\}, \{\text{Tails}\}, \{\text{Heads}, \text{Tails}\}\}$$

4. **Power Set 2^Ω**
   The ***power set*** 2^Ω is the set of all possible subsets of Ω. For a finite or countably infinite set Ω, the power set represents all the events we could possibly assign probabilities to. For example, if $\Omega = \{1, 2\}$, then the power set is:
   $$2^\Omega = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$$
   Here, $\emptyset$ is the empty set, representing no outcomes.

Finite or countably infinite Ω
If Ω is *finite* (e.g., outcomes from rolling a die) or *countably infinite* (e.g., the set of natural numbers $N = \{1, 2, 3, \dots\}$), then we can use the full power set 2^Ω. This means we can assign probabilities to every possible subset of Ω.
For example, if you're rolling a die:
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
The power set 2^Ω includes subsets like:
$$2^\Omega = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3, 4, 5, 6\}, \dots\}$$
You can define probability for each of these subsets.

Uncountable Ω
If Ω is *uncountable* (for example, if $\Omega = R$, the set of real numbers), it becomes impossible to define probability on the full power set 2^Ω.
Why? There are just too many subsets in the power set of an uncountable set, and assigning a probability to every single one would violate some mathematical properties of probability. So, instead, we need to work with a smaller collection $\mathcal{F}$, which is a subset of 2^Ω.

An example is when $\Omega = R$ (the real number line). In this case, we might work with a collection of "nice" subsets of R, like intervals (e.g., $[a, b]$, where a and b are real numbers), rather than the full power set.

Summary:

- P is a probability function that assigns values between 0 and 1 to subsets of a sample space Ω.
- $\mathcal{F}$ is the collection of subsets (or events) of Ω that we can assign probabilities to.
- If Ω is finite or countably infinite, we can use the full power set 2^{Ω}.
- If Ω is uncountable (like the real numbers R), we need to restrict to a smaller collection $\mathcal{F}$ of subsets to define probability properly.

2.10 Sigma-algebra properties

A sigma-algebra $\mathcal{F}$ is a collection of subsets of the sample space Ω that satisfies the following properties:

1. $\Omega \in \mathcal{F}$: The entire sample space must be in $\mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$: The complement of any event in $\mathcal{F}$ must also be in $\mathcal{F}$
3. If $A_1, A_2, A_3, \dots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$: The union of any countable collection of events in $\mathcal{F}$ must also be in $\mathcal{F}$.

Example:
Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, representing the outcomes of a dice roll.

1. Property 1: Ω must be in $\mathcal{F}$, so $\Omega = \{1, 2, 3, 4, 5, 6\} \in \mathcal{F}$.
2. Property 2: If $A = \{4\} \in \mathcal{F}$, then the complement $A^c = \{1, 2, 3, 5, 6\}$ must also be in $\mathcal{F}$.
3. Property 3: If $A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, \ldots$, then $\bigcup_{j=1}^{3} A_j = \{1, 2, 3\}$ must also be in $\mathcal{F}$.

2.11 Definition of probability

A ***probability*** P on a sample space Ω is a set function defined on a sigma-algebra $\mathcal{F}$ satisfying the following three axioms of Kolmogorov:

1. **Non-negativity**
   $$P(A) \geq 0 \quad \text{for every } A \in \mathcal{F}$$
   This means the probability of any event A cannot be negative.
2. **Normalization (the sure event)**
   $$P(\Omega) = 1$$
   The probability of the entire sample space Ω is 1, meaning something in Ω is guaranteed to happen.

3. **Countable additivity**

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

For any sequence of disjoint events $A_1, A_2, ..., A_n$, the probability of their union is the sum of their probabilities.

Example: coin toss
Consider tossing a fair coin. The sample space is:

$$\Omega = \{\text{Heads}, \text{Tails}\}$$

Let A_1 be the event of getting heads, and A_2 be the event of getting tails. These events are disjoint, meaning $A_1 \cap A_2 = \emptyset$.

- By **non-negativity**, $P(A_1) \geq 0$ and $P(A_2) \geq 0$.
- By **normalization**, $P(A_1) + P(A_2) = 1$. For a fair coin, this gives $P(A_1) = P(A_2) = 0.5$.
- If there were more disjoint events, their total probability would be the sum of their individual probabilities by **countable additivity**.

2.11.1 Probability spaces

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

- Ω is the **sample space**, which is the set of all possible outcomes of a random experiment.
- $\mathcal{F}$ is a **sigma-algebra** (a collection of subsets of Ω, called events, that includes Ω and is closed under complements and countable unions).
- P is a **probability measure**, which assigns a probability to each event in $\mathcal{F}$, satisfying the following properties:
- $0 \leq P(A) \leq 1$ for every event $A \in \mathcal{F}$
- $P(\Omega) = 1$
- If $A_1, A_2, \ldots$ are disjoint events, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

Example: rolling a fair die
Consider a fair six-sided die. The sample space is:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

An example of an event is rolling an even number:

$$A = \{2, 4, 6\}$$

Assuming the die is fair, the probability of each outcome is $P(\{i\}) = \frac{1}{6}$, where $i \in \{1, 2, 3, 4, 5, 6\}$. The probability of rolling an even number is:

$$P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

2.12 Conditional probability

Conditional probability refers to the probability of an event A occurring given that another event B has already occurred. We denote this as $P(A|B)$, which is read as "the probability of A given B.".
If Ω is the sample space (the set of all possible outcomes), then the formula for conditional probability is given by:

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

Where:

- $P(A \mid B)$ is the probability of event A occurring given that event B has occurred.
- $P(A \cap B)$ is the probability of both events A and B happening (i.e., the intersection of A and B).
- $P(B)$ is the probability of event B occurring.

Important: this formula is valid only if $P(B) > 0$, because we cannot condition on an event that has no chance of happening.

This formula can also be expressed in terms of the number of outcomes:

$$P(A|B) = \frac{\#(A \cap B)}{\#B}$$

Where:

- $\#(A \cap B)$ is the number of outcomes in both A and B.
- $\#B$ is the total number of outcomes in B.

Example 1: probability with a deck of cards
Imagine we have a standard deck of 52 playing cards. Let's define two events:

- Event A: drawing a heart (there are 13 hearts in a deck).
- Event B: drawing a red card (there are 26 red cards in a deck: 13 hearts and 13 diamonds).

Step 1: find $P(A \cap B)$

- $A \cap B$ means drawing a card that is both a heart and red. Since all hearts are red, we have:

$$\#(A \cap B) = 13$$

Step 2: find $P(B)$

- $\#B$ (the total number of red cards) is:

$$\#B = 26$$

Step 3: calculate $P(A|B)$
Now, we can use the conditional probability formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\#(A \cap B)}{\#B} = \frac{13}{26} = \frac{1}{2}$$

This means that if you draw a red card, the probability that it is a heart is $\frac{1}{2}$ or 50%.

Example 2: flipping a coin three times
When flipping a coin three times, the possible outcomes are:

$$\Omega = \{\text{TTT}, \text{TTH}, \text{THT}, \text{HTT}, \text{THH}, \text{HHT}, \text{HTH}, \text{HHH}\}$$

This is the sample space, which represents all possible outcomes of the experiment (flipping the coin three times). Each outcome is a sequence of heads (H) and tails (T).
Now, let's define the events from the problem:

- **Event A**: Tails occur exactly twice.

$$A = \{\text{TTT}, \text{TTH}, \text{THT}, \text{HTT}\}$$

- **Event C**: Heads appear at least once.

$$C = \{\text{TTH}, \text{THT}, \text{HTT}, \text{THH}, \text{HHT}, \text{HTH}, \text{HHH}\}$$

- **Event D**: Heads appear in the first two tosses.

$$D = \{\text{HHT}, \text{HHH}\}$$

Conditional probability: event B
We are given that **the first toss is tails**, so we are now considering a reduced sample space. This means the new event B is:

$$B = \{\text{TTT}, \text{TTH}, \text{THT}, \text{THH}\}$$

These are all the outcomes where the first toss is tails.

Intersection of Events

- Intersection of A and B: We want outcomes where there are two tails and the first toss is tails:
$$A \cap B = \{\text{TTT}, \text{TTH}, \text{THT}\}$$
These are the outcomes common to both A and B.

- Intersection of C and B: We want outcomes where there is at least one head, and the first toss is tails:
$$C \cap B = \{\text{TTH}, \text{THT}, \text{THH}\}$$
These are the outcomes common to both C and B.

- Intersection of D and B: We want outcomes where the first two tosses are heads, and the first toss is tails. But this is impossible, so:
$$D \cap B = \emptyset$$
The intersection is empty.

The conditional probability of an event A, given that B has occurred, is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Where:

- $P(A \cap B)$ is the probability of both A and B occurring.
- $P(B)$ is the probability that B occurs.

Since B has 4 possible outcomes, $P(B) = \frac{4}{8} = \frac{1}{2}$.
Now, let's calculate the conditional probabilities:

1. $P(A|B)$: There are 3 outcomes in $A \cap B$, so:
$$P(A|B) = \frac{3}{4} = 0.75$$
2. $P(C|B)$: There are 3 outcomes in $C \cap B$, so:
$$P(C|B) = \frac{3}{4} = 0.75$$
3. $P(D|B)$: Since $D \cap B$ is empty:
$$P(D|B) = 0$$

2.13 Independence of events

Two events, A and B, are said to be ***independent*** if the occurrence of one event does not affect the probability of the other event happening.
In other words, knowing that B happens does not change the probability of A, and vice versa.
The formal definition of independence is:

$$A \perp\!\!\!\perp B \iff P(A \cap B) = P(A) \cdot P(B)$$

Where:

- $P(A \cap B)$ is the probability that both events A and B occur.
- $P(A)$ is the probability that event A occurs.
- $P(B)$ is the probability that event B occurs.

2.13.1 Conditional probability with independent events

If two events are independent, then their ***conditional probabilities*** simplify.
The conditional probability of A given B, denoted as $P(A \mid B)$, is the probability that A occurs given that B has occurred.
For independent events, we can use:

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

Similarly:

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

This tells us that the probability of A does not change even if we know that B has happened, which is the hallmark of independence.

2.13.2 Disjoint (mutually exclusive) events

If two events are ***disjoint*** (mutually exclusive), they cannot both happen at the same time. This means:

$$A \cap B = \emptyset \quad \Rightarrow \quad P(A \cap B) = 0$$

So, **if A and B are disjoint, they cannot be independent**. This is because if one event occurs, the probability of the other occurring must be zero. For independent events, we expect both events to have positive probabilities.

Example: coin flip and die roll
Let's now move to the example of flipping a coin and rolling a die:

- Let A be the event "we get tails when flipping the coin"
- Let B be the event "we get a 5 or 6 when rolling the die"

Here, the sample space Ω consists of all possible outcomes of both events.

Step 1: compute individual probabilities
The probability of getting tails in a coin flip (event A) is:

$$P(A) = \frac{1}{2}$$

The probability of rolling a 5 or 6 on a die (event B) is:

$$P(B) = \frac{2}{6} = \frac{1}{3}$$

Step 2: compute the joint probability $P(A \cap B)$
Since the coin flip and die roll are independent events, we can compute the joint probability using the formula for independent events:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

This shows that the probability of getting tails on the coin and rolling a 5 or 6 on the die is $\frac{1}{6}$.

Step 3: verify by counting
Alternatively, we can count the number of favorable outcomes directly:

- The total number of possible outcomes (both flipping the coin and rolling the die) is $2 \times 6 = 12$.
- There are 2 favorable outcomes where we get tails and a 5 or 6 (specifically, Tails, 5 and Tails, 6).

Thus, we can compute:

$$P(A \cap B) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes}} = \frac{2}{12} = \frac{1}{6}$$

Both methods give the same result, confirming that the events are independent.

Conclusion
The key takeaway is that ***independent events*** are those where the occurrence of one does not impact the occurrence of the other, and **the probability of their intersection is the product of their individual probabilities**.

2.13.3 Independence of three events

The definition of independence for three events requires two conditions to be met:

- **Joint independence**

  $$P(A \cap B \cap C) = P(A)P(B)P(C)$$

  This means that the probability of all three events A, B, and C happening together (i.e., their intersection) is equal to the product of the probabilities of each event happening individually.

- **Pairwise independence**
  In addition to joint independence, the events must be ***pairwise independent***. This means that any two events from the set $\{A, B, C\}$ are also independent of each other. The conditions for pairwise independence are:

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

  $$P(A \cap C) = P(A)P(C)$$
  $$P(B \cap C) = P(B)P(C)$$

  Together, these conditions ensure that no matter how you group these events, they behave independently.

Example for three events
Suppose we roll a fair six-sided die twice. Define the following events:

- A: The first roll is a 1.
- B: The second roll is an even number.
- C: The sum of both rolls is greater than 4.

We will check if these events are independent:

- $P(A) = \frac{1}{6}$ (since there's one way to roll a 1 out of 6 possible outcomes).
- $P(B) = \frac{1}{2}$ (since 3 out of 6 numbers are even: 2, 4, 6).
- $P(C)$: The probability that the sum of two rolls is greater than 4 is $P(C) = \frac{21}{36} = \frac{7}{12}$.

Now, let's check joint and pairwise independence:

- $P(A \cap B) = P(\text{first roll is 1 and second roll is even}) = \frac{1}{12}$, and
- $P(A)P(B) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$. Thus, A and B are pairwise independent.

Similarly, you would check $P(A \cap C)$, $P(B \cap C)$, and $P(A \cap B \cap C)$ to confirm full independence.

2.13.4 Independence of n events

Now let's extend this concept to n events $A_1, A_2, \ldots, A_n$. These events are independent if the following conditions hold for every subset of the events:

- For every pair A_i and A_j:

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

- For every triplet A_i, A_j, A_k:

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$$

- In general, for any subset of k events $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$, we must have:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

In simple terms, n **events are independent if the probability of the intersection of any subset of the events is equal to the product of the probabilities of the individual events in that subset**. This means that independence is not just about all events together but also about every possible combination of them.

Example for multiple events
Consider five coin tosses. Define A_i as the event that the i-th toss results in heads. Each toss is independent of the others. The probability of any subset of events (for example, getting heads on toss 1 and toss 3) is simply the product of the individual probabilities, since each toss is independent.
For instance:

$$P(A_1 \cap A_3) = P(A_1)P(A_3) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P(A_1 \cap A_2 \cap A_5) = P(A_1)P(A_2)P(A_5) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

2.13.5 Independence of infinitely many events

For infinitely many events $A_1, A_2, A_3, \ldots$, the definition of independence is similar to the finite case. We say that $A_1, A_2, \ldots$ are independent if for every finite subset of events $A_1, \ldots, A_n$, the events are independent. That is:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

for every finite n and for every subset $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$.
This means that even when we deal with infinitely many events, the independence condition still only requires us to check finite subsets of the events.

2.14 Law of total probability

Given a set of ***exhaustive*** and ***mutually exclusive*** events $E_1, E_2, \ldots, E_k$, the law of total probability allows us to calculate the probability of an event B by considering all possible ways B can happen, weighed by the probability of each scenario. Mathematically:

$$P(B) = P(B \cap E_1) + P(B \cap E_2) + \cdots + P(B \cap E_k)$$

In words, this is the sum of the probabilities of B happening together with each event E_j.
By using conditional probabilities, the formula becomes:

$$P(B) = P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \cdots + P(B|E_k)P(E_k)$$

Where:

- $P(E_j)$ is the probability that event E_j happens.
- $P(B|E_j)$ is the conditional probability that B occurs given E_j.

Example: urns and balls problem
Let's apply the law of total probability to a practical scenario: you have k urns, and each urn contains a mix of black and white balls. You randomly select one urn, then draw a ball from that urn, and we want to know the probability of drawing a white ball.
Define:

- E_j: the event that the j-th urn is chosen.
- B: the event that a white ball is drawn.

The probability of drawing a white ball, $P(B)$, can be expressed using the law of total probability:

$$P(B) = \sum_{j=1}^{k} P(E_j) \cdot P(B|E_j)$$

Here:

- $P(E_j) = \frac{1}{k}$, since each urn is equally likely to be chosen.
- $P(B|E_j) = \frac{w_j}{w_j + b_j}$, where w_j is the number of white balls and b_j is the number of black balls in the j-th urn.

Thus, the overall probability of drawing a white ball is:

$$P(B) = \sum_{j=1}^{k} \frac{1}{k} \cdot \frac{w_j}{w_j + b_j}$$

This formula considers the likelihood of each urn being chosen and the chance of drawing a white ball from each urn.

2.15 Bayes Theorem

2.15.1 What is Bayes' Theorem?

Bayes' Theorem helps us find out the probability of something happening given new information. It updates the probability of an event based on evidence. For example, after taking a test for HIV, you want to know: *If the test is positive, what is the probability that I actually have HIV?* In other words, it calculates the probability of one event occurring, knowing that another event has already happened.

2.15.2 The formula of Bayes' Theorem

Bayes' Theorem can be written as:

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

Where:

- $P(A|B)$ is the probability of A happening, given B has occurred. This is the **conditional probability** we want to find.
- $P(B|A)$ is the probability of B happening, given that A is true. For example, it's the probability that the test is positive if someone has HIV.
- $P(A)$ is the **prior probability** of A happening before considering the new information. In our case, this is the probability of someone having HIV before taking the test.
- $P(B)$ is the total probability of B happening, regardless of whether A is true or not. In this case, it's the overall probability of testing positive.

2.15.3 Bayes' Theorem with expanded formula

To calculate $P(B)$, we use the **law of total probability**:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

Thus, Bayes' Theorem becomes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Here, A^c represents the event that A does not happen (i.e., the person does not have HIV).

General case with partition of the sample space
In a more general scenario, imagine that the sample space Ω is divided into several events $E_1, E_2, \ldots, E_k$. If we know $P(E_i)$ and $P(B|E_i)$, the probability of any specific event E_i happening given B occurs is:

$$P(E_i|B) = \frac{P(B|E_i)P(E_i)}{P(B)}$$

The denominator $P(B)$ is found using the law of total probability:

$$P(B) = P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \cdots + P(B|E_k)P(E_k)$$

Example: HIV testing
Let's use Bayes' Theorem to understand a real-world example:
Problem: Suppose in a population, only 1 in 10,000 people is infected with HIV. A test is conducted, which is:

- 99% accurate if the person has HIV.
- 5% inaccurate for someone who does not have HIV (false positive).

Given that the test result is positive, what is the probability that the person actually has HIV?

Solution
Define:

- H: the person has HIV.
- T: the test is positive.

We know:

- $P(H) = \frac{1}{10{,}000} = 0.0001$ (the prior probability of having HIV).
- $P(T|H) = 0.99$ (the likelihood of testing positive if the person has HIV).
- $P(T|H^c) = 0.05$ (the probability of testing positive if the person does not have HIV).
- $P(H^c) = 1 - P(H) = 0.9999$ (the probability of not having HIV).

Now, apply Bayes' Theorem:

$$P(H|T) = \frac{P(T|H)P(H)}{P(T|H)P(H) + P(T|H^c)P(H^c)}$$

Substitute the values:

$$P(H|T) = \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.05 \cdot 0.9999}$$

Simplify:

$$P(H|T) = \frac{0.000099}{0.000099 + 0.049995} = \frac{0.000099}{0.050094} \approx 0.002$$

So, even though the test is 99% accurate, the probability that the individual actually has HIV after testing positive is only **0.2%**. This result may be surprising, but it happens because HIV is very rare in the population, and even a small error rate can lead to many false positives.

2.15.4 Why Bayes' Theorem is important

Bayes' Theorem is crucial because it allows us to update what we know about an event after observing new evidence. In the HIV test example, we start with a very low probability of having HIV, but we revise this probability upward when the test comes back positive — even though the final probability remains low due to the rarity of HIV in the population.
This theorem is widely used in many areas, including:

- **Medical testing**: updating probabilities of diseases after diagnostic tests.
- **Spam filtering**: predicting whether an email is spam based on the presence of certain words.
- **Machine learning**: Bayesian algorithms use Bayes' Theorem to improve predictions.

Key takeaways

1. Bayes' Theorem helps us update probabilities based on new information.
2. It's useful for decision-making when we have incomplete information.
3. It can help avoid incorrect conclusions by accounting for the rarity of certain events.

3 Random variables

3.1 Definition of random variable

A ***random variable*** is a function that assigns a real number to each outcome of a random experiment. This function is defined over a ***probability space***, denoted as $(\Omega, \mathcal{F}, P)$, where:

- Ω is the set of all possible outcomes (sample space),
- $\mathcal{F}$ is a set of events (a collection of subsets of Ω),
- P is the probability measure that assigns probabilities to the events in $\mathcal{F}$.

A random variable is denoted as $X : \Omega \to R$. This means that for each outcome $\omega \in \Omega$, the random variable $X(\omega)$ gives a real number $x \in R$.

Example: coin toss
Imagine an experiment where you flip a coin. The possible outcomes (sample space) Ω are:

$$\Omega = \{\text{Heads}, \text{Tails}\}$$

Let's define a random variable X as the number of heads observed in one flip. We can represent this mathematically:

- $X(\text{Heads}) = 1$
- $X(\text{Tails}) = 0$

Here, X maps each outcome to a real number. So, $X : \Omega \to R$ where R is the set of real numbers (in this case, 0 and 1).

