MILA UNIVERSITY CENTER
L I B R A R Y

INFFERENTIAL STATISTICS

Yakoub BOULAROUK
University center of MILA

Parametric statistics

Yakoub BOULAROUK

November 30, 2024

Preface

Statistics is a powerful tool that enables us to unravel the mysteries hidden within data. It empowers us to make informed decisions, draw meaningful conclusions, and uncover valuable insights about the world around us. Two fundamental branches of statistics, estimation and inferential statistics, play a pivotal role in this process.

This book serves as a comprehensive guide to understanding and utilizing estimation and inferential statistics. Whether you are a student embarking on a statistical journey or a seasoned data analyst seeking to enhance your statistical knowledge, this book aims to equip you with the necessary tools and concepts to navigate the intricate realm of statistical inference.

Estimation, the first pillar of this book, explores the art of making accurate predictions about population parameters using sample data. We delve into the different methods of estimation, from point estimates that provide single-value approximations to interval estimates that provide ranges of likely values. Through practical examples and step-by-step explanations, we will demystify the process of estimation and impart the skills needed to confidently estimate population parameters.

Inferential statistics, the second pillar, transports us into the captivating realm of drawing conclusions about populations based on sample data. We embark on a journey of hypothesis testing, a cornerstone of inferential statistics, where we formulate hypotheses, analyze data, and interpret results to make decisions. We explore the intricacies of p-values, significance levels, and confidence intervals, unraveling the art of drawing meaningful inferences from data.

To facilitate understanding, this book adopts a hands-on approach, intertwining theory with practical applications. Real-world examples and case studies serve as beacons, illuminating the concepts and showing how they can be applied to a variety of fields, including social sciences, business, healthcare, and more.

Throughout this book, we encourage critical thinking and provide opportunities to apply the concepts through exercises and problems. We also provide supplementary resources, including datasets and software tutorials, to enhance your learning experience and empower you to apply the knowledge gained.

As we embark on this statistical journey together, I invite you to embrace the power of estimation and inferential statistics. Let us unlock the vast potential hidden within data, unravel the mysteries that surround us, and make confident,

evidence-based decisions. May this book serve as your trusted companion in your quest for statistical mastery.

Syllabus for Parametric statistics course

Instructor: Yakoub BOULAROUK
Contact Email: y.boularouk@centre-univ-mila.dz

Course Information

Target Audience: Third-year students in Applied Mathematics
Semester: 6
Course Title: Inferential Statistics
Credits: 9
Coefficient: 5

Course Objectives

Inferential Statistics is designed to equip students with the foundational concepts and theorems essential for understanding classical inferential statistical methods. Through this course, students will:

1. Gain a comprehensive understanding of fundamental concepts and theorems in inferential statistics.
2. Develop the ability to apply inferential statistical techniques to real-world problems and data sets.
3. Enhance critical thinking skills through analysis and interpretation of statistical results.
4. Prepare for advanced coursework and professional endeavors in applied mathematics and related fields.

Recommended Prerequisites

To successfully engage with Inferential Statistics, students are advised to have completed coursework in Analysis and Probability. It is strongly recommended that students possess a solid understanding of basic analysis and algebraic methods, along with proficiency in fundamental techniques of probability calculation.

Course Overview

Inferential Statistics builds upon the theoretical foundations established in previous mathematics courses and applies them to practical statistical analysis. Topics covered include but are not limited to:

- Probability Distributions
- Estimation Theory
- Hypothesis Testing
- Confidence Intervals
- Regression Analysis
- Analysis of Variance (ANOVA)

Assessment

The evaluation will consist of a combination of assignments, quizzes, examinations, and possibly projects, designed to assess understanding, application, and critical thinking skills in inferential statistics.

Communication and Collaboration

Students are encouraged to actively participate in discussions, seek clarification on concepts, and engage in collaborative learning opportunities both within and outside the classroom environment.

Please note that the syllabus is subject to modification as deemed necessary by the instructor to enhance the learning experience and meet the course objectives effectively.

Your active engagement and commitment to learning are vital for your success in this course. Looking forward to a productive semester!

