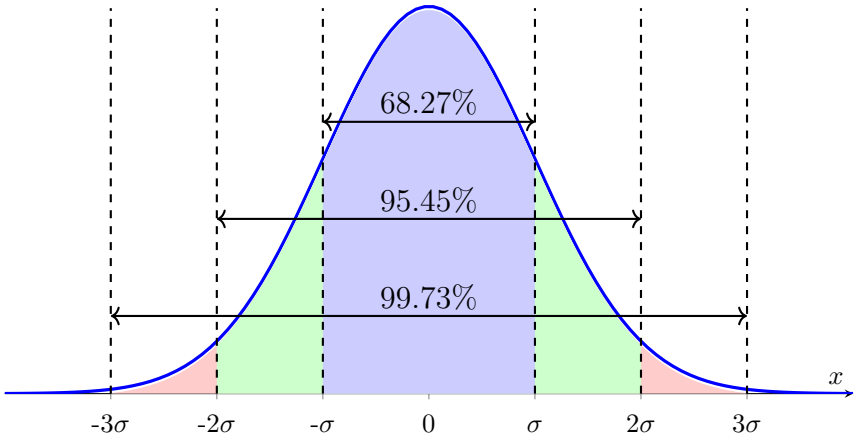
Basics of Probability and Statistics

An In-Depth Exploration of Core Concepts and Methods

Sandip Karar

June 14, 2025



Contents

1 Descriptive Statistics **7**

1.1 Introduction 7

1.2 Measures of Central Tendency 10

1.3 Partition Values: Quartiles, Deciles, and Percentiles 23

1.4 Measures of Dispersion 24

1.5 Moments, Skewness and Kurtosis 29

2 Theory of Probability **35**

2.1 Some Notation and Terminology 35

2.2 Definition of Probability 37

2.3 Axioms of Probability 38

2.4 Conditional Probability 43

2.5 Rule of Total Probability 44

2.6 Bayes' Theorem 46

2.7 Statistical Independence of Events 48

3 Random Variables and Probability Distributions **53**

3.1 What is a Random Variable? 53

3.2 Probability Distribution 54

3.3 Mean and Variance of a Random Variable 57

3.4 Joint Distribution of Two random Variables 60

3.5 Conditional Probability Distribution 66

3.6 Functions of a Random Variable 68

3.7 Standardized Random Variable 71

3.8 Chebyshev's Inequality 71

3.9 Moments and Moment Generating Function 73

4 Common Distributions **77**

4.1 Bernoulli Distribution 77

4.2 Binomial Distribution 78

4.3 Poisson Distribution 81
4.4 Uniform Distribution 85
4.5 Normal Distribution 87

5 Sampling Theory **99**
5.1 Introduction 99
5.2 Sampling Methods 99
5.3 Sample Mean, Sample Variance and Sample Proportion 101
5.4 Sampling Distributions 102
5.5 The Sampling Distribution of the Sample Mean 103
5.6 The Sampling Distribution of the Sample Variance 108
5.7 Distribution of the Ratio of Two Sample Variances 110
5.8 The Sampling Distribution of the Sample Proportion 112

6 Theory of Estimation **113**
6.1 Introduction 113
6.2 Point Estimation 113
6.3 Maximum Likelihood Estimation (MLE) 119
6.4 Bayesian Estimation 122
6.5 Estimation using Method of Moments 126
6.6 Interval Estimation 128
6.7 Determination of Sample Size 136

7 Test of Hypothesis **141**
7.1 What is Hypothesis Testing? 141
7.2 Test Statistic 144
7.3 Type I and Type II Error 146
7.4 General Procedure for Hypothesis Testing 149
7.5 Statistical Test for a Normally Distributed Population Mean μ 150
7.6 The 5 Steps of Hypothesis Testing 152
7.7 Power of a Test 153
7.8 Most Powerful Test, Neyman–Pearson Lemma, and Likelihood Ratio Test 155

8 Correlation and Regression **157**
8.1 Correlation 157
8.2 Pearson Correlation Coefficient 158
8.3 Effect of Linear Transformations on Correlation Coefficient 160
8.4 Correlation Is Not Causation 161
8.5 Regression Analysis 162

8.6 The Simple Linear Regression Model 162
8.7 Estimating Parameters Using Least Squares 163

Chapter 1

Descriptive Statistics

1.1 Introduction

https://bookdown.org/egarpor/inference/estmeth-mm.html

Statistics is a branch of mathematics that deals with the collection, organization, analysis, interpretation, and presentation of data. It provides tools for making informed decisions in the presence of uncertainty.

1.1.1 Uses of Statistics

Statistics is widely used in various fields such as:

- **Education**: Analyzing student performance and improving teaching methods.
- **Business**: Making informed decisions based on market trends and consumer behavior.
- **Healthcare**: Understanding the effectiveness of treatments and tracking disease outbreaks.
- **Government**: Planning and policy-making based on population data.

1.1.2 Types of Data

Data can be categorized based on their nature and measurement levels.

1. Qualitative (Categorical) Data

These are non-numeric data that describe categories or groups.

- **Nominal**: Categories without any inherent order (e.g., colors, gender).
- **Ordinal**: Categories with a meaningful order but no fixed interval between them (e.g., rankings).

2. Quantitative (Numerical) Data

These are numeric data representing counts or measurements.

- **Discrete**: Countable values (e.g., number of students).
- **Continuous**: Measurable quantities that can take any value within a range (e.g., height, weight).

1.1.3 Data Collection Methods

Collecting accurate data is crucial for meaningful analysis. Common methods include:

- **Surveys**: Gathering information through questionnaires.
- **Experiments**: Conducting controlled studies to observe outcomes.
- **Observations**: Recording data based on direct observation.
- **Existing Records**: Utilizing previously collected data.

1.1.4 Organizing Data

Once data is collected, organizing it helps in understanding patterns and trends.

1. Frequency Distribution Table

A frequency distribution table lists data values and their corresponding frequencies.

Number of Books	Number of Students
0–2	5
3–5	12
6–8	17
9–11	8
12–14	3

Table 1.1: Number of Books Read by Students

2. Relative Frequency

Relative frequency represents the proportion of observations within each category.

Number of Books	Relative Frequency
0–2	0.10
3–5	0.24
6–8	0.34
9–11	0.16
12–14	0.06

Table 1.2: Relative Frequency of Books Read

1.1.5 Displaying Data

Visual representations make it easier to comprehend data.

1. Bar Graph

Bar graphs are used for categorical data.

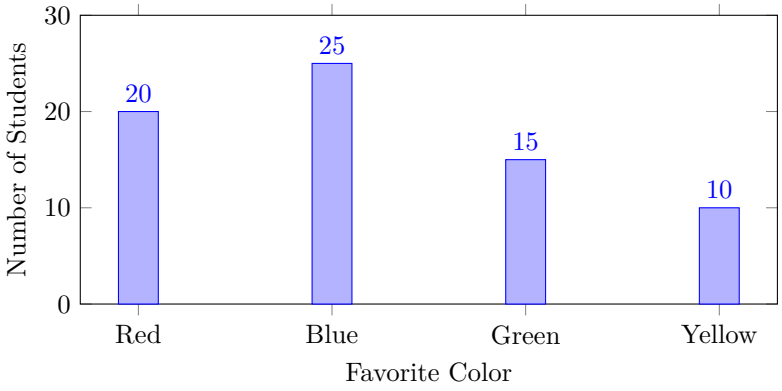


Figure 1.1: Favorite Colors of Students

2. Pie Chart

Pie charts show the proportion of categories within a whole.

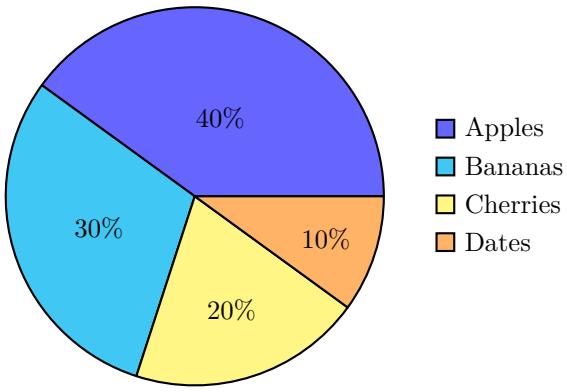


Figure 1.2: Fruit Preferences

3. Histogram

Histograms are used for continuous numerical data.

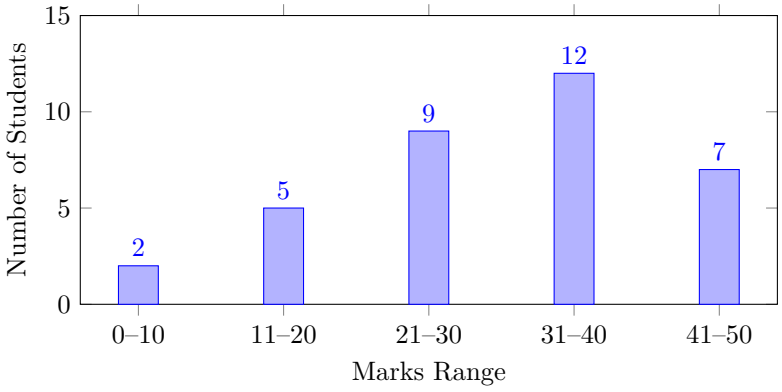


Figure 1.3: Distribution of Marks

4. Line Graph

Line graphs depict data trends over time.

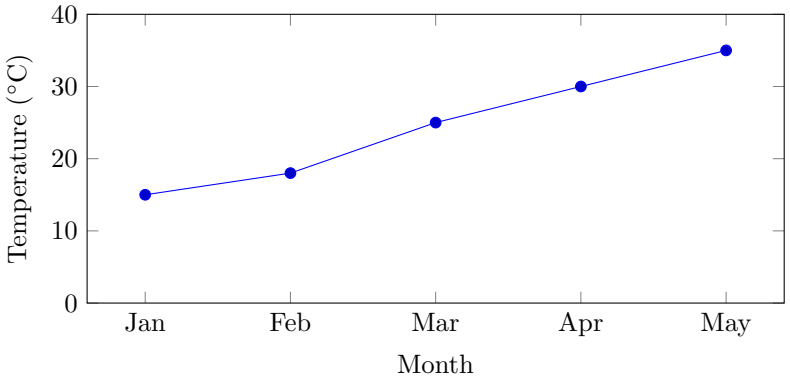
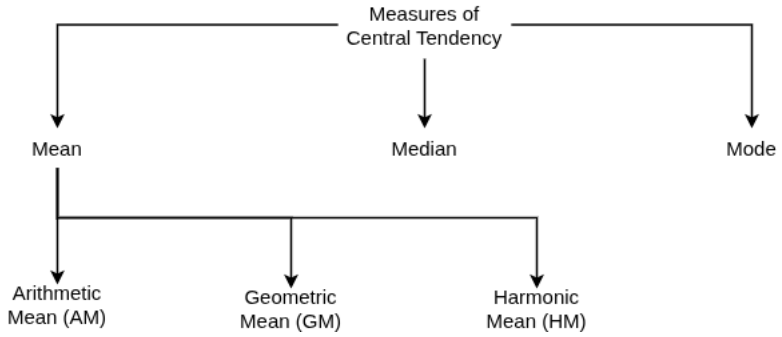


Figure 1.4: Monthly Average Temperature

1.2 Measures of Central Tendency

Quite often, data exhibit a tendency to cluster around a central value. Measures of central tendency are numerical indicators that describe this central value of a data set. The most common measures include the mean, median, and mode. Each measure tells us something different about our data, and knowing when to use each one can really help us make sense of the numbers.



1.2.1 Mean

The mean (or average) is the most commonly used measure of central tendency and is defined in several forms:

- **Arithmetic Mean (AM):**
  - Simple AM: For a dataset $x_1, x_2, \ldots, x_n$, the arithmetic mean is given by:

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

  - Weighted AM: When each value x_i has an associated frequency f_i, the weighted arithmetic mean is:

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

Theorem: If $x_i = a$ (constant) for all i, then the arithmetic mean is also a, that is,

$$\overline{x} = a.$$

Proof:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n} \cdot na = a$$

■

Theorem: If $y_i = a + x_i$, then the mean of y is given by:

$$\overline{y} = a + \overline{x}.$$

Proof:

$$\begin{aligned}\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + x_i) = \frac{1}{n}\left(\sum_{i=1}^{n} a + \sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}\left(na + \sum_{i=1}^{n} x_i\right) = a + \overline{x}.\end{aligned}$$

■

Theorem: Let a dataset be composed of two distinct groups of observations:

- Group 1 consists of n_1 observations with arithmetic mean $\overline{x}_1$,
- Group 2 consists of n_2 observations with arithmetic mean $\overline{x}_2$.

Then, the arithmetic mean $\overline{x}$ of the combined dataset (of size $n_1 + n_2$) is given by:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}.$$

Proof: Total sum of group 1 is $n_1\overline{x}_1$ and total sum of group 2 is $n_2\overline{x}_2$.

Then total sum $= n_1\overline{x}_1 + n_2\overline{x}_2$

Total number of observations $= n_1 + n_2$

Therefore, combined AM $= \overline{x} = \dfrac{\text{Total sum}}{\text{Total number of observations}} \dfrac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}$

■

- **Geometric Mean (GM):** Geometric mean of a set of n observation is the nth root of their product. It is only defined for positive values.
  - Simple GM:

$$x_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 x_2 \dots x_n}$$

– Weighted GM:

$$x_G = \left(\prod_{i=1}^{n} x_i^{f_i}\right)^{1/N} = \sqrt[N]{x_1^{f_1} x_2^{f_2} \dots x_n^{f_n}}$$

Where

$$N = \sum_{i=1}^{n} f_i$$

Theorem: The GM of a set of positive values $x_1, x_2, \dots, x_n$ is equal to the antilogarithm (exponential) of the AM of their logarithms:

$$\text{GM} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right)$$

Proof:

$$\begin{aligned}\text{GM} &= \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \exp\left(\log\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right) \\ &= \exp\left(\frac{1}{n}\log\left(\prod_{i=1}^{n} x_i\right)\right) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right)\end{aligned}$$

■

Theorem: Suppose we have two groups:

– Group 1 has N_1 positive values with geometric mean x_{G_1},

– Group 2 has N_2 positive values with geometric mean x_{G_2}.

Then the combined geometric mean GM of all $N_1 + N_2$ values is:

$$\text{GM} = \left(x_{G_1}^{N_1} \cdot x_{G_2}^{N_2}\right)^{1/(N_1+N_2)}$$

Proof: Let the product of values in group 1 be $P_1 = \prod_{i=1}^{N_1} x_i$, so that:

$$x_{G_1} = (P_1)^{1/N_1} \Rightarrow P_1 = x_{G_1}^{N_1}$$

Similarly, for group 2:

$$P_2 = \prod_{j=1}^{N_2} y_j = x_{G_2}^{N_2}$$

Then the overall product:

$$P = P_1 \cdot P_2 = x_{G_1}^{N_1} \cdot x_{G_2}^{N_2}$$

The combined GM is:

$$\text{GM} = (P)^{1/(N_1+N_2)} = \left(x_{G_1}^{N_1} \cdot x_{G_2}^{N_2}\right)^{1/(N_1+N_2)}$$

■

- **Harmonic Mean (HM):** Harmonic Mean of a set of observations is the reciprocal of the arithmetic mean of the reciprocals.
  - Simple HM:

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

  - Weighted HM:

$$x_H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{x_i}} = \frac{N}{\sum_{i=1}^{n}\frac{f_i}{x_i}}$$

It is important to note where to use GM and where to use HM. GM is useful for averaging **ratios**, **rates** and **percentages**. As an illustration, we consider the following example.

Example: The ratio of the prices in 1994 and to those in 1982 for four commodities are $0.92, 1.25, 1.75$ and 0.85. To get the average price ratio use geometric mean

$$\begin{aligned}\log x_G &= \frac{1}{n}\left(\log 0.92 + \log 1.25 + \log 1.75 + \log 0.85\right)\\ &= 0.5829 = \log 1.1436\\ \Rightarrow x_G &= 1.1436\end{aligned}$$

GM is also useful if one wants to determine the values of a variable at the midpoint of a time interval when the variable changes over time exponentially. Thus if the value at two points 0 and t be a and ar^t, then its value at the midpoint $\frac{t}{2}$ is $(a \times ar^t)^{1/2} = ar^{t/2}$.

Now consider the following example:

Example: A person goes from X to Y on cycle at 20 km/h and returns at 24 km/h. What is the average speed for the entire trip?

If we use AM, then the average speed is

$$\frac{1}{2}(20 + 24) = 22 \text{ km/h}$$

But is this correct?

Consider the total distance between X and Y is 1 km for the sake of simplicity. So the total distance covered = 2 km. The time taken for the person to go from X to Y is $\frac{1}{20} = 0.05$ hr and the time taken to return is $\frac{1}{24} = 0.04166$ hr.

Therefore average speed $= \dfrac{\text{Total distance}}{\text{Total time}} = \dfrac{2}{0.05 + 0.04166} = 21.8$ km/h.

Clearly, the AM value of 22 km/h overestimates the actual average speed. Now consider the harmonic mean (HM) of the two speeds:

$$\frac{2}{\frac{1}{20} + \frac{1}{24}} = 21.8 \text{ km/h}$$

This matches the correct value.

Now where to use the HM? The harmonic mean is particularly useful when dealing with quantities expressed in the form "x per unit y", such as "km per hour", "rupees per kg", and similar rates. Here x and y are unit of measures, not numeric variables.

Rule of Thumb:

- Use the **harmonic mean (HM)** when equal quantities of x are involved.
- Use the **arithmetic mean (AM)** when equal quantities of y are involved.

This principle can be illustrated with the following example.

Example: Suppose a train covers n **equal distances**, each of s kilometers, with speeds $v_1, v_2, \ldots, v_n$ km/h. The average speed is the total distance divided by the total time taken. Thus,

$$\text{Average speed} = \frac{ns}{\frac{s}{v_1} + \frac{s}{v_2} + \cdots + \frac{s}{v_n}} = \frac{n}{\frac{1}{v_1} + \frac{1}{v_2} + \cdots + \frac{1}{v_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{v_i}}$$

This is the **harmonic mean (HM)** of the given speeds.

On the other hand, if the train travels for n **equal time intervals**, each of duration t hours, at speeds $v_1, v_2, \ldots, v_n$ km/h, then the total distance covered is:

$$\text{Total distance} = v_1 t + v_2 t + \cdots + v_n t = t(v_1 + v_2 + \cdots + v_n)$$

and the total time is nt. So, the average speed is:

$$\text{Average speed} = \frac{t(v_1 + v_2 + \cdots + v_n)}{nt} = \frac{v_1 + v_2 + \cdots + v_n}{n} = \frac{1}{n} \sum_{i=1}^{n} v_i$$

which is the **arithmetic mean (AM)** of the given speeds.

Theorem: The sum of squared deviations from a constant A is minimized when A equals the arithmetic mean $\overline{x}$, i.e.,

$$\sum_{i=1}^{n} (x_i - A)^2 \text{ is minimized when } A = \overline{x}$$

Proof: Let $S(A) = \sum_{i=1}^{n} (x_i - A)^2$. Expand this:

$$S(A) = \sum_{i=1}^{n} (x_i^2 - 2Ax_i + A^2) = \sum x_i^2 - 2A \sum x_i + nA^2$$

To minimize $S(A)$, take derivative with respect to A and set it to zero:

$$\frac{dS}{dA} = -2 \sum x_i + 2nA = 0 \quad \Rightarrow \quad A = \frac{1}{n} \sum x_i = \overline{x}$$

Now, take the second derivative:

$$\frac{d^2 S}{dA^2} = 2n > 0$$

Since the second derivative is positive, the function $S(A)$ has a minimum at $A = \overline{x}$.

■

Theorem: For two observations,

$$\text{GM}^2 = \text{AM} \times \text{HM}$$

Proof: Let a and b be two observations (positive numbers).

Compute left-hand side:

$$\text{GM}^2 = (\sqrt{ab})^2 = ab$$

Compute right-hand side:

$$\text{AM} \times \text{HM} = \left(\frac{a+b}{2}\right)\left(\frac{2ab}{a+b}\right) = ab$$

Hence,

$$\text{GM}^2 = \text{AM} \times \text{HM}$$

■

Theorem: For any set of n positive real numbers $x_1, x_2, \ldots, x_n$, the following inequality holds:

$$\text{AM} \geq \text{GM} \geq \text{HM}$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

Proof: Let $x_1, x_2, \ldots, x_n$ be positive real numbers.

- **Step 1: Proving AM $\geq$ GM**

  We can prove this using the method of induction.

  - **Base Case**: Let us first consider two observations $x_1 = a > 0$, $x_2 = b > 0$. We have to prove:

    $$\frac{a+b}{2} \geq \sqrt{ab}$$

    Consider the square of the difference:

    $$\left(\frac{a-b}{2}\right)^2 \geq 0 \quad \Rightarrow \quad \frac{a^2 - 2ab + b^2}{4} \geq 0$$

    $$\Rightarrow a^2 + b^2 \geq 2ab \quad \Rightarrow \quad (a+b)^2 \geq 4ab$$

    $$\Rightarrow \left(\frac{a+b}{2}\right)^2 \geq ab \quad \Rightarrow \quad \frac{a+b}{2} \geq \sqrt{ab}$$

    Equality holds if and only if $a = b$.

  - **Inductive Step**: Assume the inequality holds for $n = k$, i.e., for all positive $x_1, \ldots, x_k$:

    $$\frac{x_1 + x_2 + \cdots + x_k}{k} \geq (x_1 x_2 \ldots x_k)^{1/k}$$

We must show it holds for $n = k + 1$ too.
Let $x_1, x_2, \ldots, x_k, x_{k+1}$ be positive numbers. Define:

$$A = \frac{x_1 + x_2 + \cdots + x_k}{k}, \quad G = (x_1 x_2 \ldots x_k)^{1/k}$$

By the inductive hypothesis, $A \geq G$.
Now apply the $n = 2$ case to the numbers A and x_{k+1}:

$$\frac{A + x_{k+1}}{2} \geq \sqrt{A x_{k+1}} \geq \sqrt{G x_{k+1}}$$

Now note:

$$\frac{x_1 + \cdots + x_k + x_{k+1}}{k+1} = \frac{kA + x_{k+1}}{k+1}$$

We now want to show:

$$\frac{kA + x_{k+1}}{k+1} \geq (x_1 x_2 \ldots x_k x_{k+1})^{1/(k+1)}$$

Let us define:

$$B = (x_1 x_2 \ldots x_k x_{k+1})^{1/(k+1)} = \left(G^k \cdot x_{k+1}\right)^{1/(k+1)}$$

Use the inequality between arithmetic and geometric mean on A and x_{k+1}:

$$\frac{kA + x_{k+1}}{k+1} \geq \left(A^k \cdot x_{k+1}\right)^{1/(k+1)}$$

Since $A \geq G$, and exponentiation preserves the inequality for positive values:

$$A^k \geq G^k \quad \Rightarrow \quad \left(A^k \cdot x_{k+1}\right)^{1/(k+1)} \geq \left(G^k \cdot x_{k+1}\right)^{1/(k+1)} = B$$

Therefore,

$$\frac{x_1 + \cdots + x_{k+1}}{k+1} \geq B = (x_1 x_2 \ldots x_k x_{k+1})^{1/(k+1)}$$

This therefore proves that:

$$\text{AM} \geq \text{GM}$$

- **Step 2: Proving GM $\geq$ HM**
  Recall the definition of the harmonic mean:

$$\text{HM} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Now consider the reciprocals $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}$, which are also positive. Thus we can apply the AM–GM inequality to the reciprocals:

$$\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right) \geq \left(\frac{1}{x_1 x_2 \ldots x_n}\right)^{1/n}$$

Taking reciprocals of both sides:

$$\frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \leq (x_1 x_2 \ldots x_n)^{1/n}$$

$$\Rightarrow \text{HM} \leq \text{GM}$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

Combining both steps:

$$\text{AM} \geq \text{GM} \geq \text{HM}$$

with equality throughout if and only if all the x_i are equal. ■

1.2.2 Median

> The **median** of a set of observation is the middlemost value when the observations are arranged in increasing or decreasing order of magnitude.

It is denoted by M_i or $\tilde{x}$. It divides the dataset into two equal halves: 50% of the values lie below the median and 50% lie above.

1. **Median in a Simple Series (Ungrouped Data)**

   For a dataset with n observations arranged in ascending order:

   - If n is odd, the median is the value at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position.
   - If n is even, the median is the arithmetic mean of the values at the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ positions.

   Example: Find the median of the dataset:

   $$7, 2, 5, 9, 4$$

   Arranging in ascending order: $2, 4, 5, 7, 9$. Since there are 5 values (odd), the median is the 3rd value:

   $$\text{Median} = 5$$

2. **Median in a Simple Frequency Distribution**

   In a simple frequency distribution, each data value is associated with a frequency. The procedure is identical to that of a simple frequency distribution:

   - Arrange the data in ascending order.
   - Compute cumulative frequencies based on weights.
   - Find total frequency N, then find the smallest value for which the cumulative frequency is greater than or equal to $\frac{N}{2}$.

   Example: Consider the following table containing the values, frequencies and cumulative frequencies.

Value	Frequency	Cumulative Frequency
2	3	3
4	5	8
6	7	15
8	5	20

Table 1.3: *Simple frequency distribution table to calculate median.*

$$N = 3 + 5 + 7 + 5 = 20 \quad \Rightarrow \frac{N}{2} = 10$$

Since 10 is between 8 and 15, the Median is 6. This works regardless of whether N is odd or even.

3. **Median in a Grouped Frequency Distribution**

For a grouped frequency distribution, the cumulative frequencies are used to locate the median.

- Compute cumulative frequencies.
- Find $N = \sum f_i$, the total number of observations.
- Find $\dfrac{N}{2}$.
- Locate the median class (the class whose cumulative frequency is greater than or equal to $\dfrac{N}{2}$).
- Use the formula:

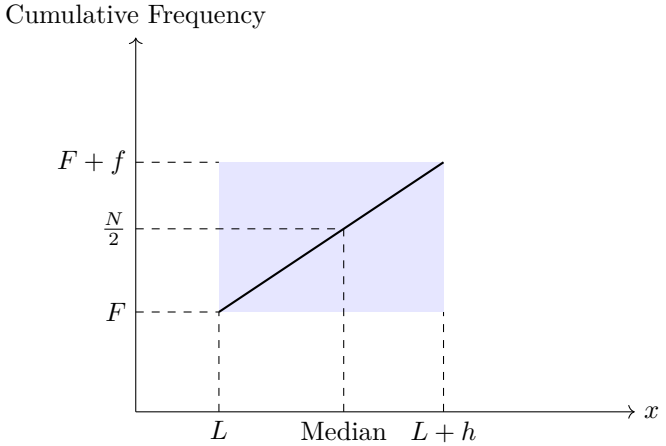
$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{f} \right) \cdot h$$

where:

- L: lower boundary of the median class
- N: total frequency
- F: cumulative frequency before the median class
- f: frequency of the median class
- h: width of the class interval

To arrive at this formula we assume that the cumulative frequency if a linear function of x within the class L and $L + h$. Then

$$\frac{\text{Median} - L}{h} = \frac{\frac{N}{2} - F}{f} \quad \Rightarrow \text{Median} = L + \left(\frac{\frac{N}{2} - F}{f} \right) \cdot h$$



Example: Consider the following table containing the class values, frequencies and cumulative frequencies.

Class Interval	Frequency	Cumulative Frequency
0–10	5	5
10–20	8	13
20–30	12	25
30–40	6	31

Table 1.4: *Grouped frequency distribution table for calculating the median.*

$$N = 5 + 8 + 12 + 6 = 31, \quad \frac{N}{2} = 15.5$$

Since 15.5 is in between 13 and 25, the Median class is 20–30. Thus,

$$L = 20, F = 13, f = 12, h = 10$$

$$\text{Median} = 20 + \left(\frac{15.5 - 13}{12}\right) \cdot 10 = 20 + \left(\frac{2.5}{12}\right) \cdot 10 = 20 + 2.08 = 22.08$$

Theorem: Let $x_1, x_2, \ldots, x_n$ be a set of observations. Define the function:

$$S(A) = \sum_{i=1}^{n} |x_i - A|$$

Then $S(A)$ is minimized when $A = \text{Median}$.

Proof: Let us arrange the observations $x_1, x_2, \ldots, x_n$ in increasing order and denote the ordered sequence by $y_1, y_2, \ldots, y_n$. Since this is just a rearrangement of the original data, we have:

$$\sum_{i=1}^{n} |x_i - A| = \sum_{i=1}^{n} |y_i - A|.$$

We now analyze the behavior of this sum in two cases:

- **Case 1: n is odd (say $n = 2m + 1$)**

$$\begin{aligned}\sum_{i=1}^{n} |x_i - A| &= \sum_{i=1}^{2m+1} |y_i - A| \\ &= |y_1 - A| + |y_2 - A| + \cdots + |y_m - A| + |y_{m+1} - A| \\ &\quad + |y_{m+2} - A| + \cdots + |y_{2m+1} - A|.\end{aligned}$$

  There are $2m + 1$ terms in the sum. We consider them in symmetric pairs from both ends:

  - The sum $|y_1 - A| + |y_{2m+1} - A|$ is minimized when $A \in [y_1, y_{2m+1}]$.
  - The sum $|y_2 - A| + |y_{2m} - A|$ is minimized when $A \in [y_2, y_{2m}]$.
  - Continuing this way, the sum $|y_m - A| + |y_{m+2} - A|$ is minimized when $A \in [y_m, y_{m+2}]$.

The unpaired central term is $|y_{m+1} - A|$, which attains its minimum value (zero) when $A = y_{m+1}$.

Since all these intervals of minimization overlap at y_{m+1}, we conclude that:

$$\sum_{i=1}^{n} |x_i - A| \text{ is minimized when } A = y_{m+1} = \text{Median}.$$

- **Case 2: n is even (say $n = 2m$)**

$$\begin{aligned}\sum_{i=1}^{n} |x_i - A| &= \sum_{i=1}^{2m} |y_i - A| \\ &= |y_1 - A| + |y_2 - A| + \cdots + |y_m - A| + |y_{m+1} - A| + \cdots + |y_{2m} - A|.\end{aligned}$$

Now there are $2m$ terms, which can be grouped into m symmetric pairs:

- $|y_1 - A| + |y_{2m} - A|$ is minimized when $A \in [y_1, y_{2m}]$,
- $|y_2 - A| + |y_{2m-1} - A|$ is minimized when $A \in [y_2, y_{2m-1}]$, and so on.
- The final pair $|y_m - A| + |y_{m+1} - A|$ is minimized when $A \in [y_m, y_{m+1}]$.

Thus, the entire sum is minimized when $A \in [y_m, y_{m+1}]$. A natural choice is:

$$A = \text{Median} = \frac{y_m + y_{m+1}}{2},$$

which lies within the minimizing interval and hence ensures that the sum is minimized.

In both cases—odd and even number of observations—the value of A that minimizes $\sum_{i=1}^{n} |x_i - A|$ is the **median** of the dataset.

■

The median is a better measure of central tendency than the mean (AM) in the presence of outliers in the observations.

The **mean** (AM) is sensitive to extreme values or outliers, while the **median** (the middle value) is more robust and resistant to such anomalies. This makes the median a better measure of central tendency in the presence of outliers or skewed data.

Consider the marks obtained by 5 students:

$$\text{Scores} = \{10, 70, 75, 80, 90\}$$

$$\textbf{Mean} = \frac{70 + 75 + 80 + 85 + 90}{5} = \frac{400}{5} = 80; \quad \textbf{Median} = \text{Middle value} = 75$$

Here, both the mean and the median are equal and representative of the data, as there are no extreme values.

Now suppose one student scored unusually low:

$$\text{Scores} = \{10, 70, 75, 80, 90\}$$

$$\textbf{Mean} = \frac{10 + 70 + 75 + 80 + 90}{5} = \frac{325}{5} = 65; \quad \textbf{Median} = \text{Middle value} = 75$$

The mean drops to 65 due to the outlier (10), even though most students scored 70 or above. The median stays at 75 and gives a better picture of the typical student performance.

1.2.3 Mode

The **mode** of a given set of observations is the value which occurs with maximum frequency. It represents the highest peak in the frequency distribution.

The mode is generally denoted by M_o.

1. **Mode in a Simple Series (Ungrouped Data)**

   To determine the mode:

   - Count the frequency of each data value.
   - The mode is the value with the highest frequency.

   Example: Find the mode of the dataset:

   $$3, 7, 2, 3, 9, 3, 5$$

   The number 3 occurs most frequently (3 times), so:

   $$\text{Mode} = 3$$

2. **Mode in a Simple Frequency Distribution**

   In a simple frequency distribution, the mode is the data value corresponding to the maximum frequency.

   Example: Consider the table:

Value	**Frequency**
4	3
5	7
6	5
7	2

   Table 1.5: *Simple frequency distribution table for calculating mode.*

   The highest frequency is 7, corresponding to the value 5. Hence:

   $$\text{Mode} = 5$$

3. **Mode in a Grouped Frequency Distribution**

   When the data is grouped into intervals, it is very difficult to find the mode accurately. However if all the classes are of equal width, then it is possible to approximately calculate the mode using the formula:

   $$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \cdot h$$

   where:

   - L: lower boundary of the modal class
   - f_1: frequency of the modal class
   - f_0: frequency of the class preceding the modal class

- f_2: frequency of the class succeeding the modal class
- h: class width

How do we arrive at the formula for mode? In addition to the modal class frequency f_1, mode also depends on f_0 (the frequency of the class preceding the modal class) and f_2 (the frequency of the class following the modal class). If they are equal, then one would take the midpoint of the modal class $L + \dfrac{h}{2}$ as the mode. However, if $f_0 - f_1$ is greater (smaller) than $f_1 - f_2$, then one would suppose that the mode is nearer to (further from) the lower boundary (L) of the modal class than the upper boundary ($L + h$). Mathematically, if we assume the proportion is same, then

$$\frac{d}{f_1 - f_0} = \frac{h - d}{f_1 - f_2}$$

Cross-multiplying and simplifying:

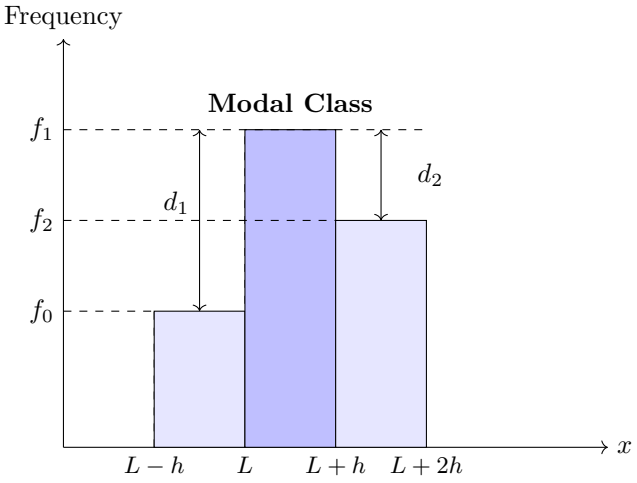
$$d \cdot (2f_1 - f_0 - f_2) = (f_1 - f_0) \cdot h$$

Solving for d:

$$d = \frac{(f_1 - f_0) \cdot h}{2f_1 - f_0 - f_2}$$

Hence, the mode is:

$$\text{Mode} = L + d = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \cdot h$$



Example: Consider the following grouped data:

Class Interval	**Frequency**
0–10	4
10–20	6
20–30	10
30–40	8

Table 1.6: *Grouped frequency distribution table for calculating the mode.*

Here, the modal class is 20–30 because it has the highest frequency ($f_1 = 10$). The required values are:

$$L = 20, \quad f_0 = 6, \quad f_1 = 10, \quad f_2 = 8, \quad h = 10$$

$$\text{Mode} = 20 + \left(\frac{10-6}{2(10)-6-8}\right) \cdot 10 = 20 + \left(\frac{4}{6}\right) \cdot 10 = 20 + 6.67 = 26.67$$

1.2.4 Comparison and When to Use Each

- **Mean** is sensitive to outliers and skewed data. It is best used for symmetric, continuous data without extreme values.
- **Median** is more robust to outliers and skewed distributions. It is ideal when the data contain extreme values or are not symmetrically distributed.
- **Mode** is useful for categorical or discrete data, especially when identifying the most frequent category is of interest.
- **Geometric Mean** is appropriate when dealing with ratios, growth rates, or multiplicative processes (e.g., population growth, interest rates).
- **Harmonic Mean** is best for averaging rates, such as speed or price per unit when quantities vary.

Each measure gives a different perspective of the 'center' of the data. The choice of measure should be guided by the nature and scale of the data, and the specific analysis objective.

1.3 Partition Values: Quartiles, Deciles, and Percentiles

Just as the median divides a data set into two equal parts, there are other statistical measures that partition the data into a fixed number of equal segments — such as 4, 10, or 100 parts when the data is arranged in increasing order of magnitude. These measures are collectively referred to as **partition values** or **quantiles**. The most commonly used partition values are the **quartiles**, **deciles**, and **percentiles**, which divide the data into four, ten, and one hundred equal parts, respectively.

Partition values are useful in identifying the spread and concentration of data. For instance, if a student scores at the 90th percentile, they performed better than 90% of the population.

1.3.1 Quartiles

Quartiles divide a ordered data set into four equal parts. There are three quartiles:

- Q_1 (First Quartile): 25% of the data falls below Q_1.
- Q_2 (Second Quartile or Median): 50% of the data falls below Q_2.
- Q_3 (Third Quartile): 75% of the data falls below Q_3.

$$Q_k = \left(\frac{k(n+1)}{4}\right) \text{th value}, \quad k = 1, 2, 3$$

Example: Consider the ordered data: {5, 7, 8, 12, 13, 15, 16, 20, 21}.

- Number of observations $n = 9$
- $Q_1 = \left(\frac{1(9+1)}{4}\right) = 3\text{rd value} = 8$
- $Q_2 = \left(\frac{2(9+1)}{4}\right) = 5\text{th value} = 13$
- $Q_3 = \left(\frac{3(9+1)}{4}\right) = 7\text{th value} = 16$

Interquartile Range (IQR): The Interquartile Range (IQR) is a measure of statistical dispersion, which describes the spread of the middle 50% of a data set. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

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

1.3.2 Deciles

Deciles divide the ordered data into ten equal parts. There are nine deciles (D_1 to D_9), such that:

$$D_k = \left(\frac{k(n+1)}{10}\right) \text{th value}, \quad k = 1, 2, \ldots, 9$$

1.3.3 Percentiles

Percentiles divide the ordered data into one hundred equal parts. There are 99 percentiles (P_1 to P_{99}), commonly used to interpret standardized test scores and similar metrics.

$$P_k = \left(\frac{k(n+1)}{100}\right) \text{th value}, \quad k = 1, 2, \ldots, 99$$

1.4 Measures of Dispersion

Measures of dispersion describe the spread or variability within a data set. While measures of central tendency (such as the mean or median) indicate the typical value, measures of dispersion indicate how much the values in the dataset differ from the central value. A small dispersion means the data points are clustered close to the center, while a large dispersion indicates data points are spread out over a wide range.