3.2 Measurability condition

The measurability condition ensures that we can assign probabilities to events related to the random variable. Specifically, for any ***Borel-measurable*** set $B \subseteq R$, we require that:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

This means that the set of outcomes ω that map to B must be an event in $\mathcal{F}$, so we can assign a probability to it.

k coin flips						k draws from an urn containing white and black balls					
X = number of tails						X = number of white balls drawn					
$k = 1$		$k = 2$		$k = 3$		$k = 1$		$k = 2$		$k = 3$	
Ω	X	Ω	X	Ω	X	Ω	X	Ω	X	Ω	X
$H \longrightarrow$	0	$HH \rightarrow$	0	$HHH \rightarrow$	0	$B \longrightarrow$	0	$BB \rightarrow$	0	$BBB \rightarrow$	0
$T \longrightarrow$	1	$TH \rightarrow$	1	$HHT \searrow$		$W \longrightarrow$	1	$BW \rightarrow$	1	$BBW \searrow$	
		$HT \nearrow$		$HTH \rightarrow$	1			$WB \nearrow$		$BWB \rightarrow$	1
		$TT \rightarrow$	2	$THH \nearrow$				$WW \rightarrow$	2	$WBB \nearrow$	
				$HTT \searrow$						$BWW \searrow$	
				$TTH \rightarrow$	2					$WWB \rightarrow$	2
				$THT \nearrow$						$WBW \nearrow$	
				$TTT \rightarrow$	3					$WWW \rightarrow$	3

3.3 Probability of events for a random variable

Now, the probability that the random variable X takes values in a set B (which is Borel-measurable) is:

$$P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

In the coin toss example, if $B = \{1\}$ (we want to find the probability of getting heads), then:

$$P(X \in \{1\}) = P(\{\text{Heads}\}) = \frac{1}{2}$$

Similarly, if $B = \{0\}$ (the probability of getting tails):

$$P(X \in \{0\}) = P(\{\text{Tails}\}) = \frac{1}{2}$$

3.4 Distribution of a random variable

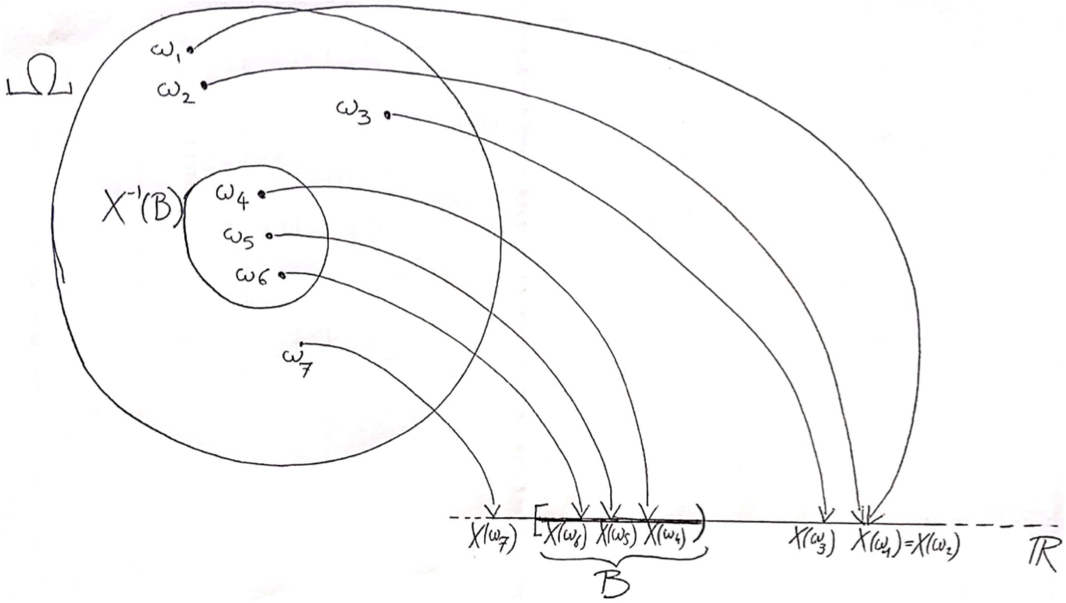
The ***distribution*** of a random variable refers to how the probabilities are distributed over different possible values of X. This distribution can be described by a function P_X, which tells us the probability that X takes on values in a set $B \subseteq R$ (the real numbers).
The function $P_X(B)$ gives the probability that $X \in B$, i.e., that X takes a value within the set B.
Mathematically, we express this as:

$$P_X(B) = P(X \in B) = P(X^{-1}(B))$$

This means the probability that X is in B is equivalent to the probability that the inverse of X is in B, connecting the distribution to the underlying probability space.

3.4.1 Kolmogorov's axioms

The function P_X must satisfy the three Kolmogorov's axioms of probability:



1. Non-negativity: for any set $B \subseteq R$, $P_X(B) \geq 0$.
2. Normalization: the probability of the entire sample space is 1, i.e., $P_X(R) = 1$
3. Additivity: if B_1 and B_2 are disjoint sets, then $P_X(B_1 \cup B_2) = P_X(B_1) + P_X(B_2)$.

This makes P_X a valid probability distribution.

3.4.2 Equality of random variables

If two random variables X and Y are ***almost surely equal***, meaning they are equal for all practical purposes except on a set of outcomes with probability zero, then they have the same distribution. We write this as:

$$P(X = Y) = 1$$

This is often denoted by:

$$X \sim Y$$

If $X \sim Y$, and if f is a Borel measurable function (a function that preserves the structure of the probability space), then applying f to both random variables gives:

$$f(X) \sim f(Y)$$

In simple terms, this means that if two random variables have the same distribution, applying the same function to them preserves that distribution.

Example: coin flips
You are flipping two coins, and X is the number of tails. The possible outcomes are:

1. HH (no tails, $X = 0$),

2. HT or TH (one tail, $X = 1$),

3. TT (two tails, $X = 2$).

The probabilities for each value of X are:

$$P(X = 0) = \frac{1}{4}, \quad P(X = 1) = \frac{1}{2}, \quad P(X = 2) = \frac{1}{4}$$

This gives the **distribution of** X, which can be written as:

$$P_X(B) = \frac{1}{4}\delta_0(B) + \frac{1}{2}\delta_1(B) + \frac{1}{4}\delta_2(B)$$

Here, $\delta_x(B)$ is called a *Dirac delta function* or ***point mass probability***. It indicates that P_X assigns a probability to specific values:

$$\delta_x(B) = \begin{cases} 1, & \text{if } x \in B \\ 0, & \text{if } x \notin B \end{cases}$$

For example, $\delta_0(B) = 1$ if $0 \in B$, and 0 otherwise.

3.5 Cumulative distribution function of a random variable

3.5.1 Definition of CDF

The CDF of a random variable X gives the probability that X takes on a value less than or equal to some real number x. Formally, it is written as:

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

This function $F(x)$ accumulates probabilities as x increases, hence the name ***cumulative distribution function***.

Practical example
Suppose X represents the score on a test (out of 100), and $F(x)$ gives the probability that a student scores x or less. For example:

- $F(50) = 0.75$ means that 75% of students scored 50 or lower.
- $F(90) = 0.95$ means that 95% of students scored 90 or lower.

3.5.2 Relationship between intervals and the CDF

The probability that X lies between two values a and b (where $a < b$) is given by the difference in the CDF values at b and a:

$$P(a < X \leq b) = F(b) - F(a)$$

This property is derived from the ***additivity*** of probability.

Example
Let's say we want to know the probability that a student scored between 60 and 80 on the test. If $F(80) = 0.90$ and $F(60) = 0.60$, the probability would be:

$$P(60 < X \leq 80) = F(80) - F(60) = 0.90 - 0.60 = 0.30$$

So, 30% of students scored between 60 and 80.

3.5.3 Properties of the CDF

A function $F(x)$ is a valid CDF if it satisfies the following three properties:

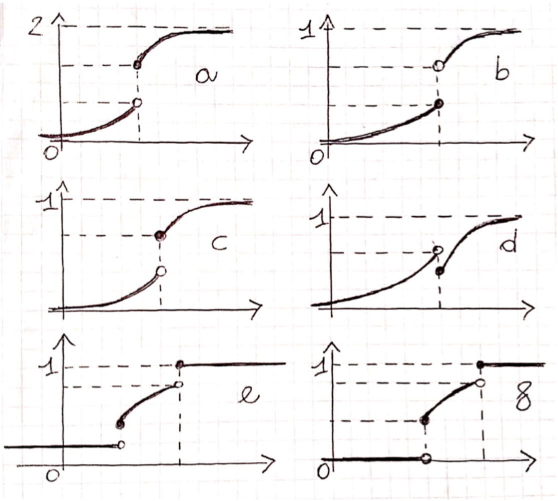
1. **Monotonicity**
$$x_1 < x_2 \implies F(x_1) \leq F(x_2)$$
This means the CDF never decreases as x increases. The probability that X is less than or equal to x can only stay the same or increase.

2. **Right-continuity**
$$\lim_{x \to x_0^+} F(x) = F(x_0)$$
The CDF is continuous from the right, meaning that as you approach a point from values greater than it, the CDF at that point equals the limit from the right.

3. **Boundary conditions**
$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$$
As x goes to negative infinity, the probability approaches zero, and as x goes to positive infinity, the probability approaches one. This reflects the fact that the probability of the random variable taking on a value within the entire real line is 1.

Example
Imagine X represents the height of people in a population. As you consider smaller and smaller values for x (approaching negative infinity), the probability that someone's height is less than x goes to 0. As x approaches very large values (positive infinity), the probability approaches 1 because the entire population's height falls within some reasonable range.

According to these properties, we can say which of the following functions are cumulative distribution functions:



- a) NO, as the maximum value of P should be 1 instead of 2

- b) NO, as it is not right-continuous
- c) YES
- d) NO, as it is decreasing
- e) NO, as the limit for $lim_{x \to -\infty} F(x) \neq 0$
- f) YES

3.6 Expected value

The expected value of a random variable X is essentially the **average or mean value** you would expect to get if you repeated an experiment many times. It takes into account all the possible values X can take, weighted by the probabilities of those values.
This "average" is defined mathematically, and it's crucial for understanding the long-term behavior of random phenomena.

3.6.1 Expected value for random variables with finitely many values

If the random variable X can take a finite number of possible values, say $x_1, x_2, \ldots, x_k$, then the expected value of X is given by:

$$E(X) = \sum_{j=1}^{k} x_j \cdot P(X = x_j)$$

This formula means we multiply each value x_j by its corresponding probability $P(X = x_j)$, and then sum the results.

Example
Consider a random variable X that can take the values $0, 1, 2$ with the following probabilities:

- $P(X = 0) = \frac{1}{4}$
- $P(X = 1) = \frac{1}{2}$
- $P(X = 2) = \frac{1}{4}$

To find the expected value, we use the formula:

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 0 + \frac{1}{2} + \frac{2}{4} = 1$$

So, the expected value $E(X) = 1$, which means that, on average, the value of X will be 1 in the long run.

3.6.2 Expected value for discrete random variables (countable infinite values)

When the random variable X takes an infinite (but countable) number of values, the expected value is given by a similar formula, but we sum over all possible values:

$$E(X) = \sum_{j=1}^{\infty} x_j \cdot P(X = x_j)$$

However, the expected value is only defined if the following condition holds:

$$\sum_{j=1}^{\infty} |x_j| \cdot P(X = x_j) < \infty$$

This condition ensures that the sum is finite and that the expected value is meaningful.

Example
Let's consider a scenario where a random variable X represents the number of times you flip a coin until you get a heads. The values of X can be $1, 2, 3, \ldots$ with probabilities:

- $P(X = 1) = \frac{1}{2}$
- $P(X = 2) = \frac{1}{4}$
- $P(X = 3) = \frac{1}{8}$
- $P(X = k) = \frac{1}{2^k}$

The expected value in this case is:

$$E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + \cdots$$

Using some algebraic techniques (like recognizing this as a geometric series), we can compute that the expected value is $E(X) = 2$.
This means that, on average, it will take 2 flips to get heads.

3.6.3 Expected value for absolutely continuous random variables

For continuous random variables, the expected value is computed using an integral rather than a sum. If the random variable X has a probability density function $f(x)$, the expected value is given by:

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

Again, this expected value is well-defined if the following condition is satisfied:

$$\int_{-\infty}^{+\infty} |x| \cdot f(x)\, dx < \infty$$

This ensures that the expected value exists.

Example
Let's consider a continuous random variable X that follows a uniform distribution over the interval $[0, 1]$. The probability density function is:

$$f(x) = 1 \quad \text{for} \quad 0 \leq x \leq 1$$

To find the expected value, we use the formula:

$$E(X) = \int_0^1 x \cdot 1\, dx = \int_0^1 x\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

So, the expected value $E(X) = \frac{1}{2}$, meaning that, on average, the values of X are centered around 0.5 in the interval.

3.6.4 Fundamental properties of the expectation

1. **Linearity**
   The linearity of expectation means that if you have two random variables, X and Y, both of which have finite expectations, and constants a and b that belong to the set of real numbers, then the expectation of the linear combination of X and Y, i.e., $aX + bY$, can be calculated using:
   $$E(aX + bY) = aE(X) + bE(Y)$$
   Example: Imagine you are playing two games. In the first game, you win an amount of money that is represented by the random variable X, and in the second game, the amount is represented by Y. Now, let's say that every time you play the first game, you multiply your winnings by a constant a, and for the second game, by a constant b. The linearity property tells us that if you want to calculate the average total winnings across both games, you can first find the expected value of your winnings from each game separately, multiply by the constants, and then add them together.

2. **Positivity**
   This property states that if a random variable X is always positive, i.e., $X > 0$, then its expectation must also be positive:
   $$E(X) > 0$$
   Example: Suppose you are measuring the height of people in a room (in meters), and heights cannot be negative. If X represents the height of a randomly chosen person from the room, the positivity property assures us that the average height (expected value) is also positive.

3. **Monotonicity**
   The monotonicity property tells us that if one random variable X is always less than or equal to another random variable Y (i.e., $0 \leq X \leq Y$), and Y has a finite expectation, then X also has a finite expectation, and moreover:
   $$E(X) \leq E(Y)$$
   Example: Consider two random variables representing the scores of two students on a test. If student A's score (represented by X) is always less than or equal to student B's score (represented by Y), then the expected score of A is less than or equal to the expected score of B.

4. **Absolute expectation**
   If the expectation of a random variable X exists, then the expectation of the absolute value of X, denoted $|X|$, also exists. Specifically, the following holds:
   $$|E(X)| \leq E(|X|)$$
   This means that the expectation of the absolute value of X is greater than or equal to the absolute value of the expectation of X.

   Example: If X represents your earnings from a game, the absolute value of your earnings cannot be smaller than the average (expected) value. This ensures that extreme fluctuations in your earnings don't make the expected value too unreliable.

5. **Bounded random variables**
   If there is a constant k such that $|X| \leq k$, meaning that the random variable X is bounded, then the expectation of X exists, is finite, and satisfies:
   $$|E(X)| \leq k$$
   Example: Suppose you have a dice game where the winnings can never exceed a fixed value, say $k = 10$. The boundedness property tells us that the expected value of your winnings will be less than or equal to 10.

6. **Same law (distribution)**
   If two random variables X and Y follow the same probability distribution, i.e., $X \sim Y$, and one of them admits finite expectation, then:
   $$E(X) = E(Y)$$
   Example: Imagine you flip two different coins, but both coins are fair, meaning they have the same probability of landing heads or tails. If the random variable X represents the outcome of the first coin and Y represents the outcome of the second, since both follow the same distribution, their expected values (e.g., the number of heads) will be the same.

3.7 Indicator function

Given a set $A \subset \Omega$, we define the ***indicator function*** of A on the set Ω as:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

The function I_A assigns a value of 1 if an element ω belongs to the set A, and 0 if it does not.

Example: Imagine rolling a fair six-sided die, where:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Let A be the set of even outcomes: $A = \{2, 4, 6\}$.

The indicator function $I_A(\omega)$ will work as follows:

- $I_A(2) = 1$, since 2 is in A.
- $I_A(3) = 0$, since 3 is not in A.
- $I_A(6) = 1$, since 6 is in A.

So, $I_A(\omega)$ helps us identify if an outcome ω belongs to the set A or not.

3.7.1 Indicator function as a random variable

Given a probability space $(\Omega, \mathcal{F}, P)$, I_A is a *random variable* if and only if $A \in \mathcal{F}$.
Here:

- Ω is the sample space
- $\mathcal{F}$ is a collection of subsets of Ω (called *events*)
- P is the probability measure.

The condition $A \in \mathcal{F}$ ensures that A is an event with a well-defined probability.

3.7.2 Expectation of the indicator function

If $A \in \mathcal{F}$, the expectation $E(I_A)$ exists and is given by:

$$E(I_A) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

where:

- $P(A)$ is the probability that an outcome is in A,
- $P(A^c)$ is the probability of the complement of A, i.e., $\omega \notin A$.

Example: Continuing with our die-rolling example, if the die is fair:

- $P(A) = \frac{3}{6} = 0.5$, since $A = \{2, 4, 6\}$ contains three outcomes out of six.
- Then $E(I_A) = P(A) = 0.5$.

The expectation $E(I_A)$ represents the probability of A occurring, confirming that I_A reflects the likelihood of ω being in A.

3.8 Bernoulli random variable

If $A \in \mathcal{F}$ (i.e., A is an event), then the indicator function I_A is a random variable that takes only two values: 0 and 1.
A random variable X that takes values only 0 or 1 is called a ***Bernoulli random variable***. This type of random variable models situations where there are two possible outcomes, typically associated with *success* and *failure*.

3.8.1 Bernoulli random variable with parameter p

A Bernoulli random variable X has a parameter p, where $p \in [0, 1]$. The parameter p represents the probability of X taking the value 1. Specifically:

- $P(X = 1) = p$: probability that $X = 1$.
- $P(X = 0) = 1 - p$: probability that $X = 0$.

In this setting, the expected value $E(X)$ of a Bernoulli random variable X is calculated as follows:

$$E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Thus, $E(X) = p$, meaning the expectation of a Bernoulli random variable is simply the probability of it being 1.

3.8.2 Connection with the indicator function

If A is an event, then the indicator function I_A is a Bernoulli random variable with parameter $P(A)$. This is because:

- $I_A(\omega) = 1$ with probability $P(A)$
- $I_A(\omega) = 0$ with probability $1 - P(A)$

Examples:

- **Number of divorced persons in a single-component family:** Define $X = 1$ if a person in a family is divorced; otherwise, $X = 0$.
- **Result of a single coin flip:** Define $X = 1$ if the outcome is tails, and $X = 0$ if it is heads. If the coin is fair, $p = 0.5$.

In each example, X can only be 0 or 1, fitting the definition of a Bernoulli random variable with parameter p.

3.9 Variance of a random variable

Let X be a random variable such that $E(X^2) < \infty$. Then, the expected value $\mu = E(X)$ also exists and is finite. We can define the ***variance*** of X as:

$$\operatorname{var}(X) := E\left((X-\mu)^2\right).$$

The ***standard deviation*** of X is the square root of the variance:

$$\sigma_X = \sqrt{\operatorname{var}(X)}.$$

Variance $\operatorname{var}(X)$ is a measure of how spread out the values of X are around the mean μ. A small variance means values are close to the mean, while a large variance indicates values are more spread out.

3.9.1 Properties of variance

1. **Zero variance condition**: $\operatorname{var}(X) = 0$ if and only if there exists a constant $c \in R$ such that $P(X = c) = 1$.
   In this case, X is a *degenerate random variable*, meaning it always takes a single value c with probability 1, so there's no variability.
2. **Alternative formula**: variance can also be computed as:

   $$\operatorname{var}(X) = E(X^2) - (E(X))^2.$$

   This formula is often useful because it allows us to calculate the variance without needing to explicitly compute $E((X-\mu)^2)$.

3.9.2 Variance of linear transformations

If X has finite variance and $a, b \in R$, then:

$$\operatorname{var}(aX + b) = a^2 \operatorname{var}(X).$$

Proof:

1. First, calculate the expected value of $aX + b$:

   $$E(aX + b) = aE(X) + b.$$

2. Now, find $\operatorname{var}(aX + b)$:

   $$\operatorname{var}(aX + b) = E\left((aX + b - E(aX + b))^2\right).$$

3. Substitute $E(aX + b) = aE(X) + b$:

$$= E\left((aX + b - (aE(X) + b))^2\right).$$

4. Simplify the expression inside E:

$$= E\left((a(X - E(X)))^2\right).$$

5. Factor out a and square it:

$$= E\left(a^2(X - E(X))^2\right) = a^2 E((X - E(X))^2).$$

6. Conclude that:

$$\mathrm{var}(aX + b) = a^2\,\mathrm{var}(X).$$

This result shows that scaling a random variable by a scales its variance by a^2, **while adding a constant b does not change the variance**.

3.10 Symmetry of the distribution of a random variable

The probability distribution of a random variable X is said to be *symmetric around* $c \in R$ if:

$$X - c \sim c - X,$$

meaning that the distribution of $X - c$ is the same as that of $c - X$. This implies that X has the same probability behavior on either side of c.

Example:
Suppose we have a random variable X with the following probability distribution:

$$P(X = 7) = \frac{1}{4}, \quad P(X = 10) = \frac{1}{2}, \quad P(X = 13) = \frac{1}{4}.$$

To check if this distribution is symmetric around $c = 10$, observe that:

- The values 7 and 13 are equidistant from 10, with $P(X = 7) = P(X = 13) = \frac{1}{4}$.
- The probability $P(X = 10) = \frac{1}{2}$ lies exactly at $c = 10$, indicating symmetry around $c = 10$.

Thus, this distribution is symmetric around $c = 10$.

3.10.1 Expectation of a symmetric distribution

If the distribution of X is symmetric around c and X has a finite expectation, then:

$$E(X) = \mu_X = c.$$

If X is symmetric around c, then $X - c$ and $c - X$ have the same distribution, and therefore the same expectation. Let's see why:

1. By the properties of expectation:

$$E(c - X) = E(X - c).$$

2. Expanding the terms and using the linearity of expectation, we get:
$$c - \mu_X = \mu_X - c.$$
3. Solving for μ_X, we obtain:
$$2c = 2\mu_X \Rightarrow \mu_X = c.$$

This result tells us that the mean μ_X of X lies at the center of the symmetry c.

3.10.2 Odd moments of a symmetric distribution

If X is symmetric around μ_X (the mean of X) and k is an odd integer, then:
$$E((X - \mu_X)^k) = 0,$$
provided that $E((X - \mu_X)^k$ exists.
This result follows because, for symmetric distributions, the odd-powered deviations from the mean $(X - \mu_X)^k$ are balanced on either side of μ_X, causing their average value (expectation) to be zero.

3.11 Markov and Chebychev's inequalities

3.11.1 Markov's inequality

If $X \geq 0$ (i.e., X is a non-negative random variable), then for any $x > 0$,
$$x\, P(X \geq x) \leq E(X).$$

Interpretation: Markov's inequality gives an upper bound on the probability that a non-negative random variable X is at least as large as some positive value x.

Proof of Markov's inequality
We can decompose X as:
$$X = XI_{\{X \geq x\}} + XI_{\{X < x\}}.$$
Since $I_{\{X \geq x\}}$ and $I_{\{X < x\}}$ are indicator functions, we have:
$$X \geq xI_{\{X \geq x\}}.$$
Now, by taking the expectation on both sides and using the monotonicity of expectation, we get:
$$E(X) \geq E\left(xI_{\{X \geq x\}}\right) = x\, E\left(I_{\{X \geq x\}}\right).$$
Since $E(I_{\{X \geq x\}}) = P(X \geq x)$, we have:
$$E(X) \geq x\, P(X \geq x),$$
which is Markov's inequality.

Example:
Suppose X is the time (in minutes) spent on a task, and the expected time $E(X) = 20$ minutes. Using Markov's inequality, we can find an upper bound for the probability that X takes 40 minutes or more:
$$P(X \geq 40) \leq \frac{E(X)}{40} = \frac{20}{40} = 0.5.$$
Thus, there's at most a 50% chance that the task will take 40 minutes or more.

3.11.2 Chebychev's inequality

For any $x > 0$,

$$P(|X - E(X)| \geq x) \leq \frac{\mathrm{Var}(X)}{x^2}.$$

Interpretation: Chebyshev's inequality provides an upper bound on the probability that the value of a random variable X deviates from its mean $E(X)$ by at least x units. This inequality applies to any random variable with finite variance, regardless of the distribution's shape.

Example:
Let X represent the score on an exam with:

- Mean $E(X) = 70$,
- Variance $\mathrm{Var}(X) = 25$.

Using Chebyshev's inequality, we can find the probability that the score deviates from the mean by at least 10 points:

$$P(|X - 70| \geq 10) \leq \frac{\mathrm{Var}(X)}{10^2} = \frac{25}{100} = 0.25.$$

Thus, there's at most a 25% chance that a score will be 10 or more points away from the mean of 70.

3.12 Independence of two random variables

Given two random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, P)$, we can consider events based on these variables. Specifically, for Borel-measurable subsets A and B of R, the event $\{X \in A, Y \in B\}$ is defined as:

$$\{X \in A, Y \in B\} = \{\omega \in \Omega : X(\omega) \in A \text{ and } Y(\omega) \in B\}.$$

Independence condition: we say that X and Y are *independent* (written $X \perp Y$) if for every $A, B \in \mathcal{B}$, the following condition holds:

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

This means that the probability of X taking a value in A and Y taking a value in B can be computed as the product of the individual probabilities.

Example:
Consider flipping a coin twice:

- Let X be the outcome of the first flip, where $X = 1$ for heads and $X = 0$ for tails.
- Let Y be the outcome of the second flip, with the same values as X.

Since the two flips do not influence each other, X and Y are independent.
For instance, if: $P(X = 1) = 0.5$ and $P(Y = 1) = 0.5$, then the probability $P(X = 1, Y = 1) = 0.5 \cdot 0.5 = 0.25$.

3.12.1 Independence condition using cumulative distribution functions (CDFs)

If F_X and F_Y are the cumulative distribution functions (CDFs) of X and Y, then X and Y are independent if and only if:

$$P(X \leq t, Y \leq s) = F_X(t) \cdot F_Y(s),$$

for every $s, t \in R$.

3.12.2 Independence of n random variables

For n random variables $X_1, X_2, \ldots, X_n$ defined on the same probability space, they are said to be *independent* if for every n-tuple $(B_1, B_2, \ldots, B_n)$ of Borel-measurable subsets of R, we have:

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

In other words, the probability of all events occurring can be computed as the product of the individual probabilities.

3.12.3 Condition using CDFs for multiple variables

The random variables $X_1, X_2, \ldots, X_n$ are independent if and only if:

$$P(X_1 \leq t_1, \ldots, X_n \leq t_n) = F_{X_1}(t_1) \cdots F_{X_n}(t_n),$$

for every $(t_1, \ldots, t_n) \in R^n$, where $F_{X_i}(t_i)$ is the CDF of X_i for each i.

3.12.4 Independence of countably infinite random variables

An infinite sequence $(X_n)_{n=1}^{\infty} = (X_1, X_2, \ldots)$ of random variables is said to be *independent* if any finite subset $X_1, X_2, \ldots, X_n$ is independent for every $n \geq 2$.

Example:
Consider an infinite sequence of coin flips:
Let $X_n = 1$ if the n-th flip is heads, and $X_n = 0$ if the n-th flip is tails.
Since each flip is independent, the sequence $(X_n)_{n=1}^{\infty}$ represents an infinite collection of independent random variables.

3.13 Covariance

Let X and Y be two random variables defined on the same probability space, with the following conditions:

- $E(X^2) < \infty$,
- $E(Y^2) < \infty$,
- $\mu_X = E(X)$ (the expectation of X),
- $\mu_Y = E(Y)$ (the expectation of Y).

3.13.1 Cauchy–Schwarz inequality

The ***Cauchy–Schwarz inequality*** provides an upper bound on $|E(XY)|$:

$$|E(XY)| \leq \sqrt{E(X^2) \cdot E(Y^2)}.$$

3.13.2 Definition of covariance

The ***covariance*** of X and Y, denoted $\operatorname{cov}(X, Y)$, measures the degree to which X and Y vary together and is defined as:

$$\operatorname{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

An alternative expression for covariance, which is often easier to calculate, is:

$$\operatorname{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Example:
Consider two random variables X and Y representing the scores on two different exams taken by a group of students:

- If high scores in one exam tend to correspond with high scores in the other, $\operatorname{cov}(X, Y) > 0$.
- If there's no consistent relationship between the scores on the two exams, $\operatorname{cov}(X, Y) \approx 0$.
- If high scores in one exam tend to correspond with low scores in the other, $\operatorname{cov}(X, Y) < 0$.

3.13.3 Properties of covariance

1. **Cauchy–Schwarz Bound:** By applying the Cauchy–Schwarz inequality to the covariance (using $X - E(X)$ and $Y - E(Y)$ instead of X and Y), we obtain:

   $$|\operatorname{cov}(X, Y)| \leq \sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)},$$

   where $\operatorname{var}(X)$ and $\operatorname{var}(Y)$ are the variances of X and Y, respectively.

2. **Variance as covariance:** The covariance of a variable with itself is equal to its variance:

   $$\operatorname{cov}(X, X) = \operatorname{var}(X).$$

3. **Symmetry:** Covariance is symmetric with respect to X and Y:

   $$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X).$$

4. **Bilinearity:** Covariance is bilinear, meaning that for any real numbers a and b, and any random variable Z,

   $$\operatorname{cov}(aX + bY, Z) = a \cdot \operatorname{cov}(X, Z) + b \cdot \operatorname{cov}(Y, Z).$$

Example: Suppose X represents the number of hours studied and Y represents exam scores. If both X and Y increase together, we would find that $\operatorname{cov}(X, Y) > 0$, indicating a positive relationship between study hours and exam performance.

These properties help us understand the behavior of covariance in different scenarios, particularly when assessing relationships between multiple random variables.

3.14 Correlation coefficient

If X and Y are two random variables with positive variance (i.e., they are not constant or degenerate), we define the ***correlation coefficient*** between X and Y as:

$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where:

- $\operatorname{cov}(X,Y)$ is the covariance between X and Y,
- $\operatorname{var}(X)$ and $\operatorname{var}(Y)$ are the variances of X and Y,
- σ_X and σ_Y are the standard deviations of X and Y,
- ρ_{XY} ranges from -1 to 1 by the Cauchy-Schwartz inequality.

3.14.1 Properties of the correlation coefficient

The value of ρ_{XY} provides insight into the relationship between X and Y:

1. $\rho_{XY} = 1$: ***Perfect positive linear dependence*** between X and Y. This means there exist constants $a > 0$ and $b \in R$ such that $Y = aX + b$.
2. $\rho_{XY} = -1$: ***Perfect negative linear dependence*** between X and Y. This means there exist constants $a < 0$ and $b \in R$ such that $Y = aX + b$.
3. $\rho_{XY} = 0$: X and Y are ***uncorrelated***, meaning that there is no linear relationship between them. This implies:

$$\operatorname{cov}(X,Y) = 0 \quad \text{and} \quad E(XY) = E(X)E(Y).$$

Examples:

1. **Example of perfect positive correlation ($\rho_{XY} = 1$):**
   Suppose $Y = 3X + 2$ and X is a random variable. Here, any increase in X leads to a proportional increase in Y. This results in $\rho_{XY} = 1$.
2. **Example of perfect negative correlation ($\rho_{XY} = -1$):**
   Suppose $Y = -2X + 5$. Here, an increase in X results in a proportional decrease in Y, which gives $\rho_{XY} = -1$.
3. **Example of Zero Correlation ($\rho_{XY} = 0$):**
   Let X represent a person's age and Y the number of books they read last year. Age may not influence the number of books read, so X and Y may have $\rho_{XY} = 0$, indicating they are uncorrelated.

The correlation coefficient ρ_{XY} provides insight into the linear relationship between X and Y, helping to determine the strength and direction of this relationship.

3.15 Correlation coefficient and variance

3.15.1 Variance of the sum of uncorrelated random variables

Think of a random variable, X, as a variable that can take on different values, each with a certain probability.
For a random variable X with mean μ:

$$\mathrm{var}(X) = E[(X-\mu)^2]$$

For random variables X and Y with means μ_X and μ_Y:

$$\mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

Proof

$$\begin{aligned}
\mathrm{var}\left(\sum_{j=1}^{k} X_j\right) &= \mathrm{cov}\left(\sum_{j=1}^{k} X_j, \sum_{l=1}^{k} X_l\right) \\
&= \sum_{j=1}^{k}\sum_{l=1}^{k} \mathrm{cov}(X_j, X_l) \\
&= \sum_{j=1}^{k}\sum_{\substack{l=1 \\ l\neq j}}^{k} \mathrm{cov}(X_j, X_l) + \sum_{j=1}^{k} \mathrm{cov}(X_j, X_j) \\
&= 0 + \sum_{j=1}^{k} \mathrm{var}(X_j) \\
&= \sum_{j=1}^{k} \mathrm{var}(X_j)
\end{aligned}$$

Example
Imagine three uncorrelated random variables:

- X_1: Rolling a die (variance = 2.92)
- X_2: Flipping a coin (variance = 0.25)
- X_3: Picking a card from a deck (variance = 6.72)

Variance of the total outcome:

$$\mathrm{var}(X_1 + X_2 + X_3) = 2.92 + 0.25 + 6.72 = 9.89$$

3.15.2 Variance of the linear combination of uncorrelated random variables

$\sum_{j=1}^{k} a_j X_j$ is a linear combination of random variables $X_1, X_2, \ldots, X_k$ with constants $a_1, a_2, \ldots, a_k$.
If X is a random variable and a is a constant:

$$\mathrm{var}(aX) = a^2\mathrm{var}(X)$$

and $\text{cov}(X_j, X_l) = 0 \quad \text{for } j \neq l$ for uncorrelated random variables

Proof

$$\text{var}\left(\sum_{j=1}^{k} a_j X_j\right) = \sum_{j=1}^{k} \text{var}(a_j X_j)$$
$$= \sum_{j=1}^{k} a_j^2 \text{var}(X_j)$$

Example

Consider three uncorrelated random variables:

- X_1: Rolling a die (variance = 2.92)
- X_2: Flipping a coin (variance = 0.25)
- X_3: Height of a person (variance = 15.6)

Form the linear combination:

$$Y = 2X_1 - 3X_2 + 0.5X_3$$

Calculate each variance:

$$\text{var}(2X_1) = 4 \cdot 2.92 = 11.68$$
$$\text{var}(-3X_2) = 9 \cdot 0.25 = 2.25$$
$$\text{var}(0.5X_3) = 0.25 \cdot 15.6 = 3.9$$

Total variance:

$$\text{var}(Y) = 11.68 + 2.25 + 3.9 = 17.83$$

3.15.3 Variance of the linear combination of any two random variables

For any random variable X and constant a:

$$\text{var}(aX) = a^2 \text{var}(X)$$

Covariance $\text{cov}(X, Y)$ measures how X and Y change together.

Formula

$$\text{var}(W) = a^2 \text{var}(X) + 2ab \, \text{cov}(X, Y) + b^2 \text{var}(Y)$$

Example

Let X be the temperature (variance = 4), Y be the humidity (variance = 9), and $\text{cov}(X, Y) = 1.5$. Consider $W = 3X - 2Y$:

- $a = 3$, so $3^2 \text{var}(X) = 9 \times 4 = 36$
- $b = -2$, so $(-2)^2 \text{var}(Y) = 4 \times 9 = 36$
- Covariance term: $2 \cdot 3 \cdot (-2) \cdot 1.5 = -18$

Total variance:

$$\text{var}(W) = 36 + 36 - 18 = 54$$

3.15.4 Relationship between independence and uncorrelation

Uncorrelation
X and Y are uncorrelated if:

$$\text{cov}(X, Y) = 0$$

Using the definition of covariance:

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y]$$

Therefore, X and Y are uncorrelated if:

$$E[XY] = E[X]E[Y]$$

Independence
X and Y are independent if:

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B)$$

Independence implies $E[XY] = E[X]E[Y]$, so X and Y are uncorrelated.

Relationship

- If X and Y are independent, they are uncorrelated.
- If X and Y are uncorrelated, they are not necessarily independent.

Example

- **Example of independence**: Rolling a die and flipping a coin are independent events, so the random variables are uncorrelated.
- **Example of uncorrelated but dependent variables**:
  - X is uniformly distributed over $[-1, 1]$.
  - $Y = X^2$.
  - $\text{cov}(X, Y) = 0$, but X and Y are not independent.

3.16 Linear combinations of random variables

3.16.1 Expected value of a linear combination

$$E\left[b + \sum_{j=1}^{k} a_j X_j\right] = b + \sum_{j=1}^{k} a_j E(X_j)$$

3.16.2 Variance of a linear combination

- **If $X_1, X_2, \ldots, X_k$ are independent**:

$$\text{var}\left(b + \sum_{j=1}^{k} a_j X_j\right) = \sum_{j=1}^{k} a_j^2 \text{var}(X_j)$$

- **If $X_1, X_2, \ldots, X_k$ are pairwise uncorrelated**:

$$\text{var}\left(b + \sum_{j=1}^{k} a_j X_j\right) = \sum_{j=1}^{k} a_j^2 \text{var}(X_j)$$

Example
Given X and Z are independent random variables:

- $E(X) = 8$, $\text{var}(X) = 0.5$
- $E(Z) = 0.4$, $\text{var}(Z) = 0.01$

Consider $Y = 3X - 4Z + 5$:

- **Expected Value**:

$$E(Y) = 3 \times 8 - 4 \times 0.4 + 5 = 27.4$$

- **Variance**:

$$\text{var}(Y) = 3^2 \times 0.5 + (-4)^2 \times 0.01 = 4.5 + 0.16 = 4.66$$

3.17 Independence properties

3.17.1 Independence of functions of random variables

If $X_1, X_2, \ldots, X_n$ are independent random variables, and $f : R^k \to R$ is a function where $k < n$, then:

$$f(X_1, \ldots, X_k) \text{ is independent of } X_{k+1}, \ldots, X_n.$$

***Example*: coin flipping**
Consider $n = 5$ coin tosses and define:

$$X_j = \begin{cases} 1 & \text{if the } j\text{-th toss is heads} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, \ldots, 5.$$

Let $Y = X_1 + X_2 + X_3$ be the number of heads in the first three tosses.
Then Y is independent of X_4 and X_5.

3.17.2 Multiplicative property of expectations

If X_1, X_2, and X_3 are independent:

$$E(X_1 X_2 X_3) = E(X_1 X_2) E(X_3) = E(X_1) E(X_2) E(X_3)$$

3.17.3 General case

For $X_1, X_2, \ldots, X_n$ independent:

$$E\left(\prod_{j=1}^{n} X_j\right) = \prod_{j=1}^{n} E(X_j)$$

4 Discrete random variables

A random variable X is said to be **discrete** if it takes (with probability 1) only a finite number or a countably infinite set of values. Formally, there exists a set C such that:

$$P(X \in C) = 1$$

where C is **at most countable**. This means:

- C can be finite (e.g., $\{1, 2, 3, 4, 5, 6\}$).
- Or C can be countably infinite, where the elements of C can be listed in a sequence (e.g., $\{0, 1, 2, 3, \dots\}$).

Examples

- **Number of tails in three coin tosses**: X can take values $\{0, 1, 2, 3\}$.
- **Number of white balls drawn from an urn**: The possible values of X are finite, depending on the number of draws.
- **Number of defective items produced in a day**: X can be any non-negative integer.
- **Number of cars at a tollbooth in a day**: X is a non-negative integer.
- **Number of heart attacks in Italy in a month**: X is a non-negative integer.

4.1 Probability mass function (PMF)

For a discrete random variable X, the **probability mass function** (PMF) $p(x)$ is given by:

$$p(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

where $x \in R$ represents the values that X can take.

4.1.1 Properties of PMF

- $p(x) \geq 0$ for all $x \in R$
- $\sum_x p(x) = 1$

Example
Let X be the number of children in a family sampled at random from a population.
Possible values of X are $0, 1, 2, 3, \dots$.
The PMF $p(x)$ gives the probability of each possible number of children:

Table 2: Number of kids in a family sampled at random from a population

x	0	1	2	3	4	tot
p(x)	0.24	0.47	0.17	0.08	0.04	1.00

4.1.2 Properties of the PMF

Let X be a discrete random variable that takes values $x_1, x_2, \dots$ with probability one. The PMF $p(x)$ has the following properties:

- **Non-negativity**:
$$0 \leq p(x) \leq 1 \quad \text{for every } x \in R$$
- **Support of the PMF**:
$$p(x) = 0 \quad \text{for every } x \notin \{x_1, x_2, \dots\}$$
- **Normalization**:
$$\sum_j p(x_j) = 1$$

4.1.3 Recovering the probability distribution from the PMF

For any Borel-measurable subset B of R, the probability that X belongs to B is:

$$P(X \in B) = \sum_{j : x_j \in B} p(x_j) = \sum_j p(x_j) \delta_{x_j}(B)$$

where $\delta_x(B)$ is defined as:

$$\delta_x(B) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

The PMF characterizes the probability distribution of a discrete random variable. If two discrete random variables have the same PMF, they have the same **law** (probability distribution).

4.2 Cumulative distribution function (CDF) for discrete random variables

The **cumulative distribution function (CDF)** $F(x)$ of a discrete random variable X is:

$$F(x) = P(X \leq x) = \sum_{j : x_j \leq x} p(x_j)$$

where $p(x_j)$ is the probability mass function (PMF) of X.

4.2.1 Properties of the CDF

- $F(x)$ is a **piecewise constant** function.
- The points where $F(x)$ **jumps** are the values in the support of X.
- If x is a point where $F(x)$ jumps, the **height of the jump** is $P(X = x)$.
- $F(x)$ is **non-decreasing**: if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
- $F(x)$ is **right-continuous**:
$$\lim_{x \to x_0^+} F(x) = F(x_0)$$
- **Limits at infinity**:
$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$$

Example: **number of kids**

x	0	1	2	3	4
p(x)	0.24	0.47	0.17	0.08	0.04
F(x)	0.24	0.71	0.88	0.96	1.00

Table 3: PMF and CDF

4.3 Discrete rvs - Expectation

The **expected value** (or **mean**) of a discrete random variable X is defined as:

$$E(X) = \sum_{j=1}^{\infty} x_j p(x_j) = \sum_{j=1}^{\infty} x_j P(X = x_j)$$

where $x_1, x_2, \ldots$ are the values that X can take and $p(x_j) = P(X = x_j)$ is the probability of X taking the value x_j.

4.3.1 Condition for finite expectation

The expectation $E(X)$ is **finite** if:

$$\sum_{j=1}^{\infty} x_j p(x_j) < +\infty$$

If the series diverges, $E(X)$ is not finite.

4.3.2 Special case: non-negative values

Even if X takes only non-negative values, the expectation is still given by:

$$E(X) = \sum_{j=1}^{\infty} x_j p(x_j)$$

However, $E(X)$ is not necessarily finite in this case.

4.4 Discrete rvs - Variance

The **variance** of a discrete random variable X is given by:

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

where $E[X]$ is the expected value of X.

Expanded form

$$\text{var}(X) = \sum_{j=1}^{\infty} (x_j - E[X])^2 p(x_j)$$

where $x_1, x_2, \ldots$ are the values that X can take, and $p(x_j) = P(X = x_j)$.

4.4.1 Condition for finite variance

The variance var(X) is finite if:

$$E[X^2] = \sum_{j=1}^{\infty} x_j^2 p(x_j) < +\infty$$

4.4.2 Alternative formula for variance

$$\mathrm{var}(X) = E[X^2] - (E[X])^2$$

where:

- $E[X^2] = \sum_{j=1}^{\infty} x_j^2 p(x_j)$
- $(E[X])^2 = \left(\sum_{j=1}^{\infty} x_j p(x_j)\right)^2$

4.5 Examples

4.5.1 1) Standard deviation of discrete rvs

Table 4: Number of children

x	0	1	2	3	4
p(x)	0.24	0.47	0.17	0.08	0.04
(x-E[X]ˆ2)	1.46	0.04	0.62	3.20	7.78
x^2	0	1	4	9	16

- The expectation of X is:

$$E(X) = \mu_X = 0 \times 0.24 + 1 \times 0.47 + 2 \times 0.17 + 3 \times 0.08 + 4 \times 0.04 = 1.21$$

- The variance of X is:

$$\mathrm{Var}(X) = 1.46 \times 0.24 + 0.04 \times 0.47 + 0.62 \times 0.17 + 3.20 \times 0.08 + 7.78 \times 0.04 = 1.04$$

The variance can also be computed as follows:

$$\mathrm{var}(X) = E(X^2) - \mu_X^2 = \sum_{j=1}^{5} x_j^2 p(x_j) - \mu_X^2$$

Calculating:

$$\mathrm{var}(X) = 0 \times 0.24 + 1 \times 0.47 + 4 \times 0.17 + 9 \times 0.08 + 16 \times 0.04 - 1.21^2 = 1.04$$

- The standard deviation is:

$$\mathrm{SD}(X) = \sqrt{\mathrm{var}(X)} = \sqrt{1.04} = 1.01$$

Event	x	$P(x = X)$	$xP(x = X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			E(X) $= 42 \approx 0.81$

4.5.2 2) Expected value of a discrete random variable

In a game of cards, you win 1£ if you draw a heart, 5£ if you draw an ace (including the ace of hearts), 10£ if you draw the king of spades and nothing for any other card you draw.
Write the probability model for your winnings, and calculate your expected winning.

4.6 Variability

We are often interested in the **variability** in the values of a random variable. Two important measures of variability are:

- **Variance**

$$\sigma^2 = \mathrm{Var}(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

where:

- σ^2 is the variance of X.
- $x_1, x_2, \ldots, x_k$ are the values that X can take.
- $E(X)$ is the expected value (mean) of X.
- $P(X = x_i)$ is the probability that X takes the value x_i.

- **Standard Deviation**

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

The standard deviation σ provides a measure of spread in the same units as the random variable X.

For the previous card game example, how much would you expect the winnings to vary from game to game?

x	$P(x = X)$	$xP(x = X)$	$(x - E(X))^2$	$P(x = X)(x - E(X))^2$
1	$\frac{12}{52}$	$\frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$\frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$\frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{10}{52} \times 84.4561 = 1.6242$
0	$\frac{35}{52}$	0	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		E$(X) = 0.81$		V$(X) = 3.4246$ SD$(X) = \sqrt{3.4246} = 1.85$

4.7 Joint and marginal probability mass functions

Given two discrete random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, P)$, the **joint probability mass function (pmf)** $p_{XY}(x, y)$ is defined as:

$$p_{XY}(x, y) = P(X = x, Y = y) = P\left(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}\right)$$

for every $x, y \in R$.