Consider the following two data sets, each containing five values: $A = \{4, 5, 5, 5, 6\}$ and $B = \{1, 3, 5, 7, 9\}$. Both sets have the same mean, which is 5. For set A, the mean is

$$\frac{4+5+5+5+6}{5} = \frac{25}{5} = 5,$$

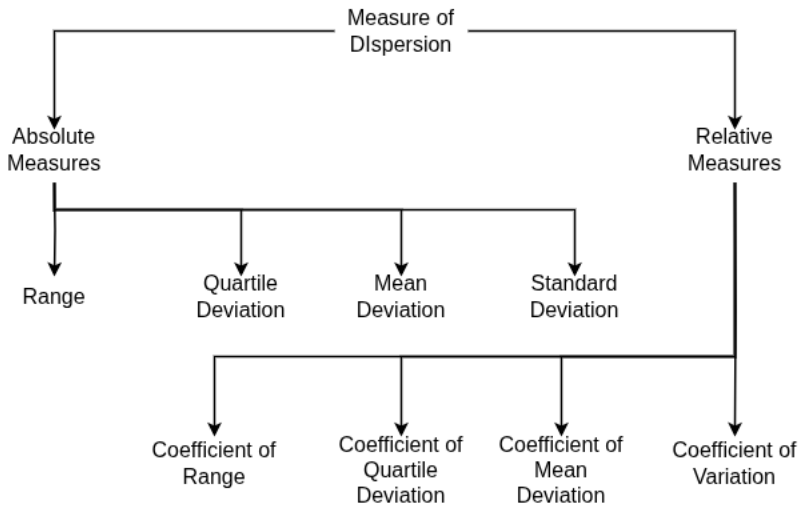
and for B, the mean is

$$\frac{1+3+5+7+9}{5} = \frac{25}{5} = 5.$$

However, their dispersions are quite different. The maximum value of A is 6 and minimum value is 4, whereas the maximum value of B is 9 and minimum value is 1. This means that although both sets center around the same average value, the values in set B are spread out much more widely around the mean compared to set A. Therefore, we say that the dispersion of B is greater than that of A.

Measures of dispersion are broadly classified into two types:

- **Absolute Measures of Dispersion**: These express dispersion in the same units as the original data.
- **Relative Measures of Dispersion**: These express dispersion as a ratio or percentage and are unit-free. They are useful for comparing variability between datasets with different units or magnitudes.



1.4.1 Absolute Measures of Dispersion

1. **Range**: Difference between the largest and smallest observations.

   $$\text{Range} = L - S$$

   where L is the largest value and S is the smallest value.

   Consider the data set: $\{10, 15, 18, 22, 25\}$. Here, the largest value $L = 25$ and the smallest value $S = 10$.

   $$\text{Range} = 25 - 10 = 15$$

2. **Quartile Deviation (Semi-Interquartile Range)**: Quartile deviation is defined as half the difference between the lower and upper quartiles.

   $$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

   where Q_1 and Q_3 are the first and third quartiles.

Consider the ordered data set: {10, 15, 20, 25, 30, 35, 40}.

Here,

$$Q_1 = 15, \quad Q_3 = 35$$

$$\text{Quartile Deviation} = \frac{35 - 15}{2} = \frac{20}{2} = 10$$

3. **Mean Deviation**: Mean deviation is the arithmetic mean of absolute deviations from mean or any other specified value.

$$\text{Mean Deviation about } A = \frac{1}{n}\sum_{i=1}^{n} |x_i - A|$$

Generally mean deviation is taken from the arithmetic mean $\overline{x}$.

$$\text{Mean Deviation about mean} = \frac{1}{n}\sum_{i=1}^{n} |x_i - A|$$

Consider the data set: {2, 4, 6, 8, 10}. First we calculate the mean from the data:

$$\overline{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

Then we compute the absolute deviations from the mean:

$$|2 - 6| = 4, \quad |4 - 6| = 2, \quad |6 - 6| = 0, \quad |8 - 6| = 2, \quad |10 - 6| = 4$$

Finally we calculate the Mean Deviation about the mean:

$$\text{Mean Deviation about mean} = \frac{4 + 2 + 0 + 2 + 4}{5} = \frac{12}{5} = 2.4$$

For weighted data, the Mean Deviation about the mean is given by:

$$\text{Mean Deviation about mean} = \frac{\sum_{i=1}^{n} f_i \, |x_i - \overline{x}|}{\sum_{i=1}^{n} f_i}$$

where:

- x_i are the data values,
- f_i are their corresponding frequencies (weights),
- $\overline{x} = \dfrac{\sum_i f_i x_i}{\sum_i f_i}$ is the weighted mean.

4. **Standard Deviation**: In considering the deviations $x_i - A$ for obtaining a measure of dispersion, we may also get rid of their signs by taking their squares $(x_i - A)^2$, instead of taking their absolute values $|x_i - A|$. The square root of the arithmetic mean of these squares i.e. $\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - A)^2}$ which is called the **root mean square deviation** about A, may be accepted as a measure of dispersion. When $A = \overline{x}$, the measure of dispersion is called the standard deviation.

$$\text{Standard deviation} = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

The square of the standard deviation i.e. σ^2 is known as **variance**.

As an example, consider the data set: $\{2, 4, 4, 4, 5, 5, 7, 9\}$. First we calculate the mean:

$$\overline{x} = \frac{2+4+4+4+5+5+7+9}{8} = \frac{40}{8} = 5$$

Then we find the Squared Deviations from the mean:

$$(2-5)^2 = 9, \quad (4-5)^2 = 1, \quad (4-5)^2 = 1, \quad (4-5)^2 = 1,$$

$$(5-5)^2 = 0, \quad (5-5)^2 = 0, \quad (7-5)^2 = 4, \quad (9-5)^2 = 16$$

Finally compute the Standard Deviation:

$$\sigma = \sqrt{\frac{9+1+1+1+0+0+4+16}{8}} = \sqrt{\frac{32}{8}} = \sqrt{4} = 2$$

For weighted data, the standard deviation is calculated as:

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \overline{x})^2}{\sum_{i=1}^{n} f_i}}$$

where,

- x_i are the data values,
- f_i are the corresponding frequencies (weights),
- $\overline{x} = \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$ is the weighted mean,

Theorem: If $x = a$ (a constant), then $\sigma_x = 0$.

Proof: Since all observations are equal to a, the mean is

$$\overline{x} = a.$$

The standard deviation is

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a-a)^2} = \sqrt{0} = 0.$$

■

Theorem: If $y = a + bx$, where a, b are constants, then

$$\sigma_y = |b|\sigma_x$$

Proof: The mean of y is

$$\overline{y} = a + b\overline{x}$$

The standard deviation of y is

$$\begin{aligned}\sigma_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a+bx_i-(a+b\overline{x}))^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(b(x_i-\overline{x})\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}b^2(x_i-\overline{x})^2} \\ &= |b|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} = |b|\sigma_x\end{aligned}$$

■

Theorem (Pooled standard deviation): Let a dataset be composed of two groups:

- Group 1: n_1 observations, mean $\overline{x}_1$, standard deviation σ_1,
- Group 2: n_2 observations, mean $\overline{x}_2$, standard deviation σ_2.

Then the combined (pooled) standard deviation σ of the dataset (size $n = n_1 + n_2$) is given by:

$$\sigma = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1(\overline{x}_1-\overline{x})^2 + n_2(\overline{x}_2-\overline{x})^2}{n_1+n_2}}$$

where

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1+n_2}$$

Proof: Let us denote $x_{1j}, (j = 1, 2, \ldots, n_1)$ and $x_{2j}, (j = 1, 2, \ldots, n_1)$ the values of the two sets. Then

$$\sigma_1^2 = \sum_{j=1}^{n_1}(x_{1j}-\overline{x}_1), \quad \sigma_2^2 = \sum_{j=1}^{n_1}(x_{2j}-\overline{x}_2)$$

The variance of the combined data set is

$$\sigma^2 = \frac{1}{n_1+n_2}\left(\sum_{j=1}^{n_1}(x_{1j}-\overline{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\overline{x})^2\right)$$

Expanding the first term:

$$\begin{aligned}\sum_{j=1}^{n_1}(x_{ij}-\overline{x})^2 &= \sum_{j=1}^{n_1}\left[(x_{1j}-\overline{x}_1)+(\overline{x}_1-\overline{x})\right]^2 \\ &= \underbrace{\sum_{j=1}^{n_1}(x_{1j}-\overline{x}_1)^2}_{=n_1\sigma_1^2} + 2(\overline{x}_1-\overline{x})\underbrace{\sum_{j=1}^{n_1}(x_{1j}-\overline{x}_1)}_{=0} + \sum_{j=1}^{n_1}(\overline{x}_1-\overline{x})^2 \\ &= n_1\sigma_1^2 + n_1(\overline{x}_1-\overline{x})^2,\end{aligned}$$

Similarly,

$$\sum_{j=1}^{n_2}(x_{2j}-\overline{x})^2 = n_2\sigma_2^2 + n_2(\overline{x}_2-\overline{x})^2.$$

Thus,

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2}{n_1 + n_2}$$

■

1.4.2 Relative Measures of Dispersion

Relative measures of dispersion are defined as ratio of absolute measures of dispersion to the corresponding measure of central tendency. The ratio is expressed in terms of a percentage.

1. **Coefficient of Range**:

$$\text{Coefficient of Range} = \frac{L - S}{L + S} \times 100\%$$

   where L is the largest value and S is the smallest value.

2. **Coefficient of Quartile Deviation**:

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\%$$

3. **Coefficient of Mean Deviation**:

$$\text{Coefficient of M.D.} = \frac{\text{Mean Deviation}}{\overline{x}} \times 100\%$$

4. **Coefficient of Variation (CV)**:

$$\text{CV} = \frac{\sigma}{\overline{x}} \times 100\%$$

1.5 Moments, Skewness and Kurtosis

1.5.1 Raw Moments and Central Moments

In descriptive statistics, **moments** are used to describe various characteristics of a dataset's distribution. Two important types of moments are:

- **Raw moments** (moments about the origin): The r-th **raw moment** of a dataset $x_1, x_2, \ldots, x_n$ is given by:

$$\mu'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

  - First raw moment: $\mu'_1 = \overline{x}$ (sample mean)

- Second raw moment: $\mu_2' = \frac{1}{n}\sum x_i^2$

- **Central moments** (moments about the mean): The r-th **central moment** is the average of the r-th powers of deviations from the mean:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$$

  - First central moment: $\mu_1 = 0$ (since the mean deviation is zero)
  - Second central moment: $\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is the variance.

Example: Consider the dataset:

$$\{2,\ 4,\ 6,\ 8\}$$

- Raw moments:

$$\mu_1' = \frac{1}{4}(2+4+6+8) = 5$$
$$\mu_2' = \frac{1}{4}(2^2+4^2+6^2+8^2) = \frac{120}{4} = 30$$

- Central moments:

$$\mu_1 = 0$$
$$\mu_2 = \frac{1}{4}[(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2]$$
$$= \frac{1}{4}(9+1+1+9) = \frac{20}{4} = 5$$

1.5.2 Relationship Between Raw and Central Moments

The r-th central moment μ_r can be expressed in terms of raw moments μ_k' and powers of the mean $\overline{x}$:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r = \frac{1}{n}\sum_{i=1}\left[\sum_{k=0}^{r}\binom{r}{k}\overline{x_i}^{r-k}\overline{x}^k\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[x_i^r - \binom{r}{1}x_i^{r-1}\overline{x} + \binom{r}{2}x_i^{r-2}\overline{x}^2 + \cdots - \overline{x}^r\right]$$
$$= \mu_r' - \binom{r}{1}\mu_{r-1}'\overline{x} + \binom{r}{2}\mu_{r-2}'\overline{x}^2 + \cdots - n\overline{x}^r$$

Since $\mu_1' = \overline{x}$, we can rewrite the above expression as:

$$\mu_r = \mu_r' - \binom{r}{1}\mu_{r-1}'\mu_1' + \binom{r}{2}\mu_{r-2}'{\mu_1'}^2 + \cdots - n{\mu_1'}^r$$

- $r = 1$:

$$\mu_1 = \mu_1' - \binom{1}{1}\mu_0'\,\mu_1'$$
$$= \mu_1' - \mu_1' = 0$$

- $r = 2$:

$$\begin{aligned}\mu_2 &= \mu'_2 - \binom{2}{1}\mu'_1\,\mu'_1 + \binom{2}{2}\mu'_0\,(\mu'_1)^2 \\ &= \mu'_2 - 2(\mu'_1)^2 + (\mu'_1)^2 = \mu'_2 - (\mu'_1)^2\end{aligned}$$

- $r = 3$:

$$\begin{aligned}\mu_3 &= \mu'_3 - \binom{3}{1}\mu'_2\,\mu'_1 + \binom{3}{2}\mu'_1\,(\mu'_1)^2 - \binom{3}{3}\mu'_0\,(\mu'_1)^3 \\ &= \mu'_3 - 3\,\mu'_2\,\mu'_1 + 3(\mu'_1)^3 - (\mu'_1)^3 = \mu'_3 - 3\,\mu'_2\,\mu'_1 + 2(\mu'_1)^3\end{aligned}$$

- $r = 4$:

$$\begin{aligned}\mu_4 &= \mu'_4 - \binom{4}{1}\mu'_3\,\mu'_1 + \binom{4}{2}\mu'_2\,(\mu'_1)^2 - \binom{4}{3}\mu'_1\,(\mu'_1)^3 + \binom{4}{4}\mu'_0\,(\mu'_1)^4 \\ &= \mu'_4 - 4\,\mu'_3\,\mu'_1 + 6\,\mu'_2\,(\mu'_1)^2 - 4(\mu'_1)^4 + (\mu'_1)^4 \\ &= \mu'_4 - 4\,\mu'_3\,\mu'_1 + 6\,\mu'_2\,(\mu'_1)^2 - 3(\mu'_1)^4\end{aligned}$$

Now let's derive the inverse relationship. Using the binomial expansion on $x_i = (x_i - \overline{x}) + \overline{x}$, the r-th raw moment can be written as

$$\begin{aligned}\mu'_r &= \frac{1}{n}\sum_{i=1}^{n}\left[\overline{x} + (x_i - \overline{x})\right]^r = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=0}^{r}\binom{r}{k}\overline{x}^{\,r-k}(x_i - \overline{x})^k\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\overline{x}^r + \binom{r}{1}\overline{x}^{\,r-1}(x_i - \overline{x}) + \binom{r}{2}\overline{x}^{\,r-2}(x_i - \overline{x})^2 + \cdots + (x_i - \overline{x})^r\right] \\ &= \overline{x}^r + \binom{r}{1}\overline{x}^{\,r-1}\mu_1 + \binom{r}{2}\overline{x}^{\,r-2}\mu_2 + \cdots + \mu_r\end{aligned}$$

where by convention $\mu_1 = 0$.

$$\mu'_r = \overline{x}^r + \binom{r}{1}\overline{x}^{\,r-1}\mu_1 + \binom{r}{2}\overline{x}^{\,r-2}\mu_2 + \cdots + \mu_r$$

- $r = 1$:

$$\begin{aligned}\mu'_1 &= \overline{x}^1 + \binom{1}{1}\overline{x}^0\,\mu_1 \\ &= \overline{x} + 0 = \overline{x}\end{aligned}$$

- $r = 2$:

$$\begin{aligned}\mu'_2 &= \overline{x}^2 + \binom{2}{1}\overline{x}^1\,\mu_1 + \binom{2}{2}\overline{x}^0\,\mu_2 \\ &= \overline{x}^2 + 0 + \mu_2 = \mu_2 + \overline{x}^2\end{aligned}$$

- $r = 3$:

$$\begin{aligned}\mu'_3 &= \overline{x}^3 + \binom{3}{1}\overline{x}^2\,\mu_1 + \binom{3}{2}\overline{x}^1\,\mu_2 + \binom{3}{3}\overline{x}^0\,\mu_3 \\ &= \overline{x}^3 + 0 + 3\,\overline{x}\,\mu_2 + \mu_3 = \mu_3 + 3\,\overline{x}\,\mu_2 + \overline{x}^3\end{aligned}$$

- $r = 4$:

$$\begin{aligned}\mu_4' &= \overline{x}^4 + \binom{4}{1}\overline{x}^3\,\mu_1 + \binom{4}{2}\overline{x}^2\,\mu_2 + \binom{4}{3}\overline{x}^1\,\mu_3 + \binom{4}{4}\overline{x}^0\,\mu_4 \\ &= \overline{x}^4 + 0 + 6\,\overline{x}^2\,\mu_2 + 4\,\overline{x}\,\mu_3 + \mu_4 \\ &= \mu_4 + 4\,\overline{x}\,\mu_3 + 6\,\overline{x}^2\,\mu_2 + \overline{x}^4\end{aligned}$$

1.5.3 Skewness

Skewness is a measure of the asymmetry of a frequency distribution about its mean. The frequency distribution of a dataset is called symmetrical about the value x_0 if the frequency of $x_0 - h$ is same as the frequency of $x_0 + h$, whatever h may be.

The sample skewness is defined as:

$$\text{Skewness}(\gamma_1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^3 = \frac{\mu_3}{\sigma^3}$$

where $\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$ is the third central moment.

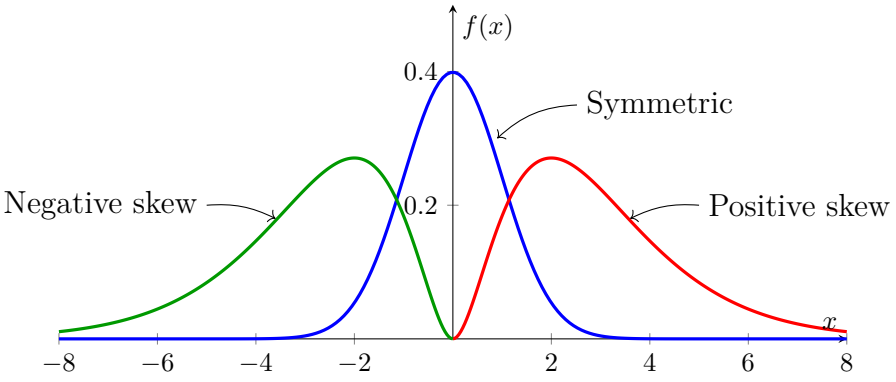
The value of skewness determines the shape of the frequency curve:

- If skewness $= 0$, the distribution is **symmetric**.
- If skewness > 0, the distribution is **positively skewed** (long right tail).
- If skewness < 0, the distribution is **negatively skewed** (long left tail).

How to interpret the formula?

Skewness uses cubed deviations $(x_i - \overline{x})^3$. Cubing serves two purposes: it preserves the sign of the deviation — meaning values greater than the mean contribute positively and those less than the mean contribute negatively — and it exaggerates the impact of larger deviations, making the measure sensitive to extreme values in the tails. This helps identify whether the data are stretched more to the right or left.

Dividing by the cube of the standard deviation σ^3 standardizes the measure, removing units and allowing for meaningful comparisons across datasets with different scales. The result is a dimensionless quantity: positive skewness indicates a long right tail, negative skewness signals a long left tail, and zero skewness implies symmetry around the mean.



Example: Given data: $x = \{2, 3, 4, 5, 8\}$

- Mean: $\overline{x} = \dfrac{2+3+4+5+8}{5} = 4.4$
- Standard deviation: $s = \sqrt{\dfrac{1}{5}\sum_i (x_i - \overline{x})^2} \approx 2.058$
- Third central moment:
$$\mu_3 = \frac{1}{5}\left[(-2.4)^3 + (-1.4)^3 + (-0.4)^3 + 0.6^3 + 3.6^3\right] \approx 3.232$$
- Skewness:
$$\gamma_1 = \frac{3.232}{(2.058)^3} \approx 0.37$$

Since skewness > 0, the distribution is **positively skewed**.

In most unimodal distributions[1], the following "rule of thumb" holds regarding the ordering of Mean, Median, and Mode under skewness:

- **Positive skew (right-tailed):**
$$\text{Mode} < \text{Median} < \text{Mean}.$$
Extreme values on the right pull the mean farther out than the median, while the mode remains at the peak of the bulk of the data.
- **Negative skew (left-tailed):**
$$\text{Mean} < \text{Median} < \text{Mode}.$$
Extreme values on the left drag the mean below the median, and the mode stays at the highest-density point on the right.

1.5.4 Kurtosis

Kurtosis measures the degree of 'peakedness' of a frequency distribution curve. Two frequency distributions may have the same mean, dispersion, and skewness, yet differ in how concentrated the values are around the mode. One distribution may have a sharper peak due to a higher concentration of values near the center, while the other appears flatter. This characteristic of a frequency distribution is known as kurtosis. It is calculated and reported either as an absolute or as a relative value. The absolute kurtosis is always a positive number.

$$\text{Absolute Kurtosis} = \frac{\mu_4}{\sigma^4}$$

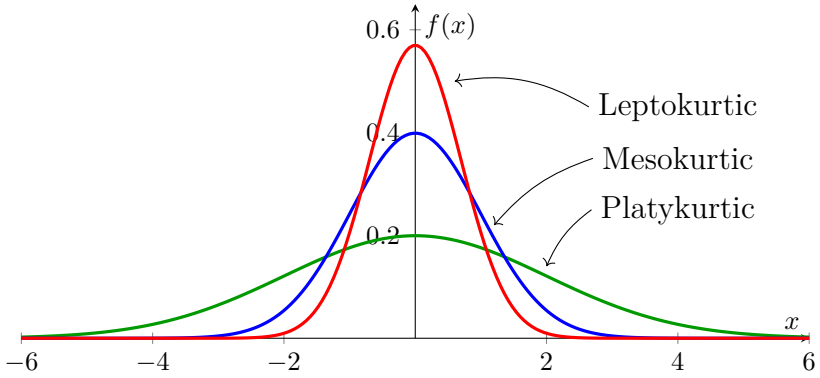
where $\mu_4 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$ is the fourth central moment.

The absolute kurtosis of a normal distribution, which we will learn in later chapter, is 3. The value 3 is taken as a datum (reference point) to calculate the relative kurtosis.

$$\text{Absolute Kurtosis}(\gamma_2) = \text{Relative Kurtosis} - 3$$

[1] A dataset is said to have a **unimodal distribution** if its values tend to cluster around a single (not multiple) central peak when plotted as a histogram or a frequency curve.

- Relative kurtosis $= 0$: **Mesokurtic** (e.g. Normal).
- Relative kurtosis > 0: **Leptokurtic** (heavy tails, sharp peak).
- Relative kurtosis < 0: **Platykurtic** (light tails, flat top).



How to interpret the formula?

Kurtosis raises deviations from the mean to the fourth power. This has a distinct purpose: it emphasizes extreme values far from the mean far more than values closer to it. Unlike cubing (used in skewness), which preserves the sign of deviations to detect asymmetry, raising to the fourth power removes the sign, treating all deviations equally, but magnifying larger ones disproportionately.

Before applying the fourth power, each deviation is **first divided by the standard deviation** σ. This step is crucial: it standardizes the scale of deviations, ensuring that the measure reflects the *relative extremity* of values, not just their raw magnitude. Even if two distributions seem to have similar tail weights, the distribution with a *smaller standard deviation* (i.e., a tighter central cluster) will yield *larger standardized deviations*, which get exaggerated further by the fourth power.

Also, dividing the fourth central moment by σ^4 makes kurtosis a **dimensionless** and **scale-invariant** quantity, allowing meaningful comparisons across datasets.

Example: Using the same data $x = \{2, 3, 4, 5, 8\}$:

- Mean $\overline{x} = 4.4$, Standard deviation $\sigma \approx 2.058$.
- Fourth central moment:

$$\mu_4 = \frac{1}{5}\left[(-2.4)^4 + (-1.4)^4 + (-0.4)^4 + 0.6^4 + 3.6^4\right] \approx 41.03.$$

- Kurtosis$(\gamma_2) = \dfrac{41.03}{(2.058)^4} \approx 2.29$, Relative Kurtosis ≈ -0.71. This dataset is **platykurtic**.

Chapter 2

Theory of Probability

2.1 Some Notation and Terminology

2.1.1 Random Experiment

An **experiment** is generally defined as one or more actions that result in a specific outcome.

An experiment E is called a **random experiment** if it satisfies the following conditions:

- All possible outcomes of E are known in advance.
- It is not possible to predict with certainty which specific outcome will occur in any given trial[a] of E.
- The experiment E can be repeated, at least conceptually, under identical conditions an infinite number of times.

[a] A **trial** is a single performance or execution of an experiment. Tossing a coin once is a trial of the coin-tossing experiment.

A common example of a random experiment is the tossing of a coin. The possible outcomes—'Head' and 'Tail'—are known in advance, but it is impossible to determine with certainty which of the two outcomes will occur on any single toss.

2.1.2 Event Space (a.k.a Sample Space)

The set of all possible outcomes of a random experiment E is called the **sample space** or **event space**, and it is denoted by S.

Each outcome, also known as an **elementary event point**, is an element of S.

For example, in the experiment of tossing a coin, the sample space is

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

where H represents 'Head' and T represents 'Tail'.

If E is the experiment of rolling a pair of dice, the sample space consists of all ordered

pairs of numbers from 1 to 6:

$$S = \{(1,1),(1,2),(1,3),\ldots,(6,6)\}$$

This sample space contains 36 distinct outcomes, as each die can show any of 6 faces independently.

A sample space is **discrete** if it consists of a finite or countable infinite set of outcomes.
A sample space is **continuous** if it contains an interval (either finite or infinite) of real numbers.

The sample space $S = \mathbb{R}^+$ is an example of a continuous sample space, whereas $S = \{H, T\}$ is a discrete sample space.

2.1.3 Events

We often focus on groups of related outcomes from a random experiment, which are represented as subsets of the sample space.

A subset of a sample space is called an **event**.

Consider the random experiment of rolling a die. The sample space is

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

Let

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

be an event, which can be described as "an even number appears when the die is rolled."

There are various types of events:

- **Impossible Event:** An event that contains no outcomes from the sample space is called an impossible event. For example, $A = \emptyset$ is an impossible event.
- **Certain Event:** An event that contains all outcomes of the sample space is called a certain or sure event. For example, $A = S$ is an impossible event.
- **Simple (Elementary) Event:** An event consisting of exactly one outcome of the sample space. For example, $A = \{4\}$ is a simple event when rolling a die.
- **Composite (Compound) Event:** An event that consists of more than one outcome of the sample space. For example, $A = \{2,4,6\}$ is a composite event when rolling a die.
- **Dependent and Independent Events:** Two events are considered dependent if the occurrence of one event influences the probability of the other event occurring. Conversely, they are independent if the occurrence of one event does not affect the probability of the other event.

2.1.4 Mutually Exclusive Events

Two events are said to be **mutually exclusive** if they cannot occur at the simultaneously. Mathematically, if events A_1 and A_2 are exhaustive, then:

$$A_1 \cap A_2 = \emptyset$$

When tossing a coin, the events 'Head' and 'Tail' are mutually exclusive because both cannot occur at the same time. If we get a head, we cannot get a tail in that toss.

2.1.5 Exhaustive Set of Events

A set of events is said to be **exhaustive** at least one of the events in the set must occur. The union of all the events in the set equals the entire sample space S. Mathematically, if events $A_1, A_2, \ldots, A_n$ are exhaustive, then:

$$A_1 \cup A_2 \cup \cdots \cup A_n = S$$

When rolling a die, the events {1}, {2}, {3}, {4}, {5}, and {6} form an exhaustive set because they cover all possible outcomes of the die roll. One of these events must occur when the die is thrown.

2.2 Definition of Probability

2.2.1 A Priori Probability

A priori probability, also known as **classical probability**, is the probability that is determined before an experiment is conducted. It is based on the knowledge of the system or experiment and is calculated using the total number of equally likely outcomes.

If there are n mutually exclusive, exhaustive and equally likely[a] outcomes of a random experiment and out of them m outcomes are favorable to an event A, then the probability of the event A is defined as

$$P(A) = \frac{m}{n}$$

[a]By the phrase '**equally likely**' it is meant that none of the outcomes is expected to occur in preference to other in any trial of the given random experiment.

For example, the probability of getting a head in a fair coin toss is

$$P(\text{Head}) = \frac{1}{2}$$

based on the assumption of equal likelihood of heads and tails.

2.2.2 A Posteriori Probability

A posteriori probability, also known as **empirical probability**, is the probability that is determined after an experiment is conducted. It is based on observed data or information obtained from the experiment.

Let A be an event of a given random experiment. Let event A occurs $N(A)$ number of times when the random experiment is repeated N times under identical conditions. The probability of the event A is defined as

$$P(A) = \lim_{N \to \infty} \frac{N(A)}{N}$$

A posteriori probability can be updated as new evidence becomes available. For example, after observing several rolls of a die, you may update the probability of rolling a particular number based on the outcomes observed.

2.3 Axioms of Probability

The subject of probability is based on three commonsense rules, known as axioms. They are:

1. $P(S) = 1$ where S is the sample space.
2. $0 \leq P(E) \leq 1$ for any event E.
3. For two mutually exclusive events E_1 and E_2,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

More generally, if $E_1, E_2, \ldots E_n$ are mutually exclusive events,

$$P(E_1 \cup E_2 \cup \ldots E_n) = P(E_1) + P(E_2) + \cdots + E_n$$

The first axiom states that the probability of the entire sample space S, which represents all possible outcomes of an experiment, is 1. It reflects the certainty that something in the sample space will occur. For example, when flipping a fair coin, the sample space is $S = \{\text{Head}, \text{Tail}\}$, and the probability that the outcome is either 'Head' or 'Tail' is 1.

The second axiom ensures that probabilities are valid numerical values between 0 and 1. A probability of 0 means the event is impossible (e.g., rolling a 7 on a standard six-sided die), while a probability of 1 means the event is certain to happen. All other events fall somewhere in between these two extremes.

This axiom applies when two events E_1 and E_2 are mutually exclusive—they cannot both happen at the same time. In such cases, the probability that either event occurs is the sum of their individual probabilities. For instance, when rolling a die, the probability of getting a 2 or a 5 is $P(2) + P(5)$, since a single die roll cannot result in both values.

These axioms imply the following theorems.

Theorem: $P(\overline{E}) = 1 - P(E)$ for any event E

Proof: Let S be a sample space and let E be an event. Then E and $\overline{E}$ are mutually exclusive. So by axiom 3,

$$P(E \cup \overline{E}) = P(E) + P(\overline{E})$$

But $E \cup \overline{E} = S$, and by axiom 1, $P(S) = 1$. Therefore,

$$P(E) + P(\overline{E}) = 1$$

which implies

$$P(\overline{E}) = 1 - P(E)$$

■

Theorem: $P(\emptyset) = 0$ where $\emptyset$ denotes the empty set.

Proof: Let S be a sample space. Then $\emptyset = \overline{S}$. Therefore

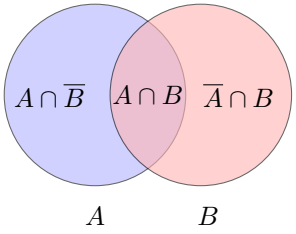
$$P(\emptyset) = 1 - P(S) = 1 - 1 = 0$$

■

Theorem: For any two events A and B (not necessarily mutually exclusive),

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

Proof: From the Venn diagram, we can see that the event $A \cup B$ consists of three mutually exclusive events (subsets) $A \cap \overline{B}$, $A \cap B$ and $\overline{A} \cap B$.



The event A consists of two mutually exclusive events $A \cap \overline{B}$ and $A \cap B$. Therefore

$$P(A) = P(A \cap \overline{B}) + P(A \cap B)$$

Similarly,

$$P(B) = P(\overline{A} \cap B) + P(A \cap B)$$

Now,

$$\begin{aligned} P(A \cup B) &= P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B) \\ &= [P(A) - P(A \cap B)] + P(A \cap B) + [P(B) - P(A \cap B)] \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

■

Example: Consider a fair six-sided die, and define two events:

- A: The event that the die shows an odd number (i.e., $A = \{1, 3, 5\}$)
- B: The event that the die shows a number greater than or equal to 4 (i.e., $A = \{4, 5, 6\}$)

The union of A and B is the event that either event A or event B occurs (or both). The union of A and B is denoted by:

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

To find the probability of the union $P(A \cup B)$, we use the formula:

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

- $P(A) = \frac{3}{6} = 0.5$ (since there are 3 odd numbers: 1, 3, 5)
- $P(B) = \frac{3}{6} = 0.5$ (since there are 3 numbers ≥ 4: 4, 5, 6)
- $P(A \cap B) = \frac{1}{6} = \frac{1}{3}$ (since 5 is the only number that is both odd and ≥ 4)

So, the probability of $A \cup B$ is:

$$P(A \cup B) = 0.5 + 0.5 - \frac{1}{3} = 1 - \frac{1}{3} = \frac{2}{3}$$

Thus, the probability of rolling a die and getting either an odd number or a number greater than or equal to 4 is $\frac{2}{3}$.

Theorem: For any three events A, B and C (not necessarily mutually exclusive),

$$\begin{aligned} P(A \cup B \cup C) =& P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) \\ & - P(A \cap C) + P(A \cap B \cap C) \end{aligned}$$

Proof: We begin by applying the two-event formula to $A \cup (B \cup C)$:

$$P(A \cup B \cup C) = P(A) + P(B \cup C) - P(A \cap (B \cup C))$$

Now, apply the two-event formula to $P(B \cup C)$:

$$P(B \cup C) = P(B) + P(C) - P(B \cap C)$$

Also, apply distributivity to expand $P(A \cap (B \cup C))$:

$$\begin{aligned} P(A \cap (B \cup C)) &= P((A \cap B) \cup (A \cap C)) \\ &= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \end{aligned}$$

Substituting back into the original expression:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + \big[P(B) + P(C) - P(B \cap C)\big] \\ &\quad - \big[P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)\big] \\ &= P(A) + P(B) + P(C) - P(B \cap C) \\ &\quad - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C) \end{aligned}$$

■

Theorem: Let $A_1, A_2, \ldots, A_n$ be n number of events of a random experiment. Then the probability of their union is given by:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) =& \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) \\ & + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots \\ & + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned}$$

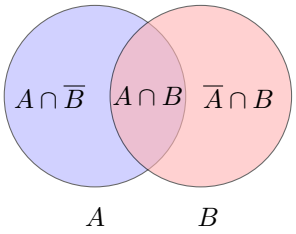
The proof of this theorem is left as an exercise.

Theorem: If A and B are two events of a random experiment, then the probability that exactly one of them occurs is given by

$$P(\text{exactly one of } A \text{ or } B) = P(A) + P(B) - 2P(A \cap B)$$

Proof: The event "exactly one of A or B occurs" means either A happens and B doesn't, or B happens and A doesn't. That means the event:

$$(A \cap \overline{B}) \cup (\overline{A} \cap B)$$



We can see it from the Venn diagram:

$$\begin{aligned} P((A \cap \overline{B}) \cup (\overline{A} \cap B)) &= P(A \cup B) - P(A \cap B) \\ &= (P(A) + P(B) - P(A \cap B)) - P(A \cap B) \\ &= P(A) + P(B) - 2P(A \cap B) \end{aligned}$$

Therefore,

$$P(\text{exactly one of } A \text{ or } B) = P(A) + P(B) - 2P(A \cap B)$$

■

Boole's Inequality: Let $A_1, A_2, \ldots, A_n$ be n events of a random experiment. Then:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Proof: We prove the inequality by induction on n.

Base Case: For $n = 1$,

$$P(A_1) = P(A_1)$$

so the inequality holds with equality.

Inductive Step: Assume the inequality holds for $n = k$, i.e.,

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i)$$

Now consider $n = k + 1$. Let $B = \bigcup_{i=1}^{k} A_i$. Then:

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) = P(B \cup A_{k+1})$$

Using the formula for the union of two events:

$$P(B \cup A_{k+1}) = P(B) + P(A_{k+1}) - P(B \cap A_{k+1}) \leq P(B) + P(A_{k+1})$$

Applying the induction hypothesis to $P(B)$:

$$P(B \cup A_{k+1}) \leq \sum_{i=1}^{k} P(A_i) + P(A_{k+1}) = \sum_{i=1}^{k+1} P(A_i)$$

Thus, the inequality holds for $n = k + 1$. By the principle of mathematical induction, the inequality holds for all $n \in \mathbb{N}$. ■

Boole's inequality provides a simple and conservative upper bound for the probability of the union of multiple events. This is important because, without detailed knowledge of the relationships between the events (e.g., how much they overlap), we can still estimate the probability that at least one of the events occurs by adding their individual probabilities.

Example: If you have three events with probabilities:

$$P(A_1) = 0.3, \quad P(A_2) = 0.5, \quad P(A_3) = 0.7,$$

but you don't know the intersections between them, Boole's inequality will tell you that the probability of at least one occurring is at most:

$$P(A_1 \cup A_2 \cup A_3) \leq 0.3 + 0.5 + 0.7 = 1.5$$

Since probabilities cannot exceed 1, this shows that the bound is very loose, but it is still useful for getting a rough estimate.

Bonferroni's inequality: Let $A_1, A_2, \ldots, A_n$ be events of a random experiment. Then:

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Proof: We proceed by induction on n.

Base case: For $n = 2$, we begin with the identity:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Using the axiom that $P(A_1 \cup A_2) \leq 1$, we substitute:

$$P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq 1$$

Rearranging:

$$P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$$

So the inequality holds for $n = 2$.

Inductive step: Assume the result holds for $n = k$, i.e.,

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq \sum_{i=1}^{k} P(A_i) - (k-1)$$

Let $B = \bigcap_{i=1}^{k} A_i$. Then:

$$P\left(\bigcap_{i=1}^{k+1} A_i\right) = P(B \cap A_{k+1}) \geq P(B) + P(A_{k+1}) - 1$$

Using the induction hypothesis:

$$P(B \cap A_{k+1}) \geq \left(\sum_{i=1}^{k} P(A_i) - (k-1)\right) + P(A_{k+1}) - 1 = \sum_{i=1}^{k+1} P(A_i) - k$$

Therefore, the inequality holds for $n = k + 1$. By induction, it holds for all $n \in \mathbb{N}$.

The Bonferroni inequality for intersections provides a lower bound for the probability of the intersection of multiple events. ■

Example: Consider three events A_1, A_2, and A_3 with probabilities:

$$P(A_1) = 0.6, \quad P(A_2) = 0.5, \quad P(A_3) = 0.7$$

Using Bonferroni's inequality, the lower bound for the probability that all three events occur is:

$$P(A_1 \cap A_2 \cap A_3) \geq 0.6 + 0.5 + 0.7 - 2 = 0.8$$

Thus, the probability that all three events occur simultaneously is at least 0.8.

2.4 Conditional Probability

Conditional probability is the probability of an event occurring given that another event has already occurred.

Let A and B be two events of a random experiment. The **conditional probability** of event A given that event B has already occurred is defined as:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) \neq 0$$

This formula tells us how likely event A is, given that we know event B has happened. The idea is that we restrict our sample space to the outcomes where B occurs, and then we compute the probability of A within this restricted sample space.

Example: Let's define two events when rolling a fair six-sided die:

- A: The event that the die shows an *even face*, i.e., $A = \{2, 4, 6\}$.
- B: The event that the die shows a *multiple of 3*, i.e., $B = \{3, 6\}$.

$$P(A) = P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2}$$
$$P(B) = P(\{3, 6\}) = \frac{2}{6} = \frac{1}{3}$$
$$P(A \cap B) = P(\{6\}) = \frac{1}{6}$$

The conditional probability of A given B is:

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

The conditional probability of B given A is:

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

Multiplication Rule of Probabilities: If A and B are any events in the sample space S, then:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B \mid A), \quad \text{if } P(A) \neq 0 \\ &= P(B) \cdot P(A \mid B), \quad \text{if } P(B) \neq 0 \end{aligned}$$

The second rule follows directly from the definition of conditional probability by multiplying both sides by $P(B)$. The first rule is obtained from the second by simply switching the roles of A and B.