The **marginal pmfs** $p_X(x)$ and $p_Y(y)$ are given by:

$$p_X(x) = P(X = x) \quad \text{and} \quad p_Y(y) = P(Y = y)$$

where $x, y \in R$.

4.7.1 How to get the marginal PMFs from the joint pmf

- The marginal pmf of X is:

$$p_X(x) = \sum_{y \in \text{range of } Y} p_{XY}(x, y)$$

- The marginal pmf of Y is:

$$p_Y(y) = \sum_{x \in \text{range of } X} p_{XY}(x, y)$$

Note:
The **range of** X and the **range of** Y are at most countable sets, meaning they can be finite or countably infinite.

4.8 Independence of discrete random variables

Let X and Y be two discrete random variables with marginal pmfs p_X and p_Y, respectively. The random variables X and Y are **independent** if:

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$$

for every $x \in \{x_1, \ldots, x_k\}$ and every $y \in \{y_1, \ldots, y_h\}$, where k and h can be infinite.

***Example 1*: tossing a biased coin twice**
Let $P(T) = 0.2$ and $P(H) = 0.8$. Define:

$$X = \begin{cases} 1 & \text{if tails on the first toss} \\ 0 & \text{if heads on the first toss} \end{cases}, \quad Y = \begin{cases} 1 & \text{if tails on the second toss} \\ 0 & \text{if heads on the second toss} \end{cases}$$

Since the tosses are independent:

$$p_{XY}(1, 0) = 0.2 \times 0.8 = 0.16, \quad p_{XY}(0, 1) = 0.8 \times 0.2 = 0.16$$

$$p_{XY}(1, 1) = 0.2 \times 0.2 = 0.04, \quad p_{XY}(0, 0) = 0.8 \times 0.8 = 0.64$$

***Example 2*: rolling two dice**
Let X_1 and X_2 represent the outcomes of the first and second die, respectively. The outcomes are independent, so:

$$p_{X_1,X_2}(j,l) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = p_{X_1}(j) \cdot p_{X_2}(l)$$

for $j, l \in \{1,2,3,4,5,6\}$.

Note:

- We cannot conclude independence by checking only one pair of values.
- We must verify that $p_{XY}(x,y) = p_X(x) \cdot p_Y(y)$ for all combinations.

We want to determine if two discrete random variables X and Y are independent. Recall that X and Y are **independent** if:

$$p_{XY}(x,y) = p_X(x) \cdot p_Y(y)$$

for every x and y in their respective ranges.

Given probabilities

- **Joint probability:**

$$p_{XY}(2,6) = P(X=2, Y=6) = P\left((X_1,X_2) \in \{(2,4),(4,2)\}\right) = \frac{2}{36} = \frac{1}{18}$$

- **Marginal probability of X at 2:**

$$p_X(2) = P(X=2) = P(\text{both } X_1 \text{ and } X_2 \text{ are even}) = \frac{3 \cdot 3}{36} = \frac{1}{4}$$

- **Marginal probability of Y at 6:**

$$p_Y(6) = P(Y=6) = \frac{8}{36} = \frac{2}{9}$$

Checking for independence

$$p_{XY}(2,6) = \frac{1}{18} \quad \text{and} \quad p_X(2) \cdot p_Y(6) = \frac{1}{4} \times \frac{2}{9} = \frac{1}{18}$$

Since these are equal, the condition holds for $(2,6)$.

Additional check for independence
To conclude independence, we need to verify:

$$p_{XY}(x,y) = p_X(x) \cdot p_Y(y) \quad \text{for all } x \text{ and } y.$$

Example check for $(2,12)$:

- Joint probability: $p_{XY}(2,12) = \frac{1}{36}$
- Marginal probabilities: $p_X(2) = \frac{1}{4}, \quad p_Y(12) = \frac{1}{9}$

- Product: $p_X(2) \cdot p_Y(12) = \frac{1}{4} \times \frac{1}{9} = \frac{1}{36}$

Conclusion
Since $p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$ holds for all tested pairs, X and Y are independent.

Another example
Let X and Y be two discrete random variables with joint pmf:

		Y		
		0	1	p_X
	-1	0	$\frac{1}{4}$	$\frac{1}{4}$
X	0	$\frac{1}{2}$	0	$\frac{1}{2}$
	1	0	$\frac{1}{4}$	$\frac{1}{4}$
	p_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

Are they uncorrelated? Are they dependent?

- Given: $p_{XY}(-1, 0) = 0$ and $p_X(-1) \cdot p_Y(0) = \frac{1}{8}$.
- Since $p_{XY}(-1, 0) \neq p_X(-1) \cdot p_Y(0)$, X and Y are **not independent**.
- Calculating expectations:

$$E[XY] = 0 \cdot p_{XY}(0, 0) + (-1) \cdot p_{XY}(-1, 1) + 1 \cdot p_{XY}(1, 1) = 0 \cdot \frac{1}{2} + \frac{1}{4} - \frac{1}{4} = 0$$

$$E[X] = 0, \quad E[Y] = \frac{1}{2}$$

  Since $E[XY] = E[X]E[Y]$, X and Y are **uncorrelated**.

- **Uncorrelation does not imply independence**: we showed that X and Y are uncorrelated but not independent.
- **Independence implies uncorrelation**: if X and Y are independent, they are always uncorrelated.

Some laws of discrete random variables:

- **Bernoulli**: $X \sim \text{Be}(p)$, $X \in \{0, 1\}$, $p \in [0, 1]$
- **Binomial**: $X \sim \text{Binom}(n, p)$, $X \in \{0, 1, \ldots, n\}$, $p \in (0, 1)$, $n \in \{1, 2, \ldots\}$
- **Poisson**: $X \sim \text{Po}(\lambda)$, $X \in \{0, 1, 2, 3, \ldots\}$, $\lambda > 0$

a, b, n, p, and λ are the parameters that characterize the probability distribution.

4.9 Bernoulli distribution

A random variable X has a **Bernoulli distribution**, denoted by:

$$X \sim \text{Bernoulli}(p) \quad \text{or} \quad X \sim \text{Be}(p)$$

where:

- p is the probability of success ($0 \leq p \leq 1$).
- $X = 1$ (success) with probability p.
- $X = 0$ (failure) with probability $1 - p$.

Probability mass function (PMF)

$$p_X(x) = P(X = x) = p^x(1-p)^{1-x}, \quad \text{for } x = 0, 1$$

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Example: **tossing a biased coin**
Let $p = 0.7$ be the probability of success (getting heads). Calculate:

$$P(X = 1) = 0.7^1 \times 0.3^0 = 0.7$$

$$P(X = 0) = 0.7^0 \times 0.3^1 = 0.3$$

4.9.1 Mean and variance of a Bernoulli distribution

If $X \sim \text{Bernoulli}(p)$, then the **pmf** of X is:

$$p_X(x) = P(X = x) = p^x(1-p)^{1-x}, \quad \text{for } x = 0, 1$$

which can be written as:

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean of a Bernoulli random variable

$$E(X) = \mu_X = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot p + 0 \cdot (1-p) = p$$

Variance of a Bernoulli random variable

$$\text{var}(X) = E(X^2) - (E(X))^2$$

Since $X^2 = X$, we have:

$$E(X^2) = E(X) = p$$

Therefore:

$$\text{var}(X) = p - p^2 = p(1-p)$$

4.10 A short digression on combinatorics

4.10.1 Permutations of n distinct objects

The number of ways to arrange n distinct objects is:

$$n! = n \cdot (n-1) \cdot \cdots \cdot 1$$

By convention, $0! = 1$.

Example: the set $\{1, 2, 3\}$ has 6 (3!) permutations:

$$(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)$$

4.10.2 k-Permutations of n elements

The number of ways to select and order k objects from n elements is:

$$P(n,k) = \frac{n!}{(n-k)!} = n \cdot (n-1) \cdots (n-k+1)$$

4.10.3 k-Combinations of n elements

The number of ways to select k elements from n without regard to order is:

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

Key difference:

- **Permutations**: order matters.
- **Combinations**: order does NOT matter.

4.10.4 Binary sequences of length n with k ones

The number of such sequences is:

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

There is a one-to-one correspondence between these sequences and subsets of k elements from $\{1, 2, \ldots, n\}$.

4.11 Binomial distribution

4.11.1 Binomial theorem

The **Binomial theorem** states that for any $a, b \in R$ and $n \in \{0, 1, 2, \ldots\}$:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}$$

where:

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

is the **binomial coefficient**.

4.11.2 Binomial distribution

Consider a biased coin with the probability of a tail (success) in a single toss being p, where $0 < p < 1$. If X is the number of tails in n flips, then:

$$X \sim \text{Binom}(n,p)$$

Probability mass function (PMF) If $X \sim \text{Binom}(n,p)$, the pmf of X is:

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \quad 0 \le p \le 1$$

where:

- $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways to arrange x successes in n trials.
- p^x: Probability of x successes.
- $(1-p)^{n-x}$: probability of $n-x$ failures.

Interpretation

- $\binom{n}{x}$ gives the number of ways to arrange x successes and $n-x$ failures.
- $p^x(1-p)^{n-x}$ gives the probability of a specific sequence with x successes and $n-x$ failures.

4.11.3 Additive property of the binomial distribution

Let $Y_1 \sim \text{Binom}(n_1, p)$ and $Y_2 \sim \text{Binom}(n_2, p)$, where Y_1 and Y_2 are independent. Then:

$$Y_1 + Y_2 \sim \text{Binom}(n_1 + n_2, p)$$

Proof:
We can represent Y_1 as:

$$Y_1 = \sum_{j=1}^{n_1} X_j \quad \text{where } X_j \sim \text{Bernoulli}(p)$$

Similarly, Y_2 can be written as:

$$Y_2 = \sum_{j=n_1+1}^{n_1+n_2} X_j \quad \text{where } X_j \sim \text{Bernoulli}(p)$$

Here, $X_1, X_2, \ldots, X_{n_1+n_2}$ are i.i.d. $\sim$ Bernoulli(p).

Combining the trials:

$$Y_1 + Y_2 = \sum_{j=1}^{n_1} X_j + \sum_{j=n_1+1}^{n_1+n_2} X_j = \sum_{j=1}^{n_1+n_2} X_j$$

Since $X_1, X_2, \ldots, X_{n_1+n_2}$ are i.i.d. $\sim$ Bernoulli(p), we have:

$$Y_1 + Y_2 \sim \text{Binom}(n_1 + n_2, p)$$

4.11.4 Binomial distribution in R

1. **Generating random samples**

   ```
   rbinom(n = m, size = n, prob = p)
   ```

   This function yields m realizations $x_1, \ldots, x_m$ of X, where $X \sim \text{Binom}(n, p)$.

2. **Calculating the PMF**

   ```
   dbinom(x = x, size = n, prob = p)
   ```

   This yields $p_X(x) = P(X = x)$ for a given x.

3. **Calculating the CDF (Cumulative Distribution Function)**

   ```
   pbinom(x = x, size = n, prob = p)
   ```

   This yields $F_X(x) = P(X \leq x)$.

4. **Calculating the quantile function**

   ```
   qbinom(p = p, size = n, prob = p)
   ```

   This yields $F^{-1}(p)$, the smallest integer x such that $F_X(x) \geq p$.

In the following R code, which value(s) should be assigned to j to obtain zero?

```
dbinom(0:6, 15, 0.4) |> sum() - pbinom(j, 15, 0.4)
```

1. **Summing PMF values and comparing to CDF**

   ```
   dbinom(0:6, 15, 0.4) |> sum() - pbinom(j, 15, 0.4)
   ```

   To make the difference zero, set $j = 6$ or any value in $[6, 7)$.

2. **Calculating the expected value using PMF**

   ```
   sum(0:15 * dbinom(0:15, 15, 1/3))
   ```

   This yields $\frac{15}{3} = 5$.

3. **Calculating a quantile and comparing to a CDF**

   ```
   qbinom(0.3, 50, 0.4) |> pbinom(50, 0.4)
   ```

   The result is greater than or equal to 0.3.

4.12 Relationship between Bernoulli and binomial distribution

Suppose $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with the same parameter p:

$$X_j \sim \text{Bernoulli}(p) \quad \text{for each } j$$

Define:

$$X = \sum_{j=1}^{n} X_j$$

Then X follows a **Binomial distribution**:

$$X \sim \text{Binom}(n, p)$$

Mean of a binomial random variable

- By the **linearity of expectation**:

$$E(X) = E\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} E(X_j)$$

- Since $E(X_j) = p$:

$$E(X) = \sum_{j=1}^{n} p = n \cdot p$$

Variance of a binomial random variable

- Since X_j's are **independent**:

$$\text{var}(X) = \text{var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \text{var}(X_j)$$

- With $\text{var}(X_j) = p(1-p)$:

$$\text{var}(X) = \sum_{j=1}^{n} p(1-p) = n \cdot p(1-p)$$

Final results

- **Mean**: $E(X) = n \cdot p$
- **Variance**: $\text{var}(X) = n \cdot p(1-p)$

4.13 Poisson distribution

The **Poisson distribution** is useful for modeling the number of rare events in a fixed, large population over a short unit of time, under the condition that the events occur independently.

Parameter: λ
λ (lambda) is the average number of occurrences in a fixed interval of time or space. It represents the **rate** of the distribution.

Probability mass function (PMF)

$$P(\text{observe } k \text{ rare events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where:

- $k = 0, 1, 2, \ldots$
- $k!$ is the factorial of k, given by $k \times (k-1) \times \cdots \times 1$.
- $e \approx 2.718$ is the base of the natural logarithm.

Mean and standard deviation

- Mean: λ
- Standard deviation: $\sqrt{\lambda}$

4.13.1 Additive property of Poisson distribution

Let $X_1, X_2, \dots, X_n$ be independent random variables such that:

$$X_j \sim \text{Poisson}(\lambda_j) \quad \text{for } j = 1, \dots, n.$$

Then, the sum of these random variables:

$$S = \sum_{j=1}^{n} X_j$$

is also a Poisson random variable with parameter:

$$S \sim \text{Poisson}\left(\sum_{j=1}^{n} \lambda_j\right).$$

Example:
Consider receiving emails at different rates:

- First hour: $\lambda_1 = 3$ emails/hour
- Second hour: $\lambda_2 = 5$ emails/hour
- Third hour: $\lambda_3 = 2$ emails/hour

The total number of emails in three hours:

$$\lambda_{\text{total}} = \lambda_1 + \lambda_2 + \lambda_3 = 10$$

follows:

$$S \sim \text{Poisson}(10).$$

The Poisson distribution is often used for modeling the number of rare events in a fixed, large population over a short unit of time. It is useful when the events are independent of each other. The **rate parameter** λ represents the average number of occurrences in a given time period. Using λ, we can describe the probability of observing exactly k rare events in a unit of time.

4.13.2 Probability mass function

$$P(\text{observe } k \text{ rare events}) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where:

- $k \in \{0, 1, 2, \dots\}$
- $k!$ represents k-factorial: $k! = k \times (k-1) \times \dots \times 1$ (and $0! = 1$)
- $e \approx 2.718$, the base of the natural logarithm

4.13.3 Mean and standard deviation

$$\text{Mean} = \lambda, \quad \text{Standard Deviation} = \sqrt{\lambda}.$$

Example
A bakery receives an average of $\lambda = 3$ orders per hour. The number of orders in any hour follows a Poisson distribution with $\lambda = 3$.

- Probability of receiving exactly 2 orders:

$$P(k=2) = \frac{3^2 e^{-3}}{2!} = \frac{9 \cdot e^{-3}}{2} \approx 0.224.$$

- Probability of receiving no orders:

$$P(k=0) = \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.050.$$

4.14 Relationship between Poisson and binomial distribution

Let $(X_n)_{n=1}^{\infty}$ be a sequence of binomial random variables where:

$$X_n \sim \text{Bin}(n,p) \quad \text{with} \quad p = \frac{\lambda}{n}.$$

We have:

$$E(X_n) = n \cdot p = \lambda.$$

To find the limit:

$$\lim_{n\to\infty} P(X_n = j) \quad \text{for } j \in \{0,1,2,\ldots\},$$

we use the binomial probability mass function:

$$P(X_n = j) = \binom{n}{j} p^j (1-p)^{n-j},$$

where:

$$\binom{n}{j} = \frac{n!}{j!(n-j)!} \quad \text{and} \quad p = \frac{\lambda}{n}.$$

As $n \to \infty$:

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}.$$

Thus:

$$\lim_{n\to\infty} P(X_n = j) = \frac{\lambda^j e^{-\lambda}}{j!}.$$

This shows that we can approximate a binomial distribution with parameters n and p by a Poisson distribution with parameter $\lambda = n \cdot p$, provided that n is large and p is small.

4.14.1 Rule of Thumb:

$$\frac{n}{p} > 500 \quad \text{for a good approximation.}$$

4.15 Example: Poisson distribution for power failures

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

4.15.1 Problem setup

We are given that electricity power failures in a rural region follow a Poisson distribution with an average rate of:

$$\lambda = 2 \text{ failures per week.}$$

We want to calculate the probability that the electricity fails only once in a given week, i.e., $k = 1$.

4.15.2 Poisson distribution formula

The probability of observing k events in a fixed time period is given by:

$$P(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

4.15.3 Plugging in the values

$$P(k = 1; \lambda = 2) = \frac{2^1 \cdot e^{-2}}{1!}.$$

Calculating each term:

- $2^1 = 2$
- $e^{-2} \approx 0.1353$
- $1! = 1$

4.15.4 Final calculation

$$P(k = 1; \lambda = 2) = \frac{2 \times 0.1353}{1} = 0.2706 \approx 0.27.$$

Now, suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given ***day*** the electricity fails three times.

To calculate the probability of three failures on a given day, we first determine the daily failure rate:

$$\lambda_{\text{day}} = \frac{\lambda_{\text{week}}}{7} = \frac{2}{7} \approx 0.2857.$$

We assume that the probability of a power failure is the same on any day of the week.

Poisson distribution formula:

$$P(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Plugging in the values:

$$P(k = 3; \lambda = 0.2857) = \frac{0.2857^3 \cdot e^{-0.2857}}{3!}.$$

Calculating each term:

- $0.2857^3 \approx 0.0233$
- $e^{-0.2857} \approx 0.7513$
- $3! = 6$

Final calculation:

$$P(k = 3; \lambda = 0.2857) = \frac{0.0233 \times 0.7513}{6} \approx 0.0029.$$

4.16 Poisson distribution conditions

A random variable may follow a Poisson distribution if:

- The event being considered is **rare**. This means the probability of the event occurring is small relative to the size of the population or the time interval.
- The population or interval is **large**, or the time or space interval should be substantial compared to the rarity of the event.
- The events occur **independently** of each other. The occurrence of one event should not affect the likelihood of another event occurring.

However, there are situations where the independence assumption may not hold. For example, consider the probability of a certain number of weddings over one summer. Since weddings are more likely to occur on weekends, the events are not entirely independent.
In such cases, a Poisson model may still be reasonable if we adjust the rate parameter λ to be different for different times. For instance, we can model:

- A higher λ for weekends.
- A lower λ for weekdays.

This idea of modeling the rate λ for a Poisson distribution against a second variable (like the day of the week) forms the basis of more advanced methods called **generalized linear models (GLMs)**. These methods are beyond the scope of this course.

4.17 Poisson in R

If $X \sim \text{Poisson}(\lambda)$, then the following R functions are used:

- `rpois(n=n, lambda=lambda)`: Generates n random realizations $x_1, \ldots, x_n$ from X.
- `dpois(x=x, lambda=lambda)`: Yields $p_X(x) = P(X = x)$, the probability mass function.
- `ppois(x=x, lambda=lambda)`: Yields $F_X(x) = P(X \leq x)$, the cumulative distribution function.

- `qpois(p=p, lambda=lambda)`: Yields $F^{-1}(p) = \inf\{x \in R : F_X(x) \geq p\}$, the quantile function.

Problem
We are given the R code:

```
dbinom(0:10, 500, 3/500) - dpois(0:10, lambda)
```

The `dbinom` function returns the binomial probabilities for $x = 0, 1, \ldots, 10$ with:

- Number of trials: $n = 500$
- Probability of success: $p = \frac{3}{500}$

We want to approximate the binomial distribution $\text{Bin}(500, \frac{3}{500})$ with a Poisson distribution. Using the approximation $\lambda = n \cdot p$:

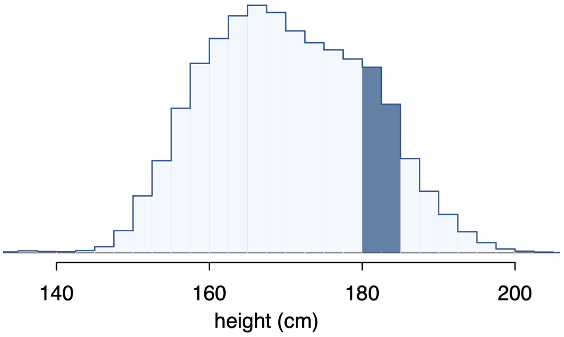
$$\lambda = 500 \times \frac{3}{500} = 3.$$

Thus, to get approximately a sequence of zeroes when subtracting the Poisson probabilities from the binomial probabilities, assign:

```
lambda <- 3
```

5 Absolutely continuous random variables

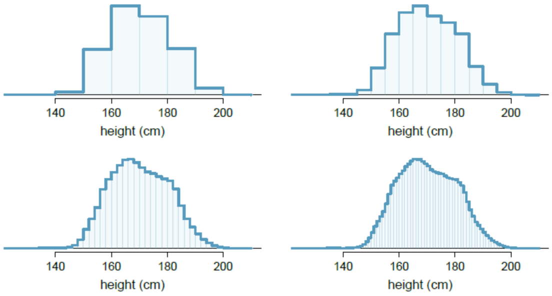
Imagine we are interested in understanding the distribution of heights among adult males in the United States. When we think about how likely it is for a randomly chosen adult male to have a height within a specific range, we use the concept of **continuous probability distributions**. The histogram shown below represents the distribution of heights. Each bar (or "bin") in the histogram shows how many people fall into that height range. The area of each bar corresponds to the proportion of people whose heights are in that interval. If we add up the areas of all the bars, we get the total probability, which is 1 (since everyone has some height).



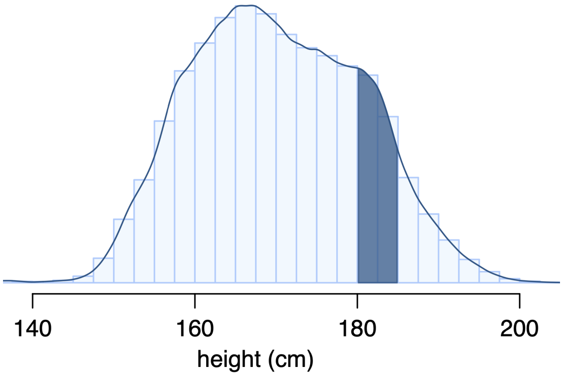
Now, suppose we want to calculate the probability that a randomly chosen male adult is between 180 cm and 185 cm. The shaded area under the histogram between these two heights represents this probability.

5.1 From histograms to absolutely continuous distributions

Adding more bins to a histogram provides greater detail. Since there are many data points, much smaller bins still work well to give us an accurate representation of the distribution. In the bottom right graph, we observe that the contour of the histogram starts to approximate a smooth curve. This suggests that the probability distribution of heights can be described by a smooth curve.



This smooth curve, as we shall see, is called a **density function**. The density function provides a continuous representation of the probability distribution and can be approximated by the histogram obtained from a large sample of data. As the number of data points increases and the bins become smaller, the histogram becomes a better approximation of this continuous density function.
Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can be modeled as the shaded area under the curve.



Since continuous probabilities are estimated as "the area under the curve", the probability of a person being exactly 180 cm (or any exact value) is defined as 0.

5.2 Absolutely continuous random variables: definition

A random variable X is said to be **absolutely continuous** if there exists a function $f : R \to R$ such that the cumulative distribution function (CDF) of X can be expressed as:

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

where $f(t)$ is called the **probability density function** (PDF) of X.

Consequences:

- The CDF $F(x)$ is **continuous**, meaning it does not have any jumps or breaks.
- If the density function f is continuous at a point x, then the CDF F is **differentiable** at x, and the derivative of F at x is given by:

$$\frac{d}{dx}F(x) = F'(x) = f(x)$$

5.3 Density function

Given an absolutely continuous random variable X with density f, the probability that X belongs to an interval $[a, b]$ is given by the area under the curve of $f(x)$ in that interval:

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

Since $P(X = x) = 0$ for every $x \in R$, we have:

$$P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = P(a \leq X \leq b)$$

5.3.1 Properties of the density function:

- $f(x) \geq 0$ for every $x \in R$. This ensures that probabilities are non-negative.
- The total area under the curve of $f(x)$ over the entire real line is 1:

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

Note: $f(x)$ is not necessarily less than or equal to 1; it can be greater than 1, but the total area under $f(x)$ must always equal 1.

5.3.2 Non-uniqueness of the density function

The probability that X falls within the interval $(a, b]$ for an absolutely continuous random variable is given by:

$$P(a < X \leq b) = \int_a^b f(x)\, dx \quad \text{for } b \geq a$$

We can also express this in terms of the cumulative distribution function (CDF):

$$\int_a^b f(x)\, dx = F(b) - F(a)$$

where $F(x)$ is the CDF of X.

Non-negativity of f: It can be shown that $f(x) \geq 0$ for every $x \in R$ because the integral of f over any interval represents a probability, which must be non-negative.

Total probability:

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

This equality indicates that the probability of X taking any value from $-\infty$ to $+\infty$ is 1:

$$1 = P(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(x)\,dx$$

Non-uniqueness of f: Strictly speaking, the density function f of X is not unique. If a function g differs from f at only a finite number of points, g is also a valid density function of X.

Equality in distribution: If two random variables X and Y have the same density function, we denote this as:

$$X \sim Y$$

This means that X and Y have the same distribution.

5.4 ACD - Cumulative distribution function

For an absolutely continuous random variable X, the cumulative distribution function (CDF), denoted as $F(x)$, is defined as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

where $f(t)$ is the probability density function (PDF) of X.

5.4.1 Properties of the cdf:

- $F(x)$ is **non-decreasing**, meaning $F(x_1) \leq F(x_2)$ for $x_1 < x_2$.
- $F(x)$ is **continuous**, reflecting the smooth nature of absolutely continuous distributions.
- Limits:
$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F(x) = 1$$
- Relationship with the PDF:
$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

5.5 ACD - Expectation

An absolutely continuous random variable X with density $f(x)$ admits a finite expectation if and only if:

$$\int_{-\infty}^{+\infty} |x| f(x)\,dx < \infty$$

In this case, the expectation of X, denoted $E(X)$ or μ_X, is given by:

$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x f(x)\,dx$$

5.5.1 Expectation of a function of X

If $Y = g(X)$, where $g : R \to R$ is a Borel-measurable function, then Y admits a finite expectation if and only if:

$$\int_{-\infty}^{+\infty} |g(x)| f(x)\, dx < \infty$$

In such a case, the expectation of Y is given by:

$$E(Y) = \int_{-\infty}^{+\infty} g(x) f(x)\, dx$$

5.6 ACD - Variance

The variance of an absolutely continuous random variable X exists and is finite if:

$$\int_{-\infty}^{+\infty} x^2 f(x)\, dx < \infty$$

In such a case, the variance $\mathrm{Var}(X)$ is given by:

$$\mathrm{Var}(X) = E\big((X - \mu_X)^2\big) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x)\, dx$$

Alternatively, it can be expressed as:

$$\mathrm{Var}(X) = E(X^2) - \mu_X^2$$

where:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx, \quad \mu_X = E(X)$$

5.6.1 Standard deviation:

The standard deviation $\mathrm{SD}(X)$ is the square root of the variance:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

5.7 ACD - Independence

Let X and Y be two absolutely continuous random variables with marginal density functions $f_X(x)$ and $f_Y(y)$, respectively. The joint density function $f_{XY}(x, y)$ satisfies:

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y) \quad \text{for all } x, y \in R.$$

This condition is equivalent to saying that X and Y are **independent**, denoted as $X \perp\!\!\!\perp Y$.

Key Interpretation: Two absolutely continuous random variables X and Y are independent if and only if their joint density is the product of their marginal densities. Otherwise, they are dependent.

5.8 Standardized random variables

If X is a random variable with finite mean $E(X)$ and finite and positive variance $\text{Var}(X) > 0$, then the standardized version of X, denoted by Z, is defined as:

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

5.8.1 Properties of the standardized random variable Z

- Mean:

$$E(Z) = E\left(\frac{X - E(X)}{\sqrt{\text{Var}(X)}}\right) = 0$$

- Variance:

$$\text{Var}(Z) = \text{Var}\left(\frac{X - E(X)}{\sqrt{\text{Var}(X)}}\right) = 1$$

A random variable Z such that $E(Z) = 0$ and $\text{Var}(Z) = 1$ is said to be **standardized**.

5.9 Joint distribution in the absolutely continuous case

Given two absolutely continuous random variables X and Y defined on the same probability space $(\Omega, \mathcal{F}, P)$, they have a joint density function $f_{XY}(x, y)$ if:

- $f_{XY}(x, y) \geq 0 \quad \text{for all } x, y \in R.$
- $f_{XY}(x, y)$ is integrable, with:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(u, v)\, du\, dv = 1$$

- The cumulative probability is given by:

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v)\, du\, dv$$

5.9.1 Joint distribution for n random variables:

The joint density function $f_{X_1 \dots X_n}(x_1, \dots, x_n)$ must satisfy the following conditions:

1. **Non-negativity:** $f_{X_1 \dots X_n}(x_1, \dots, x_n) \geq 0 \quad \text{for all } x_1, \dots, x_n \in R.$
2. **Normalization:** The total integral of the joint density over R^n equals 1:

$$\int_{R^n} f_{X_1 \dots X_n}(u_1, \dots, u_n)\, du_1 \cdots du_n = 1.$$

5.10 Uniform continuous distribution

5.10.1 UCD - Definition

The continuous uniform distribution evenly distributes mass over an interval $[a, b]$, where $a, b \in R$ and $a < b$. An absolutely continuous random variable X is said to have a uniform distribution if its probability density function (PDF) is:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{if } x < a \text{ or } x > b \end{cases}$$

where $-\infty < a < b < \infty$.

If X follows a uniform distribution on $[a, b]$, we write:

$$X \sim \text{Unif}(a, b).$$

Example: Suppose X represents the waiting time at a bus stop for the next vehicle, and t is the time between one vehicle and the next. Then $X \sim \text{Unif}(0, t)$, meaning X is equally likely to be any value between 0 and t.

5.10.2 UCD - Cumulative distribution function

The CDF of a uniform continuous random variable X on the interval $[a, b]$ is:

$$F(x) = \begin{cases} 0, & \text{if } x < a \\ \frac{x-a}{b-a}, & \text{if } a \leq x \leq b \\ 1, & \text{if } x > b \end{cases}$$

Explanation:

- $F(x) = 0$: No probability mass lies below a.
- $F(x) = \frac{x-a}{b-a}$: Probability accumulates evenly between a and b.
- $F(x) = 1$: All probability mass is accounted for when $x > b$.

Example: If $X \sim \text{Unif}(0, 10)$:

- For $x = 5$, $F(5) = \frac{5-0}{10-0} = 0.5$.
- For $x = 15$, $F(15) = 1$, since $x > 10$.

5.10.3 UCD - Mean and variance

Mean:

$$E(X) = \frac{a+b}{2}.$$

Variance:

$$\text{Var}(X) = E(X^2) - \big(E(X)\big)^2.$$

First, compute $E(X^2)$:

$$E(X^2) = \int_a^b x^2 f(x)\, dx = \frac{a^2 + ab + b^2}{3}.$$

Substituting $E(X^2)$ and $E(X)$, we get:

$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

Key Formulas:

- $E(X) = \frac{a+b}{2}$
- $\text{Var}(X) = \frac{(b-a)^2}{12}$

Example: If $X \sim \text{Unif}(0, 10)$:

- Mean: $E(X) = \frac{0+10}{2} = 5$
- Variance: $\text{Var}(X) = \frac{(10-0)^2}{12} = \frac{100}{12} \approx 8.33$

5.10.4 Uniform distribution in R

For $X \sim \text{Unif}(a, b)$, the following R functions are available:

- `runif(n=n, min=a, max=b)`: Generates n random realizations $x_1, \ldots, x_n$ of X.
- `dunif(x=x, min=a, max=b)`: Evaluates the density $f_X(x)$ of X at a given x.
- `punif(x=x, min=a, max=b)`: Evaluates the cumulative distribution function (CDF) $F_X(x) = P(X \leq x)$.
- `qunif(p=p, min=a, max=b)`: Computes the quantile $F_X^{-1}(p) = \inf\{x \in R : F_X(x) \geq p\}$.

Default values:
The default values for `min` and `max` are $a = 0$ and $b = 1$, respectively.

Examples:

- `runif(n=5, min=0, max=10)`: Generates 5 random numbers from $X \sim \text{Unif}(0, 10)$.
- `dunif(x=5, min=0, max=10)`: Returns $f_X(5) = \frac{1}{10-0} = 0.1$.
- `punif(x=5, min=0, max=10)`: Returns $F_X(5) = \frac{5-0}{10-0} = 0.5$.
- `qunif(p=0.5, min=0, max=10)`: Returns $x = 5$, since $F_X(5) = 0.5$.

5.11 Gaussian distribution

5.11.1 The Gaussian integral

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}.$$

Standard normal distribution:
The standard normal distribution has the probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

and its integral over the real line equals 1:

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

Normal distribution with mean μ and variance σ^2:
For $X \sim \mathcal{N}(\mu, \sigma^2)$, the PDF is:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Using the substitution $x = \frac{t-\mu}{\sqrt{2\sigma^2}}$ with $dx = \frac{dt}{\sqrt{2\sigma^2}}$, it can be shown that:

$$\int_{-\infty}^{+\infty} f(t)\, dt = 1.$$

Key result:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = 1.$$

This ensures that the normal distribution is a valid probability density function.

5.11.2 Normal (or Gaussian) distribution

A random variable X is said to have a Gaussian (Normal) distribution with parameters μ (mean) and σ^2 (variance) if its probability density function (PDF) is:

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

Properties

- **Non-Negativity:** $f(x; \mu, \sigma^2) \geq 0$ for all $x \in R$.
- **Normalization:** The total integral of the density over R equals 1:

$$\int_{-\infty}^{+\infty} f(x; \mu, \sigma^2) dx = 1.$$

- **Moments:**
  - Mean: $E(X) = \mu$.
  - Variance: $\text{Var}(X) = \sigma^2$.
- **Shape:**
  - The density is unimodal, symmetric, and bell-shaped.
  - The distribution is entirely characterized by μ and σ^2.

Probability of an interval
The probability that X lies in an interval $[a, b]$ is:

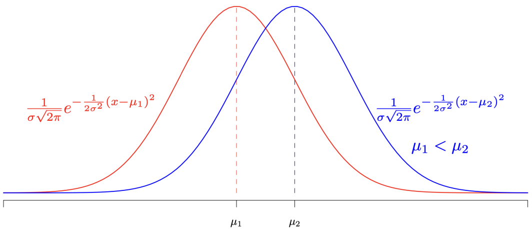
$$P(a \leq X \leq b) = \int_a^b f(x; \mu, \sigma^2) dx.$$

Since this integral cannot be calculated analytically, it is evaluated numerically (e.g., using R or other software).

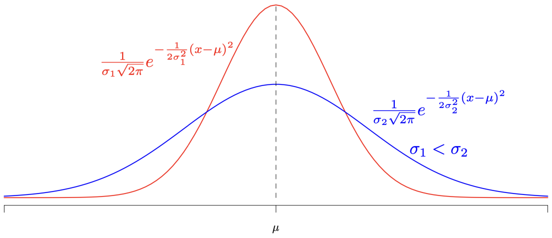
Notation: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then X has a Gaussian (Normal) distribution with mean μ and variance σ^2.

5.11.3 Parameters of the Gaussian (Normal) Distribution

- **Mean (μ):** The parameter μ characterizes the **location** of the Gaussian curve. It represents the center of the distribution and is the point of symmetry for the bell-shaped curve.



- **Variance (σ^2):** The parameter σ^2 quantifies the **dispersion** or **spread** of the curve:
  - A larger σ^2 results in a wider and flatter curve, indicating greater variability.
  - A smaller σ^2 results in a narrower and taller curve, indicating less variability.



5.11.4 Symmetry of the Gaussian distribution ($Z \sim \mathcal{N}(0,1)$)

The probability density function (PDF) of Z is:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, \quad -\infty < z < +\infty.$$

Even function
The PDF is an even function:

$$f_Z(z) = f_Z(-z), \quad \text{for all } z \in R.$$

This means the density is symmetric around zero.

Symmetry around zero
The symmetry implies:

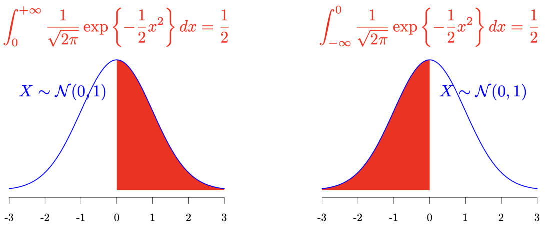
$$Z \sim -Z.$$

This means the distribution of Z is identical to the distribution of $-Z$.

Implications for probabilities:
The symmetry around zero ensures:

$$P(Z \leq 0) = P(Z \geq 0) = \frac{1}{2}.$$

Half of the probability mass lies below $Z = 0$, and the other half lies above $Z = 0$.



5.11.5 From the standardized Normal to any Normal

If $Z \sim \mathcal{N}(0,1)$, then $X = \sigma Z + \mu$ follows $\mathcal{N}(\mu, \sigma^2)$.
From any Normal to the standard Normal:
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the standardized variable:

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

follows $\mathcal{N}(0,1)$.

Proof:
The CDF of X is:

$$P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, dt.$$

Using the substitution $z = \frac{t-\mu}{\sigma}$ ($t = \sigma z + \mu$, $dt = \sigma dz$), this becomes:

$$P(X \leq x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Moments of X: Using $X = \sigma Z + \mu$ and $E(Z) = 0$, $\text{Var}(Z) = 1$:

- **Expected Value:**
$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu.$$
- **Variance:**
$$\text{Var}(X) = \text{Var}(\sigma Z + \mu) = \sigma^2 \text{Var}(Z) = \sigma^2.$$

Key results:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $E(X) = \mu$, $\text{Var}(X) = \sigma^2$, and:
$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$
- If $Z \sim \mathcal{N}(0,1)$, then $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

5.11.6 Properties of the Gaussian distribution

1. Scaling and shifting: If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in R$, then:

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

2. Linear combination of independent Gaussian random variables:
If $X_1, \ldots, X_n$ are independent Gaussian random variables, where:

$$X_j \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad j = 1, \ldots, n,$$

and $a_1, \ldots, a_n \in R$, then the linear combination:

$$Y = \sum_{j=1}^{n} a_j X_j$$

follows a Gaussian distribution:

$$Y \sim \mathcal{N}\left(\sum_{j=1}^{n} a_j \mu_j, \sum_{j=1}^{n} a_j^2 \sigma_j^2\right).$$

3. Dependence:
If $X_1, \ldots, X_n$ are dependent, their linear combination is **not necessarily Gaussian**.

5.11.7 Gaussian distribution in R

For $X \sim \mathcal{N}(\mu, \sigma^2)$, the following R functions are available:

- `rnorm(n=n, mean=mu, sd=sigma)`: Generates n realizations $x_1, \ldots, x_n$ from $X \sim \mathcal{N}(\mu, \sigma^2)$.
- `dnorm(x=x, mean=mu, sd=sigma)`: Evaluates the PDF $f_X(x)$ of X at a given x.
- `pnorm(x=x, mean=mu, sd=sigma)`: Evaluates the CDF $F_X(x) = P(X \leq x)$ of X.
- `qnorm(p=p, mean=mu, sd=sigma)`: Computes the quantile $F_X^{-1}(p) = \inf\{x \in R : F_X(x) \geq p\}$.

Default values:
The default values are `mean=0` and `sd=1`, corresponding to the standard normal distribution ($\mathcal{N}(0, 1)$).

Examples:

- `rnorm(n=5, mean=0, sd=1)`: Generates 5 random values from $\mathcal{N}(0, 1)$.
- `dnorm(x=0, mean=0, sd=1)`: Returns $\frac{1}{\sqrt{2\pi}} \approx 0.3989$, the density at $x = 0$.
- `pnorm(x=1.96, mean=0, sd=1)`: Returns approximately 0.975, the cumulative probability up to $x = 1.96$.
- `qnorm(p=0.975, mean=0, sd=1)`: Returns approximately 1.96, the quantile corresponding to $p = 0.975$.

5.12 Chi-squared distribution

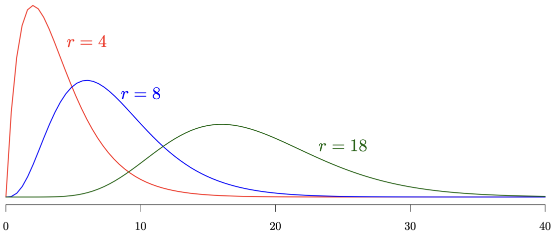
A random variable $X \sim \chi_r^2$ with r degrees of freedom has the PDF:

$$f(x; r) = a_r x^{r/2-1} e^{-x/2} I_{(0,\infty)}(x),$$

where:

$$a_r = \frac{1}{2^{r/2}\Gamma(r/2)}.$$

Graph of the density function of the chi–squared distribution with r degrees of freedom:



It is a right-skewed distribution, but the asymmetry decreases as the degrees of freedom increases.

5.12.1 Properties

- Mean: $E(X) = r$.
- Variance: $\text{Var}(X) = 2r$.

Additive property
If $X_1, \dots, X_n$ are independent with $X_j \sim \chi^2_{r_j}$, then:

$$\sum_{j=1}^{n} X_j \sim \chi^2_r, \quad \text{where } r = \sum_{j=1}^{n} r_j.$$

5.12.2 Relationship with the Gaussian distribution

- If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_1$.
- If $Z_1, \dots, Z_n$ are i.i.d. $\mathcal{N}(0,1)$, then:

$$\sum_{j=1}^{n} Z_j^2 \sim \chi^2_n.$$

5.12.3 Chi-squared distribution in R

- `rchisq(n=n, df=r)`: Generates n random values from χ^2_r.
- `dchisq(x=x, df=r)`: Evaluates the PDF of X at x.
- `pchisq(x=x, df=r)`: Computes the CDF $F_X(x) = P(X \leq x)$.
- `qchisq(p=p, df=r)`: Computes the quantile $F_X^{-1}(p)$.

5.13 Student's t-distribution

If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_r$, and Z and V are independent, then:

$$T_r = \frac{Z}{\sqrt{V/r}} \sim t_r,$$

where T_r follows a Student's t-distribution with r degrees of freedom.

5.13.1 Moments

- Mean:

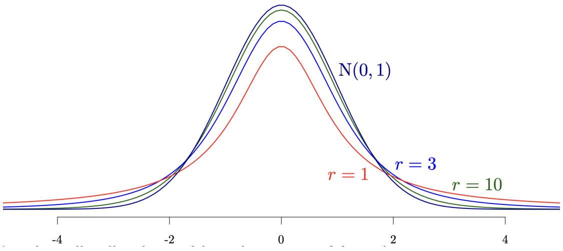
$$E(T_r) = \begin{cases} 0, & \text{if } r > 1, \\ \text{undefined}, & \text{if } r = 1. \end{cases}$$

- Variance:

$$\text{Var}(T_r) = \begin{cases} \frac{r}{r-2}, & \text{if } r > 2, \\ \text{undefined}, & \text{if } r \in \{1, 2\}. \end{cases}$$

5.13.2 Properties

- The t_r-distribution is symmetric around 0.
- The density is bell-shaped but has heavier tails than the Gaussian distribution.
- As $r \to \infty$, $t_r \to \mathcal{N}(0,1)$ (the standard normal distribution).



5.13.3 Student's t-distribution in R

- `rt(n=n, df=r)`: Generates n realizations from t_r.
- `dt(x=x, df=r)`: Evaluates the PDF at x.
- `pt(x=x, df=r)`: Computes the CDF $F_X(x) = P(X \le x)$.
- `qt(p=p, df=r)`: Computes the quantile $F_X^{-1}(p)$.

5.13.4 The Quantile function

The quantile function $F^{-1}(p)$ of a CDF $F(x)$ is defined as:

$$F^{-1}(p) = \inf\{x \in R : F(x) \ge p\}, \quad 0 < p < 1.$$

- If F is invertible (e.g., Gaussian CDF), F^{-1} is its usual inverse.
- For discrete random variables, $F^{-1}(p)$ is the smallest x such that $P(X \le x) \ge p$.

5.14 Probability distributions in R

For each distribution, the following functions are available:

- `d-`: Density or probability mass function (PMF).
- `p-`: Cumulative distribution function (CDF).
- `q-`: Quantile function.
- `r-`: Random number generator.