2.5 Rule of Total Probability

The Rule of Total Probability allows us to compute the probability of an event based on a partition of the sample space.

2.5.1 For Two Events

If B and its complement $\overline{B}$ form a partition of the sample space (i.e., mutually exclusive and collectively exhaustive events), then for any event A:

$$P(A) = P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$$

Proof: Let B and $\overline{B}$ form a partition (i.e., mutually exclusive and exhaustive events) of the sample space. Then any event A can be expressed as:

$$A = (A \cap B) \cup (A \cap \overline{B})$$

Since the sets $A \cap B$ and $A \cap \overline{B}$ are disjoint, the two parts are mutually exclusive. So:

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

Using the definition of conditional probability:

$$P(A \cap B) = P(A \mid B)P(B), \quad P(A \cap \overline{B}) = P(A \mid \overline{B})P(\overline{B})$$

Substituting:

$$P(A) = P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$$

■

2.5.2 For Multiple Events

Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space (i.e., mutually exclusive and collectively exhaustive events). Then for any event A:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

Proof: Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space, i.e.,

- The events B_i are mutually exclusive: $B_i \cap B_j = \emptyset$ for $i \neq j$
- The events B_i are collectively exhaustive: $\bigcup_{i=1}^{n} B_i = S$

Then for any event $A \subseteq S$:

$$A = A \cap S = A \cap \left(\bigcup_{i=1}^{n} B_i \right) = \bigcup_{i=1}^{n} (A \cap B_i)$$

Since the B_i are disjoint, so are the $A \cap B_i$, so:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

Using conditional probability:

$$P(A \cap B_i) = P(A \mid B_i) P(B_i)$$

Therefore:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

■

Example: Suppose a factory has three machines:

- Machine M_1 produces 30% of the items, with a defect rate of 1%.
- Machine M_2 produces 50% of the items, with a defect rate of 2%.
- Machine M_3 produces 20% of the items, with a defect rate of 3%.

Let D be the event that an item is defective. We are asked to find $P(D)$, the total probability that a randomly selected item is defective.

Using the Rule of Total Probability:

$$P(D) = P(D \mid M_1)P(M_1) + P(D \mid M_2)P(M_2) + P(D \mid M_3)P(M_3)$$

Substituting the known values:

$$\begin{aligned} P(D) &= (0.01 \times 0.30) + (0.02 \times 0.50) + (0.03 \times 0.20) \\ &= 0.003 + 0.010 + 0.006 = 0.019 \end{aligned}$$

The probability that a randomly chosen item is defective is 0.019 or 1.9%.

2.6 Bayes' Theorem

Bayes' Theorem is a fundamental result in probability theory that arises directly from the definition of conditional probability and the Rule of Total Probability.

Bayes' Theorem: Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space (i.e., mutually exclusive and collectively exhaustive events) and none of which has zero probability i.e. $P(B_i) > 0$ for all i, then for any event A with $P(A) > 0$, the probability of B_r given A is:

$$P(B_r \mid A) = \frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

for $r = 1, 2, \ldots, n$.

Proof: Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space S such that:

- The events B_i are mutually exclusive: $B_i \cap B_j = \emptyset$ for $i \neq j$,
- The union of the B_i's covers the whole sample space: $\bigcup_{i=1}^{n} B_i = S$,
- $P(B_i) > 0$ for all i.

Let A be any event with $P(A) > 0$. By the definition of conditional probability:

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

We apply the Multiplication Rule of Probabilities:

$$P(B_r \cap A) = P(B_r) \cdot P(A \mid B_r)$$

So:

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

Now, using the Rule of Total Probability:

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)$$

Substitute this into the denominator:

$$P(B_r \mid A) = \frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

■

Example: Suppose in a dataset of 1000 emails, 200 are identified as spam and 800 as non-spam. Among the spam emails, 80 contain the word 'discount', while 80 of the non-spam emails also contain this word. We need to calculate the probability that an email is spam given that it contains the word 'discount'.

- $P(S) = \dfrac{200}{1000} = 0.2$: Probability that an email is spam.

- $P(\overline{S}) = \dfrac{800}{1000} = 0.8$: Probability that an email is not spam.
- $P(D \mid S) = \dfrac{80}{200} = 0.4$: Probability that the word "discount" appears in a spam email.
- $P(D \mid \overline{S}) = \dfrac{80}{800} = 0.1$: Probability that "discount" appears in a non-spam email.

We want to find the probability that an email is spam given that it contains the word "discount", i.e., $P(S \mid D)$.

Using Bayes' Theorem:

$$P(S \mid D) = \frac{P(S) \cdot P(D \mid S)}{P(S) \cdot P(D \mid S) + P(\overline{S}) \cdot P(D \mid \overline{S})}$$

Substituting the values:

$$P(S \mid D) = \frac{0.2 \times 0.4}{0.2 \cdot 0.4 + 0.8 \cdot 0.1} = \frac{0.08}{0.08 + 0.08} = \frac{0.08}{0.16} = 0.5$$

So, the probability that the email is spam given it contains the word "discount" is 50%. Even though only 20% of all emails are spam, once we see the word "discount" the chance the email is spam rises to 50%, because that word is much more common in spam than in legitimate messages.

2.6.1 Importance of Bayes' Theorem and Updating Probability

Bayes' Theorem is important because it provides a mathematical framework for updating probabilities when new information becomes available. In real-world terms, it helps us refine our beliefs or predictions as we gather more data.

Suppose we want to determine the probability of an event A, such as an email being spam. Initially, we rely on prior knowledge, which is represented by the prior probability $P(A)$. Now, imagine we observe some new evidence B, such as the presence of a specific word like "discount" in the email.

Bayes' Theorem helps us compute the updated probability $P(A \mid B)$, known as the posterior probability, using the following formula:

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

Here:

- $P(A)$ is the **prior probability** — our initial belief (prediction) about the event A.
- $P(B \mid A)$ is the **likelihood** — the probability of observing evidence B given that A is true.
- $P(B)$ is the **marginal probability** of observing evidence B under all possible conditions.
- $P(A \mid B)$ is the **posterior probability** — our updated belief (prediction) about A after observing B.

This formula enables us to revise our estimate of the probability of A whenever new information B becomes available. Depending on the relationship between A and B, the posterior probability may be higher or lower than the prior, reflecting the impact of the new evidence.

2.7 Statistical Independence of Events

If A and B are any two events in a sample space S, we say that "A is independent of B" if:

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

This means that knowing whether or not B has occurred "does not change" the probability of A occurring.

From the definition of conditional probability:

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

if we substitute and rearrange the condition $P(A \mid B) = P(A)$, we get:

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

Thus, another equivalent definition of independence is:

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

Now, to check whether B is independent of A, we look at:

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

Since we just showed that $P(A \cap B) = P(A) \cdot P(B)$, substitute:

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

Thus, B is also independent of A.

If A is independent of B, then B is also independent of A. Therefore, we say that A and B are **mutually independent**.

Multiplication Rule for Independent Events: Two events A and B are (mutually) independent events if and only if

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

Example: Consider the experiment of rolling two fair six-sided dice, and consider the following events:

- A: The first die shows a 1.

$$A = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$$

- B: The second die shows a 2.

$$B = \{(1,2), (2,2), (3,2), (4,2), (5,2), (6,2)\}$$

The total number of outcomes in the sample space is $6 \times 6 = 36$.

We compute:

$$P(A) = \frac{6}{36} = \frac{1}{6}, \quad P(B) = \frac{6}{36} = \frac{1}{6}$$
$$A \cap B = \{(1,2)\} \quad \Rightarrow \quad P(A \cap B) = \frac{1}{36}$$

Since

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36},$$

we conclude that the events A and B are **independent**.

Theorem: If the events A and B are independent, then the following pairs of events are also independent:

$$(1)\ A \text{ and } \overline{B}, \quad (2)\ \overline{A} \text{ and } B, \quad (3)\ \overline{A} \text{ and } \overline{B}$$

Proof:

1. Since A and B are independent, we have:
$$P(A \cap B) = P(A) \cdot P(B)$$
Then,
$P(A \cap \overline{B}) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A)(1 - P(B)) = P(A) \cdot P(\overline{B})$
So, A and $\overline{B}$ are independent.
2. Similarly,
$P(\overline{A} \cap B) = P(B) - P(A \cap B) = P(B) - P(A) \cdot P(B) = P(B)(1 - P(A)) = P(\overline{A}) \cdot P(B)$
So, $\overline{A}$ and B are independent.
3. Finally,
$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)]$$
$$= 1 - [P(A) + P(B) - P(A) \cdot P(B)] = (1 - P(A))(1 - P(B)) = P(\overline{A}) \cdot P(\overline{B})$$
So, $\overline{A}$ and $\overline{B}$ are also independent.

■

Theorem: If A and B are independent events, then:

$$P(A \cup B) = 1 - P(\overline{A}) \cdot P(\overline{B})$$

Proof: Using the formula for the union of two events:

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

Since A and B are independent, $P(A \cap B) = P(A) \cdot P(B)$, so:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A) \cdot P(B) \\ &= 1 - (1 - P(A) - P(B) + P(A) \cdot P(B)) \\ &= 1 - (1 - P(A))(1 - P(B)) \\ &= 1 - P(\overline{A})P(\overline{B}) \end{aligned}$$

■

2.7.1 Pairwise vs. Mutual Independence of Multiple Events

Let A, B, and C be three events in a sample space S. The events A, B, and C are said to be **pairwise independent** if:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B), \\ P(A \cap C) &= P(A) \cdot P(C), \\ P(B \cap C) &= P(B) \cdot P(C) \end{aligned}$$

The events A, B, and C are said to be **mutually independent** if:

- They are pairwise independent, and
- The joint probability satisfies:

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

Note: Mutual independence *implies* pairwise independence, but the converse is not necessarily true.

2.7.2 Mutually Exclusive and Independent Events

Let A and B be two events in a sample space.

- If A and B are **mutually exclusive**, then:

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

- If A and B are **independent**, then:

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

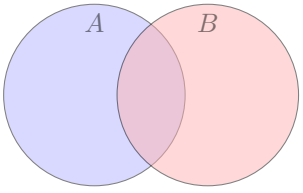
If both conditions are true, then:

$$P(A) \cdot P(B) = 0$$

This implies that either $P(A) = 0$, $P(B) = 0$, or both.

> Two events A and B cannot be both mutually exclusive and independent unless at least one of them has probability zero.

In terms of a Venn diagram, the independence of events implies that the overlap between sets A and B (i.e., $A \cap B$) should be such that its area (probability) equals the product of the areas (probabilities) of each individual circle.



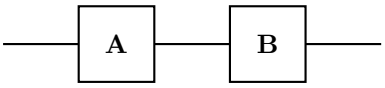
This is different from mutually exclusive events, where the sets do not overlap at all. Both the condition will be satisfied if the area of at least one of the circle is zero.

2.7.3 Reliability Analysis using Statistical Independence

Reliability analysis is the branch of engineering concerned with estimating the failure rates of systems. If a machine has a reliability of 0.95 over 1 year, it means there is a 95% chance that it will work without failure for the entire year. In many systems, components are arranged either in *series* or in *parallel*, and the system's reliability depends on the configuration.

1. **System with Components in Series**:

   Consider a system with two components, A and B, connected in **series**. In this setup, the system will work only if **both** components function properly.

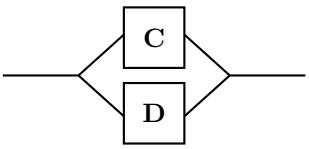
   Suppose the probability that A functions is given by $P(A) = 0.96$, and the probability that B functions is given by $P(B) = 0.92$. Assuming that the components function independently, the probability that the system works is:

   $$\begin{aligned} P(\text{System functions}) &= P(A \cap B) \\ &= P(A) \cdot P(B) \\ &= 0.96 \times 0.92 = 0.8832 \end{aligned}$$

   Since both components must function, the system's reliability is lower than that of either component individually. The more components in series, the more chances for failure.

2. **System with Components in Parallel**:

   Now consider a system with two components, C and D, connected in **parallel**. In this case, the system will function as long as **at least one** component functions.

   Suppose the probabilities that the components function are:

   $$P(C) = 0.88, \quad P(D) = 0.85$$

   Assuming independence, the probability that the system functions is given by:

   $$\begin{aligned} P(\text{System functions}) &= P(C \cup D) \\ &= P(C) + P(D) - P(C \cap D) \\ &= P(C) + P(D) - P(C).P(D) \\ &= 0.88 + 0.85 - (0.88 \times 0.85) \\ &= 1.73 - 0.748 = 0.982 \end{aligned}$$

   In parallel systems, the system is more reliable than the individual components. This configuration adds redundancy, improving fault tolerance.

Chapter 3

Random Variables and Probability Distributions

3.1 What is a Random Variable?

In most cases, we can associate a real number with each elementary event in a sample space. For example, in a coin toss, we may assign the number 1 to the outcome 'Head' and 0 to the outcome 'Tail'. Similarly, when a die is thrown, the outcomes correspond to the numbers $1, 2, 3, \ldots, 6$, depending on which face appears on top.

This assignment of numerical values to outcomes allows us to define a function on the sample space. A real-valued function defined on the sample space is called a **random variable** (also referred to as a **stochastic variable**).

Let S be a sample space of a random experiment. A **random variable** is a function

$$X : S \to \mathbb{R}$$

where each outcome $s \in S$ is mapped to a real number $X(s)$.

Note: A random variable is denoted by an uppercase letter such as X and Y. After experiment is conducted, the measured value of the random variable is denoted by a lowercase letter such as x and y.

3.1.1 Random Variable Types

There are two main types:

1. **Discrete Random Variable:**
   A discrete random variable takes on a countable[1] number of distinct values. Let X be the number of heads obtained when a fair coin is tossed three times. The possible values of X are $0, 1, 2, 3$. Since these values are countable and finite, X is a discrete random variable.

2. **Continuous Random Variable:**
   A continuous random variable takes any value within a certain range of real numbers. The possible values are uncountable and include fractions and decimals. Let Y be

[1] This means the values can be finite (like a die roll) or countably infinite (like the number of coin tosses until the first head).

the amount of time (in minutes) a customer waits in a queue at a bank. The variable Y can take any real value within a range, such as $0 \leq Y \leq 30$, including fractions like 3.5 or 12.75. Hence, Y is a continuous random variable.

3.2 Probability Distribution

To each value of a random variable X, there corresponds a definite probability. Let $x_1, x_2, \ldots, x_n$ be the possible values of X, and let $p_1, p_2, \ldots, p_n$ be the corresponding probabilities such that:

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

A statement of these values along with their associated probabilities defines the probability distribution of the random variable X.

3.2.1 Probability Mass Function (PMF)

The Probability Mass Function (PMF) is used for **discrete random variables**. It gives the probability that a discrete random variable X takes a specific value x_k. The PMF is defined as:

$$p_X(x_k) = P(X = x_k)$$

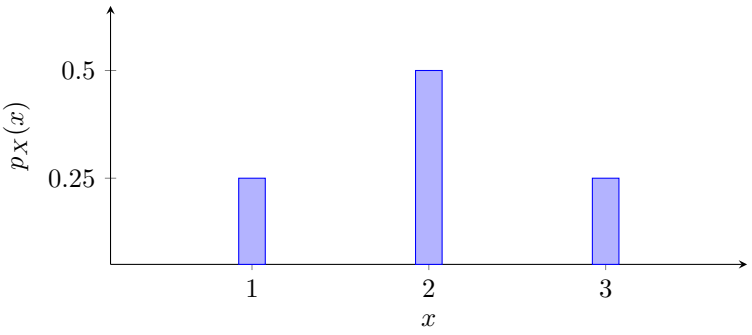
where x_k is a specific value of the random variable X. The PMF satisfies the following properties:

1. $0 \leq p_X(x_k) \leq 1$ for all k
2. $\sum_k p_X(x_k) = 1$

Example: Consider a discrete random variable X with the following distribution:

$$p_X(x) = P(X = x) = \begin{cases} \frac{1}{4}, & \text{if } x = 1 \\ \frac{1}{2}, & \text{if } x = 2 \\ \frac{1}{4}, & \text{if } x = 3 \\ 0, & \text{otherwise} \end{cases}$$

The PMF is visualized as follows:



3.2.2 Probability Density Function (PDF)

The Probability Density Function (PDF) is used for **continuous random variables**. It is defined as a function f_X such that the probability that a continuous random variable X lies within an interval $[a, b]$ is given by the integral of f_X over that interval:

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

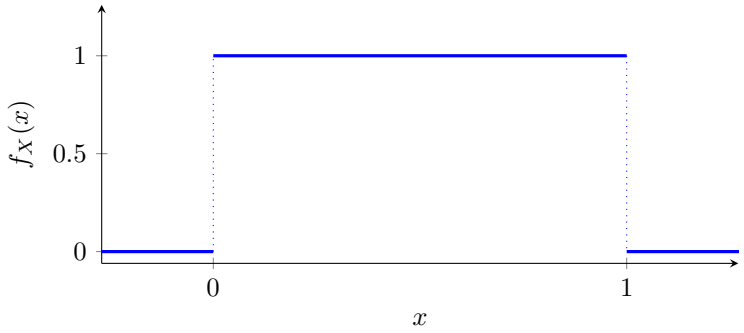
where $f_X(x)$ is the PDF of X. The PDF satisfies the following properties:

1. $f_X(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Example: Consider a continuous random variable with the following distribution:

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

The PDF can be visualised as follows:



3.2.3 Cumulative Distribution

The cumulative distribution gives the probability that a random variable takes a value less than or equal to a specified value. Mathematically **Cumulative Distribution Function** (CDF) of a random variable X is defined as:

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

This function gives the probability that X takes a value less than or equal to x.

CDF for Discrete Random Variables

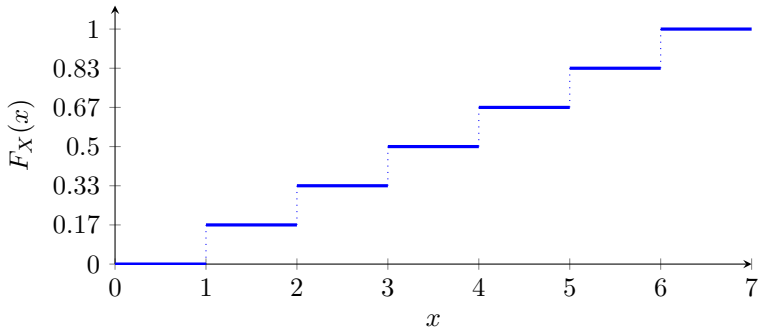
For a discrete random variable X, the CDF is given by the sum of the probabilities for all values less than or equal to x. If X takes the values $x_1, x_2, \ldots, x_n$, the CDF is:

$$F_X(x) = \sum_{x_k \leq x} P(X = x_k)$$

This sum includes all the probabilities up to and including x.

Example: Consider the random variable X representing the outcome of a fair six-sided die roll. The CDF for X is:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 1 \\ \frac{1}{6}, & \text{if } 1 \leq x < 2 \\ \frac{2}{6}, & \text{if } 2 \leq x < 3 \\ \frac{3}{6}, & \text{if } 3 \leq x < 4 \\ \frac{4}{6}, & \text{if } 4 \leq x < 5 \\ \frac{5}{6}, & \text{if } 5 \leq x < 6 \\ 1, & \text{if } x \geq 6 \end{cases}$$



CDF for Continuous Random Variables

For a continuous random variable X, the CDF is obtained by integrating the probability density function (PDF) up to x:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

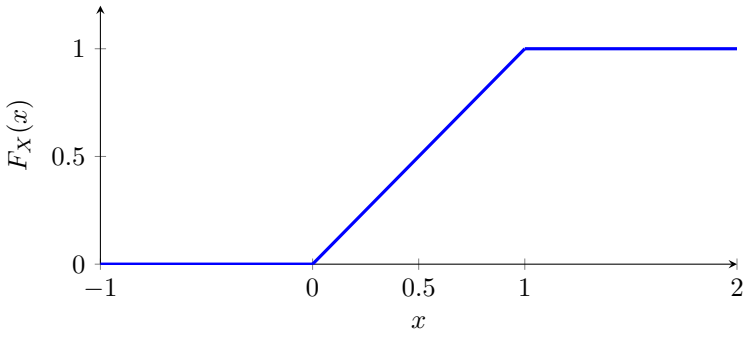
This represents the area under the PDF curve from $-\infty$ to x, giving the cumulative probability up to x.

Example: Consider a continuous random variable X with a probability density function (PDF) $f_X(x) = 1$ for $0 \leq x \leq 1$ (uniform distribution). The CDF is given by:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x \leq 1 \\ 1, & \text{if } x > 1 \end{cases}$$

This is obtained by integrating the PDF:

$$F_X(x) = \int_{0}^{x} 1 \cdot dt = x \quad \text{for} \quad 0 \leq x \leq 1$$



Properties of the CDF

1. The CDF $F_X(x)$ is a non-decreasing function, meaning:

$$F_X(x_1) \leq F_X(x_2) \quad \text{for} \quad x_1 \leq x_2$$

2. The CDF is bounded between 0 and 1:

$$0 \leq F_X(x) \leq 1 \quad \text{for all } x$$

3. The CDF approaches 1 as $x \to \infty$ and 0 as $x \to -\infty$:

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to \infty} F_X(x) = 1$$

4. For a continuous random variable, the CDF is continuous, while for a discrete random variable, the CDF has jumps at the values taken by the random variable.

5. For continuous random variable, the derivative of the CDF gives the PDF (if the derivative exists):

$$f_X(x) = \frac{d}{dx} F_X(x)$$

3.3 Mean and Variance of a Random Variable

The mean and variance are important measures that describe the central tendency and spread of a random variable, respectively.

3.3.1 Mean of a Random Variable

The **mean** or **expected value** of a random variable X, denoted as $\mathbb{E}(X)$, provides the long-run average of the outcomes of the random variable. It is defined as the weighted sum of all possible values of X, weighted by their probabilities.

1. **Discrete Random Variable**: For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p_X(x_1), p_X(x_2), \ldots, p_X(x_n)$, the expected value (mean) is denoted by is given by:

$$\mu = \mathbb{E}(X) = \sum_{i=1}^{n} x_i \cdot p_X(x_i)$$

   Example: Suppose X is the outcome of a roll of a fair die. The possible values of X are $1, 2, 3, 4, 5, 6$, and each value has a probability of $\frac{1}{6}$. The expected value is:

$$\mathbb{E}(X) = \sum_{i=1}^{6} x_i \cdot \frac{1}{6} = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$$

2. **Continuous Random Variable**: For a continuous random variable X with probability density function $f_X(x)$, the expected value is given by the integral of x weighted by the probability density function:

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

Example: Suppose X is a continuous random variable with a uniform distribution between 0 and 1. The probability density function is:

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

The expected value is:

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

Theorem If $X = a$, where $a \in \mathbb{R}$ is a constant, then

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

Proof. Since X is always equal to a,

- In the discrete case:

$$\mathbb{E}(X) = \sum_x x \cdot P(X = x) = a \cdot P(X = a) = a$$

- In the continuous case:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{-\infty}^{\infty} a \cdot \delta(x - a) \, dx = a$$

■

Theorem: If $Y = bX$, where $b \in \mathbb{R}$, then

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

Proof.

- Discrete case:

$$\mathbb{E}(bX) = \sum_x bx \cdot P(X = x) = b \sum_x x \cdot P(X = x) = b \cdot \mathbb{E}(X)$$

- Continuous case:

$$\mathbb{E}(bX) = \int_{-\infty}^{\infty} bx \cdot f_X(x) \, dx = b \int_{-\infty}^{\infty} x f_X(x) \, dx = b \cdot \mathbb{E}(X)$$

■

3.3.2 Variance of a Random Variable

The **variance** of a random variable X, denoted as σ_X^2 or $\text{Var}(X)$, measures the spread or dispersion of the random variable around its mean. It is defined as the expected squared deviation from the mean:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}(X) + (\mathbb{E}(X))^2\right] \\
&= \mathbb{E}(X^2) - 2(\mathbb{E}(X)^2 + (\mathbb{E}(X))^2 \\
&= \mathbb{E}(X^2) - (\mathbb{E}(X))^2
\end{aligned}$$

1. **Discrete Random Variable**: For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p_X(x_1), p_X(x_2), \ldots, p_X(x_n)$, the variance is given by:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} (x_i - \mathbb{E}(X))^2 \cdot p_X(x_i)$$

Alternatively, it can be computed as:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} x_i^2 \cdot p_X(x_i) - (\mathbb{E}(X))^2$$

Example: For the fair die roll example where $\mathbb{E}(X) = 3.5$, the variance is:

$$\mathrm{Var}(X) = \sum_{i=1}^{6} (x_i - 3.5)^2 \cdot \frac{1}{6} = \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2\right)$$

$$= \frac{1}{6} \cdot (6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) = \frac{17.5}{6} \approx 2.92$$

2. **Continuous Random Variable**: For a continuous random variable X with probability density function $f_X(x)$, the variance is given by:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 \cdot f_X(x)\, dx$$

Alternatively, it can be computed as:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx - (\mathbb{E}(X))^2$$

Example: For the continuous uniform random variable X between 0 and 1, $\mathbb{E}(X) = \frac{1}{2}$, the variance is:

$$\mathrm{Var}(X) = \int_0^1 x^2 \cdot 1\, dx - \left(\frac{1}{2}\right)^2 = \left[\frac{x^3}{3}\right]_0^1 - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Theorem: If $X = a$, where $a \in \mathbb{R}$ is constant, then

$$\mathrm{Var}(X) = 0$$

Proof.

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[(a-a)^2] = \mathbb{E}[0] = 0$$

■

Theorem: If $Y = bX$, where $b \in \mathbb{R}$, then

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

Proof.

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(bX) = \mathbb{E}[(bX - \mathbb{E}(bX))^2] \\ &= \mathbb{E}[(bX - b\mathbb{E}(X))^2] = \mathbb{E}[b^2(X - \mathbb{E}(X))^2] \\ &= b^2 \cdot \mathrm{Var}(X)\end{aligned}$$

■

3.4 Joint Distribution of Two random Variables

Let X and Y be two random variables defined on the same probability space. The joint distribution of X and Y describes the probability behavior of the pair (X, Y). It can be either discrete or continuous depending on the nature of X and Y.

- **Discrete Case**:

  If X and Y are discrete random variables, then their joint distribution is defined by the **joint probability mass function (PMF)**:

  $$p_{X,Y}(x, y) = P(X = x, Y = y)$$

  which gives the probability that $X = x$ and $Y = y$ simultaneously.

  The PMF must satisfy:

  - $p_{X,Y}(x, y) \geq 0$ for all x, y
  - $\sum_x \sum_y p_{X,Y}(x, y) = 1$

  The **marginal distributions** can be obtained by summing out the other variable:

  $$p_X(x) = \sum_y p_{X,Y}(x, y), \qquad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

  The value of $p_X(x)$ gives the probability $P(X = x)$ irrespective of Y. Similarly, $p_Y(y)$ gives the probability $P(Y = y)$ irrespective of X.

  Suppose the possible values of X are $x_1, x_2, \ldots, x_m$ and possible values of Y are $y_1, y_2, \ldots, y_n$. The joint distribution can be depicted in a table format as shown Table 3.1.

$\mathbf{Y \backslash X}$	x_1	x_2	$\cdots$	x_m	$p_Y(y_j)$
y_1	$p_{X,Y}(x_1, y_1)$	$p_{X,Y}(x_2, y_1)$	$\cdots$	$p_{X,Y}(x_m, y_1)$	$p_Y(y_1)$
y_2	$p_{X,Y}(x_1, y_2)$	$p_{X,Y}(x_2, y_2)$	$\cdots$	$p_{X,Y}(x_m, y_2)$	$p_Y(y_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_n	$p_{X,Y}(x_1, y_n)$	$p_{X,Y}(x_2, y_n)$	$\cdots$	$p_{X,Y}(x_n, y_m)$	$p_Y(y_n)$
$p_X(x_i)$	$p_X(x_1)$	$p_X(x_2)$	$\cdots$	$p_X(x_m)$	1

Table 3.1: *Joint probability distribution of discrete random variables X and Y.*

The **joint cumulative probability distribution function (CDF)** can be obtained from joint probability mass function:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{X,Y}(x_i, y_j)$$

It sums over all values (x_i, y_j) such that $x_i \leq x$ and $y_j \leq y$. This gives the total probability that the random pair (X, Y) falls within the region $X \leq x, Y \leq y$.

- **Continuous Case**:

  If X and Y are continuous random variables, then their joint distribution is described by the **joint probability density function**:

  $$f_{X,Y}(x, y)$$

which satisfies:

$$P((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,dx\,dy,$$

for any region $A \subset \mathbb{R}^2$.

The joint PDF must satisfy:

- $f_{X,Y}(x,y) \geq 0$ for all x, y
- $\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dx\,dy = 1$

The **marginal PDFs** are obtained by integrating out the other variable:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

The term $f_X(x)$ gives the PDF of X irrespective of Y. Similarly, $f_Y(y)$ gives the PDF of Y irrespective of X.

The **joint cumulative distribution function (CDF)** is defined as:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du$$

This function gives the probability that the random vector (X,Y) falls within the region defined by $X \leq x, Y \leq y$.

Theorem: If $Z = X + Y$, then

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

Proof.

- Discrete case:

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{x,y}(x+y)\cdot p_{X,Y}(x,y) \\
&= \sum_{x,y} x \cdot p_{X,Y}(x,y) + \sum_{x,y} y \cdot p_{X,Y}(x,y) \\
&= \sum_{x} x \cdot p_X(x) + \sum_{y} y \cdot p_Y(y) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

- Continuous case:

$$\begin{aligned}
\mathbb{E}(X+Y) &= \int_{\infty}^{\infty}\int_{\infty}^{\infty}(x+y)f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{\infty}^{\infty}\int_{\infty}^{\infty} x f_{X,Y}(x,y)\,dx\,dy + \int_{\infty}^{\infty}\int_{\infty}^{\infty} y f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{\infty}^{\infty} x f_X(x)\,dx + \int_{\infty}^{\infty} y f_Y(y)\,dy = \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

■

3.4.1 Independence of Two Random Variables

Two random variables X and Y are **independent** if and only if:

- In the discrete case:

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y) \text{ for all } (x,y)$$

- In the continuous case:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \text{ for all } (x,y)$$

Theorem: If X and Y are independent random variables, then

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Proof:

- Discrete Case:

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_x \sum_y xy \cdot p_{X,Y}(x,y) \\
&= \sum_x \sum_y xy \cdot p_X(x) \cdot p_Y(y) \\
&= \sum_x x p_X(x) \cdot \sum_y y p_Y(y) \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y)
\end{aligned}$$

- Continuous Case:

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \cdot f_X(x) \cdot f_Y(y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx \cdot \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y)
\end{aligned}$$

3.4.2 Covariance of Two Random Variables

One important feature of the joint distribution of X and Y is their covariance, which is used to measure the degree of association between X and Y. The **covariance** of two random variables X and Y, denoted by $\text{Cov}(X,Y)$ is defined as:

$$\text{Cov}(X,Y) = \mathbb{E}\big[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\big]$$

Now,

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}\big[(X-\mathbb{E}(X))\cdot(Y-\mathbb{E}(Y))\big] \\ &= \mathbb{E}\big[XY - X\cdot\mathbb{E}(Y) - \mathbb{E}(X)\cdot Y + \mathbb{E}(X)\cdot\mathbb{E}(Y)\big] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y) - \mathbb{E}(X)\cdot\mathbb{E}(Y) + \mathbb{E}(X)\cdot\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y)\end{aligned}$$

Thus we get an alternative shortcut formula for covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y)$$

Interpretation:

- $\mathrm{Cov}(X,Y) > 0$: X and Y tend to increase (or decrease) together.
- $\mathrm{Cov}(X,Y) < 0$: X increases as Y decreases (or vice versa).
- $\mathrm{Cov}(X,Y) = 0$: No linear relationship between X and Y.

Theorem: If X and Y are random variables, then

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

Proof: By definition of covariance,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

Using the commutative property of multiplication,

$$(X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) = (Y-\mathbb{E}[Y])(X-\mathbb{E}[X])$$

Therefore,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(Y-\mathbb{E}[Y])(X-\mathbb{E}[X])] = \mathrm{Cov}(Y,X)$$

■

Theorem: If X and Y are independent random variables, then

$$\mathrm{Cov}(X,Y) = 0$$

Proof: If X and Y are are independent, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\cdot\mathbb{E}(Y)$$

Thus,

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\cdot\mathbb{E}(Y) \\ &= \mathbb{E}(X)\cdot\mathbb{E}(Y) - \mathbb{E}(X)\cdot\mathbb{E}(Y) \\ &= 0\end{aligned}$$

■

Note: The converse is not necessarily true. Two random variables may have $\mathrm{Cov}(X,Y) = 0$ and yet not be independent.

Theorem: For any random variable X,

$$\text{Cov}(X, X) = \text{Var}(X)$$

Proof:

$$\begin{aligned}\text{Cov}(X, X) &= \mathbb{E}(X \cdot X) - \mathbb{E}(X) \cdot \mathbb{E}(X) \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \text{Var}(X)\end{aligned}$$

■

Theorem: Let $a, b \in \mathbb{R}$ are constants. Then for any random variables X and Y,

$$\text{Cov}(aX, bY) = ab \cdot \text{Cov}(X, Y)$$

Proof.

$$\begin{aligned}\text{Cov}(aX, bY) &= \mathbb{E}(aX \cdot bY) - \mathbb{E}(aX) \cdot \mathbb{E}(bY) \\ &= ab \cdot \mathbb{E}(XY) - ab \cdot \mathbb{E}(X) \cdot \mathbb{E}(Y) \\ &= ab \cdot (\mathbb{E}(XY) - \mathbb{E}(X) \cdot \mathbb{E}(Y)) \\ &= ab \cdot \text{Cov}(X, Y)\end{aligned}$$

■

Theorem: If X, Y are two random variables, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$$

Proof.

$$\text{Var}(X + Y) = \mathbb{E}[(X + Y - \mathbb{E}(X + Y))^2]$$

By linearity of expectation, $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, so:

$$= \mathbb{E}[(X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2]$$

Expanding the square:

$$= \mathbb{E}[(X - \mathbb{E}(X))^2 + (Y - \mathbb{E}(Y))^2 + 2(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Using linearity of expectation:

$$= \mathbb{E}[(X - \mathbb{E}(X))^2] + \mathbb{E}[(Y - \mathbb{E}(Y))^2] + 2 \cdot \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

$$= \text{Var}(X) + \text{Var}(Y) + 2 \cdot \text{Cov}(X, Y)$$

■

Theorem: For random variables $X_1, X_2, \ldots, X_n$,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{Cov}(X_i, X_j)$$

Proof: Start with the definition of variance:

$$\text{Var}(Y) = \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right]$$

Let

$$Y = \sum_{i=1}^{n} X_i$$

Then

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2\right]$$

Since expectation is linear,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

So,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left[\left(\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right)^2\right]$$

Expanding the square,

$$= \mathbb{E}\left[\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \sum_{j=1}^{n} (X_j - \mathbb{E}[X_j])\right] = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]$$

By linearity of expectation,

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]$$

Recall the definition of covariance:

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

Thus,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)$$

Split the double sum into terms where $i = j$ and $i \neq j$:

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j).$$

Since the covariance terms are symmetric, the sum over $i \neq j$ counts each pair twice, so

$$\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j) = 2 \sum_{1\leq i<j\leq n} \mathrm{Cov}(X_i, X_j)$$

Hence,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1\leq i<j\leq n} \mathrm{Cov}(X_i, X_j).$$

■

3.5 Conditional Probability Distribution

When working with two random variables X and Y, it is often important to understand the distribution and expected value of one variable given the other. This is captured by **conditional probability distributions**.

3.5.1 Conditional Probability Mass Function

If X and Y are discrete random variables with joint PMF $p_{X,Y}(x, y)$, the conditional PMF of Y given $X = x$ is:

$$p_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \quad \text{for } p_X(x) > 0$$

If X and Y are **independent**, then the joint PMF factorizes as:

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

and the conditional PMF reduces to the marginal PMF of Y:

$$p_{Y|X}(y \mid x) = p_Y(y)$$

In other words, knowing $X = x$ does not change the probability distribution of Y.

3.5.2 Conditional Probability Density Function

If X and Y are continuous random variables with joint PDF $f_{X,Y}(x, y)$, the conditional PDF of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad \text{for } f_X(x) > 0$$

If X and Y are **independent**, then the joint PDF factorizes as:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

and the conditional PDF reduces to the marginal PDF of Y:

$$f_{Y|X}(y|x) = f_Y(y)$$

Thus, knowing $X = x$ does not affect the distribution of Y.

3.5.3 Conditional Expectation

The **conditional expectation** of Y given $X = x$, denoted $\mathbb{E}[Y \mid X = x]$, is the expected value of Y calculated using the conditional distribution of Y given $X = x$. It provides the average or mean value of Y when X is known.

- **Discrete case:**
$$\mathbb{E}[Y \mid X = x] = \sum_y y \cdot P(Y = y \mid X = x)$$
- **Continuous case:**
$$\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)\, dy$$

The conditional expectation is a function of x and can be viewed as the "best guess" of Y given the value of X.

Law of Total Expectation: Let X and Y be random variables (discrete or continuous). Then

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]$$

This means that the overall expectation of X can be obtained by first computing the conditional expectation of X given Y, and then taking the expectation of that conditional expectation over the distribution of Y.

Proof:

- **Discrete case**: Assume X and Y are discrete random variables with joint PMF $p_{X,Y}(x, y)$:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x\, p_X(x) = \sum_x x \sum_y p_{X,Y}(x, y) \\
&= \sum_y \sum_x x\, p_{X,Y}(x, y) = \sum_y \left(\sum_x x\, p_{X|Y}(x|y) \right) p_Y(y) \\
&= \sum_y \mathbb{E}[X \mid Y = y]\, p_Y(y) = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]
\end{aligned}$$

- **Continuous case**: If X, Y are continuous with joint PDF $f_{X,Y}(x, y)$:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y) f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx \right) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y]\, f_Y(y)\, dy = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]
\end{aligned}$$

■

Example: Suppose two factories supply light bulbs to the market.

- Factory X produces bulbs that last on average 5500 hours.
- Factory Y produces bulbs that last on average 4200 hours.
- Factory X supplies 70% of all bulbs, and factory Y supplies 30%.

What is the expected lifetime of a randomly chosen bulb?

Let the random variable L denote the lifetime of a randomly chosen bulb. Let $F \in \{X, Y\}$ be the factory that produced the bulb.

We are asked to find the expected lifetime $\mathbb{E}[L]$.

Given,

$$\mathbb{E}[L \mid F = X] = 2000$$

$$\mathbb{E}[L \mid F = Y] = 3000$$
$$P(F = X) = \frac{70}{100} = 0.7$$
$$P(F = X) = \frac{30}{100} = 0.3$$

Using the **Law of Total Expectation**:

$$\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L \mid F]]$$

We compute:

$$\begin{aligned}\mathbb{E}[L] &= P(F = X) \cdot \mathbb{E}[L \mid F = X] + P(F = Y) \cdot \mathbb{E}[L \mid F = Y] \\ &= 0.7 \times 2000 + 0.3 \times 3000 \\ &= 1400 + 900 = 2300\end{aligned}$$

So, the expected lifetime of a randomly chosen bulb is 2300 hours.