Distribution	**Density/PMF (d-)**	**CDF (p-)**	**Quantile (q-)**	**Random (r-)**
Binomial	`dbinom()`	`pbinom()`	`qbinom()`	`rbinom()`
Poisson	`dpois()`	`ppois()`	`qpois()`	`rpois()`
Normal	`dnorm()`	`pnorm()`	`qnorm()`	`rnorm()`
Chi-squared	`dchisq()`	`pchisq()`	`qchisq()`	`rchisq()`
Student's t	`dt()`	`pt()`	`qt()`	`rt()`

Examples:

- To compute $P(X \leq 7)$ if $X \sim \chi^2_4$:

```
pchisq(7, df=4).
```

- To compute $P(X \leq 7)$ if $X \sim t_4$ (Student's t with 4 degrees of freedom):

```
pt(7, df=4).
```

- To find t such that $P(X \leq t) = 0.95$ for $X \sim t_4$:

```
qt(0.95, df=4).
```

5.15 Asymptotics

5.15.1 Convergence in distribution

A sequence of random variables $(X_n)_{n=1}^{\infty}$ with CDFs F_{X_n} converges in distribution to a random variable X with CDF F_X if:

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

for every continuity point $x \in R$ of F_X. In such a case, we write:

$$X_n \xrightarrow{d} X.$$

Characteristics:

- Convergence in distribution only considers the marginal laws of X_n and X; they do not need to be defined on the same probability space.
- The convergence is evaluated only at continuity points of F_X.

Example: Let $X = 0$ (a.s., $P(X = 0) = 1$) and $X_n = \frac{1}{n}$ (a.s., $P(X_n = 1/n) = 1$). Then:

$$X_n \xrightarrow{d} X.$$

Special cases:

- **Discrete random variables:** If X_n and X are discrete random variables taking values in $\{0, 1, 2, \dots\}$, then:

$$X_n \xrightarrow{d} X \quad \text{iff} \quad \lim_{n\to\infty} P(X_n = j) = P(X = j) \quad \text{for all } j \in \{0, 1, 2, \dots\}.$$

  Example: If $X_n \sim \text{Bin}(n, \lambda/n)$ and $X \sim \text{Poisson}(\lambda)$, then:

$$X_n \xrightarrow{d} X.$$

- **Poisson to Normal approximation:** If $X_n \sim \text{Poisson}(\lambda_n)$ with $\lambda_n \to \infty$, then:

$$\frac{X_n - \lambda_n}{\sqrt{\lambda_n}} \xrightarrow{d} Z, \quad Z \sim \mathcal{N}(0,1).$$

This implies that for large λ, the Poisson distribution can be approximated by a normal distribution with mean and variance λ.

5.15.2 Central limit theory

Let $X_1, X_2, \ldots$ be iid random variables with:

- $E(X_1) = \mu$,
- $\text{Var}(X_1) = \sigma^2$,
- $E(X_1^2) < \infty$.

The standardized sum:

$$Z_n = \frac{\sum_{j=1}^n X_j - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a standard normal random variable $Z \sim \mathcal{N}(0,1)$ as $n \to \infty$:

$$Z_n \xrightarrow{d} Z.$$

Interpretation: For large n, the sum $\sum_{j=1}^n X_j$ can be approximated by:

$$\sum_{j=1}^n X_j \sim \mathcal{N}(n\mu, n\sigma^2).$$

Examples:

- **Chi-squared distribution:** Let $X_1, \ldots, X_n \sim \chi_1^2$. Then:

$$\frac{\sum_{i=1}^n X_i - n}{\sqrt{2n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

For large n, $\chi_n^2 \sim \mathcal{N}(n, 2n)$.

- **Binomial distribution:** Let $Y_n \sim \text{Bin}(n,p)$. Then:

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

For large n, $\text{Bin}(n,p) \sim \mathcal{N}(np, np(1-p))$.

Continuity correction
When approximating a discrete distribution with a normal distribution, apply the continuity correction. For $Y_n \sim \text{Bin}(n,p)$:

$$P(k \le Y_n \le j) \approx \Phi\left(\frac{j + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

6 Inferential statistics - point estimators

6.1 Descriptive vs Inferential statistics

Statistical survey
A statistical survey refers to the process of collecting, analyzing, and interpreting data about a specific population, called the **target population**.

Examples:

- **Example 1:** In a study to measure a country's unemployment rate, the target population consists of all inhabitants aged 16–67 years.
- **Example 2:** In a study of male height in Italy, the target population consists of all adult Italian males.

Stages of a statistical survey:

1. Define the target population.
2. Collect data from the population.
3. Analyze the data.

Census vs Sampling

Census survey:
A census involves collecting data from every member of the target population.
Example: A nationwide survey collecting data from all inhabitants.
Sample survey:
A sample survey involves collecting data from a subset of the population, called the **sample**.
Reasons for using a sample survey:

- High cost of conducting a census.
- Long time required to complete a census.
- The population size may be too large to survey entirely.

Descriptive vs Inferential statistics

Descriptive statistics:

- Used in **census surveys**.
- Focuses on the graphical representation and synthesis of data.
- Provides a complete picture of the entire population since all members are surveyed.

Inferential statistics:

- Used in **sample surveys**.
- Aims to learn information about unknown quantities in the population based on the observed data.
- Estimates population parameters and measures uncertainty in those estimates.

6.2 Statistical inference

Statistical inference is based on the **inductive method**, which involves trying to establish general principles or laws from specific observations. While it does not allow one to draw conclusions with absolute certainty, it enables **evidence-based decision-making under uncertainty**.

Example: drug trials
By observing the reactions of a sample of individuals to a new drug, statistical inference is used to:

- Estimate the risk of side effects.
- Assess the drug's effectiveness.

This information helps decide whether to use the drug on a large scale.

6.2.1 Deductive vs. Inductive methods

Deductive method

- Starts from general laws, principles, or axioms.
- Draws conclusions about specific cases with **absolute certainty**.
- Example: All mathematics is based on the deductive method.

Inductive method

- Starts with specific observations or cases.
- Attempts to generalize and establish broader laws or principles.
- Example: Statistical inference, such as estimating drug effectiveness or population parameters.

6.3 Probability sample

A **probability sample** is obtained when, at the stage of designing the data collection, the statistician predetermines the probability law (**sample design**) for selecting the sample. This ensures:

- The selection process is random.
- The probability of each unit in the population being included in the sample is known and fixed in advance.

Example: To assess the quality of a school's cafeteria service:

- **Non-probability sampling:** Interview the first 100 students who show up at the cafeteria. This method is convenient but prone to selection bias.
- **Probability sampling:** Create a complete list of all enrolled students and randomly select a sample from this list, ensuring that the probability of each student being included is known and predetermined.

Non-probability sampling and selection bias
Non-probability sampling: In non-probability sampling, the choice of sample units is based on convenience or practicality, without a random mechanism.

Selection bias: This occurs when:

- The sample does not represent the population.
- Improper randomization results in skewed or unreliable conclusions.

6.3.1 Sample size

The **sample size** (n) refers to the number of units included in the sample. Let X_i represent the random variable that generates the i-th observation x_i, i.e., the value of the variable of interest at the i-th unit of the sample ($i = 1, \dots, n$).

$$X_1, X_2, \dots, X_n \quad \text{are the observed values of the sample.}$$

6.4 Infinite population

Infinite population is a mathematical model (widely used in statistics) used when considering a population consisting of a very large number of units and including all potentially observable units that do not necessarily already physically exist.

Examples:

- You want to make an investigation of the quality of electronic circuits produced by an industrial machinery. The reference population is the set of all electronic circuits that the machinery is capable of producing in the long run. The sample consists of n electronic circuits actually produced.
- In a study of the male height of adult individuals in Italy, the population consists of all Italian adult males (millions of people).

6.5 The statistical model

In infinite populations, the character of interest can be represented by a random variable X with a given probability distribution. The random variable X represents the population, and the **parameters** are the numerical constants that characterize its probability distribution.

6.5.1 Examples of parameters

- For $X \sim \mathcal{N}(\mu, \sigma^2)$, the parameters are:
  - Mean μ,
  - Variance σ^2.
- For $X \sim \text{Poisson}(\lambda)$, the parameter is:
  - Rate λ, which equals both the mean and variance.

6.5.2 Sample and observations:

- A **sample** of size n is the n-tuple $(X_1, \ldots, X_n)$ of iid random variables, where each X_i is distributed as X.
- The observed values $(x_1, \ldots, x_n)$ are a realization of the random vector $(X_1, \ldots, X_n)$.

6.5.3 Example: height survey

- Suppose we study the height of adult males in Italy.
- Assume heights follow a normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$, where:
  - μ: Mean height of adult males.
  - σ^2: Variance of the heights.

  The goal is to use the observed sample $x_1, \ldots, x_n$ to estimate the unknown parameters μ and σ^2.

6.5.4 General statistical model

Let X be a random variable with a probability distribution P_θ that depends on an unknown parameter θ. Then:

- X has a density function (or pmf) $f(x; \theta)$.
- The sample consists of n iid random variables $X_1, \ldots, X_n$, each distributed as X.

6.6 Sample statistics

A **sample statistic** $T(X_1, \ldots, X_n)$ is a random variable that is a function of the sample observations $X_1, \ldots, X_n$. It depends only on the sample and not on unknown population parameters.

Sampling distribution: The probability distribution of a statistic is called its **sampling distribution**.

6.6.1 Examples of sample statistics:

- **Sample mean:**
$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$
- **Sample variance:** The unbiased sample variance is:
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$
  The population version is:
$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$
- **Sample median:** The sample median $\widetilde{X}_n$ is the middle value of the ordered sample data. For:

  - n odd: $\widetilde{X}_n$ is the middle observation.
  - n even: $\widetilde{X}_n$ is the average of the two middle observations.

- **Other sample statistics:**
  - **Range:** Range $= \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$.
  - **Proportion:** The proportion of sample observations satisfying a given condition.

6.7 Point estimate and estimators

A **point estimate** is a single value used to approximate a population parameter. It is derived from a sample statistic, which is a function of the observed data.

6.7.1 Estimator and estimate

An **estimator** is a sample statistic $T(X_1, \ldots, X_n)$ used to estimate a population parameter θ. The value of the estimator computed for a particular sample $(x_1, \ldots, x_n)$ is called the **estimate**, denoted as:

$$t = T(x_1, \ldots, x_n).$$

6.7.2 Examples of estimators

- **Proportion:**
  - Example: Estimating the proportion p of non-defective products in a month.
  - If $X \sim \text{Bernoulli}(p)$, the **estimator** for p is:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- **Variance:**
  - Example: Estimating the population variance σ^2 when $X \sim \mathcal{N}(\mu, \sigma^2)$.
  - Estimators for σ^2:
    - ∗ **Unbiased Estimator:**

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

    - ∗ **Biased Estimator:**

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

6.7.3 Unbiasedness of an estimator

An estimator T is **unbiased** for a parameter θ if:

$$E(T) = \theta, \quad \text{for all possible values of } \theta.$$

Bias of an estimator:
The **bias** of an estimator T is:

$$B(T) = E(T) - \theta.$$

- If $B(T) = 0$, the estimator is unbiased.
- An unbiased estimator has a probability distribution centered on θ, meaning it neither systematically overestimates nor underestimates the parameter.

6.7.4 Sample mean

Let $(X_1, \ldots, X_n)$ be a random iid sample of size n from a population with mean μ and variance σ^2.
The **sample mean** is defined as:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Unbiasedness of the sample mean
The sample mean is an unbiased estimator for μ. To prove this:

$$E(\overline{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \cdot n\mu = \mu.$$

Variance of the sample mean: The variance of the sample mean is:

$$\operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

6.7.5 Sample variance

Biased sample variance
The following estimator for σ^2 is biased:

$$\widetilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Its expected value is:

$$E(\widetilde{S}_n^2) = \frac{n-1}{n}\sigma^2 < \sigma^2.$$

Thus, $\widetilde{S}_n^2$ underestimates σ^2.

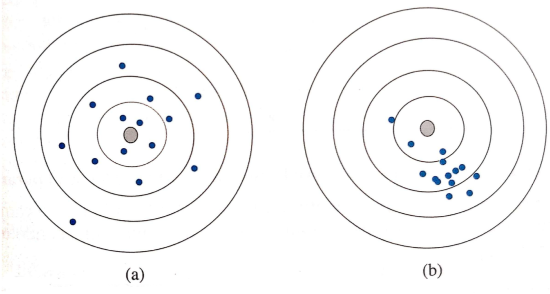
Unbiased sample variance
The unbiased estimator for σ^2 is:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Its expected value is:

$$E(S_n^2) = \sigma^2.$$

Therefore, S_n^2 is called the **unbiased sample variance**.



6.7.6 Estimates as shots on target

To estimate a parameter exactly is to hit the target.
Repeated parameter estimates can be imagined as target practice shots.
Each "shot" corresponds to the extraction of a sample from the population and the computation of the corresponding estimate.

- Fig. (a): estimates of an unbiased estimator.
  The estimates are "dispersed" around the true value of the parameter (center) with no deviations occurring in any particular direction.
- Fig. (b): estimates of a biased estimator.
  Estimates tend to be concentrated In an area below the center (bias). We have a systematic deviation.

Estimator (b) has less variance than estimator (a). Variance is not the appropriate measure for assessing the mean error of the estimate.

6.8 Mean square error

The **mean square error (MSE)** of an estimator T for a parameter θ is:

$$\text{MSE}(T) = E\left[(T-\theta)^2\right].$$

- MSE combines the variance and bias of the estimator.
- It provides a summary measure of the closeness of T to θ.

6.8.1 Decomposition of MSE

The MSE can be decomposed into:

$$\text{MSE}(T) = \text{Var}(T) + \text{Bias}(T)^2,$$

where:

- **Variance:**
  $$\text{Var}(T) = E\left[(T-E(T))^2\right].$$
- **Bias:**
  $$\text{Bias}(T) = E(T) - \theta.$$

Proof of MSE decomposition

Expanding $\mathrm{MSE}(T) = E\left[(T-\theta)^2\right]$:

$$\mathrm{MSE}(T) = E\left[(T-E(T))^2 + (E(T)-\theta)^2 + 2(T-E(T))(E(T)-\theta)\right].$$

- The first term $E\left[(T-E(T))^2\right]$ is $\mathrm{Var}(T)$.
- The second term $(E(T)-\theta)^2$ is $\mathrm{Bias}(T)^2$.
- The third term vanishes because $E(T-E(T)) = 0$.

Thus:

$$\mathrm{MSE}(T) = \mathrm{Var}(T) + \mathrm{Bias}(T)^2.$$

6.8.2 MSE for unbiased estimators

If T is unbiased ($E(T) = \theta$):

$$\mathrm{Bias}(T) = 0 \quad \Rightarrow \quad \mathrm{MSE}(T) = \mathrm{Var}(T).$$

6.8.3 Efficiency

Given two estimators T_1 and T_2 for θ:

- T_1 is **more efficient** than T_2 if:

$$\mathrm{MSE}(T_1) < \mathrm{MSE}(T_2), \quad \text{for all possible values of } \theta.$$

6.8.4 Efficiency for unbiased estimators

For two unbiased estimators T_1 and T_2, T_1 is more efficient than T_2 if:

$$\mathrm{Var}(T_1) < \mathrm{Var}(T_2), \quad \text{for all possible values of } \theta.$$

6.9 Consistency

An estimator T_n of a parameter θ is consistent in probability if:

$$\lim_{n\to\infty} P(|T_n - \theta| < \varepsilon) = 1, \quad \text{for every } \varepsilon > 0.$$

6.9.1 Consistency in mean square

T_n is consistent in mean square if:

$$\lim_{n\to\infty} \mathrm{MSE}(T_n) = \lim_{n\to\infty} E[(T_n-\theta)^2] = 0.$$

6.9.2 Relationship between consistency types

- If T_n is consistent in mean square, it is also consistent in probability.
- If T_n is consistent in probability, it may not be consistent in mean square unless $E[(T_n - \theta)^2] < \infty$.

6.9.3 Asymptotic unbiasedness

An estimator T_n is asymptotically unbiased if:

$$\lim_{n\to\infty} E(T_n) = \theta.$$

This is equivalent to:

$$\lim_{n\to\infty} \text{Bias}(T_n) = 0.$$

Examples

Sample mean: Let $X_1, \ldots, X_n$ be iid with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. The sample mean:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is consistent in mean square:

$$\lim_{n\to\infty} \text{Var}(\overline{X}_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Sample variance
Let S_n^2 be the unbiased sample variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

If $E(X^4) < \infty$, then:

$$\lim_{n\to\infty} \text{Var}(S_n^2) = 0,$$

so S_n^2 is consistent in mean square.

6.10 Distribution of the sample mean

6.10.1 Gaussian case

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $(X_1, \ldots, X_n)$ be an iid sample. The sample mean is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Since $\overline{X}_n$ is a linear combination of independent Gaussian random variables:

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

6.10.2 Non-Gaussian case

If X has an unknown distribution, the distribution of $\overline{X}_n$ is generally unknown. However, by the Central Limit Theorem:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

For large n, the distribution of $\overline{X}_n$ can be approximated by:

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

6.11 Distribution of the sample variance (Gaussian case)

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $(X_1, \ldots, X_n)$ be an iid sample. The sample variance is:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

1. The sample variance is related to the chi-squared distribution:

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

2. Variance of the sample variance:

$$\mathrm{Var}(S_n^2) = \frac{2\sigma^4}{n-1}.$$

6.12 Distribution of the sample proportion

Let $X \sim \text{Bernoulli}(p)$ and $(X_1, \ldots, X_n)$ be an iid sample. The sample proportion is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Mean and variance:

$$E(\overline{X}_n) = p, \quad \mathrm{Var}(\overline{X}_n) = \frac{p(1-p)}{n}.$$

- For large n, by the Central Limit Theorem:

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

  and:

$$\overline{X}_n \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

6.13 Point estimate and standard error

Let X be a random variable with:

- Probability mass function $p(x;\theta)$, if X is discrete, or
- Probability density function $f(x;\theta)$, if X is continuous.

Let $(X_1,\ldots,X_n)$ be an iid sample from X, and let $T(X_1,\ldots,X_n)$ be an **estimator** of the parameter θ. Once the sample is observed $(x_1,\ldots,x_n)$, the **point estimate** of θ is:

$$T(x_1,\ldots,x_n)=t.$$

- The **estimator** $T(X_1,\ldots,X_n)$ is a random variable.
- The **point estimate** t is a numerical value obtained after observing the sample.

6.13.1 Standard error (SE)

The **standard error** of an estimator T is:

$$\mathrm{SE}(T)=\sqrt{\mathrm{Var}(T)}.$$

- It measures the variability of the estimator T.
- A smaller standard error indicates a more precise estimator.

Example: sample mean
Let $(X_1,\ldots,X_n)$ be an iid sample from a population with mean μ and variance σ^2. The sample mean:

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

is an estimator of μ. The standard error of $\overline{X}_n$ is:

$$\mathrm{SE}(\overline{X}_n)=\sqrt{\mathrm{Var}(\overline{X}_n)}=\sqrt{\frac{\sigma^2}{n}}.$$

6.14 Likelihood and Maximum Likelihood Estimation (MLE)

6.14.1 Likelihood function

Let $(X_1,\ldots,X_n)$ be an iid sample with marginal PDF or PMF $f(x;\theta)$, where $\theta\in\Theta$ is the unknown parameter. The **likelihood function** is:

$$L(\theta)=L_{x_1,\ldots,x_n}(\theta)=\prod_{i=1}^{n}f(x_i;\theta).$$

- The likelihood measures how likely it is to observe the sample $(x_1,\ldots,x_n)$ given θ.

6.14.2 Log-likelihood function

The **log-likelihood function** is:

$$\ell(\theta)=\ell_{x_1,\ldots,x_n}(\theta)=\log L(\theta)=\sum_{i=1}^{n}\log f(x_i;\theta).$$

- Logarithms simplify computations since the product becomes a sum.
- The maximum of $\ell(\theta)$ occurs at the same point as the maximum of $L(\theta)$.

6.14.3 Maximum Likelihood Estimator (MLE)

The **maximum likelihood estimator** (MLE) $\hat{\theta}$ is the value that maximizes the likelihood function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Interpretation: The MLE is the parameter value that makes the observed data most likely.

6.14.4 Properties of the MLE

Under regularity conditions, the MLE has the following properties:

- **Asymptotic unbiasedness:**
$$\lim_{n \to \infty} \text{Bias}(\hat{\theta}) = 0.$$
- **Consistency:**
$$\hat{\theta} \xrightarrow{P} \theta, \quad \text{as } n \to \infty.$$
- **Asymptotic efficiency:** For large n, the MLE achieves the smallest possible variance among unbiased estimators.
- **Asymptotic Normality:**
$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta)),$$
where $I(\theta)$ is the Fisher information.
- **Invariance property:** If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$ for any function g.

7 Inferential statistics - confidence intervals and hypothesis testing

7.1 Interval estimation

Instead of estimating a parameter θ with a single value (point estimate), interval estimation provides a range of plausible values for θ, called a **confidence interval**. This range is constructed to contain θ with a certain level of confidence.

7.1.1 Components of interval estimation

- An **estimator** T for the parameter θ.
- The **probability distribution** of the estimator T.
- A **confidence level** $1-\alpha$, representing the reliability of the interval.
- The **confidence interval**, a set of values for θ.

7.1.2 Random intervals

Let X represent a random variable (rv) whose probability distribution depends on an unknown parameter θ. Let $X_1, \ldots, X_n$ be an independent and identically distributed (iid) sample from X.

Consider two statistics:

$$L_1 = L_1(X_1, \ldots, X_n), \quad L_2 = L_2(X_1, \ldots, X_n),$$

where $L_1 \leq L_2$ for every possible sample.

The interval:

$$[L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$$

is a **random interval**, as its bounds are random variables that depend on the sample.

7.1.3 Confidence interval

Let X (the population) be a random variable (rv) whose probability distribution depends on an unknown parameter θ. Let $X_1, \ldots, X_n$ be an independent and identically distributed (iid) sample from X.

The random interval:

$$[L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$$

is a **confidence interval** (CI) of level $1-\alpha$ for the parameter θ if:

$$P\left(L_1(X_1, \ldots, X_n) \leq \theta \leq L_2(X_1, \ldots, X_n)\right) = 1-\alpha.$$

Typically, α is set to 0.05 (95% confidence level) or 0.01 (99% confidence level).

The **numerical interval**:

$$[l_1, l_2] = [L_1(x_1, \ldots, x_n), L_2(x_1, \ldots, x_n)],$$

is the realization of the random interval based on the observed sample and is called the **estimated confidence interval**.

Key insight
While it is impossible to determine whether the estimated confidence interval contains the true value of θ, the probability of obtaining an interval that contains θ is $1 - \alpha$.

Pivotal quantities
A **pivotal quantity** is a random variable:

$$W = g(\theta; X_1, \ldots, X_n),$$

that depends on the sample observations and the parameter θ, but whose probability distribution is independent of θ.

Examples of pivotal quantities

1. If $X_1, \ldots, X_n \sim \mathrm{N}(\mu, 4)$, a pivotal quantity is:

$$W = \frac{\bar{X}_n - \mu}{2/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

2. If $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, a pivotal quantity is:

$$W = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

   where $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}$.