3.6 Functions of a Random Variable

In many practical scenarios, we are interested not just in a random variable X, but in some transformation or function of X, such as $Y = g(X)$. Y is also a random variable. Understanding how the distribution of X affects the distribution of Y is a key part of probability theory.

3.6.1 Discrete Case

If X is a discrete random variable with known probability mass function (PMF) $p_X(x) = P(X = x)$, and $Y = g(X)$, then the PMF of Y is computed as:

$$p_Y(y) = P(Y = y) = \sum_{\{x \mid g(x)=y\}} p_X(x)$$

That is, for each possible value y of Y, sum the probabilities of all values x of X that are mapped to y by the function g. This summation holds for any function g: whether it is one-to-one or many-to-one.

Example: Let X be the outcome of a fair six-sided die, so $X \in \{1, 2, 3, 4, 5, 6\}$ with $P(X = x) = \frac{1}{6}$. Let $Y = X \bmod 2$ (i.e., Y is the parity of X).

Then Y takes values in $\{0, 1\}$, where:

$$f_Y(0) = P(Y = 0) = P(X \in \{2, 4, 6\}) = \frac{3}{6} = 0.5$$
$$f_Y(1) = P(Y = 1) = P(X \in \{1, 3, 5\}) = \frac{3}{6} = 0.5$$

3.6.2 Continuous Case

If X is a continuous random variable with probability density function (PDF) $f_X(x)$, and $Y = g(X)$ is a function of X, then the PDF of Y depends on whether g is monotonic[2] or not.

[2] A function $g(x)$ is called **monotonic** if it never "switches direction" as x moves along its domain. In plain english it is either strictly increasing or strictly decreasing function. If g is strictly increasing or

Monotonic Transformation

Theorem: If g is a strictly monotonic and differentiable function, and $Y = g(X)$, then the PDF of Y is:

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

Or equivalently:

Theorem. If g is a strictly monotonic and differentiable function, and $Y = g(X)$, then the CDF of Y given by:

- If g is strictly increasing, then

$$F_Y(y) = F_X\left(g^{-1}(y)\right)$$

- If g is strictly decreasing, then

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$$

This formula ensures that the total probability is preserved under the transformation.

Proof: We treat separately the cases when g is strictly increasing and strictly decreasing.

- **Case 1: g is strictly increasing.**

  Since g is strictly increasing, it has a well-defined inverse

$$x = g^{-1}(y)$$

  Let $F_X(x)$ and $F_Y(y)$ be the CDFs of X and Y. Then

$$F_Y(y) = P\big(Y \leq y\big) = P\big(g(X) \leq y\big)$$

  Because g is increasing,

$$g(X) \leq y \quad \Longleftrightarrow \quad X \leq g^{-1}(y)$$

  Hence

$$F_Y(y) = P\big(X \leq g^{-1}(y)\big) = F_X\big(g^{-1}(y)\big)$$

  Now differentiate both sides using the chain rule:,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\, F_X\big(g^{-1}(y)\big) = f_X\big(g^{-1}(y)\big)\, \frac{d}{dy}\, g^{-1}(y)$$

  Since g^{-1} is increasing, its derivative is positive i.e. $\dfrac{d}{dy}\, g^{-1}(y) > 0$, and so

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\, \left| \tfrac{d}{dy}\, g^{-1}(y) \right|$$

---

decreasing, then it's **one-to-one and onto**, so there is a well-defined inverse function g^{-1} that 'undoes' g i.e.

$$g^{-1}(g(x)) = x$$

- **Case 2: g is strictly decreasing.**

  Again g^{-1} exists, but now g^{-1} is decreasing. For a decreasing g,

  $$g(X) \leq y \quad \Longleftrightarrow \quad X \geq g^{-1}(y)$$

  Thus

  $$F_Y(y) = P\big(Y \leq y\big) = P\big(X \geq g^{-1}(y)\big) = 1 - P\big(X < g^{-1}(y)\big) = 1 - F_X\big(g^{-1}(y)\big)$$

  Differentiate to get the PDF:

  $$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\Big[\, 1 - F_X\big(g^{-1}(y)\big)\Big] = -f_X\big(g^{-1}(y)\big)\, \frac{d}{dy}\, g^{-1}(y)$$

  Since g^{-1} is decreasing, $\frac{d}{dy}\, g^{-1}(y) < 0$, so the two negatives cancel:

  $$f_Y(y) = f_X\big(g^{-1}(y)\big)\, \Big|\tfrac{d}{dy}\, g^{-1}(y)\Big|$$

In both cases we arrive at the same formula:

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\, \Big|\tfrac{d}{dy}\, g^{-1}(y)\Big|$$

■

Example: Let X has the following PDF:

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

and define $Y = -\log(X)$.

To calculate the PDF of Y, note that $g(x) = -\log(x)$ is strictly decreasing on $(0, 1)$. The inverse function is

$$g^{-1}(y) = e^{-y}$$

The absolute value of the derivative is:

$$\left|\frac{d}{dy} e^{-y}\right| = e^{-y}$$

The PDF of X is $f_X(x) = 1$ for $x \in (0, 1)$, so:

$$f_Y(y) = f_X(e^{-y}) \cdot e^{-y} = 1 \cdot e^{-y} = e^{-y}, \quad y > 0$$

Non-Monotonic Transformation

If g is not monotonic, the formula for $f_Y(y)$ generalizes to:

$$f_Y(y) = \sum_{x \in g^{-1}(y)} \frac{f_X(x)}{|g'(x)|}$$

Here, the sum runs over all x values such that $g(x) = y$ (as g can be a many-to-one function such that multiple values of x can be mapped to same y). The proof is omitted as it is beyond the scope of this text.

Example: Let the random variable X follows a standard normal distribution[3] i.e. $X \sim \mathcal{N}(0,1)$ and define $Y = X^2$. The function $g(x) = x^2$ is not one-to-one, but has two inverse branches: $x = \sqrt{y}$ and $x = -\sqrt{y}$.

Then:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sqrt{y}} \exp\left(-\frac{y}{2}\right) + \frac{1}{\sqrt{y}} \exp\left(-\frac{y}{2}\right) \right) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right), \quad y > 0$$

This is the PDF of a chi-squared distribution with 1 degree of freedom.

3.7 Standardized Random Variable

Let X be a random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$, where $\sigma = \sqrt{\mathrm{Var}(X)}$ must satisfy $\sigma > 0$. The **standardized random variable** X^* is defined by

$$X^* = \frac{X - \mu}{\sigma}$$

By construction, X^* has

$$\text{Mean} = \mathbb{E}[X^*] = \mathbb{E}\left[\frac{X - \mu}{\sigma}\right] = \frac{\mathbb{E}(X) - \mu}{\sigma} = 0,$$

and

$$\text{Variance} = \mathrm{Var}(X^*) = \mathrm{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\,\mathrm{Var}(X) = 1$$

3.8 Chebyshev's Inequality

Chebyshev's Inequality is a fundamental result in probability theory that provides an upper bound on the probability that the value of a random variable deviates from its mean by more than a certain number of standard deviations. It is particularly useful when the distribution of the random variable is unknown.

Chebyshev's Inequality: Let X be a random variable with finite expected value $\mu = \mathbb{E}(X)$ and finite variance $\sigma^2 = \mathrm{Var}(X)$. Then for any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof[4]:

- **Discrete Random Variables**

  Assume X takes values in a countable set (sample space) $S \subset \mathbb{R}$ with probability mass function $p_X(x) = P(X = x)$. The variance of X is given by:

  $$\sigma^2 = \sum_{x \in S} (x - \mu)^2 p_X(x)$$

[3] The discussion on standard normal distribution will be coming in a later section.

[4] An equivalent version of Chebyshev's Inequality states that the probability that X lies within $k\sigma$ is bounded by the inequality:

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Let $A = \{x \in S : |x - \mu| \geq k\sigma\}$ and $\overline{A} = S \setminus A = \{x : |x - \mu| < k\sigma\}$. Then,

$$\sigma^2 = \sum_{x \in A} (x - \mu)^2 p_X(x) + \sum_{x \in \overline{A}} (x - \mu)^2 p_X(x)$$

On the set A, we have $(x - \mu)^2 \geq k^2\sigma^2$, hence

$$\sum_{x \in A} (x - \mu)^2 p_X(x) \geq k^2\sigma^2 \sum_{x \in A} p(x) = k^2\sigma^2 P(|X - \mu| \geq k\sigma)$$

Therefore,

$$\sigma^2 \geq k^2\sigma^2 P(|X - \mu| \geq k\sigma)$$

Dividing both sides by $k^2\sigma^2$ gives:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- **Continuous Random Variables**

  Suppose X is a continuous random variable with probability density function $f_X(x)$. The variance is:

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

  Let $B = \{x \in \mathbb{R} : |x - \mu| \geq k\sigma\}$ and $\overline{B} = \{x : |x - \mu| < k\sigma\}$. Then,

  $$\sigma^2 = \int_B (x - \mu)^2 f_X(x)\, dx + \int_{\overline{B}} (x - \mu)^2 f_X(x)\, dx$$

  On B, we have $(x - \mu)^2 \geq k^2\sigma^2$, so

  $$\int_B (x - \mu)^2 f_X(x)\, dx \geq k^2\sigma^2 \int_B f_X(x)\, dx = k^2\sigma^2 P(|X - \mu| \geq k\sigma)$$

  Thus,

  $$\sigma^2 \geq k^2\sigma^2 P(|X - \mu| \geq k\sigma),$$

  and dividing both sides by $k^2\sigma^2$ yields:

  $$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

■

Properties:

- Chebyshev's Inequality holds for any distribution with finite mean and variance, regardless of its shape.
- The bound provided by the inequality is not always tight; in many cases, the actual probability is much smaller than the upper bound.
- This inequality is particularly useful when the distribution is unknown.

Example: Suppose X is a random variable with mean $\mu = 50$ and standard deviation $\sigma = 5$. Using Chebyshev's Inequality, the probability that X lies outside the interval $[35, 65]$ (which is $\mu \pm 3\sigma$) is bounded by

$$P(|X - 50| \geq 15) \leq \frac{1}{3^2} = \frac{1}{9} \approx 0.1111.$$

Thus, we can say that at least $1 - \frac{1}{9} = \frac{8}{9} \approx 88.89\%$ of the probability mass lies within three standard deviations of the mean.

3.9 Moments and Moment Generating Function

Moments are quantitative measures that capture various aspects of the shape of a probability distribution—its central location, spread, asymmetry, and tail heaviness.

3.9.1 Raw Moments and Central Moments

Moments can be calculated about the origin (raw moments) or about the mean (central moments).

The k**-th raw moment** (also called the moment about the origin) of a random variable X is defined as:

$$\mu'_k = \mathbb{E}[X^k]$$

The k**-th central moment** of a random variable X is defined as:

$$\mu_k = \mathbb{E}[(X-\mu)^k]$$

where $\mu = \mathbb{E}(X)$ is the mean of the distribution.

Moments provide insight into the shape of a distribution. In particular:

- **Mean**: First raw moment:
$$\mu'_1 \;=\; \mathbb{E}(X) \;=\; \mu$$
- **Variance**: The second central moment:
$$\mu_2 \;=\; \mathbb{E}\big[(X-\mu)^2\big] = \sigma^2$$
- **Skewness**: Measures the asymmetry of the distribution. The coefficient of skewness as the third central moment of the standardized random variable $X^* \;=\; \dfrac{X-\mu}{\sigma}$:
$$\gamma_1 \;=\; \mathbb{E}\left[(X^*)^3\right] = \frac{\mathbb{E}\big[(X-\mu)^3\big]}{\sigma^3} \;=\; \frac{\mu_3}{\sigma^3}$$
- **Kurtosis**: Measures the 'tailedness' or 'peakedness' of the distribution. The coefficient of kurtosis is defined as the fourth central moment of the standardized random variable $X^* \;=\; \dfrac{X-\mu}{\sigma}$ minus 3:
$$\gamma_2 \;=\; \mathbb{E}\left[(X^*)^4\right] = \frac{\mathbb{E}\big[(X-\mu)^4\big]}{\sigma^4} - 3 \;=\; \frac{\mu_4}{\sigma^4} - 3$$

3.9.2 Relation Between Raw and Central Moments

Let X be a random variable with mean $\mu = \mathbb{E}(X)$. The k-th central moment of X is defined as:

$$\mu_k = \mathbb{E}[(X-\mu)^k]$$

To relate this to the raw moments $\mu'_r = \mathbb{E}[X^r]$, we expand $(X-\mu)^k$ using the binomial theorem:

$$(X-\mu)^k = \sum_{r=0}^{k} \binom{k}{r} (-\mu)^{k-r} X^r$$

Taking expectations on both sides:

$$\mu_k = \mathbb{E}[(X-\mu)^k] = \mathbb{E}\left[\sum_{r=0}^{k}\binom{k}{r}(-\mu)^{k-r}X^r\right] = \sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}\mu^{k-r}\mathbb{E}[X^r]$$

Hence, the central moment is:

$$\mu_k = \sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}\mu^{k-r}\mu'_r$$

This formula expresses the k-th central moment μ_k as a linear combination of raw moments $\mu'_r = \mathbb{E}[X^r]$ for $r = 0, 1, \ldots, k$.

- **First central moment**:
$$\mu_1 = \mu'_1$$
- **Second central moment**[5]**:**
$$\mu_2 = \mu'_2 - \mu^2$$
- **Third central moment:**
$$\mu_3 = \mu'_3 - 3\mu\mu'_2 + 3\mu^2\mu'_1 - \mu^3$$
- **Fourth central moment:**
$$\mu_4 = \mu'_4 - 4\mu\mu'_3 + 6\mu^2\mu'_2 - 4\mu^3\mu'_1 + \mu^4$$

Now to get the the expression of the raw moment μ'_k in terms of central moments μ_r, we expand $X^k = (\mu + (X-\mu))^k$ as:

$$X^k = \sum_{r=0}^{k}\binom{k}{r}\mu^{k-r}(X-\mu)^r$$

Taking expectation on both sides:

$$\mu'_k = \mathbb{E}[X^k] = \sum_{r=0}^{k}\binom{k}{r}\mu^{k-r}\mathbb{E}[(X-\mu)^r] = \sum_{r=0}^{k}\binom{k}{r}\mu^{k-r}\mu_r$$

$$\mu'_k = \sum_{r=0}^{k}\binom{k}{r}\mu^{k-r}\mu_r$$

This gives the expression of the raw moment μ'_k in terms of central moments μ_r for $r = 0, 1, \ldots, k$, where:

$$\mu_r = \mathbb{E}[(X-\mu)^r], \quad \mu_0 = 1$$

- **First raw moment:**
$$\mu'_1 = \mu$$

[5] An easier way to calculate the second central moment:

$$\mu_2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}(X) + \mu^2 = \mu'_2 - \mu^2$$

- **Second raw moment:**
$$\mu_2' = \mu^2 + \mu_2$$
- **Third raw moment:**
$$\mu_3' = \mu^3 + 3\mu\mu_2 + \mu_3$$
- **Fourth raw moment:**
$$\mu_4' = \mu^4 + 6\mu^2\mu_2 + 4\mu\mu_3 + \mu_4$$

3.9.3 Moment Generating Function

There is a clever way of organizing all the moments into one mathematical object, and that object is called the moment generating function.

The **moment generating function** (MGF) of a random variable X is a function $M_X : \mathbb{R} \to [0, \infty)$ given by
$$M_X(t) \;=\; \mathbb{E}\left(e^{tX}\right)$$
provided the expectation exists in an open neighborhood[a] of $t = 0$.

[a] An **open neighborhood** of $t = 0$ is an open interval around 0, say $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. The condition says that the expectation $\mathbb{E}[e^{tX}]$ must *converge* (i.e., be finite) for all values of t in some interval around 0.

More explicitly, the moment generating function (MGF) of a random variable X can be written as:

- If X is a **discrete random variable** with probability mass function $p_X(x_i) = P(X = x_i)$, then
$$M_X(t) = \sum_{x_i} e^{tx_i} p_X(x_i)$$
- If X is a **continuous random variable** with probability density function $f_X(x)$, then
$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

The method to generate moments is given in the following theorem.

Theorem: Let $M_X(t)$ be the moment generating function (MGF) of a random variable X. Then the kth raw moment μ_k' of X is given by the kth derivative of $M_X(t)$ evaluated at $t = 0$, i.e.,
$$\mu_k' = M_X^{(k)}(0) = \left.\frac{d^k}{dt^k} M_X(t)\right|_{t=0}$$

Proof: Let us start by expanding e^{tX} using its Taylor series about $t = 0$:
$$e^{tX} = \sum_{n=0}^{\infty} \frac{(tX)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n X^n}{n!}$$

Taking expectation on both sides:
$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right)$$

We can interchange summation and expectation[6]:

$$\begin{aligned} M_X(t) &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}(X^n) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mu'_n \\ &= 1 + \frac{t}{1}\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots + \frac{t^k}{k!}\mu'_k + \frac{t^{k+1}}{(k+1)!}\mu'_{k+1} + \dots \end{aligned}$$

This is the Taylor expansion of $M_X(t)$, and by definition of the derivative:

$$\frac{d^k}{dt^k}M_X(t) = \mu'_k + t\mu'_{k+1} + \text{terms with higher orders of } t \dots$$

Thus,

$$\mu'_k = \left.\frac{d^k}{dt^k}M_X(t)\right|_{t=0}$$

■

[6]Provided the series converges absolutely which it does in some neighborhood around $t = 0$ due to the MGF assumption.

Chapter 4

Common Distributions

4.1 Bernoulli Distribution

A **Bernoulli trial** is a random experiment that has exactly two possible outcomes:

1. **Success**, with probability p,

2. **Failure**, with probability $1 - p$.

For any Bernoulli trial, we define a random variable X such that if the experiment results in success, then $X = 1$. Otherwise, $X = 0$. It follows that X is a discrete random variable, with probability mass function $p_X(x)$ defined by

$$p_X(x) = P(X = x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

This can be compactly written as:

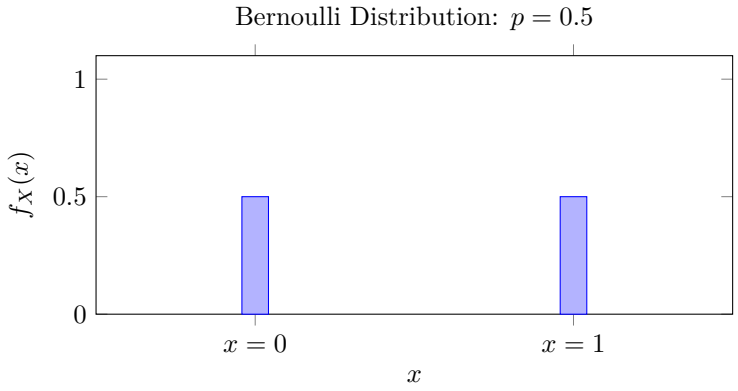
$$p_X(x) = p^x(1 - p)^{1-x}, \quad \text{for } x \in \{0, 1\}$$

The random variable X is said to follow a **Bernoulli distribution** with parameter p, written as:

$$X \sim \text{Bernoulli}(p)$$

Example: In a fair coin toss, the outcomes can be either 'Head' or 'Tail'. If we define a success as getting 'Head', then $p = 0.5$. The PMF of the distribution is then given by

$$f_X(x) = P(X = x) = \begin{cases} 0.5, & \text{if } x = 1 \\ 0.5, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$



4.1.1 Mean and Variance of the Bernoulli Distribution

Let $X \sim \text{Bernoulli}(p)$.

- **Mean**:

$$\mathbb{E}(X) = \sum_x x \cdot P(X = x) = 0 \cdot (1 - p) + 1 \cdot p = p$$

$$\boxed{\mathbb{E}(X) = p}$$

The mean of a Bernoulli distribution is simply the probability of success, p.

- **Variance**:

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

Note that for a Bernoulli variable, $X^2 = X$ (since X is either 0 or 1), so:

$$\mathbb{E}[X^2] = \mathbb{E}(X) = p$$

$$\text{Var}(X) = p - p^2 = p(1 - p)$$

$$\boxed{\text{Var}(X) = p(1 - p)}$$

The variance of a Bernoulli distribution depends on both the probability of success and failure. It is maximum when $p = 0.5$.

4.2 Binomial Distribution

The Binomial distribution arises from repeating a Bernoulli trial independently n number of times, where each trial has the same probability of success p.

Let X denote the number of successes in n independent Bernoulli trials, where each trial has two outcomes: success (with probability p) and failure (with probability $1-p$). Then the discrete random variable X follows a **Binomial distribution** with parameters n and p, written as:

$$X \sim \text{Binomial}(n, p)$$

The probability mass function (PMF) of X is given by:

$$p_X(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

To derive this PMF, consider the following:

- We want the probability of getting exactly x successes (and hence $n - x$ failures) in n trials.
- Each specific sequence of outcomes with x successes and $n-x$ failures has probability:

  $$p^x (1-p)^{n-x}$$

  because of the independence of trials.
- However, there are multiple ways (distinct sequences) to arrange x successes among n trials. The number of such arrangements is given by the binomial coefficient:

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

Thus, the total probability of getting exactly x successes is:

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

This expression defines the PMF of the binomial distribution.

A Binomial random variable is the sum of independent Bernoulli random variables i.e. if

$$X = \sum_{i=1}^{n} X_i, \text{ with } X_i \sim \text{Bernoulli}(p)$$

for all $i = 1, 2, \ldots, n$, then,

$$X \sim \text{Binomial}(n, p)$$

Example: Suppose a fair coin (with $p = 0.5$) is tossed 4 times. Let X be the number of heads observed. Then $X \sim \text{Binomial}(4, 0.5)$. The PMF is:

$$p_X(x) = \binom{4}{x} \times (0.5)^x \times (0.5)^{4-x}, \quad x = 0, 1, 2, 3, 4$$

Evaluating:

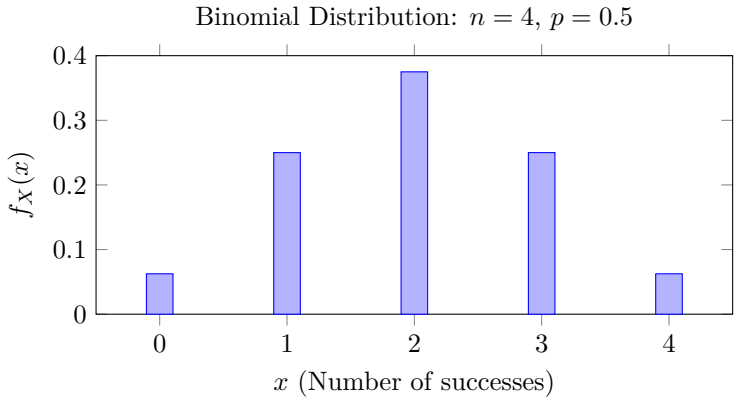
$$p_X(0) = \binom{4}{0} \times (0.5)^0 \times (0.5)^4 = 1 \times 1 \times 0.0625 = 0.0625$$

$$p_X(1) = \binom{4}{1} \times (0.5)^1 \times (0.5)^3 = 4 \times 0.5 \times 0.125 = 0.25$$

$$p_X(2) = \binom{4}{2} \times (0.5)^2 \times (0.5)^2 = 6 \times 0.25 \times 0.25 = 0.375$$

$$p_X(3) = \binom{4}{3} \times (0.5)^3 \times (0.5)^1 = 4 \times 0.125 \times 0.5 = 0.25$$

$$p_X(4) = \binom{4}{4} \times (0.5)^4 \times (0.5)^0 = 1 \times 0.0625 \times 1 = 0.0625$$



Example: A fair six-sided die is rolled 8 times. What is the probability that the number 3 or 4 appears exactly 3 times?

Let a 'success' be defined as getting either a 3 or a 4 in a single roll. The probability of success on one roll is:

$$p = P(3 \text{ or } 4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Let X be the number of successes (i.e., times 3 or 4 occurs) in $n = 8$ independent die rolls. Then X follows a binomial distribution:

$$X \sim \text{Binomial}\left(n = 8, p = \frac{1}{3}\right)$$

We want to find the probability three successes i.e. $X = 3$:

$$P(X = 3) = \binom{8}{3} \times \left(\frac{1}{3}\right)^3 \times \left(\frac{2}{3}\right)^5$$

Now, compute the values:

$$\binom{8}{3} = 56, \quad \left(\frac{1}{3}\right)^3 = \frac{1}{27}, \quad \left(\frac{2}{3}\right)^5 = \frac{32}{243}$$

$$P(X = 3) = 56 \times \frac{1}{27} \times \frac{32}{243} = \frac{1792}{6561} \approx 0.273$$

4.2.1 Mean and Variance of the Binomial Distribution

Let $X \sim \text{Binomial}(n, p)$.

- **Mean**:

  Consider X as the sum of n independent Bernoulli random variables:

  $$X = X_1 + X_2 + \cdots + X_n, \quad \text{where } X_i \sim \text{Bernoulli}(p)$$

  Since expectation is linear:

  $$\mathbb{E}(X) = \mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$$

Each X_i has expected value p, so:

$$\mathbb{E}(X) = n \cdot p$$

Thus on average, we can expect $n \cdot p$ successes in n trials.

- **Variance**:

  Since the X_i's are independent:

  $$\text{Var}(X) = \text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)$$

  Each X_i is Bernoulli with variance $p(1-p)$, so:

  $$\text{Var}(X) = n \cdot p(1-p)$$

  The variance increases with the number of trials and depends on both the probability of success and failure.

4.3 Poisson Distribution

The Poisson distribution is commonly used to model the number of occurrences of an event in a fixed interval of time or space, under the following assumptions:

- Events occur independently.
- The average rate (λ) of occurrence is constant over the interval.
- Two events cannot occur at exactly the same instant.

Let X denote the number of such events occurring in a fixed interval with an average value λ, then we say that the discrete random variable X follows a **Poisson distribution** with parameter $\lambda > 0$, and write:

$$X \sim \text{Poisson}(\lambda)$$

The probability mass function (PMF) of X is given by:

$$p_X(x) = P(X = x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & \text{for } x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases}$$

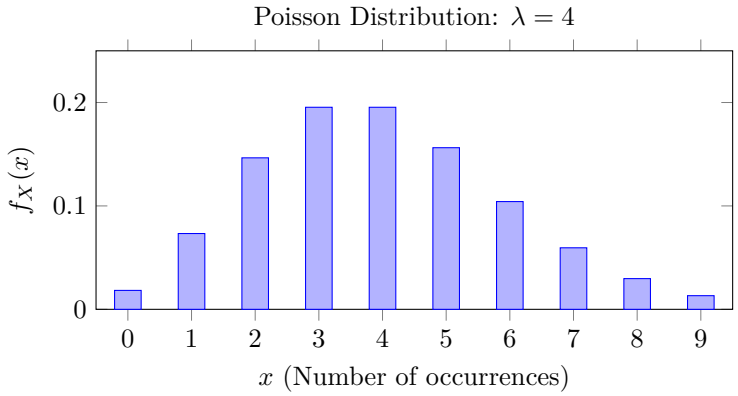
Example: Suppose a call center receives an average of 4 calls per minute. What is the probability that exactly 2 calls are received in a particular minute?

Let X be the number of calls per minute. Then:

$$X \sim \text{Poisson}(\lambda = 4)$$

We want to find $P(X = 2)$:

$$P(X = 2) = \frac{e^{-4} \cdot 4^2}{2!} = \frac{e^{-4} \cdot 16}{2} = 8e^{-4} \approx 0.1465$$



Example: In a football league, the number of goals scored by a team in a match is modeled using a Poisson distribution. Based on historical performance, Team A scores an average of 2.1 goals per match.

1. What is the probability that Team A scores exactly 3 goals in an upcoming match?
2. What is the probability that Team A scores fewer than 2 goals?
3. What is the probability that Team A scores at least 2 goals?
4. What is the expected number of goals Team A will score over their next 5 matches?

The Poisson probability mass function (PMF) is given by:

$$f_X(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{where } \lambda = 2.1, \quad x = 0, 1, 2, \ldots$$

1. Probability that Team A scores exactly 3 goals:

$$f_X(3) = \frac{e^{-2.1} \cdot 2.1^3}{3!} = \frac{e^{-2.1} \cdot 9.261}{6} \approx \frac{0.1225 \cdot 9.261}{6} \approx 0.189$$

2. Probability that Team A scores fewer than 2 goals:

$$P(X < 2) = P(X = 0) + P(X = 1) = f_X(0) + f_X(1)$$

$$f_X(0) = e^{-2.1} \approx 0.1225, \quad f_X(1) = \frac{e^{-2.1} \cdot 2.1}{1!} \approx 0.2573$$

$$P(X < 2) \approx 0.1225 + 0.2573 = 0.3798$$

3. Probability that Team A scores at least 2 goals:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - 0.3798 = 0.6202$$

4. Expected number of goals over 5 matches:

$$5 \times \mathbb{E}(X) = 5 \times \lambda = 5 \times 2.1 = 10.5$$

Theorem: The Poisson distribution can be obtained as the limiting distribution of the Binomial distribution when the number of trials $n \to \infty$, the success probability $p \to 0$, while the expected number of successes $\lambda = np$ remains constant. Formally,

$$\text{Binomial}(n, p) \longrightarrow \text{Poisson}(\lambda) \quad \text{as } n \to \infty, p \to 0 \text{ such that } np = \lambda(\text{constant})$$

Proof: Let $X \sim \text{Binomial}(n, p)$. The probability mass function (PMF) is:

$$p_X(k) = \binom{n}{k} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

We assume $n \to \infty$ and $p \to 0$ such that the product $\lambda = np$ remains fixed and finite. We can write the binomial coefficient as:

$$\begin{aligned}\binom{n}{x} &= \frac{n(n-1)\cdots(n-x+1)}{x!} \\ &= \frac{n^k}{x!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\end{aligned}$$

As $n \to \infty$, each of the product terms approaches 1: Thus,

$$\binom{n}{x} \to \frac{n^x}{x!}$$

Now using this limiting expression of $\binom{n}{x}$ and replacing p with $\dfrac{\lambda}{n}$, we get:

$$p_X(x) \approx \frac{n^x}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Rewrite the last term as:

$$\left(1-\frac{\lambda}{n}\right)^{n-x} = \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x}$$

As $n \to \infty$,

$$\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda} \quad \text{and} \quad \left(1-\frac{\lambda}{n}\right)^{-x} \to 1,$$

since x is fixed.

Therefore in the limit $n \to \infty$:

$$\lim_{n\to\infty} p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda} \cdot 1 = \frac{e^{-\lambda}\lambda^x}{x!}$$

This matches the PMF of the Poisson distribution with parameter λ.

■

4.3.1 Mean and Variance of the Poisson Distribution

Let $X \sim \text{Poisson}(\lambda)$.

- **Mean**:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^{\infty} x \cdot P(X=x) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=1}^{\infty} x \cdot \frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!} \\ &= e^{-\lambda}\cdot\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda\cdot e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \quad (\text{let } k = x-1) \\ &= \lambda\cdot e^{-\lambda}\cdot e^{\lambda} = \lambda\end{aligned}$$

$$\mathbb{E}(X) = \lambda$$

- **Variance**:

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

First we evaluate

$$\mathbb{E}[X^2] = \sum_{x=0}^{\infty} x^2 \cdot \frac{e^{-\lambda}\lambda^x}{x!}$$

We use the identity $x^2 = x(x-1) + x$, giving:

$$\mathbb{E}[X^2] = \sum_{x=0}^{\infty} [x(x-1) + x] \cdot \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \left(\sum_{x=0}^{\infty} \frac{x(x-1)\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{x\lambda^x}{x!} \right)$$

We compute each sum:

$$\sum_{x=0}^{\infty} \frac{x(x-1)\lambda^x}{x!} = \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} = \lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2 e^{\lambda}$$

And,

$$\sum_{x=0}^{\infty} \frac{x\lambda^x}{x!} == \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{\lambda}$$

Therefore,

$$\mathbb{E}[X^2] = e^{-\lambda} \left(\lambda^2 e^{\lambda} + \lambda e^{\lambda} \right) = \lambda^2 + \lambda$$

Therefore:

$$\text{Var}(X) = (\lambda + \lambda^2) - \lambda^2 = \lambda$$

$$\text{Var}(X) = \lambda$$

There is an alternative way of calculating the variance using the moment generating function (MGF).

The moment generating function (MGF) of X is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tx} \cdot \frac{e^{-\lambda}\lambda^x}{x!}$$

Factor out the constant $e^{-\lambda}$:

$$M_X(t) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!}$$

This is the exponential series:

$$\sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{\lambda e^t}$$

Therefore,

$$M_X(t) = e^{-\lambda} \cdot e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

To compute the variance $\text{Var}(X)$, we use the identity:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}(X))^2 = \mu_2' - \mu^2 = \mu_2' - (\mu_1')^2$$

We compute the first and second raw moments using derivatives of the MGF:
First raw moment (mean):

$$M_X'(t) = \frac{d}{dt}\left[e^{\lambda(e^t-1)}\right] = \lambda e^t \cdot e^{\lambda(e^t-1)}$$

Evaluating at $t = 0$:

$$\mu_1' = M_X'(0) = \lambda \cdot 1 \cdot e^{\lambda(1-1)} = \lambda$$

Second raw moment:

$$M_X''(t) = \frac{d}{dt}\left[\lambda e^t \cdot e^{\lambda(e^t-1)}\right] = \lambda e^t\left[\lambda e^t \cdot e^{\lambda(e^t-1)} + e^{\lambda(e^t-1)}\right] = \lambda e^t e^{\lambda(e^t-1)}(\lambda e^t + 1)$$

Evaluating at $t = 0$:

$$\mu_2' = M_X''(0) = \lambda \cdot 1 \cdot 1 \cdot (\lambda \cdot 1 + 1) = \lambda(\lambda + 1) = \lambda^2 + \lambda$$

Now compute the variance:

$$\text{Var}(X) = \mu_2' - (\mu_1')^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

4.4 Uniform Distribution

The Uniform distribution is the simplest continuous probability distribution, where all outcomes in a given interval are equally likely.

Let X be a continuous random variable that is uniformly distributed on the interval $[a, b]$, where $a < b$. This means that X has constant probability density over this interval. We write:

$$X \sim \text{Uniform}(a, b)$$

The probability density function (PDF) of X is given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

That is, the probability density is constant between a and b, and zero elsewhere. The total area under the curve is 1, ensuring it satisfies the definition of a probability density function.

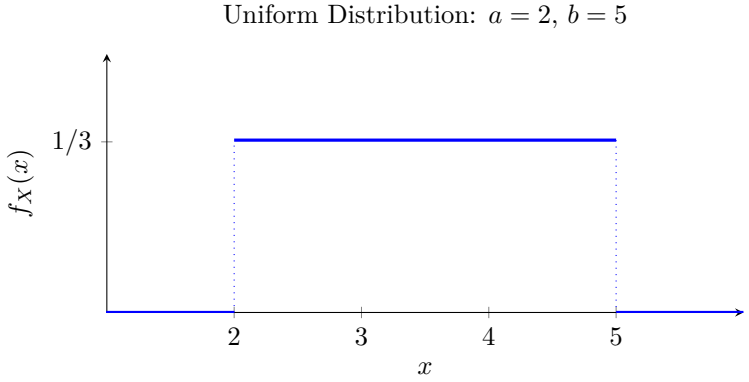
A continuous uniform distribution models situations where every outcome in an interval is equally likely—such as the exact time (within an hour) a bus arrives, or the position of a point randomly placed on a stick.

Example: Suppose that a variable X is uniformly distributed over the interval $[2, 5]$. Then:

$$f_X(x) = \begin{cases} \frac{1}{5-2} = \frac{1}{3}, & 2 \le x \le 5 \\ 0, & \text{otherwise} \end{cases}$$

We can compute probabilities over intervals by integrating the density. For example:

$$P(3 \leq X \leq 4) = \int_3^4 \frac{1}{3} dx = \frac{1}{3}(4-3) = \frac{1}{3}$$



4.4.1 Mean and Variance of the Uniform Distribution

Let $X \sim \text{Uniform}(a, b)$.

- **Mean**:

$$\begin{aligned}
\mathbb{E}(X) &= \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a} \int_a^b x\, dx \\
&= \frac{1}{b-a} \cdot \left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a} \cdot \left(\frac{b^2 - a^2}{2}\right) \\
&= \frac{1}{b-a} \cdot \frac{(b-a)(b+a)}{2} = \frac{a+b}{2}
\end{aligned}$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$

- **Variance**:

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

with:

$$\mathbb{E}(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b = \frac{b^3 - a^3}{3(b-a)}$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Example: For $X \sim \text{Uniform}(2, 5)$:

$$\mathbb{E}(X) = \frac{2+5}{2} = 3.5, \quad \text{Var}(X) = \frac{(5-2)^2}{12} = \frac{9}{12} = 0.75$$

4.5 Normal Distribution

The **Normal distribution**, also known as the **Gaussian distribution**, is one of the most important continuous probability distributions in statistics. It models many naturally occurring phenomena such as heights, test scores, measurement errors, etc. When a continuous random variable X is said to follow a Normal distribution with parameter μ and σ^2, we denote it as:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

The probability density function (PDF) of the Normal distribution is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \text{for } -\infty \leq x \leq \infty$$

The distribution is completely determined by the parameters μ and σ^2. In future, we will show that the mean and variance of the normal distribution are those parameters μ and σ^2 respectively.

Example: Let $X \sim \mathcal{N}(2, 1^2)$, i.e., a normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 1$. We want to compute the value of the probability density function (PDF) at $x = 1.5$.

The PDF of a normal distribution is:

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

Substitute $\mu = 2$, $\sigma = 1$, and $x = 1.5$:

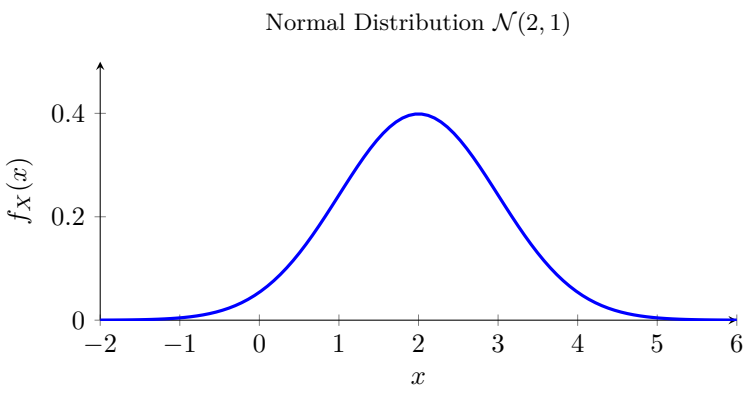
$$f_X(1.5) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1.5-2)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{0.25}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp(-0.125)$$

Now compute numerically:

$$\frac{1}{\sqrt{2\pi}} \approx 0.3989, \quad \exp(-0.125) \approx 0.8825$$

$$f_X(1.5) \approx 0.3989 \times 0.8825 \approx 0.3521$$

The plot of the PDF of $X \sim \mathcal{N}(2, 1)$ is shown in the figure below:



4.5.1 Mean and Variance of the Normal Distribution

Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

- **Mean**:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{-\infty}^{\infty} x\, \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$\text{Let } u = \frac{x-\mu}{\sigma} \quad \Longrightarrow \quad x = \sigma u + \mu, \quad dx = \sigma\, du,$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} (\sigma u + \mu)\, \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\left(-\tfrac{u^2}{2}\right) \sigma\, du$$

$$= \int_{-\infty}^{\infty} (\sigma u + \mu)\, \frac{1}{\sqrt{2\pi}} \exp\!\left(-\tfrac{u^2}{2}\right) du$$

$$= \underbrace{\int_{-\infty}^{\infty} \sigma u\, \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du}_{=0 \text{ (odd integrand)}} + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$

$$= 0 \;+\; \mu \cdot 1 \;=\; \mu.$$

$$\boxed{\mathbb{E}(X) = \mu}$$

- **Variance**:

$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big] = \int_{-\infty}^{\infty} (x-\mu)^2\, f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2\, \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

$$\text{Let } u = \frac{x-\mu}{\sigma} \quad \Longrightarrow \quad x - \mu = \sigma u, \quad dx = \sigma\, du,$$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (\sigma u)^2\, \frac{1}{\sqrt{2\pi}\,\sigma} \exp\!\left(-\tfrac{u^2}{2}\right) \sigma\, du$$

$$= \int_{-\infty}^{\infty} \sigma^2 u^2\, \frac{1}{\sqrt{2\pi}} \exp\!\left(-\tfrac{u^2}{2}\right) du$$

$$= \sigma^2 \underbrace{\int_{-\infty}^{\infty} u^2\, \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du}_{\text{We need to prove this equals 1}}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2\, e^{-u^2/2}\, du$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_{0}^{\infty} u^2\, e^{-u^2/2}\, du$$

Now, using the Gamma integral formula[1],

$$\begin{aligned}\int_0^\infty u^2\, e^{-u^2/2}\, du &= \frac{1}{2}\left(\tfrac{1}{2}\right)^{-\frac{3}{2}}\Gamma\!\left(\tfrac{3}{2}\right)\\ &= \frac{1}{2}\cdot 2^{3/2}\cdot\frac{1}{2}\sqrt{\pi} \quad \left((\tfrac{1}{2})^{-3/2}=2^{3/2},\ \Gamma(\tfrac{3}{2})=\tfrac{1}{2}\Gamma(\tfrac{1}{2})=\tfrac{1}{2}\sqrt{\pi}\right)\\ &= \frac{\sqrt{2\pi}}{2}\end{aligned}$$

Therefore,

$$\mathrm{Var}(X) = \sigma^2\frac{2}{\sqrt{2\pi}}\times\frac{\sqrt{2\pi}}{2} = \sigma^2$$

$$\boxed{\mathrm{Var}(X)=\sigma^2}$$

4.5.2 Moment Generating Function of the Normal Distribution

Let $X \sim N(\mu,\sigma^2)$ be a normal random variable with mean μ and variance σ^2. The moment generating function (MGF) of X is defined as

$$\begin{aligned}M_X(t) = \mathbb{E}\left(e^{tX}\right) &= \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left(tx-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\end{aligned}$$

Rewrite the exponent as:

$$\begin{aligned}tx-\frac{(x-\mu)^2}{2\sigma^2} &= -\frac{1}{2\sigma^2}(x-\mu)^2+tx\\ &= -\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)+tx\\ &= -\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}-\frac{\mu^2}{2\sigma^2}+tx\\ &= -\frac{x^2}{2\sigma^2}+\left(\frac{\mu}{\sigma^2}+t\right)x-\frac{\mu^2}{2\sigma^2} \quad \text{(Group terms involving } x\text{)}\\ &= -\frac{1}{2\sigma^2}\left[x^2-2(\mu+\sigma^2t)x\right]-\frac{\mu^2}{2\sigma^2}\\ &= -\frac{1}{2\sigma^2}\left[(x-(\mu+\sigma^2t))^2-(\mu+\sigma^2t)^2\right]-\frac{\mu^2}{2\sigma^2}\\ &= -\frac{(x-(\mu+\sigma^2t))^2}{2\sigma^2}+\frac{(\mu+\sigma^2t)^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}\\ &= -\frac{(x-(\mu+\sigma^2t))^2}{2\sigma^2}+\frac{\mu^2+2\mu\sigma^2t+\sigma^4t^2-\mu^2}{2\sigma^2}\\ &= -\frac{(x-(\mu+\sigma^2t))^2}{2\sigma^2}+\mu t+\frac{1}{2}\sigma^2t^2\end{aligned}$$

[1]Gamma integral formula:

$$\int_0^\infty x^n e^{-ax^2}dx = \tfrac{1}{2}\,a^{-\frac{n+1}{2}}\Gamma(\tfrac{n+1}{2})$$

$$\Gamma(n+1)=n\Gamma(n),\quad \Gamma(\tfrac{1}{2})=\sqrt{\pi}$$

Substitute this back into the integral for the MGF:

$$M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2} + \mu t + \frac{1}{2}\sigma^2 t^2\right) dx$$

Factor out terms that do not depend on x:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \cdot \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right) dx}_{=1}$$

The integral is the integral of a normal pdf with mean $\mu + \sigma^2 t$ and variance σ^2, and is equal to 1.

Hence,

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

This moment generating function exists for all real values of t and uniquely characterizes the normal distribution.

4.5.3 Properties of the Normal Distribution

1. The **support** of a normal random variable $X \sim N(\mu, \sigma^2)$ is the entire real line:

   $$(-\infty \leq X \leq +\infty)$$

   This reflects that, however unlikely, arbitrarily large positive or negative values can occur.

2. The probability density function is perfectly **symmetric** about its mean μ since,

   $$f_X(\mu + x_0) = f_X(\mu - x_0) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x_0^2}{2\sigma^2}\right) \quad \forall\, x_0 \in \mathbb{R}.$$

   As a result, the left and right tails of the distribution mirror each other.

3. Since the distribution is symmetrical about μ, its mean and median coincide.To get the mode, we need to calculate the peak point of the distribution.

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

   $$\frac{d}{dx} f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \cdot \frac{d}{dx}\left(-\tfrac{(x-\mu)^2}{2\sigma^2}\right)$$

   $$= f_X(x)\left(-\tfrac{2(x-\mu)}{2\sigma^2}\right) = -\frac{x-\mu}{\sigma^2} f_X(x)$$

   Setting the derivative to zero for a stationary point:

   $$-\frac{x-\mu}{\sigma^2} f_X(x) = 0 \quad \Longrightarrow \quad x - \mu = 0$$

   $$\Longrightarrow \quad x = \mu$$

Now,

$$\begin{aligned} f''_X(x) &= -\frac{1}{\sigma^2} f_X(x) + \left(-\frac{x-\mu}{\sigma^2}\right) f'_X(x) \\ &= -\frac{1}{\sigma^2} f_X(x) + \left(-\frac{x-\mu}{\sigma^2}\right)\left(-\frac{x-\mu}{\sigma^2} f_X(x)\right) \\ &= -\frac{1}{\sigma^2} f_X(x) + \frac{(x-\mu)^2}{\sigma^4} f_X(x) \\ &= \frac{(x-\mu)^2 - \sigma^2}{\sigma^4} f_X(x) \end{aligned}$$

Evaluating at the stationary point $x = \mu$:

$$f''_X(\mu) = \frac{(\mu-\mu)^2 - \sigma^2}{\sigma^4} f_X(\mu) = -\frac{1}{\sigma^2} f_X(\mu) < 0,$$

Hence the peak (mode) of the normal density occurs at $x = \mu$.

> For normal distribution, all measures of central tendency coincide:
>
> $$\text{Mean} = \text{Median} = \text{Mode} = \mu.$$

4. The normal distribution curve has two **points of inflection**[2] at a distance σ on either side of μ i.e. at

   $$x = \mu \pm \sigma.$$

   At these points the second derivative of $f_X(x)$ vanishes, marking the transition between "concave down" near the center and "concave up" in the tails. To show that let's take the second derivative of $f_X(x)$ and equate it to zero:

   $$\begin{aligned} f''_X(x) = \frac{(x-\mu)^2 - \sigma^2}{\sigma^4} f_X(x) &= 0 \\ \Rightarrow (x-\mu)^2 &= \sigma^2 \\ \Rightarrow x - \mu &= \pm\sigma \\ \Rightarrow x &= \mu \pm \sigma \end{aligned}$$

5. All odd central moments are zero (due to symmetry), and the even central moments have closed-form expressions:

   $$E\big[(X-\mu)^{2n+1}\big] = 0, \qquad E\big[(X-\mu)^{2n}\big] = \sigma^{2n}\,(2n-1)!!, \quad n = 1, 2, \ldots$$

   In particular, the variance is $E[(X-\mu)^2] = \sigma^2$, and the fourth central moment is $3\sigma^4$, etc.

6. The kurtosis of normal distribution is 3. (prove it)

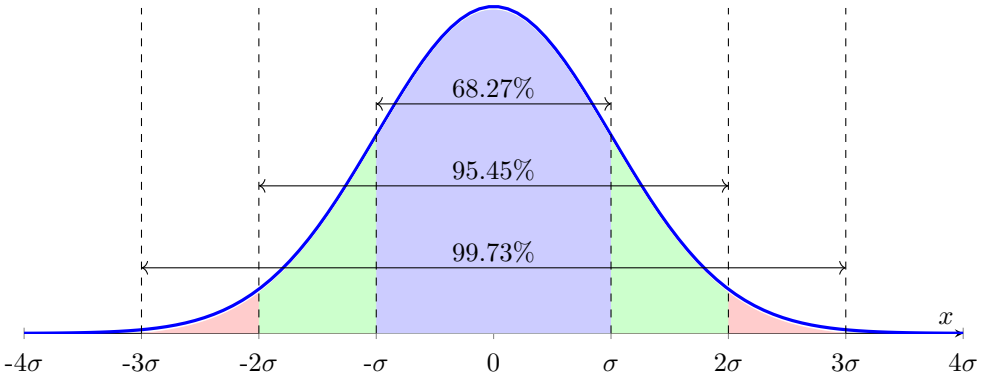
7. A normal distribution $X \sim N(\mu, \sigma^2)$ satisfies the following empirical rules:

   > **Empirical Rule (68-95-99.7 Rule)**:
   >
   > - About 68.27% of the values lie within σ of the mean ($\mu \pm \sigma$).
   > - About 95.45% of the values lie within 2σ of the mean ($\mu \pm 2\sigma$).
   > - About 99.73% of the values lie within 3σ of mean ($\mu \pm 3\sigma$).

[2] A **point of inflection** of a function $f(x)$ is a point $x = a$ such that

$$f''(a) = 0,$$

At the point of inflexion, the second derivative $f''(x)$ changes sign as x passes through a, meaning the curve switches between concave-up and concave-down at $x = a$.



8. The linear combination of two independent normal variables is also normal.

Theorem: Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be two independent normal random variables. Then for any real constants a and b, the linear combination

$$Z = aX + bY$$

is also normally distributed with

$$Z \sim N\left(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2\right)$$

Proof: Since $X \sim N(\mu_1, \sigma_1^2)$, its moment generating function (MGF) is

$$M_X(t) = \exp\left(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right)$$

Similarly, the MGF of $Y \sim N(\mu_2, \sigma_2^2)$ is

$$M_Y(t) = \exp\left(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right)$$

Consider $Z = aX + bY$. Since X and Y are independent, the MGF of Z is:

$$M_Z(t) = \mathbb{E}\left(e^{t(aX+bY)}\right) = \mathbb{E}\left(e^{taX}\right) \cdot \mathbb{E}\left(e^{tbY}\right) = M_X(at) \cdot M_Y(bt)$$

Substituting the MGFs:

$$M_Z(t) = \exp\left(a\mu_1 t + \frac{1}{2}a^2\sigma_1^2 t^2\right) \cdot \exp\left(b\mu_2 t + \frac{1}{2}b^2\sigma_2^2 t^2\right)$$

Combining the exponents:

$$M_Z(t) = \exp\left((a\mu_1 + b\mu_2)t + \frac{1}{2}(a^2\sigma_1^2 + b^2\sigma_2^2)t^2\right)$$

This is the MGF of a normal distribution with mean $a\mu_1 + b\mu_2$ and variance $a^2\sigma_1^2 + b^2\sigma_2^2$. Therefore,

$$Z \sim N\left(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2\right)$$

■

4.5.4 Standard Normal Distribution:

When $\mu = 0$ and $\sigma^2 = 1$, the normal distribution is called the **standard normal distribution**, denoted as:

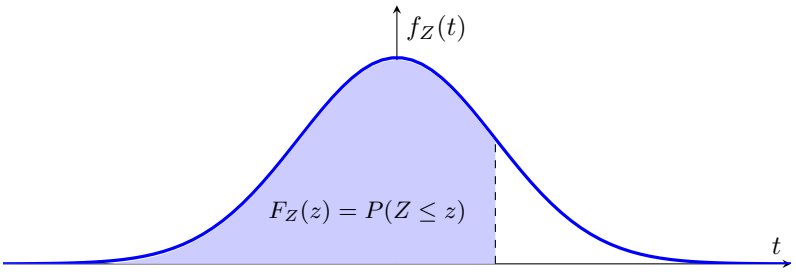
$$Z \sim \mathcal{N}(0, 1)$$

Its PDF becomes:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Cumulative PDF:

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} f_Z(t)dt$$



Theorem. Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable and let $Z \sim N(0, 1)$ be a standard normal random variable. Then, for any real number k,

$$F_X(k) = F_Z\left(\frac{k-\mu}{\sigma}\right)$$

where F_X and F_Z denote the cumulative distribution functions of X and Z, respectively.

Proof: Define the function

$$z = g(x) \;=\; \frac{x-\mu}{\sigma},$$

which is strictly increasing (since $\sigma > 0$) and differentiable, with inverse

$$g^{-1}(y) = x \;=\; \sigma z + \mu$$

Set

$$Y \;=\; g(X) \;=\; \frac{X-\mu}{\sigma}$$

Then Y has the same distribution as Z, i.e. $Y \sim N(0, 1)$. By the change-of-variable theorem for CDFs (strictly increasing case),

$$F_Z(z) \;=\; P\big(Z \leq z\big) \;=\; F_X\big(g^{-1}(z)\big)$$

Hence,

$$F_Z(z) \;=\; F_X\big(\sigma z + \mu\big)$$

Now replace z by $\frac{k-\mu}{\sigma}$. Since $\sigma \cdot \frac{k-\mu}{\sigma} + \mu = k$, we obtain

$$F_Z\left(\frac{k-\mu}{\sigma}\right) = F_X(k)$$

as required.

■

Any normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ can be converted to a **standard normal variable** using the transformation:

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

Theorem: Let $Z \sim N(0,1)$ be a standard normal random variable with CDF $F_Z(z)$. Then, for any real number k,

$$F_Z(-k) = 1 - F_Z(k)$$

Since $f_Z(z)$ is symmetrical about zero, so for any k

$$f_Z(-k) = f_Z(k)$$

Now,

$$\begin{aligned} F_Z(-k) &= P(Z \leq -k) \\ &= \int_{-\infty}^{-k} f_Z(z)dz \end{aligned}$$

Change variable $t = -z$, so when $z = -\infty \to t = +\infty$, and $z = -k \to t = k$, with $dz = -dt$:

$$\begin{aligned} F_Z(-k) &= \int_{\infty}^{k} f_Z(-t)\,(-dt) \\ &= -\int_{\infty}^{k} f_Z(t)\,dt \\ &= \int_{k}^{\infty} f_Z(t)dt \\ &= P(Z \geq k) \\ &= 1 - P(Z < k) \\ &= 1 - F_Z(k) \end{aligned}$$

■

4.5.5 Standard Normal Table

The **standard normal table** (or Z-table) shown in Table 4.1 is used to quickly find cumulative probabilities for the standard normal distribution $Z \sim \mathcal{N}(0,1)$ without evaluating the integral

$$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt$$

by hand. By converting any normal random variable $X \sim N(\mu, \sigma^2)$ into the standard form $Z = (X-\mu)/\sigma$, one can look up probabilities such as $P(X \leq x)$ in a single, universal table, greatly simplifying calculations in statistical inference and hypothesis testing.

To look up $F_Z(k)$, follow the following steps:

1. Write k to **two decimal places**, e.g. $k = 1.23$.

Table 4.1: Standard Normal CDF values $F_Z(z) = P(Z \leq z)$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999
3.1	0.99903	0.99906	0.9991	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.9994	0.99942	0.99944	0.99946	0.99948	0.9995
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.9996	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.9997	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.9998	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.9999	0.9999	0.9999	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

2. Split into

$$\text{row part} = 1.2, \quad \text{column part} = 0.03$$

3. In Table 4.1, go to the row labeled "1.2" and the column labeled "0.03". The entry at their intersection is

$$F_Z(1.23) = P(Z \leq 1.23) \; = \; 0.89065$$

4. For negative z, use symmetry:

$$F_Z(-k) = P(Z \leq -k) = 1 - P(Z \leq k) = 1 - F_Z(k)$$

5. For right-tail probabilities,

$$P(Z > k) = 1 - F_Z(k)$$

6. For probabilities within a specified interval of Z values,

$$P(k_1 \leq Z \leq k_2) = P(Z \leq k_2) - P(Z \leq k_1) = F_Z(k_2) - F_Z(k_1)$$

Example: Suppose the heights of adult males are normally distributed with mean $\mu = 175$ cm and standard deviation $\sigma = 10$ cm. Let X denote the height of a randomly chosen male. Then:

$$X \sim \mathcal{N}(175, 100)$$

What is the probability that a randomly chosen male is taller than 190 cm?

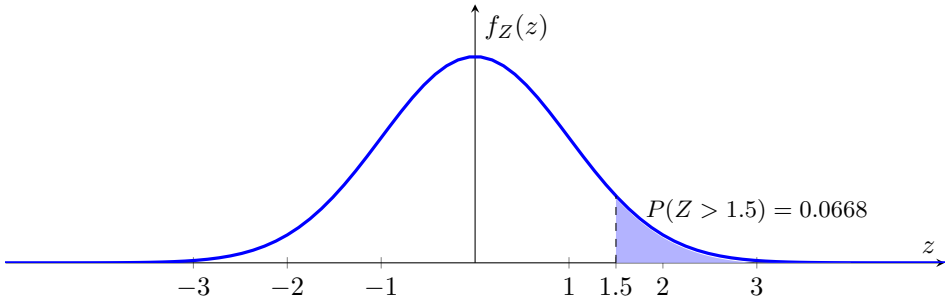
We standardize:

$$Z = \frac{190 - 175}{10} = 1.5$$

Using the standard normal table:

$$P(X > 190) = P(Z > 1.5) = 1 - F_Z(1.5) \approx 1 - 0.9332 = 0.0668$$

Thus, approximately 6.68% of adult males are taller than 190 cm.



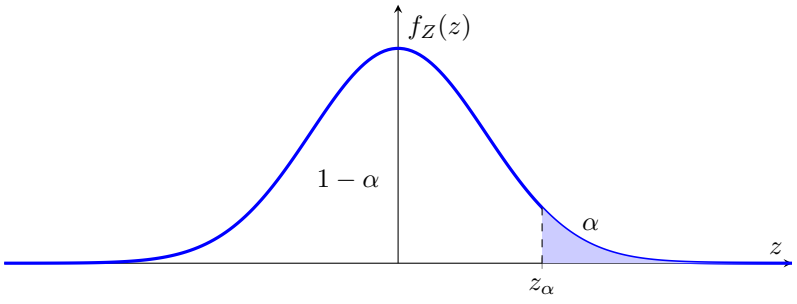
4.5.6 Critical Points of the Standard Normal Distribution

In the standard normal distribution, certain values on the horizontal axis divide the distribution into regions with specified probabilities. These special values are called **critical points**.

One-sided critical point: For a given probability α, the one-sided critical point z_α is such that

$$P(Z > z_\alpha) = \alpha$$

or equivalently $P(Z \leq z_\alpha) = 1 - \alpha$.



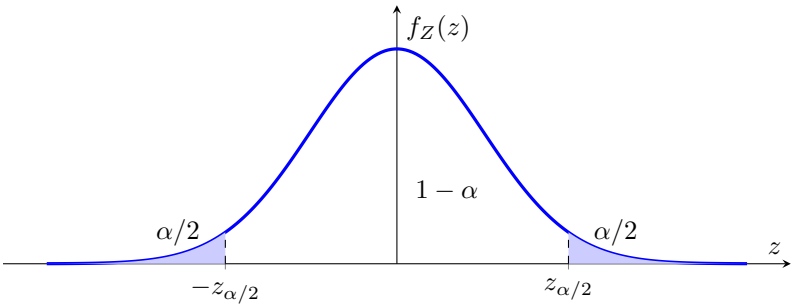
- For $\alpha = 0.05$, the one-sided critical point is approximately $z_{0.05} \approx 1.645$.
- For $\alpha = 0.01$, the one-sided critical point is approximately $z_{0.01} \approx 2.326$.

Two-sided critical points: For a given probability α, the two-sided critical points $\pm z_{\alpha/2}$ are such that

$$P(Z > z_{\alpha/2}) = \alpha/2, \quad P(Z < -z_{\alpha/2}) = \alpha/2$$

or equivalently

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \; = \; 1 - \alpha$$



- For $\alpha = 0.05$, the two-sided critical points are approximately ± 1.96.
- For $\alpha = 0.01$, the two-sided critical points are approximately ± 2.576.

These critical points help us identify how extreme a value is relative to the overall distribution, and they are fundamental when constructing intervals to estimate population parameters. These critical points are used in hypothesis tests and to construct confidence intervals which we will see in future chapters.

Chapter 5

Sampling Theory

5.1 Introduction

Sampling theory is a part of statistics that helps us understand how to learn about a large group by looking at just a small part of it. For example, imagine a company makes a new type of battery and wants to know how long the batteries last. Testing every single battery would take too much time and money, so the company picks a few batteries to test. These few batteries are called a **sample**, and all the batteries made by the company are called the **population**.

The company really wants to know the average lifetime of all batteries—this is called a **parameter**. But since they can't test them all, they use the average lifetime from the sample—this is called a **statistic**.

Sampling theory helps us understand how close this statistic is likely to be to the real average. It also helps us decide how many batteries to test and how to choose them so we get useful, reliable results.

Population: The entire group of individuals or items that we want to learn about. Example: All the batteries produced by a company.

Sample: A smaller group taken from the population, which is actually tested or studied. Example: 100 batteries chosen from the whole production batch.

Parameter: A numerical value that describes a characteristic of the population (usually unknown). Example: The true average lifetime of all the batteries.

Statistic: A numerical value that describes a characteristic of the sample (used to estimate the parameter). Example: The average lifetime of the 100 batteries tested.

5.2 Sampling Methods

5.2.1 Simple Random Sampling

A **simple random sampling** is one in which every member of the population has an equal chance of being selected in the sample.

Mathematically, this means that each possible sample of size n from a population has the same probability of being chosen. For example, suppose a factory produces 10,000 batteries in a day. To estimate the average lifespan, 100 batteries are selected randomly so that each battery has the same chance of inclusion. An important advantage of simple random sampling is that it is straightforward to analyze using statistical theory, which makes inference about the population simpler

In this text, we will limit our discussion to **simple random sampling**. Before a random sample of size n is selected, the observations are modeled as the random variables $X_1, X_2, \ldots, X_n$. For example, if we randomly select 5 light bulbs from a production batch, their lifespans can be represented by the random variables X_1, X_2, X_3, X_4, X_5, each denoting the lifespan (in hours) of a selected bulb.

$$\begin{aligned} X_1 &= \text{Lifespan (in hours) of 1st selected bulb} \\ X_2 &= \text{Lifespan (in hours) of 2nd selected bulb} \\ &\ldots \\ X_5 &= \text{Lifespan (in hours) of 5th selected bulb} \end{aligned}$$

Assume a first draw yields the following lifespans (in hours) for $n = 5$ randomly selected light bulbs:

$$\text{Draw 1:} \quad \{X_1 = 1200, X_2 = 1140, X_3 = 1180, X_4 = 1300, X_5 = 1250\}$$

Because each sample is chosen at random, a fresh draw of five bulbs would almost surely yield different numerical values for $X_1, X_2, \ldots, X_5$. Assume a second draw produces:

$$\text{Draw 2:} \quad \{X_1 = 1400, X_2 = 1550, X_3 = 1200, X_4 = 1420, X_5 = 1380\}$$

In this way, each X_i behaves as a genuine random variable, capturing the uncertainty inherent in the sampling process.

There are two main types of simple random sampling:

1. **Simple Random Sampling With Replacement (SRSWR)**: This is a method of selecting a sample of size n from a population of size N one by one such that after each stage of selection, the element is returned to the population before the next draw. Because each selection is made from the full population, the sample observations $X_1, X_2, \ldots, X_n$ are *independent and identically distributed (i.i.d.)*[1] random variables following the population distribution.
2. **Simple Random Sampling Without Replacement (SRSWOR)**: This is a method of selecting a sample of size n from a population of size N one by one such that after each stage of selection, the element is not returned to the population. So there is no chance of a particular item being selected twice in the sample. Although the sample observations $X_1, X_2, \ldots, X_n$ are identically distributed (each has the same marginal distribution), they are *not independent*, due to the changing composition of the population after each draw.

5.2.2 Other Sampling Methods

1. **Stratified Sampling**: In stratified sampling, the population is divided into distinct subgroups or strata based on a specific characteristic (e.g., age, income, region), and

[1] **Independent and Identically Distributed (i.i.d.)** is a fundamental assumption in statistics. *Identically distributed* means that each random variable X_i follows the same probability distribution (e.g., normal, binomial). *Independent* means the outcome of one observation does not influence or provide information about the others; knowing X_1 gives no information about X_2, X_3, etc.

a random sample is drawn from each stratum. This method ensures representation from all key subgroups.

Example: A company wants to sample employee opinions. Employees are divided into departments (e.g., HR, Sales, R&D), and a random sample is taken from each department.

Advantages: Increases accuracy by reducing variability; ensures important groups are represented.

Disadvantages: Requires knowledge of strata and population characteristics in advance.

2. **Systematic Sampling**: Systematic sampling selects every k-th individual from a population list after a random starting point. The interval k is calculated by dividing the population size by the desired sample size.

   Example: If a company has a list of 1,000 employees and wants to survey 100, it selects a random starting point between 1 and 10, then picks every 10th employee on the list.

   Advantages: Simple and quick to implement; useful when population is ordered.

   Disadvantages: Can introduce bias if there is a hidden pattern in the population that coincides with the sampling interval.

3. **Cluster Sampling**: In cluster sampling, the population is divided into clusters (often based on geography or natural groupings). A few clusters are randomly selected, and then all individuals within those clusters are included in the sample.

   Example: A research team wants to survey households in a city. The city is divided into neighborhoods (clusters), a few neighborhoods are selected at random, and all households in those neighborhoods are surveyed.

   Advantages: Cost-effective and practical for large, dispersed populations.

   Disadvantages: Can lead to higher sampling error if clusters are not homogeneous.

4. **Multistage Sampling**: Multistage sampling combines several sampling techniques. Typically, it begins with cluster sampling to select large groups, and then simple random or stratified sampling is used within those groups.

   Example: In a national education survey, schools are randomly selected (cluster sampling), then students within each selected school are randomly chosen (simple random sampling).

   Advantages: Flexible and practical for large-scale surveys; reduces cost and time.

   Disadvantages: More complex design and analysis; potential for increased sampling error if stages are not carefully planned.

5.3 Sample Mean, Sample Variance and Sample Proportion

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population which are modeled as random variables. The **sample mean** is defined as:

$$\text{Sample Mean} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

It represents the average of the observed sample values.

The **sample variance** is defined as:

$$\text{Sample Variance} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

This measures the spread or variability of the sample values around the sample mean. The denominator $n-1$ (instead of n) ensures that S^2 is an *unbiased estimator* of the population variance σ^2. We will discussed the concept of unbiased estimator in later chapter.

The **sample standard deviation** is the positive square root of the sample variance:

$$\text{Sample Standard Deviation} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Now let's suppose X_i is modeled as a binary indicator variable where

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th observation has the characteristic of interest} \\ 0, & \text{otherwise} \end{cases}$$

The **sample proportion** for the the characteristic of interest is defined as:

$$\text{Sample Proportion} = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X}{n}$$

It represents the fraction of the sample exhibiting the characteristic of interest and serves as an estimator of the population proportion p.

5.4 Sampling Distributions

The value of any statistic (e.g. sample mean) will vary from sample to sample.

> The **sampling distribution** of a statistic is the probability distribution of the statistic's values computed from all possible random samples of the same size taken from a given population.

Suppose a factory produces thousands of batteries, and the lifetimes of these batteries follow a distribution with a population mean $\mu = 100$ hours and a population standard deviation $\sigma = 20$ hours.

Now, imagine taking a random sample of 5 batteries and computing the average lifetime (sample mean). You repeat this process many times—each time taking a new random sample of 5 batteries and calculating its mean. Each of these sample means will be a bit different due to natural variation in the samples. If you plot all these sample means on a graph, the result is the **sampling distribution of the sample mean**.

The standard deviation of the sampling distribution of a statistic is given a specific name — it is called the **standard error** of that sample statistic.

The **standard error** of a sample statistic is the standard deviation of its sampling distribution. It measures how much the statistic is expected to vary from sample to sample due to random chance.

5.5 The Sampling Distribution of the Sample Mean

Theorem: Let $X_1, X_2, \ldots, X_n$ be random samples of size n *chosen with replacements* from a population with mean μ and variance σ^2, then

$$\mathbb{E}(\overline{X}) = \mu, \text{ and } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Proof: We assume the population consists of N elements $\{y_1, y_2, \ldots, y_N\}$. The population mean and population variance are defined as:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} y_j, \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \mu)^2$$

The sample mean is defined as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

By the linearity of expectation:

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$$

Since each X_i is drawn from the population $\{y_1, y_2, \ldots, y_N\}$ each with probability $\frac{1}{N}$. Hence, we have for all i:

$$\mathbb{E}(X_i) = \sum_{j=1}^{N} y_j \cdot \underbrace{(X_i = y_j)}_{1/N} = \frac{1}{N}\sum_{j=1}^{N} y_j = \mu$$

So:

$$\mathbb{E}(\overline{X}) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

Using the formula for the variance of a sum of independent random variables:

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$

Now,

$$\mathrm{Var}(X_i) = \sum_{j=1}^{N}(y_j - \mu)^2 \cdot \underbrace{P(X_i = y_j)}_{1/N} = \frac{1}{N}\sum_{j=1}^{N}(y_j - \mu)^2 = \sigma^2$$

Therefore,

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

■

Theorem: Let $X_1, X_2, \ldots, X_n$ be random samples of size n *chosen without replacement* from a population of size N with mean μ and variance σ^2, then

$$\mathbb{E}(\overline{X}) = \mu, \text{ and } \text{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

We skip the proof as it is beyond the scope of this text. The term

$$\frac{N-n}{N-1}$$

is often called **finite population correction factor**, is close to 1 (and can be omitted for most practical purposes) unless the sample constitutes a substantial portion of the population.

5.5.1 Sampling from a Normal Distribution

In the preceding discussion, no specific assumptions were made about the actual distribution of the population from which the observations $X_1, X_2, \ldots, X_n$ were sampled. Nevertheless, we know two key characteristics of the sampling distribution of the sample mean $\overline{X}$:

- Its expected value: $\mathbb{E}(\overline{X})$
- Its variance: $\text{Var}(\overline{X})$

But what about the shape of the sampling distribution? If the population itself is normally distributed, then the sampling distribution of $\overline{X}$ is also normal, regardless of the sample size.

Theorem: When sampling is done from a normal distribution with mean μ and standard deviation σ, the sample mean $\overline{X}$ follows a normal distribution:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof: Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Define the sample mean as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Since each X_i is normally distributed and is independent, the sample mean $\overline{X}$ is a linear combination of independent normal random variables:

$$\overline{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

A linear combination of independent normal variables is also normally distributed. Therefore, $\overline{X} \sim \mathcal{N}(\mathbb{E}(\overline{X}), \text{Var}(\overline{X})$.

From the previous theorem, we already know,

$$\mathbb{E}[\overline{X}] = \mu \quad \text{and} \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Thus we conclude:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

■

5.5.2 Central Limit Theorem (CLT)

In the previous section, we saw that when we are sampling from a normal distribution, X is also normally distributed. However, there are many situations where we cannot determine the exact form of the distribution of X. In such circumstances, we may appeal to the central limit theorem and obtain an approximate distribution.

Central Limit Theorem: If $\overline{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2, then $\overline{X}$ approximately follows $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ as $n \to \infty$.
In other words, the statistic

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a random variable whose distribution function approaches to that of the standard normal distributions as $n \to \infty$.

Proof: Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = \text{Var}(X_i) < \infty$. Define the standardized sum:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)$$

Then, as $n \to \infty$, we have to prove[2]

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1)$$

Now define

$$Y_i = \frac{X_i - \mu}{\sigma} \quad \text{so that} \quad \mathbb{E}[Y_i] = 0, \quad \text{Var}(Y_i) = 1$$

Then we can write:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Let $M_Y(t)$ be the moment generating function (MGF) of Y_i. Then, since $Y_1, \ldots, Y_n$ are i.i.d., the MGF of Z_n is:

$$M_{Z_n}(t) = \mathbb{E}\left(e^{tZ_n}\right) = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Using a Taylor expansion of $M_Y(t)$ around $t = 0$, we have:

$$M_Y(t) = 1 + \frac{t^2}{2} + \frac{\kappa_3 t^3}{6} + \cdots$$

where κ_3 is the third central moment of Y_i.

[2]The notation $Z_n \xrightarrow{d} Z$ (read as "converges in distribution") means: the distribution of a sequence of random variables Z_n converges to the distribution of another random variable Z.

Substituting $t/\sqrt{n}$ into this expansion:

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{\kappa_3 t^3}{6n^{3/2}} + o\left(\frac{1}{n}\right)$$

Therefore,

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \xrightarrow{n\to\infty} e^{t^2/2}$$

But $e^{t^2/2}$ is the MGF of the standard normal distribution $\mathcal{N}(0,1)$. By the uniqueness theorem for MGFs, this implies:

$$Z_n \xrightarrow{d} \mathcal{N}(0,1)$$

■

The Central Limit Theorem says that even if the population distribution is not normal, the sampling distribution of the sample mean will be approximately normal when the sample size is sufficiently large.

A common question is "how large does n have to be before the normality of $\overline{X}$ is reasonable?" The answer depends on the degree of non-normality of the underlying distribution from which the sample has been drawn. The more non-normal the population distribution is, the larger n needs to be.

> A useful **rule-of-thumb** is that n should be at least 30 for the central limit theorem to take effect.

If the number of observations n increases, the expected value of the sample mean $\overline{X}$ remains fixed at μ, but its variance decreases, approaching zero. In other words,

$$\mathrm{Var}(\overline{X}) = \mathbb{E}\left[(\overline{X} - \mu)^2\right] \to 0 \quad \text{as} \quad n \to \infty.$$

This means the sample mean $\overline{X}$ converges to μ in mean square, since the average squared deviation from the true mean diminishes with larger samples. This result represents one form of the **Law of Large Numbers (LLN)**, which formalizes the idea that

$$\overline{X} \to \mu \quad \text{as} \quad n \to \infty.$$

Together with the Central Limit Theorem, the Law of Large Numbers assures us that the sample mean not only becomes approximately normally distributed but also increasingly concentrates around the true mean μ.

5.5.3 The Sampling Distribution of the Sample Mean When σ is Unknown

In the preceding subsections, we assume that the population variance σ^2 is known. If n is large, this does not pose any problems even when σ is unknown, as it is reasonable in that case to substitute for it the sample standard deviation S. However, for small sample sizes, the distribution of the sample mean $\overline{X}$ is not known unless we assume that the sample comes from a normal population. Under this assumption, one can prove the following:

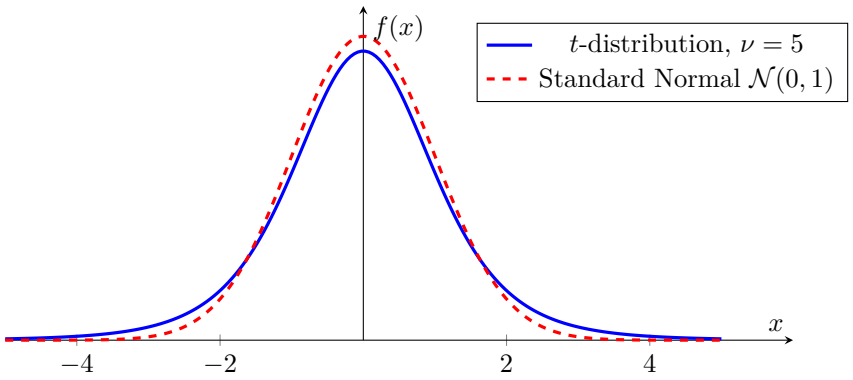
Theorem: Let $\overline{X}$ be the sample mean of a random sample of size n drawn from a normal population with mean μ. Define the sample variance as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then the statistic (standardized sample mean)

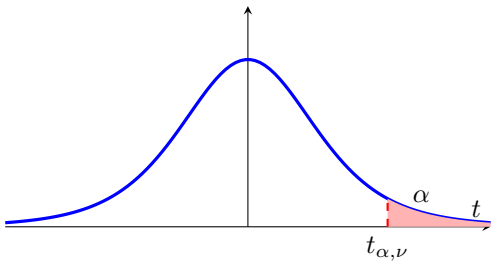
$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a t**-distribution** with $\nu = n - 1$ degrees of freedom.



As illustrated in the figure above, the overall shape of the t-distribution closely resembles that of the standard normal distribution: both are bell-shaped and symmetric about the mean. Like the standard normal distribution, the t-distribution has a mean of 0. However, its variance depends on the parameter ν, known as the **degrees of freedom**. The variance of the t-distribution is greater than 1 but decreases as ν increases, approaching 1 in the limit.

The t-distribution with ν degrees of freedom converges to the standard normal distribution as $\nu \to \infty$. As a general rule of thumb, the standard normal distribution provides a good approximation to the t-distribution when the sample size is 30 or larger.



The critical point $t_{\alpha,\nu}$ is defined as the point so that the area to its right under the t-distribution with ν degrees of freedom equals α i.e.

$$P\left(t > t_{\alpha,\nu}\right) = \alpha$$

By symmetry,

$$t_{1-\alpha,\nu} = -\,t_{\alpha,\nu}$$

So the critical point for a left-tail area of α is $-t_{\alpha,\nu}$.

Example: Suppose we take a random sample of $n = 10$ measurements of battery lifespans (in hours) from a normally distributed population. The data are:

$$\{42,\ 38,\ 41,\ 39,\ 40,\ 37,\ 44,\ 36,\ 38,\ 40\}$$

We want to estimate the population mean μ and test whether the mean battery life is significantly different from 40 hours.

This is a case where the population standard deviation σ is unknown, so we use the sample standard deviation S, and apply the following test statistic:

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

where:

- $\overline{X}$ is the sample mean,
- S is the sample standard deviation,
- $\mu_0 = 40$ is the hypothesized population mean,
- $n = 10$ is the sample size.

The statistic t follows a t-distribution with $\nu = n - 1 = 9$ degrees of freedom under the assumption that the population is normal.

Sample mean:

$$\overline{X} = \frac{1}{10}(42 + 38 + 41 + 39 + 40 + 37 + 44 + 36 + 38 + 40) = \frac{395}{10} = 39.5$$

Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{10}(X_i - \overline{X})^2 = \frac{1}{9}\sum_{i=1}^{10}(X_i - 39.5)^2 \approx 6.17$$

$$S = \sqrt{6.17} \approx 2.48$$

Test statistic:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{39.5 - 40}{2.48/\sqrt{10}} \approx \frac{-0.5}{0.784} \approx -0.637$$

The computed t-value is approximately -0.637, and it follows a t-distribution with $\nu = 9$ degrees of freedom. We can compare this value to critical values from the t-table or compute a p-value to make inference about μ.