3. If $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$, a pivotal quantity is:

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

   where S^2 is the sample variance.

Constructing a confidence interval

1. Consider a pivotal quantity:

$$W = g(\theta; X_1, \ldots, X_n),$$

   that depends on the observations through a proper point estimator of θ, and ensure that it is a monotonic function of θ.

2. Choose two values $a, b \in R$ such that:

$$P\left(a \leq g(\theta; X_1, \ldots, X_n) \leq b\right) = 1 - \alpha.$$

   This is possible because the distribution of W is known and does not depend on θ.

3. Find two statistics $L_1 = L_1(X_1, \ldots, X_n)$ and $L_2 = L_2(X_1, \ldots, X_n)$ such that $L_1 \leq L_2$ and:

$$a \leq g(\theta; X_1, \ldots, X_n) \leq b \iff L_1(X_1, \ldots, X_n) \leq \theta \leq L_2(X_1, \ldots, X_n).$$

Applications of confidence intervals
We will construct confidence intervals for:

- The mean of a Normal population with known variance.
- The mean of a Normal population with unknown variance.
- The variance of a Normal population with unknown mean.

7.2 Confidence intervals of a Gaussian population

Mean (known variance)

Let $X \sim N(\mu, \sigma_0^2)$ be a Gaussian population with known variance σ_0^2. Let $(X_1, \ldots, X_n)$ be an independent and identically distributed (iid) sample from X. The pivotal quantity is:

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1).$$

Let $z_{\alpha/2}$ be the value such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 \quad \text{and} \quad P(Z < -z_{\alpha/2}) = \alpha/2.$$

By the symmetry of the normal distribution:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Rewriting in terms of μ, we obtain:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

Thus, the confidence interval (CI) for the mean μ is:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right).$$

Key points:

- α is the probability that the CI does not contain μ.
- If $\alpha = 0.05$, approximately 95% of the CIs will contain μ across many samples.
- Increasing the confidence level $1 - \alpha$ (e.g., using $\alpha = 0.01$) leads to a wider CI.

Symmetry of the CI The symmetric CI is chosen because it minimizes the width of the interval. For a symmetric normal distribution:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Other asymmetric intervals are possible but do not minimize the width.

7.2.1 Mean (unknown variance)

If the variance σ^2 is unknown, we use the sample variance S^2. Let:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} follows a Student's t-distribution with $n-1$ degrees of freedom.
Let $t_{n-1,\alpha/2}$ satisfy:

$$P(T > t_{n-1,\alpha/2}) = \alpha/2 \quad \text{and} \quad P(T < -t_{n-1,\alpha/2}) = \alpha/2.$$

By the symmetry of the t-distribution:

$$P(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}) = 1 - \alpha.$$

Rewriting in terms of μ, we obtain:

$$P\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Thus, the confidence interval for μ is:

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Key points:

- The t-distribution accounts for the extra uncertainty due to estimating σ^2.
- For large n, the t-distribution approximates the normal distribution.
- The CI widens as n decreases, reflecting increased variability in small samples.

7.2.2 Variance (unknown mean

Consider a Gaussian population $X \sim N(\mu, \sigma^2)$, where both the mean μ and the variance σ^2 are unknown. Let $(X_1, \ldots, X_n)$ be an independent and identically distributed (iid) sample from X. The pivotal quantity for estimating σ^2 is:

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where S^2 is the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

and χ^2_{n-1} represents a chi-squared distribution with $n-1$ degrees of freedom.

Confidence interval derivation
Let $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ denote the critical values of the chi-squared distribution such that:

$$P(W > \chi^2_{n-1,\alpha/2}) = \alpha/2 \quad \text{and} \quad P(W < \chi^2_{n-1,1-\alpha/2}) = \alpha/2.$$

From the properties of the chi-squared distribution:

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq W \leq \chi^2_{n-1,\alpha/2}\right) = 1-\alpha.$$

Rewriting in terms of σ^2, we have:

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha.$$

Thus, the confidence interval for σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right).$$

Estimated confidence interval
Given a sample $(x_1, \ldots, x_n)$ with sample variance s^2, the estimated confidence interval is:

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right).$$

Example
Let $n = 10$ and $\alpha = 0.05$. Using statistical software (e.g., R), compute:

- $\chi^2_{n-1,\alpha/2} = \chi^2_{9,0.025}$: `qchisq(1 - 0.05/2, df = 9)`.
- $\chi^2_{n-1,1-\alpha/2} = \chi^2_{9,0.975}$: `qchisq(0.05/2, df = 9)`.

Remarks

- The chi-squared distribution is not symmetric, so $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ must be computed separately.
- For simplicity, an interval with equiprobable tails is used, but this does not guarantee minimum width.
- As $n \to \infty$, the asymmetry of the chi-squared distribution diminishes, and the interval approximates one with minimum width.

7.2.3 Variance (known mean)

Let $X \sim N(\mu, \sigma^2)$, where the mean μ is known, and the variance σ^2 is to be estimated. Given a sample $(X_1, \ldots, X_n)$, the estimator for σ^2 is:

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The pivotal quantity based on S_0^2 is:

$$W = \frac{nS_0^2}{\sigma^2} \sim \chi^2_n,$$

where χ^2_n denotes the chi-squared distribution with n degrees of freedom.

Confidence interval derivation
Let $\chi^2_{n,\alpha/2}$ and $\chi^2_{n,1-\alpha/2}$ denote the critical values of the chi-squared distribution such that:

$$P(W > \chi^2_{n,\alpha/2}) = \alpha/2 \quad \text{and} \quad P(W < \chi^2_{n,1-\alpha/2}) = \alpha/2.$$

From the properties of the chi-squared distribution:

$$P\left(\chi^2_{n,1-\alpha/2} \leq W \leq \chi^2_{n,\alpha/2}\right) = 1 - \alpha.$$

Rewriting in terms of σ^2, we have:

$$P\left(\frac{nS_0^2}{\chi^2_{n,\alpha/2}} \leq \sigma^2 \leq \frac{nS_0^2}{\chi^2_{n,1-\alpha/2}}\right) = 1 - \alpha.$$

Thus, the confidence interval for σ^2 is:

$$\left(\frac{nS_0^2}{\chi^2_{n,\alpha/2}}, \frac{nS_0^2}{\chi^2_{n,1-\alpha/2}}\right).$$

Estimated confidence interval
Given a sample $(x_1, \ldots, x_n)$ and the observed value of S_0^2:

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

the estimated confidence interval is:

$$\left(\frac{nS_0^2}{\chi^2_{n,\alpha/2}}, \frac{nS_0^2}{\chi^2_{n,1-\alpha/2}}\right).$$

Example
Let $n = 10$ and $\alpha = 0.05$. Using statistical software (e.g., R), compute:

- $\chi^2_{n,\alpha/2} = \chi^2_{10,0.025}$: `qchisq(1 - 0.05/2, df = 10)`.
- $\chi^2_{n,1-\alpha/2} = \chi^2_{10,0.975}$: `qchisq(0.05/2, df = 10)`.

Remarks

- The chi-squared distribution is not symmetric, so the bounds $\chi^2_{n,\alpha/2}$ and $\chi^2_{n,1-\alpha/2}$ must be computed separately.
- This CI assumes the tails of the distribution are equiprobable, resulting in equal probability in the left and right tails.

7.2.4 Unknown mean and variance

If neither the mean μ nor the variance σ^2 is known, consider the pivotal quantity:

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where χ^2_{n-1} follows a chi-squared distribution with $n-1$ degrees of freedom.

Let $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ satisfy:

$$P(W > \chi^2_{n-1,\alpha/2}) = \alpha/2 \quad \text{and} \quad P(W < \chi^2_{n-1,1-\alpha/2}) = \alpha/2.$$

The confidence interval for σ^2 is:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right).$$

Key points:

- The chi-squared distribution is not symmetric, so the CI is not symmetric either.
- For large n, the asymmetry diminishes, and the CI approximates symmetry.

7.3 Asymptotic confidence intervals

- Asymptotic confidence intervals (CIs) are based on approximations of the distribution of pivotal quantities for large sample sizes (n).
- These intervals are used when the exact distribution of the pivotal quantity is unknown or too complicated to work with.
- The confidence level for an asymptotic CI is approximately calculated.
- Asymptotic CIs are valid only for large sample sizes ($n > 120$).

7.3.1 Asymptotic CI for the mean μ (unknown variance)

Let $X_1, \ldots, X_n$ be an iid sample. For large n, whether X is Gaussian or non-Gaussian:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} Z \sim N(0,1), \quad n \to \infty,$$

where $\bar{X}$ is the sample mean and S is the sample standard deviation.

Thus, the asymptotic CI for μ is:

$$\left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} \right].$$

Example: If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$, and the CI becomes:

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}} \right].$$

7.3.2 Asymptotic CI for a proportion p

Let $X \sim \text{Ber}(p)$, where $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Let $X_1, \dots, X_n$ be an iid sample of size n, and let $\bar{X}$ be the sample mean. Then:

$$E(\bar{X}) = p, \quad \text{Var}(\bar{X}) = \frac{p(1-p)}{n}.$$

For large n, we have:

$$\frac{\bar{X} - p}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}} \xrightarrow{d} Z \sim N(0, 1).$$

Thus, the **Wald asymptotic** CI for p is:

$$\left[\bar{x} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right].$$

Example: If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$, and the CI becomes:

$$\left[\bar{x} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right].$$

Condition for validity:

$$n \cdot \bar{x} \geq 5 \quad \text{and} \quad n \cdot (1 - \bar{x}) \geq 5.$$

7.3.3 Score confidence interval for p

Starting from the approximation:

$$P\left(\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2} \right) \approx 1 - \alpha,$$

we obtain the score confidence interval:

$$\left[\frac{\bar{x} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x}) + \frac{z_{\alpha/2}^2}{4n}}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}}, \frac{\bar{x} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x}) + \frac{z_{\alpha/2}^2}{4n}}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}} \right].$$

Advantage: The score CI generally performs better than the Wald CI, particularly when p is near 0 or 1.

7.4 Decisions under conditions of uncertainty

A company manufactures airplane spare parts using a new machine with a lighter aluminum alloy. The goal is to test and evaluate the production process. The conditions are as follows:

- The average weight of the parts should be $\mu = 15\,\text{kg}$.
- If the average weight is significantly different from 15 kg, the production process must be stopped and overhauled.

The decision-making process

1. A random sample of $n = 16$ pieces is chosen, and their weights are observed: $x_1, x_2, \dots, x_{16}$.
2. The sample mean $\bar{x}$ is calculated:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
3. If $\bar{x}$ is "very different" from $\mu = 15$, the decision is made to stop the production process.

Uncertainty due to sampling

- The entire population is not known; the decision is based on the sample mean $\bar{x}$.
- Even if $\bar{x} \neq 15$ for the sample, the true population mean μ might still be 15 kg.
- The uncertainty arises due to **sampling error**, defined as:
$$\text{Sampling Error} = \bar{x} - \mu.$$

Defining a decision rule

To decide whether to stop production based on the sample mean, it is necessary to account for the sampling error. This involves:

- Using statistical methods to determine whether the observed $\bar{x}$ is significantly different from $\mu = 15$.
- Defining a threshold or confidence interval around μ that accounts for variability in the sample.
- Making a decision:
  - If $\bar{x}$ falls within the threshold, continue production.
  - If $\bar{x}$ falls outside the threshold, stop production.

Statistical tools for decision-making

- Hypothesis testing can be used to determine whether the observed $\bar{x}$ deviates significantly from μ:
  - Null Hypothesis (H_0): $\mu = 15$.
  - Alternative Hypothesis (H_1): $\mu \neq 15$.
- A confidence interval can provide a range of plausible values for μ based on the sample mean:
$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Conclusion

By considering sampling error and statistical tools, the company can make an informed decision about whether the new machine meets the desired production standards.

7.5 Hypothesis testing

Let X represent the population of interest (e.g., weights of spare parts). The probability distribution of X is assumed to belong to a parametric family (e.g., Gaussian distribution with known variance), but the value of the parameter θ is unknown.
A **statistical hypothesis** is a conjecture about the unknown parameter θ. Two opposite hypotheses are formulated:

- **Null Hypothesis (H_0):** The hypothesis held true until proven otherwise. It represents the status quo or baseline assumption.
- **Alternative Hypothesis (H_1):** The hypothesis that opposes H_0, in favor of which empirical evidence is sought.

Key features of hypotheses:

- They must be **incompatible** (both cannot be true).
- They must be **exhaustive** (one must be true).

Examples of hypotheses

- **Example 1:** Testing if spare parts weigh 15 kg on average.

$$H_0 : \mu = 15, \quad H_1 : \mu \neq 15.$$

- **Example 2:** Monitoring ball bearing production:

$$H_0 : \mu = 5, \quad H_1 : \mu < 5.$$

- **Example 3:** Checking if weekly sales of frozen broccoli have increased:

$$H_0 : \mu = 2400, \quad H_1 : \mu > 2400.$$

7.5.1 Types of statistical hypotheses

- **Point Hypothesis:** Specifies a single value for the parameter θ.
- **Composite Hypothesis:** Specifies a range of values for θ.
  - The null hypothesis (H_0) can be either a point or composite hypothesis.
  - The alternative hypothesis (H_1) is always a composite hypothesis.

Example:

- Point null hypothesis: $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$.
- Composite null hypothesis: $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$.

7.5.2 Directional and non-directional hypotheses

- **Right-tailed test (directional):** The alternative hypothesis considers parameter values greater than those specified by H_0:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0.$$

- **Left-tailed test (directional):** The alternative hypothesis considers parameter values less than those specified by H_0:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0.$$

- **Two-tailed test (non-directional):** The alternative hypothesis considers parameter values different from those specified by H_0:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

7.5.3 Statistical hypothesis testing

- A statistical test is a rule to decide whether to reject H_0 based on the sample.
- A significance test evaluates whether the sample data contradict H_0 and support H_1, using **proof by contradiction**:
  - Assume H_0 is true.
  - Determine if the observed data would be very unusual under H_0.
  - If the data are highly improbable under H_0, reject H_0 in favor of H_1.

7.6 Test statistics

To decide whether to reject the null hypothesis (H_0), an appropriate estimator T of the parameter θ is chosen. This estimator, called the **test statistic**, satisfies the following properties:

- The distribution of T must be known under H_0.
- The decision to reject H_0 is based on the observed value t of T, and how far t is from the hypothesized value θ_0.

7.6.1 Critical region and critical values

- The **critical region** (or rejection region) is the set of values for the test statistic for which H_0 is rejected.
- If the observed test statistic t lies in the critical region, reject H_0; otherwise, fail to reject H_0.
- The boundaries of the critical region are called **critical values**. These are determined by controlling error probabilities.

7.6.2 Types of errors in hypothesis testing

First and second kind errors

- A **Type I error** (error of the first kind) occurs when H_0 is rejected even though it is true. This is also called a **false positive**.
- A **Type II error** (error of the second kind) occurs when H_0 is not rejected even though it is false. This is also called a **false negative**.

7.6.3 Significance level (α)

- In the case of a **point null hypothesis** ($H_0 : \theta = \theta_0$), the **significance level** α is the probability of committing a Type I Error:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

- For a **composite null hypothesis** (e.g., $H_0 : \theta \leq \theta_0$), α represents the maximum probability of committing a Type I Error:

$$\alpha = \max_{\theta \in H_0} P(\text{Reject } H_0 \mid \theta).$$

Summary

- The test statistic T is used to determine whether to reject H_0.
- The critical region, defined by critical values, specifies when H_0 should be rejected.
- Errors in hypothesis testing:
  - Type I Error (false positive): Rejecting H_0 when it is true.
  - Type II Error (false negative): Failing to reject H_0 when it is false.
- The significance level α controls the probability of a Type I Error.

7.7 Two Equivalent approaches to identify the rejection zone

7.7.1 1. Critical value approach

The decision-making rule is based on comparing the test statistic with critical values, k_1, k_2, k_3, k_4, which depend on the chosen significance level α.

Rules for hypothesis testing:

- **Right-tailed test:** Reject H_0 if:

$$t > \theta_0 + k_1, \quad k_1 > 0.$$

- **Left-tailed test:** Reject H_0 if:

$$t < \theta_0 - k_2, \quad k_2 > 0.$$

- **Two-tailed test:** Reject H_0 if:

$$t > \theta_0 + k_3 \quad \text{or} \quad t < \theta_0 - k_4, \quad k_3, k_4 > 0.$$

The critical values k_1, k_2, k_3, k_4 are chosen to ensure:

$$P(\text{reject } H_0 \mid H_0) = \alpha.$$

7.7.2 2. P-value approach

The p-value represents the probability of observing a test statistic as extreme as the one obtained, assuming H_0 is true. The decision rule:

- Reject H_0 if:
$$\text{p-value} \leq \alpha.$$

7.7.3 Stages of hypothesis testing

1. Define the null (H_0) and alternative (H_1) hypotheses based on the problem.
2. Choose a significance level α (e.g., $\alpha = 0.1, 0.05, 0.01$).
3. Collect a random sample from the population.
4. Compute the appropriate test statistic.
5. Make a decision about H_0 based on the critical value or p-value.

7.7.4 Real-world example: testing spare parts weight

We test:

$$H_0 : \mu = 15, \quad H_1 : \mu \neq 15.$$

Assume $X \sim N(\mu, 4)$, and a random sample of size $n = 16$ is drawn. The sample mean $\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(\mu, \frac{4}{16}) = N(\mu, 0.25)$.

Decision rule
We need to decide:

- Do not reject H_0 and continue the production process.
- Reject H_0 and stop the production process.

To determine the decision rule:

- We reject H_0 if $\bar{x} \leq 15 - k_1$ or $\bar{x} \geq 15 + k_2$.
- For a two-tailed test, distribute the Type I error (α) equally between the two tails:

$$P(\bar{X} \leq 15 - k_1 \mid H_0) = \frac{\alpha}{2}, \quad P(\bar{X} \geq 15 + k_2 \mid H_0) = \frac{\alpha}{2}.$$

Critical values
Under H_0, $\bar{X} \sim N(15, 0.25)$. Using the standard normal distribution:

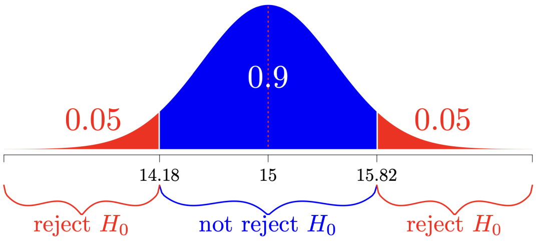
$$P\left(Z \leq -z_{\alpha/2}\right) = \frac{\alpha}{2}, \quad P\left(Z \geq z_{\alpha/2}\right) = \frac{\alpha}{2}.$$

Thus:

$$k_1 = k_2 = z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}} = z_{\alpha/2} \cdot \sqrt{0.25}.$$

If $\alpha = 0.1$, then $z_{\alpha/2} = z_{0.05} = 1.645$. Hence:

$$k_1 = k_2 = 1.645 \cdot \sqrt{0.25} = 1.645/2 = 0.8225.$$



Reject H_0 if:

$$|\bar{x} - 15| \geq 0.8225.$$

Example calculation

Suppose the sample mean is $\bar{x} = 14$. Compute:

$$|\bar{x} - 15| = |14 - 15| = 1.$$

Since $1 > 0.8225$, we reject H_0 and conclude that the mean weight is significantly different from 15.

7.8 Relationship between the two approaches

- Both approaches lead to the same decision about H_0.
- The critical value approach identifies a rejection region, and the test statistic t is compared to critical values.
- The p-value approach computes the probability of observing t or a more extreme value, given H_0, and compares it to α.

7.8.1 Steps for each approach

- **Critical value approach:**
  1. Compute critical values for the rejection region.
  2. Reject H_0 if t falls in the critical region.
  3. Otherwise, do not reject H_0.
- **P-value approach:**
  1. Compute the p-value.
  2. Reject H_0 if p-value $\leq \alpha$.
  3. Otherwise, do not reject H_0.

7.8.2 Why the p-Value approach is preferred

- The p-value is more informative than critical values:
  - It provides the smallest significance level α at which H_0 can be rejected.
  - It indicates the strength of evidence against H_0.

7.9 Logic of significance testing

Significance testing is used to decide whether to reject the null hypothesis (H_0) based on the sample data. The process is constructed with a focus on controlling the probability of committing a Type I Error (α), which is rejecting H_0 when it is true.

7.9.1 Key principles of significance testing

- **Protecting against Type I error:** The tests are designed to minimize the risk of rejecting H_0 when it is true. This ensures that the null hypothesis is only rejected if there is strong empirical evidence against it.
- **Rejection vs. non-rejection:**
  - **Reject H_0:** Indicates that the data provide sufficient evidence to contradict H_0.
  - **Do not reject H_0:** Indicates that the data do not provide sufficient evidence to reject H_0, but this does not mean H_0 is true.
- **Uncertainty in non-rejection:** Failure to reject H_0 may occur because:
  - H_0 is true.
  - The test lacks statistical power, possibly due to a small sample size (n).
- **Role of the research hypothesis:** The null hypothesis (H_0) is formulated as the negation of the research hypothesis that the researcher seeks to validate. Rejection of H_0 provides support for the research hypothesis.

7.9.2 Examples of hypotheses

- **Example 1: smoking and cancer**
  - Research hypothesis: smoking increases the risk of cancer.
  - Null hypothesis: smoking does not increase the risk of cancer (H_0 : No association).
- **Example 2: drug effectiveness**
  - Research hypothesis: a new drug is effective.
  - Null hypothesis: the new drug is not effective (H_0 : No effect).
- **Example 3: criminal trial**
  - Null hypothesis: the defendant is innocent.
  - Alternative hypothesis: the defendant is guilty.
- In these examples, rejecting H_0 has serious implications, and significance tests are designed to protect against Type I Error, such as:
  - Claiming smoking is dangerous without evidence.
  - Approving an ineffective drug for sale.
  - Convicting an innocent person.

7.9.3 Rejection and non-rejection in practice

- **Reject H_0:** Indicates strong evidence against H_0 in favor of the alternative hypothesis (H_1).
- **Do not reject H_0:** Indicates that the data do not contradict H_0, but:
  - This does not prove H_0 is true.
  - Further testing with larger samples might lead to rejection of H_0.

7.10 The power of a test

Consider the case where the null and alternative hypotheses are simple:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1.$$

The following table summarizes the decision outcomes and associated probabilities:

	H_0 is True	**H_0 is False**
Do not reject H_0	Correct decision $(1 - \alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1 - \beta)$

- α: Probability of committing a **Type I Error** (rejecting H_0 when it is true).
- β: Probability of committing a **Type II Error** (failing to reject H_0 when it is false).
- $1 - \beta$: Probability of correctly rejecting H_0 when it is false. This is called the **power of the test**.

7.10.1 Definition of the power of a test

The power of a test is:

$$\text{Power} = 1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ is false}),$$

i.e., the probability of rejecting the null hypothesis when the alternative hypothesis is true.

7.10.2 The power function

When the alternative hypothesis is composite (e.g., $H_1 : \theta \geq \theta_0$, $H_1 : \theta \leq \theta_0$, or $H_1 : \theta \neq \theta_0$), the probability of rejecting H_0 depends on the true value of the parameter θ and is described by the **power function**, denoted by $\pi(\theta)$.

$$\pi(\theta) = P(\text{Reject } H_0 \mid \theta).$$

The power function represents the probability of rejecting H_0 as θ varies within the range of values specified by H_1.

7.10.3 Unbiased tests

A significance test of level α is said to be **unbiased** if:

$$\pi(\theta) \geq \alpha \quad \text{for every value of } \theta \text{ covered by } H_1.$$

This means that the probability of rejecting H_0 when it is false is at least as large as the significance level α.

Key points

- The power of a test $(1-\beta)$ measures its ability to detect when H_0 is false.
- A higher power indicates a better test, as it reduces the probability of committing a Type II Error.
- The power function $\pi(\theta)$ varies with the true parameter θ and describes the effectiveness of the test across the parameter space.
- Unbiased tests ensure that the probability of rejecting H_0 is higher when H_0 is false than when it is true.

7.11 Two-tailed Z test for the mean (Gaussian population with known variance)

Let $X \sim N(\mu, \sigma^2)$, and assume the variance σ^2 is known. Let $(X_1, \ldots, X_n)$ be an iid sample from X. To test the population mean μ, the hypotheses are:

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

The test statistic is the sample mean $\bar{X}_n$. Under H_0, the distribution of $\bar{X}_n$ is:

$$\bar{X}_n \sim N\left(\mu_0, \frac{\sigma^2}{n}\right).$$

7.11.1 Decision rule

We reject H_0 if $\bar{x}$ is far enough from μ_0, specifically:

$$\bar{x} \leq \mu_0 - k_1 \quad \text{or} \quad \bar{x} \geq \mu_0 + k_2,$$

where $k_1, k_2 > 0$ are thresholds determined by the significance level α.

7.11.2 Type I error and thresholds

The Type I error probability (α) is split equally between the two tails:

$$P\left(\bar{X}_n \leq \mu_0 - k_1 \mid H_0\right) = \frac{\alpha}{2}, \quad P\left(\bar{X}_n \geq \mu_0 + k_2 \mid H_0\right) = \frac{\alpha}{2}.$$

Under H_0, $\bar{X}_n \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$. Using the standard normal distribution:

$$P\left(Z \leq -z_{\alpha/2}\right) = \frac{\alpha}{2}, \quad P\left(Z \geq z_{\alpha/2}\right) = \frac{\alpha}{2},$$

where:

$$z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2}),$$

is the critical value of the standard normal distribution.

The critical thresholds are:

$$k_1 = k_2 = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Thus, the rejection rule becomes:

$$\text{Reject } H_0 \text{ if } |\bar{x} - \mu_0| \geq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

7.11.3 p-Value approach

The p-value is the probability of observing a test statistic as extreme as or more extreme than the observed value under H_0. For the two-tailed test:

$$\text{p-value} = 2 \cdot P\left(\bar{X}_n \geq |\bar{x} - \mu_0| \mid H_0\right),$$

or equivalently:

$$\text{p-value} = 2 \cdot \left(1 - \Phi\left(\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right)\right).$$

Reject H_0 if:

$$\text{p-value} \leq \alpha.$$

Summary of the Z test

- **Test statistic:**

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{under } H_0.$$

- **Rejection rule (critical value approach):**

$$\text{Reject } H_0 \text{ if } |\bar{x} - \mu_0| \geq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

- **Rejection rule (p-Value Approach):**

$$\text{Reject } H_0 \text{ if p-value} \leq \alpha.$$

- **Connection with confidence intervals:** Reject H_0 if $\mu_0 \notin \left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$, which is the confidence interval for μ at level $1 - \alpha$.

Remark

The equivalence between hypothesis testing and confidence intervals arises because both are based on the same sampling distribution.

7.12 Duality between confidence intervals and tests

Consider a hypothesis test to evaluate:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

The connection between confidence intervals and hypothesis tests can be summarized as follows:

- A **confidence interval** of level $1 - \alpha$ gives the range of plausible values for θ based on the sample data.
- A **hypothesis test** of level α evaluates whether θ_0 is a plausible value for θ based on the sample data.
- These two procedures are equivalent:
    - Reject H_0 if $\theta_0 \notin [L_1, L_2]$, where $[L_1, L_2]$ is the confidence interval of level $1 - \alpha$.
    - Do not reject H_0 if $\theta_0 \in [L_1, L_2]$.

7.12.1 Confidence interval to hypothesis test

Let $[L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$ be the bounds of a confidence interval for θ at level $1 - \alpha$. By definition:

$$1 - \alpha = P(L_1 \leq \theta \leq L_2).$$

Equivalently:

$$\alpha = P(\theta_0 \notin [L_1, L_2] \mid H_0).$$

Thus, a two-tailed hypothesis test with significance level α can be constructed as follows:

- Compute the confidence interval of level $1 - \alpha$.
- **Decision Rule:**
    - Reject H_0 if $\theta_0 \notin [L_1, L_2]$.
    - Do not reject H_0 if $\theta_0 \in [L_1, L_2]$.

7.12.2 Hypothesis test to confidence interval

Given a two-tailed hypothesis test with significance level α, a confidence interval of level $1 - \alpha$ can be constructed as follows:

- Identify all values of θ_0 for which the null hypothesis $H_0 : \theta = \theta_0$ is not rejected at level α.
- The resulting set of θ_0 values forms the confidence interval at level $1 - \alpha$.

Key insight

The duality between confidence intervals and hypothesis tests means that:

- A confidence interval of level $1 - \alpha$ contains all values of θ_0 that would not be rejected by a two-tailed test with significance level α.
- A hypothesis test with significance level α can be performed by checking whether θ_0 lies inside or outside the confidence interval of level $1 - \alpha$.

7.13 Overview of hypothesis tests

7.13.1 Z Test for the mean μ (Gaussian population with known variance)

Let $X_1, \ldots, X_n$ be an iid sample from $X \sim N(\mu, \sigma^2)$, where σ^2 is known. The test statistic is:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{under } H_0 : \mu = \mu_0.$$

The observed value of Z is:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Decision rules:

- **Right-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$\text{Reject } H_0 \text{ if } z \geq z_\alpha.$$

- **Left-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$\text{Reject } H_0 \text{ if } z \leq -z_\alpha.$$

- **Two-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$\text{Reject } H_0 \text{ if } |z| \geq z_{\alpha/2}.$$

7.13.2 T Test for the mean μ (Gaussian population with unknown Vvariance)

Let $X_1, \ldots, X_n$ be an iid sample from $X \sim N(\mu, \sigma^2)$, where σ^2 is unknown. The test statistic is:

$$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0 : \mu = \mu_0,$$

where S is the sample standard deviation. The observed value of T is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Decision rules:

- **Right-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$\text{Reject } H_0 \text{ if } t \geq t_{n-1,\alpha}.$$

- **Left-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$\text{Reject } H_0 \text{ if } t \leq -t_{n-1,\alpha}.$$

- **Two-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$\text{Reject } H_0 \text{ if } |t| \geq t_{n-1,\alpha/2}.$$

7.13.3 χ^2 Test for variance σ^2 (Gaussian population with unknown mean)

Let $X_1, \ldots, X_n$ be an iid sample from $X \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. The test statistic is:

$$W = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \quad \text{under } H_0 : \sigma^2 = \sigma_0^2,$$

where S^2 is the sample variance. The observed value of W is:

$$w = \frac{(n-1)s^2}{\sigma_0^2}.$$

Decision rules:

- **Right-Tailed Test:** $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 > \sigma_0^2$

$$\text{Reject } H_0 \text{ if } w \geq \chi^2_{n-1,\alpha}.$$

- **Left-Tailed Test:** $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 < \sigma_0^2$

$$\text{Reject } H_0 \text{ if } w \leq \chi^2_{n-1,1-\alpha}.$$

- **Two-Tailed Test:** $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$

$$\text{Reject } H_0 \text{ if } w \leq \chi^2_{n-1,1-\alpha/2} \text{ or } w \geq \chi^2_{n-1,\alpha/2}.$$

7.13.4 Asymptotic Z Test for the mean μ (unknown variance, large sample)

Let $X_1, \ldots, X_n$ be an iid sample from a population with unknown mean μ, unknown variance σ^2, and large sample size n. By the Central Limit Theorem (CLT), the test statistic is:

$$Z = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim N(0,1) \quad \text{approximately under } H_0 : \mu = \mu_0.$$

Decision rules:

- **Right-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

$$\text{Reject } H_0 \text{ if } z \geq z_\alpha.$$

- **Left-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

$$\text{Reject } H_0 \text{ if } z \leq -z_\alpha.$$

- **Two-Tailed Test:** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

$$\text{Reject } H_0 \text{ if } |z| \geq z_{\alpha/2}.$$

7.13.5 Asymptotic Z Test for the proportion p (large samples)

Let $X \sim \text{Ber}(p)$ and $X_1, \dots, X_n$ be an iid sample. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. The test statistic is:

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1) \quad \text{approximately under } H_0 : p = p_0.$$

Decision rules:

- **Right-Tailed Test:** $H_0 : p = p_0$ vs $H_1 : p > p_0$

$$\text{Reject } H_0 \text{ if } z \geq z_\alpha.$$

- **Left-Tailed Test:** $H_0 : p = p_0$ vs $H_1 : p < p_0$

$$\text{Reject } H_0 \text{ if } z \leq -z_\alpha.$$

- **Two-Tailed Test:** $H_0 : p = p_0$ vs $H_1 : p \neq p_0$

$$\text{Reject } H_0 \text{ if } |z| \geq z_{\alpha/2}.$$

7.14 Confidence interval for the difference between two means

To estimate the difference between the means of two independent populations, consider:

- $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, where:

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2 \quad \text{(homoscedasticity: equal variances).}$$

- Independent samples:

$$\begin{aligned} (X_1, \dots, X_n) &\quad \text{from population } X, \\ (Y_1, \dots, Y_m) &\quad \text{from population } Y. \end{aligned}$$

7.14.1 Pooled sample variance

The pooled sample variance S_p^2 combines the variances from the two samples:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2},$$

where:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y}_m)^2.$$

7.14.2 Pivotal quantity

The following pivotal quantity has a t-Student distribution with $n + m - 2$ degrees of freedom:

$$T = \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

7.14.3 Confidence interval

The confidence interval for $\mu_X - \mu_Y$ at level $1-\alpha$ is:

$$\left[\bar{X}_n - \bar{Y}_m - t_{n+m-2,\alpha/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X}_n - \bar{Y}_m + t_{n+m-2,\alpha/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right],$$

where $t_{n+m-2,\alpha/2}$ is the critical value of the t-distribution with $n+m-2$ degrees of freedom at the level $\alpha/2$.

Summary

- Use the pooled sample variance S_p^2 to estimate the common variance.
- Construct the confidence interval using the t-Student distribution with $n+m-2$ degrees of freedom.
- The width of the confidence interval depends on the sample sizes (n,m), the pooled variance S_p^2, and the confidence level $1-\alpha$.

7.15 Test for $\mu_X - \mu_Y$: Gaussian populations with equal variances

We test the null hypothesis:

$$H_0 : \mu_X - \mu_Y = 0,$$

against an appropriate alternative, where:

- $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$,
- $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (homoscedasticity: equal variances),
- $X \perp Y$ (independent populations),
- $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ are independent samples from X and Y, respectively.

Estimator for the common variance
The common variance σ^2 is estimated using the pooled sample variance:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2},$$

where:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2.$$

Sampling distributions

- The sample means $\bar{X}_n$ and $\bar{Y}_m$ follow:

$$\bar{X}_n \sim N\left(\mu_X, \frac{\sigma^2}{n}\right), \quad \bar{Y}_m \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right).$$

- Under $H_0 : \mu_X = \mu_Y$, the test statistic:

$$T = \frac{\bar{X}_n - \bar{Y}_m}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

follows a t-Student distribution with $n+m-2$ degrees of freedom.

Decision rules for a test of level α

- **Right-Tailed Test:** $H_1 : \mu_X - \mu_Y > 0$

$$\text{Reject } H_0 \text{ if } t \geq t_{n+m-2,\alpha}.$$

- **Left-Tailed Test:** $H_1 : \mu_X - \mu_Y < 0$

$$\text{Reject } H_0 \text{ if } t \leq -t_{n+m-2,\alpha}.$$

- **Two-Tailed Test:** $H_1 : \mu_X - \mu_Y \neq 0$

$$\text{Reject } H_0 \text{ if } |t| \geq t_{n+m-2,\alpha/2}.$$

Observed value of the test statistic The observed value of the test statistic is:

$$t = \frac{x - y}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where x and y are the sample means of X and Y, respectively.

Remarks

- The test is valid only under the assumption of equal variances. If this assumption does not hold, Welch's t-test should be used.
- For a two-tailed test, H_0 is rejected if $0 \notin$ the confidence interval for $\mu_X - \mu_Y$, which is:

$$\left[\bar{X}_n - \bar{Y}_m - t_{n+m-2,\alpha/2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X}_n - \bar{Y}_m + t_{n+m-2,\alpha/2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right].$$

7.16 Test for $\mu_X - \mu_Y$: Gaussian populations, unknown but equal variances

We aim to test:

$$H_0 : \mu_X - \mu_Y = 0,$$

against an appropriate alternative, where:

- $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$,
- $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (homoscedasticity),
- $X \perp Y$ (independent populations),
- σ^2 is unknown.

Pooled variance:
The common variance σ^2 is estimated using the pooled sample variance:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2},$$

where:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2.$$

Test statistic:

$$T = \frac{\bar{X}_n - \bar{Y}_m}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where:

- $\bar{X}_n$: sample mean of X,
- $\bar{Y}_m$: sample mean of Y,
- S_p^2: pooled variance.

Under H_0, T follows a t-distribution with $n + m - 2$ degrees of freedom:

$$T \sim t_{n+m-2}.$$

Decision rules for a test of level α

- **Right-Tailed Test:** $H_1 : \mu_X - \mu_Y > 0$

$$\text{Reject } H_0 \text{ if } t \geq t_{n+m-2,\alpha}.$$

- **Left-Tailed Test:** $H_1 : \mu_X - \mu_Y < 0$

$$\text{Reject } H_0 \text{ if } t \leq -t_{n+m-2,\alpha}.$$

- **Two-Tailed Test:** $H_1 : \mu_X - \mu_Y \neq 0$

$$\text{Reject } H_0 \text{ if } |t| \geq t_{n+m-2,\alpha/2}.$$

Confidence interval relationship
For a two-tailed test, H_0 is rejected if $0 \notin \text{CI}$, where the confidence interval for $\mu_X - \mu_Y$ is:

$$\left[\bar{X}_n - \bar{Y}_m - t_{n+m-2,\alpha/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X}_n - \bar{Y}_m + t_{n+m-2,\alpha/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right].$$

Remarks

- **Two-Tailed Test:** The decision to reject H_0 is directly linked to the confidence interval for $\mu_X - \mu_Y$.
- **Alternative for Unequal Variances:** If the assumption of equal variances is questionable, Welch's t-test should be used instead.

7.17 Power of the (two-tailed) T test

Consider the hypothesis test:

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0,$$

where $X_1, \ldots, X_n$ is an iid sample from $X \sim N(\mu, \sigma^2)$ with unknown variance σ^2.

The test statistic is:

$$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}},$$

where S is the sample standard deviation.

Under H_0, $T \sim t_{n-1}$.

The power of the test measures the probability of correctly rejecting H_0 when the true population mean is μ_1 ($\mu_1 \neq \mu_0$):

$$\text{Power} = P\left(|T| \geq t_{n-1,\alpha/2} \mid \mu = \mu_1\right).$$

To evaluate this probability:

- Under H_1, the test statistic can be expressed as:

  $$T = \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} = \frac{Z + (\mu_1 - \mu_0)\sqrt{n}/\sigma}{\sqrt{V/(n-1)}},$$

  where:

  - $Z \sim N(0,1)$ (standard normal),
  - $V \sim \chi^2_{n-1}$ (chi-squared with $n-1$ degrees of freedom),
  - $Z \perp V$ (independent).

- The distribution of T under H_1 is non-central t with $n-1$ degrees of freedom and non-centrality parameter:

  $$\delta = \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}.$$

Power expression

The power of the test is:

$$\text{Power} = P\left(|T| \geq t_{n-1,\alpha/2} \mid \mu = \mu_1\right).$$

This can be expressed as:

$$\text{Power} = 1 - P\left(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2} \mid \mu = \mu_1\right),$$

or equivalently:

$$\text{Power} = P\left(T \leq -t_{n-1,\alpha/2} \mid \mu = \mu_1\right) + P\left(T \geq t_{n-1,\alpha/2} \mid \mu = \mu_1\right).$$

Key observations

- When $\mu_1 = \mu_0$, T follows the t-Student distribution with $n-1$ degrees of freedom, and the power is equal to the significance level α.
- When $\mu_1 \neq \mu_0$, the power increases as:
  - The sample size n increases,
  - The effect size $|\mu_1 - \mu_0|$ increases,
  - The variance σ^2 decreases.

7.18 Non central t-Stundent distribution

Let:

- $Z \sim N(0,1)$ (standard normal random variable),
- $V \sim \chi^2_r$ (chi-squared random variable with r degrees of freedom),
- $Z \perp V$ (independent random variables),
- $\delta \neq 0$ (non-centrality parameter).

The random variable:

$$T_{r,\delta} = \frac{Z+\delta}{\sqrt{V/r}},$$

has a **non-central t-Student distribution** with r degrees of freedom and non-centrality parameter δ.

Special case

If $\delta = 0$, the distribution reduces to the standard (central) t-Student distribution with r degrees of freedom:

$$T_{r,0} \sim t_r.$$

Properties

- The random variable $T_{r,\delta}$ has an **absolutely continuous distribution**.
- The shape of the distribution depends on both the degrees of freedom r and the non-centrality parameter δ.
- For large values of r, the distribution approximates a normal distribution with mean δ and variance 1.

Cumulative distribution function (CDF) and quantiles in R

The cumulative distribution function (CDF) and quantiles of the non-central t-Student distribution can be computed in R using the functions:

- `pt`: for the cumulative probability $P(T_{r,\delta} \leq t)$,
- `qt`: for the quantiles.

To specify the non-centrality parameter δ, use the argument `ncp`.

Example:

- To compute $P(T_{3,2} \leq 4)$:

  ```
  pt(4, df=3, ncp=2)
  ```

  yields $P(T_{3,2} \leq 4)$.

- To compute the 95th percentile of $T_{5,1.5}$:

  ```
  qt(0.95, df=5, ncp=1.5).
  ```

7.19 Asymptotic confidence interval for the difference between two proportions

Let:

- $X \sim \text{Ber}(p_X)$ and $Y \sim \text{Ber}(p_Y)$ be two independent Bernoulli random variables,
- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent iid samples from X and Y, respectively.

The sample proportions are:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \bar{Y}_m = \frac{1}{m} \sum_{i=1}^{m} Y_i.$$

We aim to construct an asymptotic confidence interval for the difference in proportions:

$$p_X - p_Y.$$

Variance of the difference in proportions
The variance of $\bar{X}_n - \bar{Y}_m$ is:

$$\text{Var}(\bar{X}_n - \bar{Y}_m) = \frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}.$$

Since p_X and p_Y are unknown, they are estimated using the sample proportions $\bar{X}_n$ and $\bar{Y}_m$:

$$\text{Var}(\bar{X}_n - \bar{Y}_m) \approx \frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}.$$

Test statistic
The following test statistic approximately follows a standard normal distribution ($N(0,1)$) for large n and m:

$$Z = \frac{\bar{X}_n - \bar{Y}_m - (p_X - p_Y)}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}}}.$$

Confidence interval
The bounds of the asymptotic confidence interval for $p_X - p_Y$ at level $1 - \alpha$ are:

$$\left[\bar{X}_n - \bar{Y}_m - z_{\alpha/2} \cdot \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}}, \bar{X}_n - \bar{Y}_m + z_{\alpha/2} \cdot \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n} + \frac{\bar{Y}_m(1-\bar{Y}_m)}{m}} \right],$$

where $z_{\alpha/2}$ is the critical value of the standard normal distribution for a significance level α.

Summary

- The asymptotic confidence interval assumes large sample sizes (n, m).
- The interval is centered on the observed difference in sample proportions $\bar{X}_n - \bar{Y}_m$.
- The width of the confidence interval depends on the variability of the sample proportions and the sample sizes.

7.20 Asymptotic test for comparing two proportions

Let:

- $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be two independent iid samples,
- $X \sim \text{Ber}(p_X)$, $Y \sim \text{Ber}(p_Y)$,

and we aim to test:

$$H_0 : p_X = p_Y \quad \text{vs} \quad H_1 : p_X \neq p_Y \quad \text{(or other directional alternatives).}$$

Test statistic
Under $H_0 : p_X = p_Y = \hat{p}$, the pooled proportion is:

$$\hat{p} = \frac{n\bar{X}_n + m\bar{Y}_m}{n+m},$$

where:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

The test statistic is:

$$Z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

which approximately follows a standard normal distribution ($N(0,1)$) for large n and m.

Decision rules
Let z be the observed value of Z:

- **Two-Tailed Test:** $H_1 : p_X \neq p_Y$

$$\text{Reject } H_0 \text{ if } |z| \geq z_{\alpha/2}.$$

- **Right-Tailed Test:** $H_1 : p_X > p_Y$

$$\text{Reject } H_0 \text{ if } z \geq z_{\alpha}.$$

- **Left-Tailed Test:** $H_1 : p_X < p_Y$

$$\text{Reject } H_0 \text{ if } z \leq -z_{\alpha}.$$

Remarks

- The test assumes large sample sizes (n, m) for the normal approximation to hold.
- The asymptotic confidence interval for $p_X - p_Y$ is linked to the test: H_0 is rejected if $0 \notin$ CI.

7.21 R functions for CIs and tests

- `t.test`: Used for:
  - Confidence intervals and tests for a single population mean.
  - Confidence intervals and tests for the difference of two means (default: Welch's test for unequal variances).
- `prop.test`: Used for:
  - Asymptotic tests and confidence intervals for a single proportion.
  - Asymptotic tests and confidence intervals for the difference of two proportions.
- `binom.test`: Used for:
  - Exact tests and confidence intervals for a single proportion.