5.6 The Sampling Distribution of the Sample Variance

When we take a random sample from a population, not only the sample mean but also the **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

behaves as a random variable. That is, the value of the sample variance S^2 will vary from one sample to another.

Theorem: If S^2 is the variance of a random sample of size n taken (with replacements) from a population of variance σ^2, then

$$\mathbb{E}(S^2) = \sigma^2$$

Proof: Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. The sample variance is defined as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Now,

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}\left((X_i - \mu) - (\overline{X} - \mu)\right)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\end{aligned}$$

Taking expectations on both sides:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n \cdot \mathbb{E}\left[(\overline{X} - \mu)^2\right]$$

Now, observe:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] = \sum_{i=1}^{n}\mathbb{E}(X_i - \mu)^2 = n\sigma^2$$

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

Therefore:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

Dividing both sides by $n-1$, we get:

$$\mathbb{E}(S^2) = \frac{1}{n-1} \cdot (n-1)\sigma^2 = \sigma^2$$

■

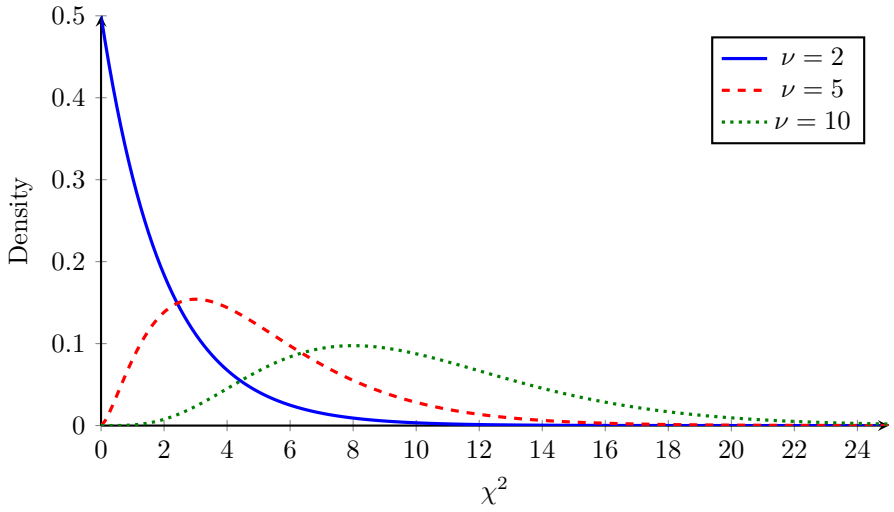
Thus the expectation value of the sample variance is the population variance. This result illustrates why the term $n-1$ is used in the denominator of the definition of sample variance, rather than n. But we still don't know the exact shape of the sampling distribution. To describe the exact sampling distribution of S^2, we require the additional assumption that the population is normally distributed. Under this assumption, the following result holds:

Theorem: If S^2 is the variance of a random sample of size n taken from a normal population of variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows a **chi-squared distribution** with $\nu = n - 1$ degrees of freedom.

This theorem tells us how the sample variance S^2 is distributed around the true population variance σ^2.



A Chi-squared distribution has the following properties:

- The chi-squared distribution is not symmetric; it is skewed to the right, especially for small degrees of freedom.
- As the sample size increases ($n \to \infty$), the distribution becomes more symmetric and approaches normality.
- The expected value of the distribution is $\mathbb{E}[\chi^2] = n - 1$.
- The critical point $\chi^2_{\alpha,\nu}$ is defined as a point such that

  $$P\left(\chi^2_\nu > \chi^2_{\alpha,\nu}\right) \;=\; \alpha$$

  where χ^2_ν denotes a chi-square random variable with ν degrees of freedom.

5.7 Distribution of the Ratio of Two Sample Variances

A problem closely related to that of finding the distribution of the sample variance is that of determining the distribution of the ratio of the variances of two independent random samples. This problem is of considerable importance because it arises in hypothesis testing situations where we want to assess whether two samples come from populations with equal variances.

Theorem: Let S_1^2 and S_2^2 be the sample variances of two independent random samples of sizes n_1 and n_2, respectively, drawn from two normal populations with equal variances σ_1^2 and σ_2^2 respectively. Then the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}$$

follows an F-distribution ($F \sim F_{\nu_1,\nu_2}$) with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom.

This theorem tells us how the ration of sample variances S_1^2/S_2^2 is distributed around the true population variance σ_1^2/σ_2^2.

Properties of F-distribution:

- The F-distribution is characterized by two parameters:
  - ν_1: **numerator degrees of freedom**,
  - ν_2: **denominator degrees of freedom**.
- Reciprocal property:

  If $X \sim F_{\nu_1,\nu_2}$, then the reciprocal of X i.e. $Y = \dfrac{1}{X} \sim F_{\nu_2,\nu_1}$.

- The critical point $F_{\alpha;\nu_1,\nu_2}$ is defined as a point such that

  $$P\big(X > F_{\alpha;\nu_1,\nu_2}\big) \;=\; \alpha$$

  Equivalently[3],

  $$P\left(\frac{1}{X} < \frac{1}{F_{\alpha;\nu_1,\nu_2}}\right) = \alpha$$

  But $\dfrac{1}{X} = Y \sim F_{\nu_2,\nu_1}$. Under the upper-tail critical point definition convention, if $F_{1-\alpha;\nu_2,\nu_1}$ is the critical value satisfying

  $$P\big(Y > F_{1-\alpha;\nu_2,\nu_1}\big) = 1 - \alpha$$

  then equivalently

  $$P\big(Y \leq F_{1-\alpha;\nu_2,\nu_1}\big) = \alpha$$

  Comparison with $P(Y < 1/F_{\alpha;\nu_1,\nu_2}) = \alpha$ shows

  $$\frac{1}{F_{\alpha;\nu_1,\nu_2}} \;=\; F_{1-\alpha;\nu_2,\nu_1}$$

  and hence

  $$F_{\alpha;\nu_1,\nu_2} \;=\; \frac{1}{F_{1-\alpha;\nu_2,\nu_1}}$$

[3]This inverse relationship is maintained because $f(x) = \dfrac{1}{x}$ is a **strictly decreasing bijection** on $(0,\infty)$. Hence the event $\{A < B\}$ is is exactly the same as the event $\{f(A) > f(B)\}$ or the event $\left\{\dfrac{1}{A} > \dfrac{1}{B}\right\}$.

5.8 The Sampling Distribution of the Sample Proportion

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli population with proportion parameter p. Then the **sample proportion** is defined as:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X}{n}$$

where X is the number of successes in the sample of size n. Since X is a binomial random variable with parameters n and p, i.e., $X \sim \text{Bin}(n, p)$, it follows that:

- The mean of $\hat{p}$ is:
$$\mathbb{E}(\hat{p}) = \frac{1}{n}\mathbb{E}(X) = \frac{np}{n} = p$$
- The variance of $\hat{p}$ is:
$$\text{Var}(\hat{p}) = \frac{1}{n^2}\text{Var}(X) = \frac{npq}{n^2} = \frac{pq}{n}, \quad \text{where } q = 1 - p$$

Theorem: For large sample size n, the distribution of sample proportion $\hat{p}$ can be approximated by a normal distribution according to the Central Limit Theorem:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{pq}{n}\right)$$

This approximation is generally considered valid when both $np \geq 5$ and $nq \geq 5$.

The standardized form of $\hat{p}$ is:

$$\frac{\hat{p} - p}{\sqrt{pq/n}}$$

which follows the standard normal distribution $\mathcal{N}(0, 1)$ for large sample.

Example: Suppose the true population proportion is $p = 0.6$ and a sample of size $n = 100$ is taken. Then:

$$\mathbb{E}(\hat{p}) = 0.6$$
$$\text{Var}(\hat{p}) = \frac{0.6 \cdot 0.4}{100} = 0.0024$$

Thus, the distribution of $\hat{p}$ can be approximated as $N(0.6, 0.0024)$.

Chapter 6

Theory of Estimation

6.1 Introduction

Estimation is a fundamental component of **statistical inference**, which deals with drawing conclusions about population parameters from the analysis of sample data. There are two primary types of statistical inference:

1. **Estimation of parameters**: The true value of a population parameter is an unknown constant. The goal of estimation is to make informed guesses about this parameter using sample data, along with an assessment of the accuracy of these guesses.

2. **Hypothesis testing**: Sometimes, preliminary or tentative information about a population parameter is available. The objective of hypothesis testing is to use sample data to either support or reject such information about the parameter.

In this chapter, we focus on the first type—estimation of parameters. Hypothesis testing will be addressed in the next chapter.

Statistical estimation techniques are broadly classified into two categories:

1. **Point estimation**: A point estimation provides a single best guess of the unknown population parameter.

2. **Interval estimation**: An interval estimation gives a range of plausible values for the parameter, along with a specified level of confidence that the interval contains the true value.

Both point and interval estimation play crucial roles in quantifying uncertainty and guiding decision-making in the presence of incomplete information.

6.2 Point Estimation

Let θ ne an unknown parameter (e.g. the population mean) associated with a particular variable. For estimating θ on the basis of random samples $X_1, X_2, \ldots, X_n$, we may use a particular statistic T. This statistic T is called the **point estimator** of θ and the value of T obtained from a given sample is referred to as an **estimate** of θ.

Example: When we estimate the population mean $\theta = \mu$, the most intuitive estimator is the sample mean $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$. Similarly sample variance (S^2) estimates population variance (σ^2) and sample proportion ($\hat{p}$) estimates population proportion (p)

6.2.1 Desirable Properties of a Good Estimator

There are often multiple point estimates available for any given parameter. So it is important to develop some evaluating criteria to judge the performance of each estimator and compare their performance. A good estimator should possess following desirable properties that make it reliable in estimating the true parameter value.

1. **Unbiasedness**:

   An estimator T is said to be an **unbiased** estimator of θ if
   $$\mathbb{E}(T) = \theta$$

   Otherwise, T is said to be biased. The **bias** ($\mathcal{B}$) is given by
   $$\mathcal{B} = \mathbb{E}(T) - \theta$$

   Example: The sample mean $\overline{X}$ and sample variance S^2 are unbiased estimator of the population mean μ and population variance σ^2 respectively, because
   $$\mathbb{E}\left(\overline{X}\right) = \mu, \quad \mathbb{E}\left(S^2\right) = \sigma^2$$

   The **mean-square-error** of the estimator T, denoted by $\text{MSE}(T)$ is defined as
   $$\text{MSE}(T) = \mathbb{E}\left[(T-\theta)^2\right]$$

   MSE measures, on average, how close an estimator comes to the true value of the parameter.

   Theorem: Let T be an estimator of a population parameter θ. Then, the Mean Squared Error (MSE) of T is given by:
   $$\text{MSE}(T) = \text{Var}(T) + \mathcal{B}^2(T)$$

   Proof:
   $$\begin{aligned}
   \text{MSE}(T) &= \mathbb{E}\left[(T-\theta)^2\right] \\
   &= \mathbb{E}\left[((T-\mathbb{E}(T)) + (\mathbb{E}(T)-\theta))^2\right] \\
   &= \mathbb{E}\left[(T-\mathbb{E}(T))^2 + 2(T-\mathbb{E}(T))(\mathbb{E}(T)-\theta) + (\mathbb{E}(T)-\theta)^2\right] \\
   &= \mathbb{E}[(T-\mathbb{E}(T))^2] + \underbrace{2(\mathbb{E}(T)-\mathbb{E}(T))}_{=0}\mathbb{E}\left(T-\mathbb{E}(T)\right) + (\mathbb{E}(T)-\theta)^2 \\
   &= \text{Var}(T) + \mathcal{B}^2(T)
   \end{aligned}$$

   ■

   For an unbiased estimator $\mathcal{B} = 0$, and therefore $\text{MSE}(T) = \text{Var}(T)$.

   In this context, the **standard error (SE)** of T is defined as the standard deviation of T i.e. $\sqrt{\text{Var}(T)}$ which is diffierent from $\text{MSE}(T)$.

2. **Consistency**:

It is desirable that the estimator should behave more and more satisfactorily as the sample size n becomes larger. Consistency provides the criteria.

> An estimator T_n (from a sample of size n) of a parameter θ is said to be **consistent** if, as the sample size n grows, T_n converges in probability to the true parameter value. Which means that for every $\varepsilon > 0$,
>
> $$P\big(|T_n - \theta| > \varepsilon\big) \to 0 \quad \text{as} \quad n \to \infty$$

Consistency as defined above is sometimes called **weak consistency**. If we replace convergence in probability with almost sure convergence, i.e.

$$P\left(\lim_{n\to\infty} T_n - \theta\right) = 1 \quad \text{as} \quad n \to \infty,$$

then the estimator is said to be **strongly consistent**[1].

> **Sufficient Conditions for Consistency**: An estimator T_n of a parameter θ is said to be consistent if it satisfies the following two conditions:
>
> (a) If T_n is an *asymptotically unbiased* estimator of θ i.e.
>
> $$\mathbb{E}(T_n) \to \theta \quad \text{as } n \to \infty$$
>
> (b) The variance of estimator T_n decreases with increasing sample size i.e.
>
> $$\text{Var}(T_n) \to 0 \quad \text{as } n \to \infty$$

Proof: By Chebyshev's inequality, if both conditions hold, then

$$\begin{aligned}
P\big(|T_n - \theta| \geq \varepsilon\big) &\leq \frac{\mathbb{E}(T_n - \theta)^2}{\varepsilon^2}, \quad \text{for every } \varepsilon > 0 \\
&= \frac{1}{\varepsilon^2}\left(\mathbb{E}\left[(T_n - \mathbb{E}(T_n)) + (\mathbb{E}(T_n) - \theta)\right]^2\right) \\
&= \frac{1}{\varepsilon^2}\left(\underbrace{\mathbb{E}\left(T_n - \mathbb{E}(T_n)\right)^2}_{\text{Var}(T_n)} + \left(\mathbb{E}(T_n) - \theta\right)^2\right) \\
&= \frac{1}{\varepsilon^2}\left(\text{Var}(T_n) + \left(\mathbb{E}(T_n) - \theta\right)^2\right) \\
&\to 0 \quad \text{as} \quad n \to \infty
\end{aligned}$$

■

An estimator can be consistent even if it is biased for each finite n, provided the bias vanishes as $n \to \infty$.

Example: Let $X_1, X_2, \ldots, X_n$ be i.i.d. with mean μ and finite variance σ^2. The sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

satisfies

$$\mathbb{E}[\overline{X}_n] = \mu, \qquad \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n} \;\longrightarrow\; 0.$$

[1] **Weak consistency** says "in the long run, most of your estimates will be good," but allows for occasional wildly bad estimates—even when n is very large.

Strong consistency rules out even those rare catastrophes: it guarantees that once you've accumulated enough data, your estimator will stay arbitrarily close to θ for every subsequent sample.

Hence, by the two conditions above, $\overline{X}_n$ is a consistent estimator of μ.

3. **Efficiency**:

> Among all unbiased estimators, the one with the smallest variance is said to be most **efficient**.

Unbiasedness is certainly a desirable property for point estimators but the criterion of unbiasedness does not generally provide a unique statistic for a given problem of estimation. For example, for symmetric population distribution, the sample median is also unbiased estimator for all sample sizes. Clearly, we need a further criterion to decide among different candidates.

One natural refinement is to compare their variances which measures the spread of the sampling distribution. Although both the sample mean and the sample median of a normal population are unbiased and have bell-shaped sampling distributions centered at μ, the variance of the sample mean is

$$\mathrm{Var}\big(\overline{X}\big) = \frac{\sigma^2}{n},$$

whereas the variance of the sample median is approximately

$$\mathrm{Var}\big(X_{\text{median}}\big) \approx 1.5708\,\frac{\sigma^2}{n}.$$

Because the mean's distribution is more concentrated around μ, it will, on average, provide estimates closer to the truth. In other words, among unbiased estimators we favor the one with the smaller variance—and we call it, the most **efficient** estimator.

This leads to the concept of the *Minimum Variance Unbiased Estimator (MVUE)*:

> The unbiased estimator T^* is a **Minimum Variance Unbiased Estimator (MVUE)** of a parameter if it has the smallest variance among all unbiased estimators of the parameter.
> Formally, if $\mathcal{U}$ is the class of all unbiased estimators of θ, then the MVUE T^* satisfies
> $$\mathrm{Var}\big(T^*\big) \;=\; \inf_{T\in\mathcal{U}} \mathrm{Var}\big(T\big)$$

If two unbiased estimators T_1 and T_2 estimate the same parameter θ, the **relative efficiency** of T_1 with respect to T_2 is defined as:

> $$\text{Relative Efficiency} = \frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1)}.$$

An estimator is more efficient if it has a smaller variance. If the relative efficiency is close to 1, both estimators are equally good in terms of variance.

4. **Sufficiency**:

> A statistic is said to be **sufficient** for a parameter if it captures all the information in the sample about that parameter.

Sufficiency is a key concept because it allows us to summarize the data without losing any relevant information about the parameter of interest.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution with conditional joint probability density function (or joint probability mass function)

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta),$$

where θ is an unknown parameter[2].

A statistic $T(\mathbf{X})$ is said to be **sufficient** for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T(\mathbf{X}) = t$ does not depend on θ. That is,

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n \mid T = t; \theta) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n \mid T = t)$$

for all θ.

In other words, once you know $T = t$, the probability (or density) of seeing any particular arrangement of the raw observations does not depends on θ. After you condition on $T = t$, you look at the probability of different possible datasets that all share that same T-value. If that conditional probability still changes with θ, then those leftover items are carrying extra clues about θ.

A useful tool to verify sufficiency is the Neyman–Fisher Factorization Theorem, which states:

Neyman–Fisher Factorization Theorem: A necessary and sufficient condition for the statistic $T(\mathbf{X})$ to be a sufficient statistic for θ is that the joint PDF (or joint PMF) function of the sample can be factorized as:

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta) = g(T; \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

where

- $g(T; \theta)$ is a function that depends on the data $(x_1, x_2, \ldots, x_n)$ only through the function $T(x_1, x_2, \ldots, x_n)$.
- $h(x_1, x_2, \ldots, x_n)$ is a function of the data that does not depend on the parameter θ.

Example: Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with common success probability p. Therefore

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1 - p \end{cases}$$

We wish to show that the total proportion of successes in the sample,

$$T(\mathbf{X}) \;=\; \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a sufficient statistic for p.

[2]**Why we write the joint PDF in the form $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta)$ and not $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n \mid \theta)$?** Because in frequentist statistics, the parameter θ is treated as a fixed (but unknown) constant, while the data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ are considered random variables. Therefore, we write the joint probability density (or mass) function as $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta)$ which emphasizes that this is a function of the data and a given fixed parameter.

On the other hand, the notation $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n \mid \theta)$ is typically reserved for conditional distributions, where θ is treated as a random variable—as in Bayesian statistics. In the frequentist context, θ is not random, so we avoid the conditional notation.

The joint probability mass function of the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n;\, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}\,(1-p)^{\,n-\sum_{i=1}^n x_i}$$

Observe that this can be written in the form

$$f(x_1, x_2, \ldots, x_n;\, p) = \underbrace{p^{nT(\mathbf{x})}\,(1-p)^{\,n-nT(\mathbf{x})}}_{g\left(T(\mathbf{x}),p\right)} \times \underbrace{1}_{h(x)}$$

where

$$T(\mathbf{x}) = \sum_{i=1}^{n} x_i$$

Here:

- $g\big(T(\mathbf{x}); p\big) = p^{T(\mathbf{x})}(1-p)^{\,n-T(\mathbf{x})}$ depends only on the statistic $T(\mathbf{x})$ and the parameter p.
- $h(x) = 1$ depends on the full data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ but not on p.

Therefore, the statistic $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient for the parameter p.

Example: Let $X_1, X_2, \ldots, X_n$ be independent random variables drawn from a normal distribution with unknown mean μ and known variance σ^2. That is,

$$X_i \sim \mathcal{N}(\mu, \sigma^2),$$

so each density is

$$f_{X_i}(x_i; \mu) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\Big(-\tfrac{1}{2\sigma^2}(x_i-\mu)^2\Big)$$

We wish to show that the sample mean

$$T(\mathbf{X}) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a sufficient statistic for μ.

The joint density of the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is

$$\begin{aligned} f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\big(-\tfrac{1}{2\sigma^2}(x_i-\mu)^2\big) \\ &= (2\pi\,\sigma^2)^{-\frac{n}{2}} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

Rewrite the exponential term:

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n} x_i^2 \;-\; 2\mu\sum_{i=1}^{n} x_i \;+\; n\,\mu^2 = \sum_{i=1}^{n} x_i^2 \;-\; n\,\mu^2$$

Thus

$$\exp\left(-\tfrac{1}{2\sigma^2}\sum_{i}(x_i-\mu)^2\right) = \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i} x_i^2\right) \;\times\; \exp\left(\tfrac{\mu}{\sigma^2}\sum_{i} x_i \;-\; \tfrac{n\mu^2}{2\sigma^2}\right)$$

Hence the joint density factors as

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \mu) = \underbrace{\exp\Big(\tfrac{n\mu}{\sigma^2}\, T(\mathbf{x}) \;-\; \tfrac{n\mu^2}{2\sigma^2}\Big)}_{g\left(T(\mathbf{x}),\mu\right)} \;\times\; \underbrace{(2\pi\,\sigma^2)^{-\frac{n}{2}} \exp\left(-\tfrac{1}{2\sigma^2}\sum_i x_i^2\right)}_{h(x)}$$

where $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$. Therefore, by the Neyman–Fisher factorization theorem, because the statistic $T(\mathbf{X}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ is sufficient for μ.

Example: Let X_1 and X_2 be two independent and identically distributed random variables from the Poisson distribution with parameter $\lambda > 0$, i.e.,

$$X_1, X_2 \overset{iid}{\sim} \text{Poisson}(\lambda)$$

The probability mass function of a Poisson random variable is given by

$$P(X_i = x_i; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, \ldots$$

Since X_1 and X_2 are independent, the joint PMF of the sample is

$$p_{X_1,X_2}(x_1, x_2; \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} = \frac{e^{-2\lambda}\lambda^{x_1+x_2}}{x_1!x_2!}$$

Now consider the statistic

$$T = X_1 + 2X_2$$

To apply the Neyman–Fisher Factorization Theorem, we attempt to factor the joint PMF in the form

$$p_{X_1,X_2}(x_1, x_2; \lambda) = g\big(T(x_1, x_2), \lambda\big) \cdot h(x_1, x_2)$$

i.e., express the λ-dependence entirely through the statistic T.

However, in our case the joint PMF is

$$p_{X_1,X_2}(x_1, x_2; \lambda) = \frac{e^{-2\lambda}\lambda^{x_1+x_2}}{x_1!x_2!} = \frac{e^{-2\lambda}\lambda^{T(x_1,x_2)-x_2}}{x_1!x_2!}$$

where the λ-dependent part is $\lambda^{T(x_1,x_2)-x_2}$, not only a function of $T(x_1, x_2)$ but also a function of x_2. Therefore, the statistic $T = X_1 + 2X_2$ is not sufficient.

6.3 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is one of the most widely used methods for estimating the parameters of a statistical model. The basic idea is to choose the parameter values that make the observed data most probable.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a population with joint probability density function (or probability mass function) $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta)$, where θ is the unknown parameter to be estimated. Given $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, it may be looked upon as a function of θ, called the **likelihood function** of θ and is denoted by $L(\theta)$.

$$L(\theta) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta)$$

The value of θ that maximizes this function is called the **maximum likelihood estimator (MLE)** of θ:

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

In practice, it is often more convenient to work with the **log-likelihood function**:

$$\log L(\theta) = \log f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta)$$

Maximizing the log-likelihood yields the same estimator as maximizing the likelihood.

Example: Poisson distribution

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Poisson distribution with parameter $\lambda > 0$. The PMF for the Poisson distribution is:

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The likelihood function is:

$$\begin{aligned} L(\lambda) &= f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \lambda) \\ &= \prod_{i=1}^{n} f(x_i; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\lambda^{\sum_i x_i} \prod_{i=1}^{n} \frac{1}{x_i!} \end{aligned}$$

The log-likelihood is:

$$\log L(\lambda) = -n\lambda + \left(\sum_i x_i\right) \log \lambda - \sum_i \log(x_i!)$$

Differentiating and setting the derivative to zero:

$$\frac{d}{d\lambda} \log L(\lambda) = -n + \frac{1}{\lambda} \sum_i x_i = 0 \Rightarrow \hat{\lambda} = \frac{1}{n} \sum_i x_i = \overline{x}$$

Hence, the MLE of λ is $\hat{\lambda} = \overline{X}$.

Example: Normal distribution

Assume $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The PDF is:

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

Likelihood function:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

Log-likelihood function:

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

(i) **Case 1: μ unknown, σ known ($= \sigma_0$)**

Log-likelihood function:

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Taking derivative and setting to zero:

$$\frac{d}{d\mu}\log L(\mu) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \Rightarrow \hat{\mu} = \overline{x}$$

Hence, the MLE of μ is $\hat{\mu} = \overline{X}$.

(ii) **Case 2: μ known ($= \mu_0$), σ unknown**

Log-likelihood:

$$\log L(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}(x_i - \mu_0)^2$$

Taking derivative w.r.t. σ^2 and setting to zero:

$$\frac{d}{d\sigma^2}\log L(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i}(x_i - \mu_0)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i}(x_i - \mu_0)^2$$

Hence, the MLE of σ^2 is $\hat{\sigma^2} = \frac{1}{n}\sum_i(x_i - \mu_0)^2$.

(iii) **Case 3: μ and σ both unknown**

Log-likelihood:

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}(x_i - \mu)^2$$

Taking partial derivatives and solving the system:

$$\frac{\partial}{\partial\mu}\log L(\mu, \sigma^2) = 0 \Rightarrow \hat{\mu} = \overline{x}$$
$$\frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2) = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i}(x_i - \overline{x})^2$$

Thus, the MLEs for μ and σ^2 are the sample mean $\overline{X}$ and sample variance $\frac{1}{n}\sum_i(x_i - \bar{x})^2$ (without Bessel's correction), respectively.

It is important to note that the maximum likelihood estimator (MLE) of the population variance σ^2 is **not an unbiased estimator**. The MLE is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X}$ is the sample mean. Buit we have already seen that the unbiased estimator of σ^2 is:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}\hat{\sigma}^2$$

The MLE estimator tends to underestimate the true variance.

6.4 Bayesian Estimation

Bayesian estimation is a method of statistical inference in which the unknown parameter θ is modeled as a random variable Θ with a probability distribution $\pi_\Theta(\theta)$, known as the **prior distribution**. It is intended to reflect our knowledge of the parameter θ, before we gather data.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be random variables representing the observations in the sample data with joint PDF (PMF) given $\Theta = \theta$ is given by $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ which is also known as the **likelihood**. When data $\mathbf{X} = \mathbf{x}$ are observed, the extra information about θ is combined with the prior distribution to obtain the **posterior distribution** $\pi_\Theta(\theta \mid \mathbf{x})$ for given $\mathbf{X} = \mathbf{x}$ using Bayes theorem as follows:

$$\pi_\Theta(\theta \mid \mathbf{X}) = \frac{f_{\mathbf{X}}(\mathbf{x} \mid \theta)\pi_\Theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

where,

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \sum\limits_{\theta} f_{\mathbf{X}}(\mathbf{x} \mid \theta)\pi_\Theta(\theta), & \text{in the discrete case,} \\ \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x} \mid \theta)\pi_\Theta(\theta)\, d\theta, & \text{in the continuous case.} \end{cases}$$

Thus,

$$\underbrace{\pi_\Theta(\theta \mid \mathbf{x})}_{\text{posterior}} \propto \underbrace{f_{\mathbf{X}}(\mathbf{x} \mid \theta)}_{\text{likelihood}} \times \underbrace{\pi_\Theta(\theta)}_{\text{prior}}$$

In practice the constant of proportionality chosen in such a way that it makes the total mass of the posterior distribution equal to one. The posterior distribution reflects our updated belief about the parameter after seeing the data.

Example: Suppose you have three coins in your pocket:

- Coin 1: Biased in favour of tails with head probability $\theta = 0.25$
- Coin 2: A fair coin with $\theta = 0.5$
- Coin 3: Biased in favour of heads with $\theta = 0.75$

You randomly select one coin and flip it once. You observe a head. What is the posterior probability that you chose Coin 3?

This is a classic example of Bayesian inference with a discrete parameter space. The **population** consists of three types of coins, each with a different probability of producing a head: $\theta \in \{0.25, 0.5, 0.75\}$.

We assume one of these coins is selected at random. The **sample** is a single coin toss from the selected coin, which results in observing a head. Using this one data point, we update our belief (prior distribution) over the possible values of θ to obtain a posterior distribution.

Let $X = 1$ denote the event that you observe a head, and $X = 0$ for a tail.

Let θ denote the probability of heads. Then $\theta \in \{0.25, 0.5, 0.75\}$.

The **prior probabilities** are:

$$P(\theta = 0.25) = P(\theta = 0.5) = P(\theta = 0.75) = \frac{1}{3}$$

Because the probability of selecting any coin at random is same before we have the knowledge of the sample observation.

The **likelihood** is given by the Bernoulli probability mass function:

$$P(X = x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

Since we observed $X = 1$, the likelihood becomes $P(X = 1 \mid \theta) = \theta$.

We now calculate the unnormalized and normalized **posterior probabilities** using Bayes' Theorem:

$$\underbrace{P(\theta \mid X = 1)}_{\text{posterior}} \quad \propto \quad \underbrace{P(X = 1 \mid \theta)}_{\text{likelihood}} \times \underbrace{P(\theta)}_{\text{prior}}$$

Coin	θ	**Prior** $P(\theta)$	**Likelihood** $P(X = 1 \mid \theta)$	**Unnorm. Posterior** $P(\theta \mid X = 1)$	**Norm. Posterior** $\frac{P(\theta \mid X = 1)}{\sum_\theta P(\theta \mid X = 1)}$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
	Sum	1.00		0.495	1.000

Table 6.1: *Table for calculating the posterior probabilities.*

The posterior probability that the coin chosen was Coin 3, given that a head was observed, is:

$$P(\theta = 0.75 \mid X = 1) = 0.5$$

This illustrates how Bayesian estimation updates our belief about which coin was selected based on the observed outcome.

6.4.1 Bayesian Approach to Point Estimation

When you have your posterior density $\pi_\Theta(\theta \mid \mathbf{X})$, you still need a single "best-guess" $\hat{\theta}$. Bayesian decision theory tells us that the choice of $\hat{\theta}$ depends on a loss function $L(\theta, \hat{\theta})$, which quantifies how "bad" it is to decide $\hat{\theta}$ when the true parameter is θ. When our estimate is $\hat{\theta}$, the **expected posterior loss** is

$$h(\hat{\theta}) = \int_\theta L(\theta, \hat{\theta}) \pi_\Theta(\theta \mid \mathbf{X}) d\theta$$

The Bayes estimator $\hat{\theta}$ minimises the expected posterior loss i.e.

$$\begin{aligned} \hat{\theta} &= \arg\min_{\hat{\theta}} h(\hat{\theta}) \\ &= \arg\min_{\hat{\theta}} \int_\theta L(\theta, \hat{\theta}) \pi_\Theta(\theta \mid \mathbf{X}) d\theta \end{aligned}$$

The form of the minimiser depends on the choice of L. Some common cases are:

1. **Squared-error loss:**

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

The posterior expected loss

$$\begin{aligned} h(\hat{\theta}) &= \int (\theta - \hat{\theta})^2 \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \\ &= \int \left(\theta^2 - 2\hat{\theta} \cdot \theta + \hat{\theta}^2\right) \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \\ &= \int \theta^2 \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \; - \; 2\hat{\theta} \int \theta \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \; + \; \hat{\theta}^2 \int \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \\ &= \mathbb{E}[\Theta^2 \mid \mathbf{X}] \; - \; 2\hat{\theta} \, \mathbb{E}[\Theta \mid \mathbf{X}] \; + \; \hat{\theta}^2, \end{aligned}$$

using $\int \pi(\theta \mid \mathbf{X}) \, d\theta = 1$

To find the minimiser, differentiate $h(\hat{\theta})$ with respect to $\hat{\theta}$:

$$\frac{dh(\hat{\theta})}{d\hat{\theta}} \; = -2 \, \mathbb{E}[\Theta \mid \mathbf{X}] \; + \; 2 \, \hat{\theta}.$$

Setting this derivative to zero gives

$$-2 \, \mathbb{E}[\Theta \mid \mathbf{X}] + 2 \, \hat{\theta} = 0 \quad \Longrightarrow \quad \hat{\theta} = \mathbb{E}[\Theta \mid \mathbf{X}]$$

Finally, check the second derivative:

$$\frac{d^2 h(\hat{\theta})}{d\hat{\theta}^2} \; = 2 > 0,$$

so this critical point is indeed a minimum.

Hence, the Bayes estimator $\hat{\theta}$ under absolute-error loss is the **mean of the posterior distribution**:

$$\hat{\theta} = \mathbb{E}[\Theta \mid \mathbf{X}]$$

2. **Absolute-error loss:**

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

The posterior expected loss

$$h(\hat{\theta}) = \int_{\Theta} |\theta - \hat{\theta}| \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta$$

Split the integral at $\hat{\theta}$:

$$h(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \; + \; \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) \, \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta$$

Differentiate with respect to $\hat{\theta}$. By Leibniz's rule[3]:

$$\frac{d \, h(\hat{\theta})}{d\hat{\theta}} = \int_{-\infty}^{\hat{\theta}} \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta \; - \; \int_{\hat{\theta}}^{\infty} \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta = \Pi_{\Theta}(\hat{\theta}) - \big(1 - \Pi_{\Theta}(\hat{\theta})\big),$$

where $\Pi_{\Theta}(t) = \int_{-\infty}^{t} \pi_{\Theta}(\theta \mid \mathbf{X}) \, d\theta$ is the posterior CDF. Setting this derivative to zero:

$$\Pi_{\Theta}(\hat{\theta}) - \big(1 - \Pi_{\Theta}(\hat{\theta})\big) = 0 \quad \Longrightarrow \quad \Pi_{\Theta}(\hat{\theta}) = \tfrac{1}{2}$$

[3]The **Leibniz integral rule** (differentiation under the integral sign) states that if

$$H(t) = \int_{a(t)}^{b(t)} g(x, t) \, dx,$$

then

$$\frac{d}{dt} \int_{a(t)}^{b(t)} g(x, t) \, dx = g\big(b(t), t\big) \, b'(t) - g\big(a(t), t\big) \, a'(t) + \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} \, g(x, t) \, dx.$$

Hence the Bayes estimator $\hat{\theta}$ under absolute-error loss is the **median (midpoint of the CDF) of the posterior distribution**, satisfying

$$\int_{-\infty}^{\hat{\theta}} \pi_\Theta(\theta \mid \mathbf{X})\, d\theta = \tfrac{1}{2}$$

Finally, the second derivative is

$$\frac{d^2\, h(\hat{\theta})}{d\hat{\theta}^2} = 2\, \pi_\Theta(\hat{\theta} \mid \mathbf{X}) \;\geq\; 0,$$

so $h(\hat{\theta})$ is convex at the solution, confirming that $\hat{\theta}$ indeed minimizes the expected loss.

3. **Zero–one loss:**

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \theta = \hat{\theta}, \\ 1, & \theta \neq \hat{\theta}. \end{cases}$$

The expected posterior loss:

$$h(\hat{\theta}) = \int_\Theta L(\theta, \hat{\theta})\, \pi_\Theta(\theta \mid \mathbf{X})\, d\theta.$$

Since $L(\theta, \hat{\theta}) = 0$ only at $\theta = \hat{\theta}$ and equals 1 elsewhere, we can simplify the expected posterior loss:

$$\begin{aligned} h(\hat{\theta}) &= \int_\Theta L(\theta, \hat{\theta})\, \pi_\Theta(\theta \mid \mathbf{X})\, d\theta \\ &= \int_{\theta \neq \hat{\theta}} \pi_\Theta(\theta \mid \mathbf{X})\, d\theta \\ &= 1 - \pi_\Theta(\hat{\theta} \mid \mathbf{X}) \end{aligned}$$

Therefore,

$$\hat{\theta} = \arg\min_{\hat{\theta}} \Big(1 - \pi_\Theta(\hat{\theta} \mid \mathbf{X})\Big) = \arg\max_{\hat{\theta}} \pi_\Theta(\hat{\theta} \mid \mathbf{X})$$

Hence, the Bayes estimator under zero–one loss is the value of θ that maximizes the posterior density, i.e., the **maximum a posteriori (MAP)** estimator. In other words, it is the **mode of the posterior distribution**.

Example: Suppose we have:

- Observations: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$,
- Prior distribution: $\mu \sim \mathcal{N}(0, \tau^{-2})$

Likelihood:

$$f_{\mathbf{X}}(\mathbf{x} \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2}\right)$$

Prior distribution:

$$\pi_M(\mu) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{\mu^2\tau^2}{2}\right)$$

Then the posterior distribution is given by

$$\pi_M(\mu \mid \mathbf{x}) \propto f_{\mathbf{X}}(\mathbf{x} \mid \mu) \cdot \pi_M(\mu)$$
$$\propto \exp\left(-\frac{1}{2}\sum_i (x_i - \mu)^2\right) \cdot \exp\left(-\frac{\mu^2\tau^2}{2}\right)$$

Now expand the squared term in the sum:

$$\sum_i (x_i - \mu)^2 = \sum_i x_i^2 - 2\mu \sum_i x_i + n\mu^2$$

Ignoring terms not involving μ, we write:

$$\pi_M(\mu \mid \mathbf{x}) \propto \exp\left(\mu \sum_i x_i - \frac{n}{2}\mu^2\right) \cdot \exp\left(-\frac{1}{2}\mu^2\tau^2\right)$$
$$\propto \exp\left[\mu \sum_i x_i - \frac{1}{2}(n + \tau^2)\mu^2\right]$$

Now,

$$\mu \sum_i x_i - \frac{1}{2}(n + \tau^2)\mu^2 = -\frac{1}{2}(n + \tau^2)\left[\mu^2 - \frac{2\sum_i x_i}{n + \tau^2}\mu\right]$$
$$= -\frac{1}{2}(n + \tau^2)\left[\left(\mu - \frac{\sum_i x_i}{n + \tau^2}\right)^2 - \left(\frac{\sum_i x_i}{n + \tau^2}\right)^2\right]$$

Drop the constant term (independent of μ):

$$\pi_M(\mu \mid \mathbf{x}) \propto \exp\left[-\frac{1}{2}(n + \tau^2)\left(\mu - \frac{\sum_i x_i}{n + \tau^2}\right)^2\right]$$

So the posterior distribution of μ given data $\mathbf{x}$ is a Normal distribution with mean and variance given by:

$$\text{Mean} = \frac{\sum_i x_i}{n^2 + \tau^2} = \frac{n\overline{x}}{n^2 + \tau^2}, \quad \text{Variance} = \frac{1}{n^2 + \tau^2}$$

The normal density is symmetric, and so the posterior mean, median and mode have the same value. Thus the optimal Bayes estimate of μ under squared, absolute and zero-one error loss is given by

$$\hat{\theta} = \frac{n\overline{X}}{n^2 + \tau^2}$$

6.5 Estimation using Method of Moments

The **method of moments** is a classical technique used to estimate unknown parameters of a probability distribution using sample data. The core idea is simple: it equates the theoretical moments of a distribution (which depend on the parameters) to the corresponding sample moments computed from the data.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with a distribution that depends on one or more parameters $\theta_1, \theta_2, \ldots, \theta_k$.

- The **r-th population moment** about the origin is defined as:

$$\mu_r' = \mathbb{E}\left(X^r\right)$$

  which is a function of the unknown parameters $\theta_1, \ldots, \theta_k$.

- The **r-th sample moment** about the origin is defined as:

$$m_r' = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

  which is computable from observed data.

The **method of moments estimator** $\hat{\theta}$ is obtained by equating the sample moment to the corresponding population moment. Let's equate the first sample moment to the first population moment:

$$m_1' = \mu(\theta)$$

Solving this equation for θ yields the estimator:

$$\hat{\theta} = \mu^{-1}(m_1')$$

assuming $\mu(\theta)$ is invertible.

This approach can be extended to multiple parameters. If the distribution depends on multiple parameters $\theta_1, \ldots, \theta_k$, then the first k theoretical moments are equated to the first k sample moments:

$$\begin{aligned} m_1' &= \mu_1'(\theta_1, \theta_2, \ldots, \theta_k), \\ m_2' &= \mu_2'(\theta_1, \theta_2, \ldots, \theta_k), \\ &\vdots \\ m_k' &= \mu_k'(\theta_1, \theta_2, \ldots, \theta_k) \end{aligned}$$

Solving these k equations yields the method of moments estimators $\hat{\theta}_1, \theta_2, \ldots, \hat{\theta}_k$.

Example: Exponential Distribution

Suppose $X_1, X_2, \ldots, X_n$ is a sample from an exponential distribution with parameter $\lambda > 0$, having density

$$f_X(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

The first population moment (mean) is:

$$\mu_1' = \mathbb{E}\left(X\right) = \frac{1}{\lambda}$$

The first sample moment is:

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

Equating the moments:

$$\overline{X} = \frac{1}{\lambda} \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{\overline{X}}$$

Hence, the method of moments estimator for λ is:

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

6.6 Interval Estimation

A **point estimate**, being a single number, gives no indication of how accurate or reliable it is. For example, suppose a sample of batteries yields an average lifetime of $\bar{X} = 218.2$ hours. While this provides a best guess for the true average lifetime μ, it says nothing about how close it is to μ. Due to sampling variability, the point estimate is almost never exactly equal to the true value. A more informative approach is to report an **interval estimate**, which provides a range of plausible values for μ and reflects the uncertainty in the estimation process.

Basically, the purpose of interval estimation for a population parameter θ is to find two values L and R from the random sample such that

$$L \leq \theta \leq R$$

with some specific probability. Information about the precision of an interval estimate is conveyed by the width of the interval $R - L$. Because L and R depend on sample values, they will be random. The interval $[L, R]$ should have the following properties:

1. The probability that θ lies within $[L, R]$, i.e., $P(L \leq \theta \leq R)$, should be high.
2. The length of the interval, $R - L$, should be as short as possible to ensure precision.

In addition to providing the interval $[L, R]$, we also specify a measure of confidence in the accuracy of the estimate. This leads to the concept of a **confidence interval (CI)**.

- The **confidence interval** is the interval estimate $[L, R]$ of the parameter θ.
- The **confidence level** is the probability that the confidence interval contains the true value of θ.
- The endpoints L and R are called the **lower** and **upper confidence limits**, respectively.

Definition:

- A **confidence interval** for a parameter is a range of values, computed from sample data, within which the true parameter is believed to lie.
- The **confidence level** is the probability that the confidence interval contains the true parameter value. It is typically chosen close to 1, such as 0.95 or 0.99.

We can write for the interval estimate of θ

$$P(L \leq \theta \leq R) = 1 - \alpha$$

We read this as we are $100(1 - \alpha)\%$ confident that θ lies within the interval $[L, R]$. The interval is called the $100(1 - \alpha)\%$ **confidence interval** or simply the $100(1 - \alpha)\%$ **CI**.

- For a 95% CI, $\alpha = 0.05$,
- For a 99% CI, $\alpha = 0.01$.

A 95% confidence interval means that if we repeated the sampling procedure many times, approximately 95% of the calculated intervals would contain the true parameter value.

6.6.1 Pivotal Method

One of the most widely used methods for constructing confidence intervals is the **pivotal quantity method**, also known as the **pivotal method**. This approach is particularly useful when we can identify a function of the sample data and the parameter, called a *pivotal quantity*, whose distribution does not depend on the unknown parameter.

> A **pivotal quantity** is a function $T(X_1, X_2, \ldots, X_n; \theta)$ of the sample data and the parameter θ, such that the probability distribution of T is independent of θ.

The steps to construct a confidence interval using the pivotal method are as follows:

1. Identify a suitable pivotal quantity $T(X_1, \ldots, X_n; \theta)$ whose distribution is known and does not depend on θ.
2. Find constants a and b such that

   $$P(a \leq T(X_1, \ldots, X_n; \theta) \leq b) = 1 - \alpha$$

   where $1-\alpha$ is the desired confidence level. The constants a and b are called **critical values**.
3. Solve the inequality to find the interval

   $$a \leq T(X_1, \ldots, X_n; \theta) \leq b$$

   for θ in terms of the sample data.
4. The resulting interval gives the $100(1-\alpha)\%$ confidence interval for θ.

6.6.2 Confidence Interval for the Mean in a Normal Population (Known Variance)

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, where the variance σ^2 is known. The sample mean $\overline{X}$ follows

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The pivotal quantity

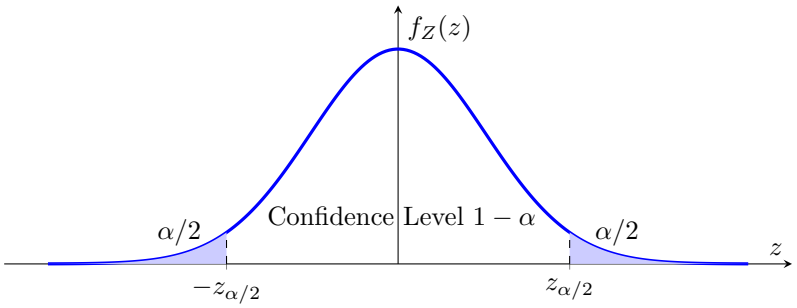
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution, $Z \sim \mathcal{N}(0, 1)$, independent of the unknown parameter μ.

To find the confidence interval, we have to find critical values a and b such that

$$P\left(a \leq Z \leq b\right) = 1 - \alpha$$

Since the distribution is symmetric about zero, we can choose the critical values $a = -q, b = q$. For standard normal distribution, $q = z_{\alpha/2}$ represents the value of z with tail area $\alpha/2$.



Thus,

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1-\alpha$$

which translates to

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Therefore, the $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- For a 95% confidence level, $1-\alpha = 0.95$. From table $z_{\alpha/2} = z_{0.025} = 1.96$.
- For a 99% confidence level, $1-\alpha = 0.99$. From table $z_{\alpha/2} = z_{0.005} = 2.576$.

Example: Suppose the lifetimes of a certain type of battery are normally distributed with unknown mean μ and known standard deviation $\sigma = 10$ hours. A random sample of $n = 25$ batteries yields a sample mean of $\overline{X} = 100$ hours. Construct a 95% confidence interval for the population mean lifetime.

- Given: $\sigma = 10$, $n = 25$, $\overline{X} = 100$, and confidence level = 95%
- For 95% confidence, $z_{\alpha/2} = z_{0.025} = 1.96$

Compute the standard error:

$$\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Compute the margin of error:

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \times 2 = 3.92$$

Construct the confidence interval:

$$[100 - 3.92, \quad 100 + 3.92] = [96.08,\ 103.92]$$

We are 95% confident that the true mean lifetime μ of the batteries lies between 96.08 and 103.92 hours.

6.6.3 Confidence Interval for the Mean in a Normal Population (Unknown Variance)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Let $\overline{X}$ be the sample mean and S^2 be sample variance.

The pivotal quantity

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom:

$$T \sim t_{n-1}$$

Let $t_{\alpha/2}, n-1$ be the critical value of the t-distribution such that $P(t > t_{\alpha/2,n-1}) = \alpha/2$. Hence, we can write

$$P\left(-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}\right) = 1 - \alpha$$

which implies

$$P\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Therefore, the $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \quad \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$$

6.6.4 Confidence Interval for the Variance in a Normal Population

Consider a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$ with unknown variance σ^2. Let S^2 be sample variance.

The pivotal quantity

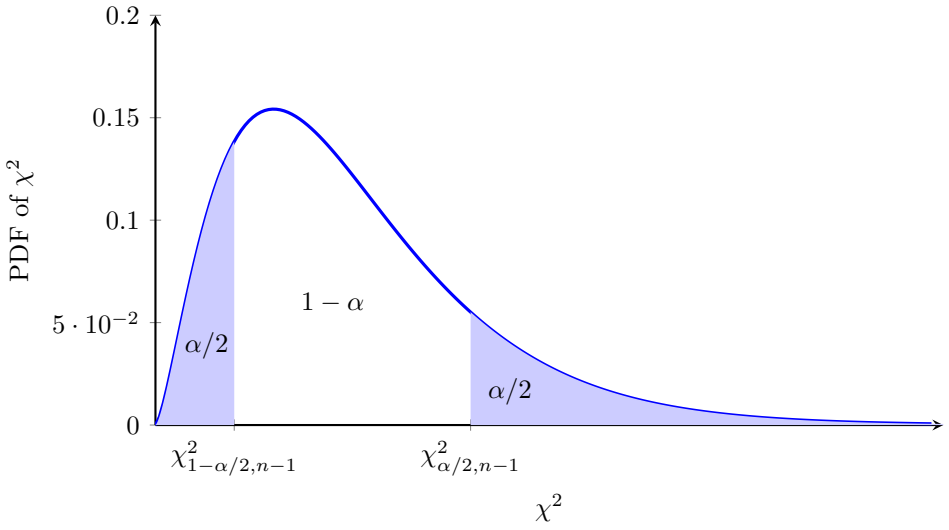
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows a chi-square distribution with $\nu = n - 1$ degrees of freedom:

$$\chi^2 \sim \chi^2_{n-1}$$

We now need to find the critical values L and U such that

$$P\left(L \leq \frac{(n-1)S^2}{\sigma^2} \leq U\right) = 1 - \alpha$$



We take areas to the left of L and to the right of R to be equal to $\alpha/2$ i.e.

$$L = \chi^2_{1-\alpha/2,n-1}, \quad R = \chi^2_{\alpha/2,n-1}$$

Thus,

$$P\left(\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha$$

Rearranging to solve for σ^2, we get

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha$$

Therefore, the $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \quad \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right]$$

6.6.5 Confidence Interval for the Difference of Two Normally Distributed Population Means

When comparing the means of two normally distributed populations, it is often of interest to estimate the difference between the population means, say $\mu_1 - \mu_2$, using data from independent random samples from each population.

Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample from population 1 with mean μ_1 and variance σ_1^2, and $Y_1, Y_2, \ldots, Y_{n_2}$ be a random sample from population 2 with mean μ_2 and variance σ_2^2. Let $\overline{X}$ and $\overline{Y}$ denote the respective sample means, and S_1^2, S_2^2 the respective sample variances.

We consider two cases:

(i) Population Variances Known

Assume that σ_1^2 and σ_2^2 are known. Then the sampling distribution of $\overline{X} - \overline{Y}$ is normal:

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Define the pivotal quantity:

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Then, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$\left[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

(ii) Population Variances Unknown

Assume the populations are normally distributed but σ_1^2 and σ_2^2 are unknown. The confidence interval depends on whether variances can be assumed equal.

(a) **Equal but Unknown Variances:**
Assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and estimate with pooled variance:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Then, the pivotal quantity

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Thus, the $100(1-\alpha)\%$ confidence limits is:

$$(\overline{X} - \overline{Y}) \pm t_{\alpha/2,\, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(b) **Unequal and Unknown Variances (Welch's Approximation):**
If equal variance cannot be assumed, use:

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

This approximately follows a t-distribution with degrees of freedom:

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Then, the approximate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$\left[(\overline{X}-\overline{Y}) - t_{\alpha/2,\nu}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\ (\overline{X}-\overline{Y}) + t_{\alpha/2,\nu}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$$

6.6.6 Confidence Interval for the Ratio of Two Normally Distributed Population Variances

Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample from population 1 with mean μ_1 and variance σ_1^2, and $Y_1, Y_2, \ldots, Y_{n_2}$ be a random sample from population 2 with mean μ_2 and variance σ_2^2. Let S_1^2, S_2^2 denote the respective sample variances.

The pivotal quantity

$$F \;=\; \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}$$

follows an F–distribution with (n_1-1, n_2-1) degrees of freedom.

$$F \sim F_{n_1-1,\ n_2-1}$$

Let $F_{1-\alpha/2;\,n_1-1,n_2-1}$ and $F_{\alpha/2;\,n_1-1,n_2-1}$ be the lower and upper $\alpha/2$–quantiles of this F–distribution:

$$P\big(F_{1-\alpha/2;\,n_1-1,n_2-1} \le F \le F_{\alpha/2;\,n_1-1,n_2-1}\big) \;=\; 1-\alpha$$

or,

$$P\left(F_{1-\alpha/2;\,n_1-1,n_2-1} \le \frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \le F_{\alpha/2;\,n_1-1,n_2-1}\right) \;=\; 1-\alpha$$

Rewriting in terms of σ_1^2/σ_2^2 gives

$$P\left(\frac{1}{F_{\alpha/2;\,n_1-1,n_2-1}}\frac{S_1^2}{S_2^2} \;\le\; \frac{\sigma_1^2}{\sigma_2^2} \;\le\; \frac{1}{F_{1-\alpha/2;\,n_1-1,n_2-1}}\frac{S_1^2}{S_2^2}\right) = 1-\alpha$$

Since for F-distribution,

$$F_{1-\alpha/2;\,n_1-1,n_2-1} = \frac{1}{F_{\alpha/2;\,n_2-1,n_1-1}}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for the ratio σ_1^2/σ_2^2 is

$$\left[\frac{1}{F_{\alpha/2;\,n_1-1,n_2-1}}\frac{S_1^2}{S_2^2}\ ,\ F_{\alpha/2;\,n_2-1,n_1-1}\frac{S_1^2}{S_2^2}\right]$$

6.6.7 Confidence Interval for a Population Proportion

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli population with unknown proportion parameter p. Define the sample proportion as

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X}{n}$$

where X is the number of successes in the sample of size n.

For large n, the sampling distribution of $\hat{p}$ is approximately normal by the Central Limit Theorem:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Since p is unknown, we approximate the standard deviation using $\hat{p}$, leading to the following pivotal quantity:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

For sufficiently large n, Z approximately follows the standard normal distribution:

$$Z \sim N(0, 1)$$

This approximation is generally considered valid when both $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Let $z_{\alpha/2}$ be the critical value from the standard normal distribution such that

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

Then,

$$P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha$$

This implies

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

Therefore, the $100(1-\alpha)\%$ confidence interval for the population proportion p is

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Example: A political poll surveys $n = 400$ registered voters to estimate the proportion who support Proposition X. Out of the 400 respondents, $X = 180$ say they support the measure. Construct a 95% confidence interval for the true support proportion p.

- Sample size: $n = 400$
- Number of successes: $X = 180$
- Sample proportion:
$$\hat{p} = \frac{X}{n} = \frac{180}{400} = 0.45$$
- Check normal approximation conditions:
$$n\hat{p} = 400 \times 0.45 = 180 \geq 5, \quad n(1-\hat{p}) = 400 \times 0.55 = 220 \geq 5.$$
- For 95% confidence, $z_{\alpha/2} = z_{0.025} = 1.96$.

Compute the margin of error:

$$z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.45 \times 0.55}{400}} \approx 1.96 \times 0.0249 \approx 0.0488$$

Thus the 95% confidence interval for p is

$$\big[\,0.45 - 0.0488,\ 0.45 + 0.0488\big] \;=\; [\,0.4012,\ 0.4988\,]$$

We are 95% confident that the true proportion of all registered voters who support Proposition X lies between 40.12% and 49.88%.

6.7 Determination of Sample Size

Sample size determination is a fundamental aspect of planning any statistical study. If the sample is too small, the results may not be trustworthy. But if the sample is too large, it can waste time, money, and effort. The goal is to find a sample size that is just right: big enough to give accurate and useful results, but not bigger than necessary. There is no universal ideal sample size for all problems. The required sample size depends on:

- The **parameter** being estimated,
- A specified **margin of error (E)**,
- A desired **confidence level** (commonly 90%, 95%, or 99%).

6.7.1 Determining Sample Size for Estimating a Population Mean

Let μ denote the population mean and σ the population standard deviation (assumed known). To estimate μ within a margin of error E at confidence level $1-\alpha$, the sample size n must satisfy:

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq E$$

Solving for n:

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Example: A researcher wants to estimate the average height of adult males in a city with a 95% confidence level and a margin of error of 2 cm. Suppose previous studies suggest $\sigma = 7$ cm.

$$z_{\alpha/2} = z_{0.025} = 1.96$$
$$n \geq \left(\frac{1.96 \cdot 7}{2}\right)^2 = \left(\frac{13.72}{2}\right)^2 = (6.86)^2 = 47.06$$

Thus, a sample size of at least 48 individuals is needed.

6.7.2 Determining Sample Size for Estimating a Population Proportion

Suppose the goal is to estimate a population proportion p. The margin of error in this case must satisfy:

$$z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq E$$

Solving for n:

$$n \geq \left(\frac{z_{\alpha/2}}{E}\right)^2 \cdot p(1-p)$$

If the value of p is known roughly in the neighbourhood of a value p^*, then n can be determined as:

$$n \geq \left(\frac{z_{\alpha/2}}{E}\right)^2 \cdot p^*(1-p^*)$$

Without prior knowledge of p, $p(1-p)$ can be replaced by its maximum possible value[4] $\frac{1}{4}$ corresponding to $p = \frac{1}{2}$ and n is determined from the relation:

$$n \geq \frac{1}{4}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Example: A political analyst wants to estimate the proportion of voters who support a candidate with 95% confidence and a margin of error of 5%. If no prior estimate for p is available, the conservative choice is $p = 0.5$:

$$z_{\alpha/2} = z_{0.025} = 1.96, \quad E = 0.05, \quad p = 0.5$$
$$n \geq \left(\frac{1.96}{0.05}\right)^2 \times 0.5 \times 0.5 = 1536.64 \times 0.25 = 384.16$$

Thus, a sample size of at least 385 respondents is needed.

6.7.3 Finite Population Correction (FPC)

When sampling without replacement from a finite population of size N, and the sample size n exceeds 5% of the population ($n > 0.05N$), the effective sample size should be adjusted using the finite population correction (FPC):

$$n_{\text{adj}} = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

where n_0 is the sample size calculated assuming an infinite population.

[4]To calculate the maximum value of $f = p(1-p) = p - p^2$, we need to differentiate it w.r.t. p and set it to zero.

$$\frac{df}{dp} = 1 - 2p = 0 \quad \Rightarrow p = \frac{1}{2}$$
$$\frac{d^2f}{dp^2} = -2$$

The double derivative is negtive indicating f is maximum at $p = \frac{1}{2}$.
The maximum value of f is

$$f_{max} = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

To prove this formula let us rewrite the formula for the variance of the sample mean:

- Finite population (without replacement):

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}$$

- Infinite population (or with replacement):

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

To achieve the same precision (equal variances), set

$$\frac{\sigma^2}{n_0} = \frac{\sigma^2}{n_{\text{adj}}} \times \frac{N-n_{\text{adj}}}{N-1}$$

where n_0 is the sample size required to achieve the variance for the infinite population and n_{adj} is the sample size required to achieve the variance for the finite population.

Cancel σ^2 and rearrange:

$$\begin{aligned}
\frac{1}{n_0} &= \frac{1}{n_{\text{adj}}} \times \frac{N-n_{\text{adj}}}{N-1} \\
\Rightarrow \quad n_{\text{adj}} \cdot \frac{N-1}{n_0} &= N - n_{\text{adj}} \\
\Rightarrow \quad n_{\text{adj}} \left(1 + \frac{N-1}{n_0}\right) &= N \\
\Rightarrow \quad n_{\text{adj}}(n_0 + N - 1) &= N n_0
\end{aligned}$$

Hence, the adjusted sample size is

$$n_{\text{adj}} = \frac{N n_0}{n_0 + N - 1} = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

Example: From the earlier example, if the population consists of only 2,000 individuals:

$$n_{\text{adj}} = \frac{384.16}{1 + \frac{384.16-1}{2000}} = \frac{384.16}{1 + 0.1916} = \frac{384.16}{1.1916} \approx 322.4$$

Thus, only about 323 individuals are needed when the population is finite.

Table 6.2: Important Confidence Interval Formulas

Parameter	**Assumptions**	**Confidence Interval Formula**
Population mean μ (known σ)	Normal population or large n (CLT)	$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
Population mean μ (unknown σ)	Normal population	$\bar{X} \pm t_{\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}}$
Difference of means $\mu_1 - \mu_2$ (equal variances)	Normal populations, independent samples	$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,\, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
Proportion p	Large n, with $np \geq 5$, $n(1-p) \geq 5$	$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference of proportions $p_1 - p_2$	Large n_1, n_2	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Population variance σ^2	Normal population	$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right)$
Ratio of variances $\frac{\sigma_1^2}{\sigma_2^2}$	Normal, independent samples	$\left(\frac{s_1^2/s_2^2}{F_{1-\alpha/2}}, \frac{s_1^2/s_2^2}{F_{\alpha/2}} \right)$

Chapter 7

Test of Hypothesis

7.1 What is Hypothesis Testing?

There are many problems in which, rather than estimating the value of a parameter, we must decide whether a statement concerning a parameter is true or false. Statistically speaking, we test a hypothesis about a parameter.

> A **statistical hypothesis** is a statement about the parameter of a population.

For example, a factory produces screws that are supposed to be 5 cm long. A quality inspector takes a sample of 50 screws and finds that the average length is 4.8 cm. Here, the hypothesis is the statement:

"The average length of the screws is 5 cm."

Based on the sample's average length, the inspector tests whether this difference is simply due to random variation or if the machine requires adjustment.

> **Hypothesis testing** is a formal procedure for testing a claim or hypothesis about a population parameter using sample data.

It helps us to determine whether the evidence in a sample supports a certain belief or hypothesis about the population.

7.1.1 Null and Alternative Hypothesis

The hypothesis that will actually be tested is called the **null hypothesis**, denoted by H_0. This is a particular claim about a population parameter. The null hypothesis is assumed to be true unless there is any strong evidence to the contrary.

The **alternative hypothesis**, denoted by H_1 or H_a, is a hypothesis that contradicts the null hypothesis. For the above example, we define the hypotheses as:

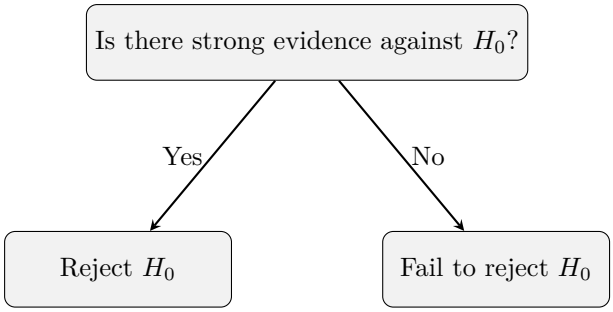
- **Null Hypothesis (H_0):** The average length of screws is 5 cm.
- **Alternative Hypothesis (H_1 or H_a):** The average length is not 5 cm.

In simple expression we write:

$$H_0 : \mu = 5$$

$$H_1 : \mu \neq 5$$

The standard procedure is to assume that H_0 is true. The burden of proof is placed on those who believe in the alternative claim[1] (alternative hypothesis). This initially favored claim (H_0) will not be rejected in favor of the alternative claim (H_1 or H_a) unless the sample evidence provides significant support for the alternative claim. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis.



Statisticians avoid saying "accept H_0" because it wrongly implies that the null hypothesis has been proven true. When there is no strong evidence against H_0, we retain it—but this does not confirm its truth; instead, we say we "fail to reject H_0". On the other hand, when the evidence is strong, we confidently "reject H_0". Thus, rejecting H_0 is a **strong decision**, supported by sufficient evidence, while retaining H_0 is a **weak decision**, reflecting inconclusive results.

To test the hypothesis in the above example, we assume that the standard deviation $\sigma = 0.3$ is known. Since the sample size is $n = 50$, the sampling distribution of the sample mean is approximately normal:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

or,

$$\overline{X} \sim N\left(50, 0.042\right)$$

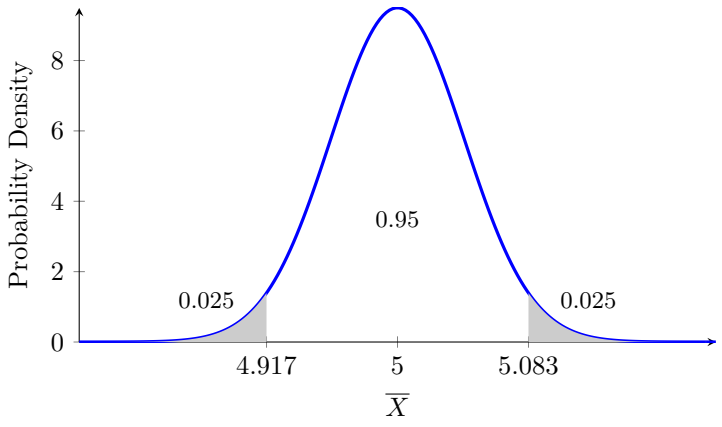


Figure 7.1: *Sampling distribution of sample mean for the screw length.*

[1] A close analogy can be made to a criminal court trial, where the jury holds to the null hypothesis of "Not guilty" unless there is convincing evidence of guilt. The purpose of the hearing is to establish the assertion that the accused is guilty rather than to prove that he or she is innocent.

The distribution of $\overline{X}$ (sampling distribution) is shown in Figure 7.1. From the figure, we can see that there is a 95% chance that the value of $\overline{X}$ measured from a random sample will lie in the region $4.917 \leq \overline{X} \leq 5.083$. However, the observed value is 4.8 cm, which falls well outside of this interval. The chance of obtaining such an extreme value when the true mean is actually 5 cm is less than 5%.

This provides strong statistical evidence against the null hypothesis. Therefore, we reject the null hypothesis $H_0 : \mu = 5$ in favor of the alternative hypothesis $H_1 : \mu \neq 5$. It suggests that the mean length of the screws is significantly different from 5 cm, and that the manufacturing process may need to be adjusted.

> The probability of the tails of the distribution, which determines the threshold for making a decision, is called the **level of significance**, denoted by α.

In our example, we chose the level of significance $\alpha = 0.05$, which is equally split between the two tails of the distribution, allocating 0.025 to each side.

> The range of values for which the null hypothesis is rejected is called the **critical region**.

For the above example, the critical region is characterized by $\overline{X} < 4.917$ and $\overline{X} > 5.083$.

7.1.2 One-Sided and Two-Sided Hypothesis Testing

In hypothesis testing, the form of the alternative hypothesis determines whether the test is **one-sided** (one-tailed) or **two-sided** (two-tailed).

- **Two-Sided Test:** Used when we are interested in detecting any difference from the null hypothesis value, whether it is an increase or a decrease. For example:

  $$H_0 : \mu = 5 \quad \text{vs.} \quad H_1 : \mu \neq 5$$

  This test considers deviations on both sides of the hypothesized mean as we saw in the previous example. The significance level α is split between the two tails of the sampling distribution.

- **One-Sided Test:** Used when we are only interested in deviations in one direction.

  - To test if the mean is *less than* 5:

    $$H_0 : \mu = 5 \quad \text{vs.} \quad H_1 : \mu < 5$$

  - To test if the mean is *greater than* 5:

    $$H_0 : \mu = 5 \quad \text{vs.} \quad H_1 : \mu > 5$$

  In this case, the entire significance level α is placed in one tail of the distribution.

The choice between one-sided and two-sided testing depends on the research question. If deviations in both directions are meaningful, a two-sided test is appropriate. If only an increase or a decrease is relevant, a one-sided test is more powerful.

7.2 Test Statistic

The **test statistic** is a function of the sample data that forms the basis for making the statistical decision to either reject or not reject the null hypothesis.

The main purpose of the test statistic is to provide a measure of how far the sample statistic (such as the sample mean) deviates from the hypothesized value under the null hypothesis. The further this value is from the hypothesized value, the stronger the evidence against the null hypothesis.

Depending on the type of hypothesis test being conducted, the test statistic can take various forms. For instance, a **z-test statistic** is used to test hypotheses about a population mean when the population standard deviation (σ) is known and the sample size is large ($n > 30$). The z-test compares the sample mean to the population mean, and the test statistic is given by:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $\overline{X}$ is the sample mean, μ_0 is the population mean under the null hypothesis, σ is the population standard deviation, and n is the sample size. If the null hypothesis is true, $\mathbb{E}(\overline{X}) = \mu_0$, and it follows that the distribution of Z is the standard normal distribution i.e. $Z \sim \mathcal{N}(0, 1)$.

For the example with a sample of 50 screws, where the sample mean is $\overline{X} = 4.8$ cm, the population mean under the null hypothesis is $\mu_0 = 5$ cm, and the population standard deviation is $\sigma = 0.3$ cm. The z-test statistic is then calculated as follows:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.3/\sqrt{50}} \approx -4.72$$

Once the test statistic is calculated, it is compared with a **critical value** calculated from the relevant probability distribution of the test statistic under null hypothesis (e.g., the standard normal distribution). The critical value c is chosen so that

$$P(\text{Test Statistic is more extreme than } c \mid H_0) = \alpha$$

where α is the level of significance. Equivalently, the set of all values of the test statistic that lead to rejection of H_0 is called the **critical region** or **rejection region**. Therefore

$$P(\text{Test statistic} \in \text{rejection region} \mid H_0) = \alpha$$

Test Type	**Critical Value(s)**	**Rejection Region**
Upper-tailed z-test	z_α	$\{Z > z_\alpha\}$
Lower-tailed z-test	$-z_\alpha$	$\{Z < -z_\alpha\}$
Two-tailed z-test	$\pm\, z_{\alpha/2}$	$\{\lvert Z\rvert > z_{\alpha/2}\}$

Table 7.1: *Critical values and rejection regions for z-tests at significance level α.*

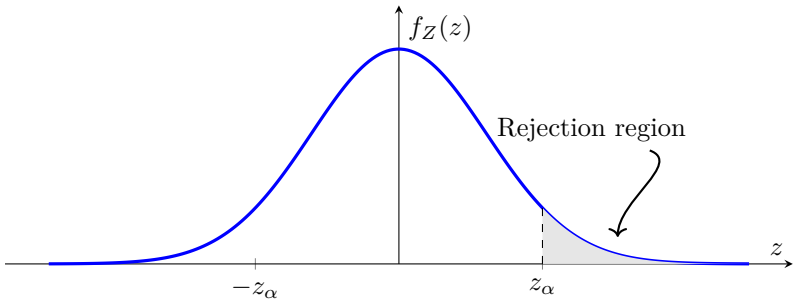


Figure 7.2: *Rejection region for upper-tailed z-test.*

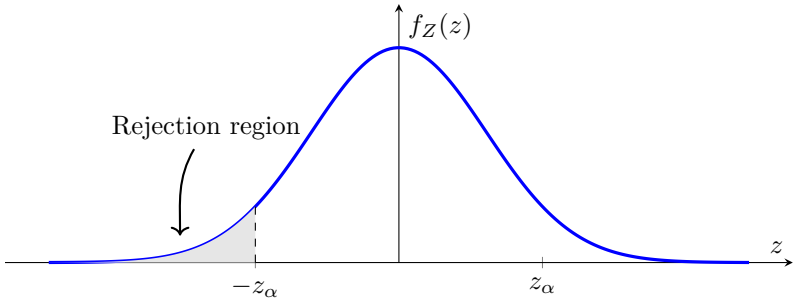


Figure 7.3: *Rejection region for lower-tailed z-test.*

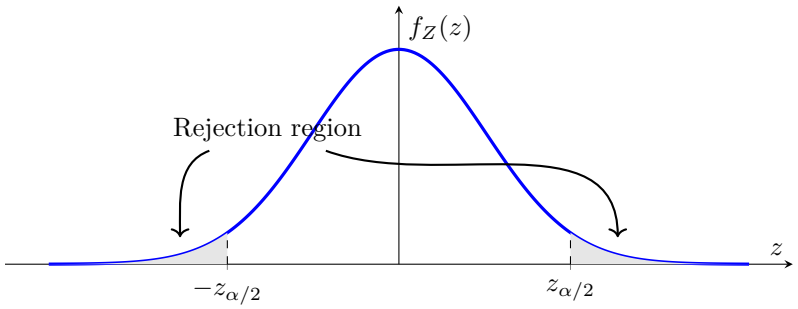


Figure 7.4: *Rejection region for two-tailed z-test.*

Now we set the decision rule:

- If the observed test statistic z_{obs} lies in the rejection region, then we reject H_0.
- Otherwise, we fail to reject H_0.

In our example, the computed z-value is -4.72, which lies in the left-hand critical region $\{Z < z_{\alpha/2} = -1.96\}$ for the two-tailed test with $\alpha = 0.05$. Therefore, we reject the null hypothesis H_0. In general, the larger the magnitude of the test statistic, the stronger the evidence against the null hypothesis.

7.3 Type I and Type II Error

This decision procedure can lead to either of two incorrect conclusions, known as Type I and Type II errors.

For example, suppose the true average length of the screws is indeed 5 cm. However, due to random variation in the sample, we might observe a test statistic that falls into the critical region. In this case, we would reject the null hypothesis H_0 in favor of the alternative H_1, even though H_0 is actually true. This mistake is called a **Type I error**. The probability of a Type I error is also called the **level of significance**, denoted by α. It is usually set at $\alpha = 0.05$ or $\alpha = 0.01$.

$$P(\text{Type I error}) = P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) = \alpha$$

On the other hand, suppose the true average length of the screws has actually changed (for example, to 4.8 cm), but the observed sample does not provide enough evidence (such as 4.92 cm) to reject H_0. In this case, we fail to reject the null hypothesis, even though it is false. This mistake is called a **Type II error**. The probability of Type II error is denoted by β.

$$P(\text{Type II error}) = P(\text{Not rejecting } H_0 \mid H_0 \text{ is false}) = \beta$$

	H_0 is True	H_1 is True
Reject H_0	**Type I Error** (α)	Correct Decision
Fail to Reject H_0	Correct Decision	**Type II Error** (β)

7.3.1 Tradeoff between Type I and Type II Errors

These error probabilities are inherently related through the choice of the rejection region. Generally, reducing α increases β, and vice versa. This trade-off can be visualized by the overlapping distributions of the test statistic (sample mean in the picture) under H_0 and H_1 as shown in Figure 7.5.

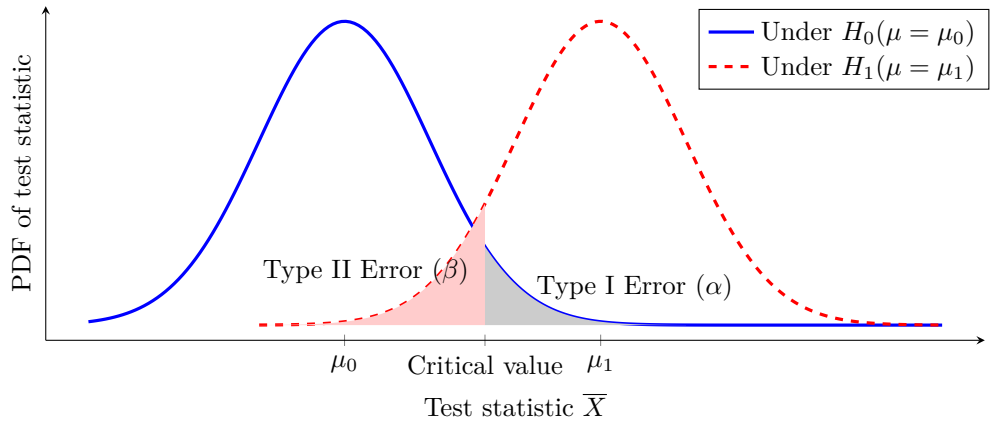


Figure 7.5: *Type I and Type II error.*

Here we assume the population parameter (i.e. population mean μ) under H_0 is μ_0 and

under H_1 is a specific value[2] μ_1.

Mathematically, for a given critical value c:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(T > c \mid H_0)$$

$$\beta = P(\text{Fail to reject } H_0 \mid H_1 \text{ true}) = P(T \leq c \mid H_1)$$

Changing the value of c affects α and β in opposite ways, showing the trade-off between making Type I and Type II errors.

- If we make the critical value c higher, the rejection region becomes smaller (more stringent), so α will **decrease**, but β will **increase**.
- If we lower the value c, the rejection region becomes larger (more liberal), so α will **increase**, but β will **decrease**.

The Figure 7.6 shows one typical example of the relationship between α and β.

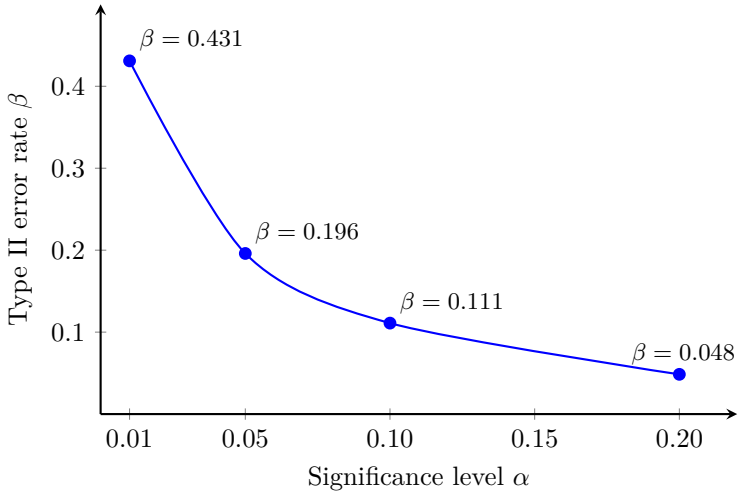


Figure 7.6: *Illustration of the inverse relationship between α and β.*

Hence, a balance must be struck depending on the consequences of making Type I vs. Type II errors. To arrive at a fair compromise we should know the cost of each type of error. In practice, these costs depend on the true (but unknown) parameter and are hard to quantify exactly, so we adopt one of three conventional values—1%, 5%, or 10%—based on the relative severity of each error.

For example, consider the design of an email spam filter. A **Type I error** here means flagging a legitimate email as spam, potentially causing the user to miss important messages. A **Type II error** would allow a spam message into the inbox, usually a minor nuisance. Thus, in this case, the cost of a Type I error is considered more serious than that of a Type II error, and the filter should be designed to minimize α, the probability of Type I error, even at the expense of a slightly higher β. A common choice might be $\alpha = 1\%$, to ensure that almost no valid emails are incorrectly filtered out.

On the other hand, imagine a public health agency conducting virus screening at airports during an outbreak of a contagious disease. A **Type I error** in this context would mean

[2]In practice, the specific value μ_1 is unknown when the alternative hypothesis is specified as $H_1 : \mu \neq \mu_0$. However, to compute the probability of a Type II error (β), we must assume a particular value of μ under the alternative hypothesis. This means that β is not a single number, but rather a function of the true value of μ under H_1. For every possible value of μ_1 under H_1, there is a corresponding value of β.

falsely identifying a healthy traveler as infected, leading to temporary quarantine and inconvenience. A **Type II error** would allow an infected traveler to enter the general population, potentially triggering a wider outbreak. In this scenario, the cost of a Type II error is far more serious. To minimize this risk, we may choose a higher significance level, such as $\alpha = 10\%$, accepting more false positives in order to reduce the probability of missing actual infections.

Finally, in cases where the consequences of Type I and Type II errors are either unknown or roughly same, such as in exploratory scientific studiess, a balanced approach is often taken. The conventional significance level $\alpha = 5\%$ serves as a compromise that moderately controls both types of errors in the absence of more specific cost information.

7.3.2 How β Depends on the True Parameter Value?

The probability of a Type II error (β) is not fixed; it varies depending on how far the actual parameter value is from the hypothesized value under H_0. Specifically:

- If the true value is close to H_0, then the evidence against H_0 is weak, and β is high (it's harder to detect the difference).
- If the true value is far from H_0, then the evidence becomes stronger, and β is lower (easier to detect the difference).

We generally write β as a function of the parameter value θ_1 under alternative hypothesis denoted by $\beta(\theta)$.

7.3.3 Dependence on Sample Size

Earlier, we discussed that one can choose a low α or a low β depending on which type of error is more serious. But what if both Type I and Type II errors are costly, and we want to minimize both? The only effective way to achieve this is by improving the reliability of the evidence—primarily by increasing the sample size. As the sample size n increases, the sampling distributions under both H_0 and H_1 become narrower (i.e., have smaller variance), which makes it easier to distinguish between them. This shrinkage leads to a reduction in both α and β. Therefore, when minimizing both types of error is important, increasing the sample size is the most practical solution.

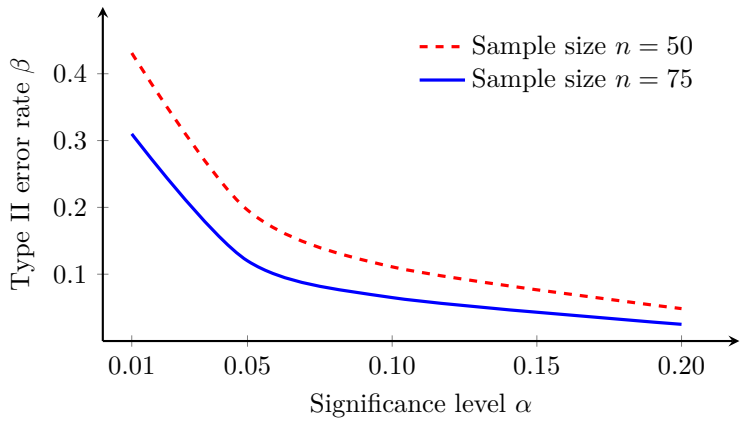


Figure 7.7: *Effect of sample size on the inverse relationship between α and β. Larger sample size reduces β for the same α.*

7.3.4 p-Value in Hypothesis Testing

> The p-**value** is the probability, under null hypothesis, that the test statistic takes a value equal to or more extreme than the value actually observed.

Let T be the test statistic used to assess H_0 versus H_1, and let t_{obs} be its observed value from the sample. Then for **one-sided test**,

$$p\text{-value} = P(T \geq t_{\text{obs}} \mid H_0)$$

For a **two-sided test** (where both large and small values of T count against H_0),

$$p\text{-value} = P(T \geq t_{\text{obs}} \mid H_0) \; + \; P(T \leq -t_{\text{obs}} \mid H_0)$$

Assuming the null distribution is symmetric:

$$p\text{-value} = 2P(T \geq t_{\text{obs}} \mid H_0)$$

- A **small** $p-$value (below the chosen significance level α) indicates that observing t_{obs} would be very unlikely if H_0 were true, so we **reject** H_0.
- A **large** $p-$value means the data are consistent with H_0, so we **fail to reject** H_0.

Condition	**Decision**
p-value $\leq \alpha$	Reject H_0
p-value $> \alpha$	Fail to reject H_0

Table 7.2: *Decision rule based on p-value.*

Example. Suppose $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ with known $\sigma = 0.3$, and we test

$$H_0 : \mu = 5 \quad \text{vs.} \quad H_1 : \mu > 5$$

using

$$T = \frac{\overline{X} - 5}{\sigma/\sqrt{n}} \; \sim \; N(0,1) \quad \text{under } H_0$$

If $\overline{X} = 5.1$, $n = 36$, and $\sigma = 0.3$, then

$$t_{\text{obs}} = \frac{5.1 - 5}{0.3/\sqrt{36}} = 2.00 \quad \Rightarrow p\text{-value} = P(Z \geq 2.00) \approx 0.0228$$

Since p-value $= 0.0228 < 0.05$, we reject H_0 at 5% level of significance.

7.4 General Procedure for Hypothesis Testing

Step 1. Formulate the hypotheses

Formulate the null hypothesis (H_0), which represents the default claim about the population parameter, and the alternative hypothesis (H_1), which represents the effect or difference you aim to detect.

Step 2. Choose a significance level

Select a level of significance α (commonly 0.01, 0.05, or 0.10) that defines the maximum probability of committing a Type I error (rejecting H_0 when it is true).

Step 3. Define and calculate the test statistic

Identify an appropriate test statistic based on the hypotheses and assumption of population distribution. Then collect a sample of size n and compute the observed value of the test statistic from the data.

Step 4. Calculate the p-value or determine the critical value(s) from data

- *p-value approach:* Compute the probability, under H_0, of observing a test statistic at least as extreme as the one obtained.
- *Critical value approach:* Identify the cutoff point(s) from the null distribution that correspond to the chosen α, and define the rejection region(s).

Step 5. Make a decision

- If the p-value is less than or equal to α, **reject** H_0; otherwise, **do not reject** H_0.
- State the result in context, indicating whether there is sufficient evidence to support the alternative hypothesis at the chosen significance level.

7.5 Statistical Test for a Normally Distributed Population Mean μ

7.5.1 Standard Deviation(σ) Known

1. Hypothesis:

$$H_0 : \mu = \mu_0$$

$$H_1 : \begin{cases} \mu \geq \mu_0, & \text{Right tail test} \\ \mu \leq \mu_0, & \text{Left tail test} \\ \mu \neq \mu_0, & \text{Two-sided test} \end{cases}$$

2. Select α (commonly 0.01, 0.05, or 0.10) as the maximum probability of a Type I error.

3. Test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

where $\overline{X}$ is the sample mean, σ is known, and n is the sample size. Under H_0, $Z \sim N(0,1)$. We then calculate the test statistic from the sample data. Let z_{obs} is the observed value of Z from sample data.

4. We now calculate the p-value from the distribution of test statistic:

$$p\text{-value} = \begin{cases} P(Z \geq z_{\text{obs}}), & \text{for } H_1 : \mu > \mu_0 \\ P(Z \leq z_{\text{obs}}), & \text{for } H_1 : \mu < \mu_0 \\ 2\,P\big(|Z| \geq |z_{\text{obs}}|\big), & \text{for } H_1 : \mu \neq \mu_0 \end{cases}$$

Alternatively, we can identify the critical values and the rejection region at level α:

H_1	**Critical Value(s)**	**Rejection Region**
$\mu > \mu_0$	z_α	$\{Z > z_\alpha\}$
$\mu < \mu_0$	$-z_\alpha$	$\{Z < -z_\alpha\}$
$\mu \neq \mu_0$	$\pm\, z_{\alpha/2}$	$\{\lvert Z\rvert > z_{\alpha/2}\}$

5. If p-value $\leq \alpha$, **reject** H_0; otherwise, **do not reject** H_0.

   Alternatively, if z_{obs} lies in the rejection region, **reject** the H_0; otherwise, **do not reject** H_0.

7.5.2 Standard Deviation(σ) Unknown

1. Hypothesis:

$$H_0 : \mu = \mu_0$$

$$H_1 : \begin{cases} \mu \geq \mu_0, & \text{Right tail test} \\ \mu \leq \mu_0, & \text{Left tail test} \\ \mu \neq \mu_0, & \text{Two-sided test} \end{cases}$$

2. Select α (commonly 0.01, 0.05, or 0.10) as the maximum probability of a Type I error.

3. Test statistic:

   - **Large sample** ($n \geq 30$):

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

     where S is the sample standard deviation. By the Central Limit Theorem, $Z \sim N(0,1)$ approximately under H_0. Let z_{obs} is the observed value of Z from sample data.

   - **Small sample** ($n < 30$):

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

     Here, T follows the t-distribution with $n-1$ degrees of freedom under H_0. Let t_{obs} is the observed value of T from sample data.

4. We now calculate the p-value from the distribution of the test statistic:

   - **Large sample** ($n \geq 30$) — Use standard normal distribution:

$$p\text{-value} = \begin{cases} P(Z \geq z_{\text{obs}}), & \text{for } H_1 : \mu > \mu_0, \\ P(Z \leq z_{\text{obs}}), & \text{for } H_1 : \mu < \mu_0, \\ 2\,P(|Z| \geq |z_{\text{obs}}|), & \text{for } H_1 : \mu \neq \mu_0 \end{cases}$$

- **Small sample** ($n < 30$) — Use t-distribution with $n-1$ degrees of freedom:

$$p\text{-value} = \begin{cases} P(T \geq t_{\text{obs}}), & \text{for } H_1 : \mu > \mu_0, \\ P(T \leq t_{\text{obs}}), & \text{for } H_1 : \mu < \mu_0, \\ 2\,P(|T| \geq |t_{\text{obs}}|), & \text{for } H_1 : \mu \neq \mu_0 \end{cases}$$

Alternatively, identify critical values and rejection regions at level α:

- **Large sample** ($n \geq 30$)

H_1	**Critical Value(s)**	**Rejection Region**
$\mu > \mu_0$	z_α	$\{Z > z_\alpha\}$
$\mu < \mu_0$	$-z_\alpha$	$\{Z < -z_\alpha\}$
$\mu \neq \mu_0$	$\pm\, z_{\alpha/2}$	$\{\lvert Z\rvert > z_{\alpha/2}\}$

- **Small sample** ($n < 30$)

H_1	**Critical Value(s)**	**Rejection Region**
$\mu > \mu_0$	$t_{\alpha,\,n-1}$	$\{T > t_{\alpha,\,n-1}\}$
$\mu < \mu_0$	$-t_{\alpha,\,n-1}$	$\{T < -t_{\alpha,\,n-1}\}$
$\mu \neq \mu_0$	$\pm\, t_{\alpha/2,\,n-1}$	$\{\lvert T\rvert > t_{\alpha/2,\,n-1}\}$

5. If p-value $\leq \alpha$, **reject** H_0; otherwise, **do not reject** H_0.

   Alternatively, if z_{obs} (for large sample) or t_{obs} (for small sample) lies in the rejection region, **reject** the H_0; otherwise, **do not reject** H_0.

7.6 The 5 Steps of Hypothesis Testing

1. **State the hypotheses**
   Example: A factory claims their lightbulbs last an average of 1,000 hours. We suspect they don't.
   $H_0 : \mu = 1000$ hours
   $H_1 : \mu \neq 1000$ hours

2. **Choose a significance level**
   Let's use $\alpha = 0.05$.

3. **Collect data and calculate the test statistic**
   Suppose we test 30 lightbulbs and find the sample mean $\overline{x} = 980$, with a standard deviation $s = 50$.

4. **Calculate the p-value or compare the test statistic to a critical value**
   Using a Z-test:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{980 - 1000}{50/\sqrt{30}} \approx -2.19$$

   From Z-tables, $p \approx 0.0286$ (two-tailed).

5. **Make a decision**
   Since $p < \alpha$, we reject H_0.
   Conclusion: There is evidence that the average bulb life is not 1,000 hours.

7. Final Example: Coin Toss

You suspect a coin is biased. You flip it 100 times and get 60 heads.
$H_0 : p = 0.5$ (fair coin)
$H_1 : p \neq 0.5$ (biased coin)

$$z = \frac{0.6 - 0.5}{\sqrt{0.5 \times 0.5/100}} = 2.0$$

From Z-tables, $p \approx 0.0455$. Since $p < 0.05$, we reject H_0.
Conclusion: There is evidence the coin might be biased.

7.7 Power of a Test

One very important concept related to error probabilities is the notion of the *power* of a test. Intuitively, power measures a test's ability to detect when the null hypothesis is false. It is the probability of rejecting H_0 given that a specific alternative is true.

The **power** of a test at a specific alternative parameter value $\theta = \theta_1$ is defined as

$$\begin{aligned} \text{Power}(\theta_1) &= P(\text{Rejecting } H_0 \mid \theta = \theta_1) \\ &= 1 - P(\text{Not rejecting } H_0 \mid \theta = \theta_1) \\ &= 1 - \beta(\theta_1) \end{aligned}$$

Here $\beta(\theta_1)$ is the probability of a Type II error when the true parameter equals θ_1.

In other words, **it quantifies your test's sensitivity to detect real departures from the null hypothesis**. The power of this two-sided Z-test ($H_0 : \mu = 5$ vs. $H_1 : \mu \neq 5$) with known $\sigma = 0.3$, $n = 50$, and $\alpha = 0.05$ is the probability it will reject H_0 when μ truly equals 4.8. Here the critical region is

$$\overline{X} < 4.9168 \quad \text{or} \quad \overline{X} > 5.0832$$

If in reality $\mu = 4.8$, then

$$\overline{X} \sim N\big(4.8,\ \text{SE}^2\big),$$

and

$$P\big(\bar{X} < 4.9168\big) = F_Z\Big(\frac{4.9168 - 4.8}{0.04243}\Big) \approx 0.9971$$

while

$$P\big(\bar{X} > 5.0832\big) = 1 - F_Z\Big(\frac{5.0832 - 4.8}{0.04243}\Big) \approx 0$$

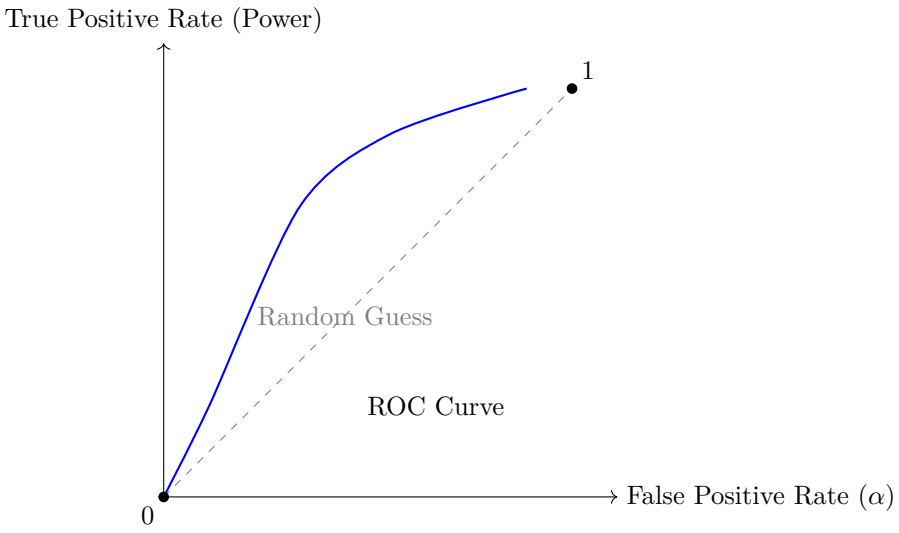
Thus the test's power at $\mu = 4.8$ is about 0.9971 ($\approx 99.7\%$), meaning there's a 99.7% chance (very sensitive) of detecting this 0.2-unit shift. Higher power can be achieved by increasing the sample size, accepting a larger α, or targeting a larger effect size. Before running an experiment, one often conducts a power analysis to choose n (or α) so that the power exceeds a desired threshold (commonly 0.8 or 0.9) for a specified effect size.

7.7.1 Receiver Operating Characteristic (ROC) Curve

The ROC curve is a graphical tool that illustrates the trade-off between sensitivity (power) and specificity (1 - α) for different thresholds of a test statistic.

- **True Positive Rate (TPR) or Sensitivity:** $1 - \beta = P(\text{Reject } H_0 \mid H_1 \text{ true})$.
- **False Positive Rate (FPR):** $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true})$.

By varying the critical value c, we obtain different pairs $(\alpha, 1-\beta)$, which can be plotted with α on the x-axis and power on the y-axis to produce the ROC curve.



The closer the ROC curve approaches the top-left corner, the better the test's ability to discriminate between H_0 and H_1.

The **Area Under the Curve (AUC)** quantifies overall test performance: an AUC of 1 indicates perfect discrimination, whereas an AUC of 0.5 corresponds to random guessing.

7.7.2 Most Powerful Tests and the Neyman–Pearson Lemma

When choosing among all tests of a fixed size (level) α, one may ask: *Which test maximizes power for a given simple alternative?*

The Neyman–Pearson lemma gives a definitive answer for testing a simple null

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta = \theta_1. \tag{7.1}$$

Define the likelihood ratio

$$\Lambda(x) = \frac{f(x;\theta_0)}{f(x;\theta_1)}. \tag{7.2}$$

The lemma states that the test which rejects H_0 when

$$\Lambda(x) \leq k \tag{7.3}$$

is the *most powerful* test of size α, where k is chosen so that $P_{\theta_0}(\Lambda(x) \leq k) = \alpha$.

Example: Exponential Lifetimes Suppose individual bulb lifetimes X_i are independent and exponentially distributed with density

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0, \tag{7.4}$$

where $\lambda = 1/\mu$ is the failure rate. We wish to test

$$H_0 : \lambda = \lambda_0 \quad \text{vs.} \quad H_1 : \lambda = \lambda_1; (\lambda_1 < \lambda_0). \tag{7.5}$$

The joint density for a sample of size n is

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n X_i}. \tag{7.6}$$

The likelihood ratio is

$$\Lambda = \frac{L(\lambda_0)}{L(\lambda_1)} = \Big(\frac{\lambda_0}{\lambda_1}\Big)^n \exp\Big[-(\lambda_0 - \lambda_1)\sum_{i=1}^n X_i\Big]. \tag{7.7}$$

Rejecting for small Λ is equivalent to rejecting when

$$\sum_{i=1}^n X_i > c, \tag{7.8}$$

where c solves

$$P_{\lambda_0}\big(\sum_{i=1}^n X_i > c\big) = \alpha. \tag{7.9}$$

This test is most powerful of level α for distinguishing λ_0 vs. λ_1.

Increasing n, choosing a larger α, or considering a larger separation between λ_0 and λ_1 will increase power.

7.8 Most Powerful Test, Neyman–Pearson Lemma, and Likelihood Ratio Test

7.8.1 Most Powerful Test

A **most powerful** test between two simple hypotheses

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta = \theta_1$$

is a test of size α whose power under H_1 is at least as large as that of any other test of the same size. Concretely, if $\phi(X) \in \{0, 1\}$ denotes our test function, then for any other test $\psi(X)$ with

$$\mathbb{E}_{\theta_0}[\psi(X)] \le \alpha,$$

we have

$$\mathbb{E}_{\theta_1}[\phi(X)] \;\ge\; \mathbb{E}_{\theta_1}[\psi(X)].$$

7.8.2 Neyman–Pearson Lemma

Theorem. (Neyman–Pearson) Let X have density or mass $f(x;\theta)$. To test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1,$$

the most powerful test of size α rejects H_0 when the likelihood ratio

$$\Lambda(x) = \frac{f(x;\theta_1)}{f(x;\theta_0)}$$

exceeds a cutoff $k > 0$ chosen so that

$$\Pr_{\theta_0}\big(\Lambda(X) > k\big) = \alpha.$$

Equivalently, the test function is

$$\phi(x) = \begin{cases} 1, & \Lambda(x) > k, \\ \gamma, & \Lambda(x) = k, \\ 0, & \Lambda(x) < k, \end{cases}$$

with $\gamma \in [0, 1]$ chosen to make $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

7.8.3 Likelihood Ratio Test

For testing a composite null

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

define the likelihood ratio statistic

$$\lambda(x) = \frac{\max_{\theta \in \Theta_0} L(\theta; x)}{\max_{\theta \in \Theta} L(\theta; x)},$$

where $L(\theta; x) = f(x; \theta)$. We reject H_0 for small values of $\lambda(x)$ (equivalently, large values of $-2 \log \lambda(x)$).

Under standard regularity conditions, as the sample size $n \to \infty$,

$$-2 \log \lambda(X) \xrightarrow{d} \chi^2_{\dim(\Theta) - \dim(\Theta_0)}.$$

7.8.4 Example: Normal Mean

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with known σ^2, and consider

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1.$$

The Neyman–Pearson ratio test rejects H_0 when

$$\frac{L(\mu_1)}{L(\mu_0)} = \exp\!\Big(\tfrac{\mu_1 - \mu_0}{\sigma^2} \sum_{i=1}^{n} X_i \; - \; \tfrac{n(\mu_1^2 - \mu_0^2)}{2\sigma^2} \Big) > k.$$

After taking logarithms and rearranging, this reduces to the usual level-α test

$$\overline{X} > \mu_0 + z_{1-\alpha} \, \frac{\sigma}{\sqrt{n}}.$$

Chapter 8

Correlation and Regression

8.1 Correlation

In statistics, **correlation** between two random variables is a measure of the degree of linear association between them.

For example, imagine you record how many hours each of the six students studies in a week and their corresponding exam scores:

Student	**Hours Studied** (X)	**Exam Score** (Y)
1	2	65
2	3	70
3	4	76
4	5	78
5	6	82
6	7	89

Table 8.1: *Study hours and exam scores of six students.*

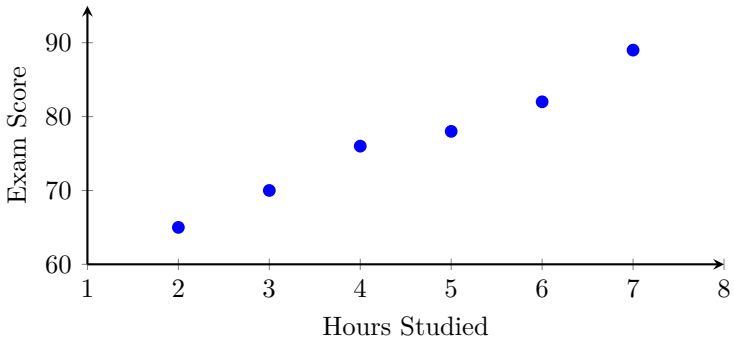


Figure 8.1: *Scatter plot of study time and exam performance.*

As illustrated in the table and the accompanying scatter plot, exam scores increase as students dedicate more hours to studying. This clear upward pattern reflects a positive correlation: students who invest more time in preparation tend to achieve higher marks. In general, the direction of correlation can be classified as follows:

1. **Positive correlation**: An increase in one variable is accompanied by an increase

in the other.

2. **Negative correlation**: An increase in one variable is accompanied by a decrease in the other.

3. **No (zero) correlation**: Changes in one variable show no consistent association with changes in the other.

8.2 Pearson Correlation Coefficient

Pearson's correlation coefficient for a population, commonly denoted by the Greek letter ρ (rho), is also known as the **population correlation coefficient**. Given a pair of random variables X and Y, the population correlation coefficient ρ_{XY} is defined as

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big]}{\sigma_X \sigma_Y}$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ are the means of X and Y, σ_X and σ_Y are their standard deviations, and $\text{Cov}(X,Y)$ is the covariance between X and Y.

Like all population parameters, The value of ρ_{XY} is not known to us. We may need to estimate it from the random sample observation pairs (X,Y). Let

$$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$$

be n pairs of observations with respective means $\overline{X},\overline{Y}$ and variances S_X^2, S_X^2 respectively. It turns out that a point estimate of $\text{Cov}(X,Y)$ is the sample covariance:

$$S_{XY} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

The point estimate of σ_X is sample standard deviation of X:

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

The point estimate of σ_Y is sample standard deviation of Y:

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

Substituting these estimates for their population counterparts, we get the formula for the **sample correlation coefficient** as

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Alternatively, r_{XY} can be expressed as

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

We can simplify the formula for r_{XY} into a form most convenient for computing from raw data by using the identities:

$$\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n\overline{x} \cdot \overline{y} = \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}$$

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2 = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}$$

$$\sum_i (y_i - \overline{y})^2 = \sum_i y_i^2 - n\overline{y}^2 = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}$$

Thus we obtain the formula for r_{XY}:

$$r_{XY} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

The value of r_{XY} lies within -1 and 1 i.e. $-1 \leq r_{XY} \leq 1$.

- $r > 0$ indicates a **positive correlation**.
- $r < 0$ indicates a **negative correlation**.
- $r = 0$ suggests **no correlation** (but not necessarily independence).
- $r = \pm 1$ indicates a **perfect linear correlation** ($+1$ for positive and -1 for negative).

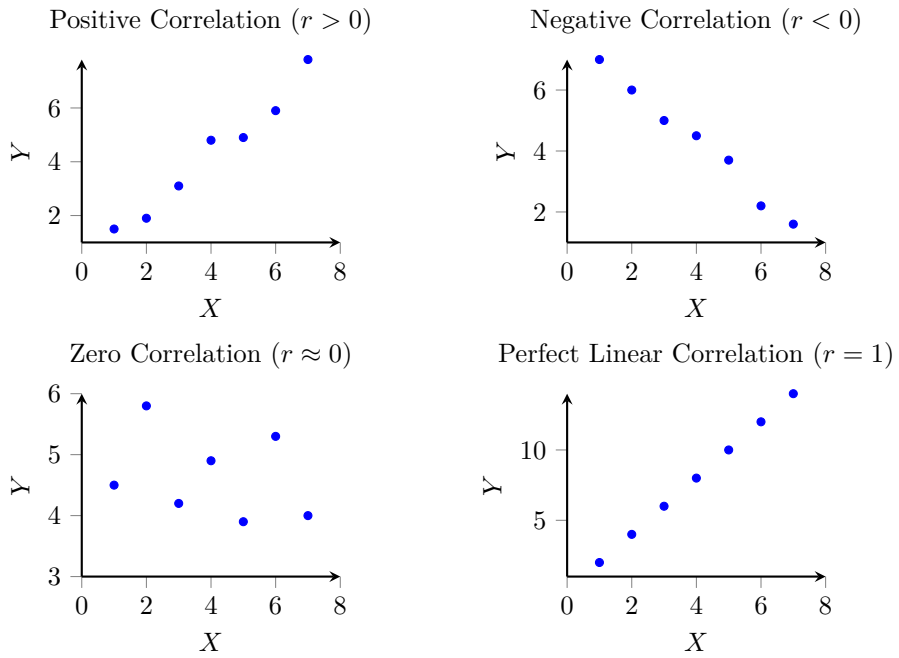


Figure 8.2: *Scatterplots illustrating positive, negative, zero, and perfect linear correlations.*

Example: Suppose we have data on students' hours studied (X) and exam scores (Y) as given in the table 8.1. We need to calculate the Pearson corelation coefficient.

The formula we will use:

$$r_{XY} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Student	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	65	4	4225	130
2	3	70	9	4900	210
3	4	76	16	5776	304
4	5	78	25	6084	390
5	6	82	36	6724	492
6	7	89	49	7921	623
$n = 6$	$\sum x_i = 27$	$\sum y_i = 460$	$\sum x_i^2 = 139$	$\sum y_i^2 = 35630$	$\sum x_i y_i = 2149$

Table 8.2: *Correlation coefficient calculation table.*

Substituting the values:

$$\begin{aligned} r_{XY} &= \frac{6 \cdot 2149 - 27 \cdot 460}{\sqrt{6 \cdot 139 - 27^2} \cdot \sqrt{6 \cdot 35630 - 460^2}} \\ &= \frac{12894 - 12420}{\sqrt{834 - 729} \cdot \sqrt{213780 - 211600}} \\ &= \frac{474}{\sqrt{105} \cdot \sqrt{2180}} \approx \frac{474}{10.247 \cdot 46.690} \\ &\approx \frac{474}{478.74} \approx 0.9907 \end{aligned}$$

This indicates a very strong positive correlation between hours studied and exam scores.

8.3 Effect of Linear Transformations on Correlation Coefficient

Theorem: Suppose we have two random variables X and Y with correlation coefficient ρ_{XY}. Define new variables

$$U = a + bX, \quad V = c + dY$$

where a, c are constants (shifts) and $b, d \neq 0$ are scaling factors. Then the population correlation coefficient between U and V is

$$\rho_{UV} = \frac{bd}{|b||d|} \cdot \rho_{XY}$$

Proof:

$$\rho_{UV} = \frac{\mathrm{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{\mathrm{Cov}(a + bX, c + dY)}{\sqrt{\mathrm{Var}(a + bX)} \cdot \sqrt{\mathrm{Var}(c + dY)}}$$

Now,

$$\begin{aligned}\mathrm{Cov}(a+bX, c+dY) &= bd\,\mathrm{Cov}(X,Y)\\ \mathrm{Var}(a+bX) &= c^2\mathrm{Var}(X)\\ \mathrm{Var}(c+dY) &= d^2\mathrm{Var}(Y)\end{aligned}$$

Thus,

$$\rho_{UV} = \frac{\mathrm{Cov}(bX, dY)}{\sqrt{\mathrm{Var}(bX)}\cdot\sqrt{\mathrm{Var}(dY)}} = \frac{bd\,\mathrm{Cov}(X,Y)}{|b|\sigma_X\cdot|d|\sigma_Y} = \frac{bd}{|b||d|}\cdot\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$$

Since

$$\frac{bd}{|b||d|} = \mathrm{sgn}(b)\cdot\mathrm{sgn}(d)$$

where $\mathrm{sgn}(\cdot)$ is the sign function[1], it follows that

$$\rho_{UV} = \frac{bd}{|b||d|}\cdot\rho_{XY} = \mathrm{sgn}(b)\cdot\mathrm{sgn}(d)\cdot\rho_{XY}$$

■

This shows that linear transformations preserve the magnitude of the correlation coefficient but may reverse its sign depending on the scaling factors.

- If b and d have the same sign, then $\rho_{UV} = \rho_{XY}$.
- If b and d have opposite signs, then $\rho_{UV} = -\rho_{XY}$.
- Adding constants a and c has no effect on the value of the correlation.

The sample correlation coefficient also satisfies the same transformation relation:

$$r_{UV} = \mathrm{sgn}(b)\cdot\mathrm{sgn}(d)\cdot r_{XY}$$

8.4 Correlation Is Not Causation

It is important to understand that a high correlation between two variables does not necessarily imply that one variable causes the other. Correlation measures the strength and direction of a linear relationship between variables, but it does not provide evidence about causality.

There are several reasons why correlation does not imply causation:

- **Confounding variables:** A third variable may influence both variables under study, creating a spurious correlation.
- **Reverse causality:** The direction of cause and effect may be opposite to what is assumed.

[1]The **sign function**, denoted by $\mathrm{sgn}(x)$, is defined as:

$$\mathrm{sgn}(x) = \begin{cases} -1, & \text{if } x < 0,\\ 0, & \text{if } x = 0,\\ 1, & \text{if } x > 0.\end{cases}$$

It extracts the sign of a real number x.

- **Coincidence:** Sometimes, correlations arise purely by chance.

For example, consider a study might that finds a positive correlation between umbrella sales and slipping accidents. This doesn't mean umbrellas cause slips or vice versa. Instead, rainy weather acts as the common factor: rain simultaneously drives people to buy umbrellas and makes sidewalks slippery, creating the illusion of a direct relationship between the two variables.

This example illustrates why it is crucial to use careful experimental design, statistical controls, and domain knowledge before concluding causal relationships from correlated data.

8.5 Regression Analysis

Regression analysis is a statistical method used to explore and model the relationship between a dependent variable and one or more independent variables.

It plays a central role in data analysis, prediction, and inference, particularly when trying to establish a functional relationship between variables.

In its simplest form—**simple linear regression**—we wish to study the relationship between two variables X and Y and use it to predict Y from X. The variable X acts as the **independent variable** (predictor, causal variable) whose values are controlled by the experimenter and Y is the **dependent variable** (response) which is also subjected to unaccountable variations (errors).

For example, a teacher wants to examine whether there's a relationship between how long students study and the scores they achieve in a test. By treating the number of hours studied as the independent variable and the test score as the dependent variable, regression helps us determine whether there is a consistent trend between the two.

8.6 The Simple Linear Regression Model

A simple linear regression model assumes the existence of a linear relationship between X (predictor variable) and Y (response) that is disturbed by a random error ϵ which can be written as an equation of the form:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- β_0: y-intercept of the line,
- β_1: slope of the line (rate of change in Y per unit increase in X),
- ϵ: random error, accounting for unexplained variation

Given a dataset of n observations, represented as pairs:

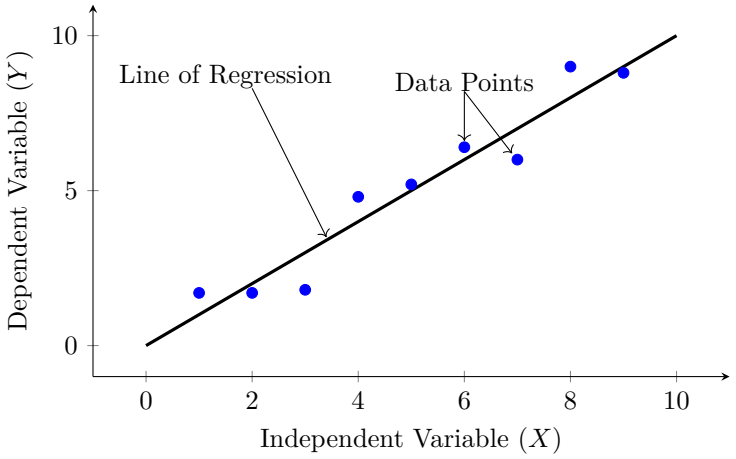
$$(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$$

the objective is to estimate the unknown parameters β_0 and β_1, and then use these estimates to define a straight line that best fits the data.

The **fitted regression line** (also called the prediction equation) is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Here, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated values of the intercept and slope, respectively, obtained from the sample data.

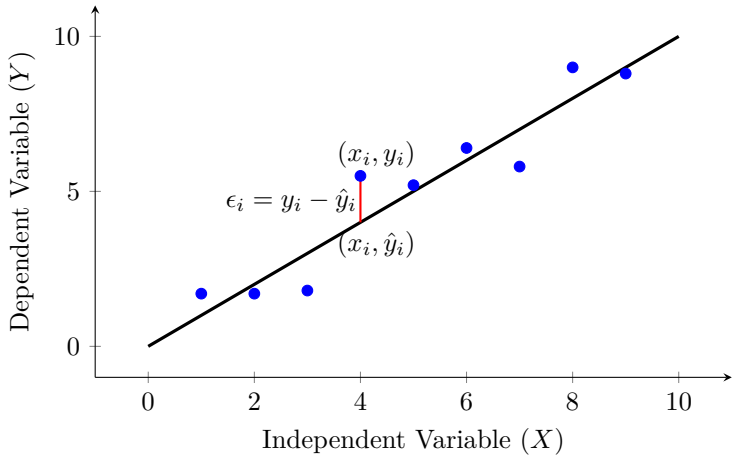


8.7 Estimating Parameters Using Least Squares

Given n data points $(x_1, y_1), \ldots, (x_n, y_n)$, the estimates of the parameters β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data. To define what we mean by a "best fit" line, consider each data point (x_i, y_i) and its corresponding prediction $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ from the regression line. This predicted value is known as the **fitted value**. The difference between the observed value y_i and the fitted value $\hat{y}_i$ is called the **residual** (error), denoted by ϵ_i:

$$\epsilon_i = y_i - \hat{y}_i$$

The residual ϵ_i represents the vertical distance between an observed data point and the regression line. A positive residual indicates that the point lies above the regression line, while a negative residual indicates that it lies below. The closer the residuals are to zero, the better the fitted values approximate the observed data. Therefore, the estimates of the parameters β_0 and β_1 should be such that these residuals (errors) are as small as possible. However, minimizing the simple sum of residuals is not appropriate, because the positive and negative errors can cancel each other out, even when individual errors are large. A more effective approach is the **method of least squares**, which involves minimizing the sum of the squares of the residuals.



The **least-squares line** is the line that minimizes the **residual sum of squares (RSS)**[2]:

$$\text{RSS} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

To derive the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$, we have to take the partial derivatives of RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them to zero.

$$\frac{\partial}{\partial \hat{\beta}_0}(RSS) = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$
$$\frac{\partial}{\partial \hat{\beta}_1}(RSS) = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Multiplying the first equation by $\sum_{i=1}^{n} x_i$ and second equation by n and then subtracting second from the first yields

$$\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - n \sum_{i=1}^{n} x_i y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i\right)^2 - n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

Thus we get the expression for $\hat{\beta}_1$

[2] In some texts, the residual sum of squares (RSS) is also called the **sum of squares of errors (SSE)**

$$\hat{\beta}_1 = \left[n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right] \Big/ \left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]$$
$$= \left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right]\right) \Big/ \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

It is convenient to introduce some notation for the sums of squared deviation from mean and sums of cross-products of deviation.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

Using these notation we can write,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

The expression for $\hat{\beta}_0$ is calculated as

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

8.7.1 Calculation of RSS

We can now calculate the value of RSS based on the value of $\hat{\beta}_0$ and $\hat{\beta}_0$ found using the method of least square:

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}\Big[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big]^2 \qquad \text{Since } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \sum_{i=1}^{n}\Big[y_i - (\overline{y} - \hat{\beta}_1\overline{x} + \hat{\beta}_1 x_i)\Big]^2 \qquad \text{Since } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \\
&= \sum_{i=1}^{n}\Big[(y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x})\Big]^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned}$$

$$
\begin{aligned}
&= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} \\
&= S_{yy} - 2\left(\frac{S_{xy}}{S_{xx}}\right) S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \qquad \text{Since } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \\
&= S_{yy} - \frac{2S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\
&= S_{yy} - \hat{\beta}_1 S_{xy}
\end{aligned}
$$

$$\text{RSS} = S_{yy} - \hat{\beta}_1 S_{xy}$$

8.7.2 Example

We aim to model the relationship between the number of hours studied (X) and the corresponding test score (Y) using simple linear regression. The goal is to estimate a linear equation that best describes this relationship based on observed data from five students. Once the model is established, we will use it to predict the expected test score for a student who studies for six hours.

The observed data are as follows:

Student	**Hours Studied** (x_i)	**Test Score** (y_i)
1	2	65
2	3	70
3	5	75
4	7	85
5	9	95

Step 1: Compute Means

$$
\begin{aligned}
\overline{x} &= \frac{2+3+5+7+9}{5} = \frac{26}{5} = 5.2, \\
\overline{y} &= \frac{65+70+75+85+95}{5} = \frac{390}{5} = 78
\end{aligned}
$$

Step 2: Compute S_{xx} and S_{xy}

$$
\begin{aligned}
S_{xx} &= \sum (x_i - \overline{x})^2 \\
&= (2-5.2)^2 + (3-5.2)^2 + (5-5.2)^2 + (7-5.2)^2 + (9-5.2)^2 \\
&= 32.8
\end{aligned}
$$

$$
\begin{aligned}
S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= (2-5.2)(65-78) + (3-5.2)(70-78) + \ldots + (9-5.2)(95-78) \\
&= 137.0
\end{aligned}
$$

Step 3: Estimate Parameters

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{137.0}{32.80} \approx 4.177$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 78 - 4.177 \times 5.2 \approx 56.28$$

Step 4: Regression Equation

$$\hat{Y} = 56.28 + 4.177X$$

Step 5: Predict Test Score for $X = 6$

$$\hat{Y} = 56.28 + 4.177 \times 6 \approx 81.34$$