Contents

Chapter 1

Sampling

Objectives

After studying this chapter, you should:

- Equip students with a comprehensive understanding of sampling and its critical role in statistical analysis.
- Foster analytical skills necessary for evaluating sampling techniques, avoiding common pitfalls, and applying statistical methods effectively.
- Prepare students to conduct independent research by providing the foundational knowledge required for proper sampling and data analysis.

Introduction

Sampling is a fundamental aspect of statistical analysis, serving as a bridge between theoretical concepts and practical applications. In many research scenarios, it is often impractical or impossible to collect data from an entire population due to constraints such as time, cost, and accessibility. Consequently, researchers rely on samples—subsets of the population—to draw conclusions and make inferences about the larger group.

This chapter provides a comprehensive overview of sampling techniques and their significance in statistical methodology. We will explore various sampling methods, including random, stratified, and systematic sampling, each with its advantages and challenges. Understanding these methods is crucial for minimizing bias and ensuring that samples accurately represent the population.

Furthermore, we will discuss common pitfalls that researchers may encounter when sampling and strategies to avoid them, enhancing the integrity of the data collected. Key statistical concepts related to samples, such as empirical mean and variance, will also be examined to equip students with the necessary tools to analyze and interpret sample data effectively.

By the end of this chapter, students will gain a solid understanding of sampling principles, enabling them to conduct independent research with confidence and precision. The knowledge acquired will not only enhance their analytical skills but also prepare them to navigate the complexities of data collection and analysis in various fields of study.

1.1 Notions of samples

Sampling is a fundamental concept in statistics, and it's essential to understand how it works, why it's used, and how to implement it. Here are more details about sampling, along with practical examples:

Why Sampling is Used: Sampling is used in statistics for several reasons:

1. **Cost-Efficiency:** Collecting data from an entire population can be time-consuming and expensive. Sampling allows researchers to gather a smaller, more manageable set of data points.

2. **Practicality:** In some cases, it's impossible to collect data from an entire population, especially if the population is large or widely dispersed.

3. **Destruction of Items:** When the data involves destructive testing or examination, such as in medical research, it's not feasible to collect data from the entire population.

4. **Accuracy:** When done correctly, sampling can provide accurate and representative information about the entire population.

1.1.1 Sampling Methods

There are different methods of sampling, depending on the research objectives and available resources. Some common sampling methods include:

1. Simple Random Sampling (SRS)

Simple Random Sampling (SRS) is a fundamental sampling technique used in statistics to select a subset of individuals from a larger population, ensuring that each individual has an equal chance of being chosen. This method helps obtain unbiased and representative data, which is crucial for making accurate inferences about the entire population.

Key Characteristics of SRS

- **Equal Probability**: Each member of the population has an identical chance of being included in the sample.
- **Random Selection**: Individuals are chosen randomly, often using random number generators or other unbiased selection methods.
- **Independence**: The selection of one individual does not affect the selection of another, ensuring true randomness.

Steps for Implementing SRS

1. **Define the Population**: Clearly outline the group from which you will draw your sample.
2. **Determine Sample Size**: Decide the number of individuals you need to represent the population accurately.
3. **Select Individuals Randomly**: Use methods like random number tables or software tools to ensure unbiased selection.

Example: If you have a population of 1,000 students and want a sample of 100, each student is assigned a number. A random number generator selects 100 numbers, ensuring each student has an equal likelihood of being chosen.

Solution: In this scenario, simple random sampling is employed to select a sample of students from a population of **1,000 students**. Each student is assigned a unique number from 1 to 1,000, and a random number generator is used to select 100 students. This method ensures that every student has an equal chance of being chosen. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,000 students**, each assigned a unique number from 1 to 1,000.

Assign Numbers

Each student is assigned a number as follows:

- Student 1: Number 1
- Student 2: Number 2
- Student 3: Number 3
- $\vdots$
- Student 1000: Number 1000

Randomly Select Sample

A random number generator selects **100 unique numbers** from the range of 1 to 1,000. For this example, assume that the selected numbers are:

- 5, 12, 37, 45, 78, 89, 123, 234, 345, 456, 567, 678, 789, 890, 901, 934, 987, 1000, etc.

Survey Selected Students

The researcher surveys the students corresponding to the selected numbers. Each of the 100 selected numbers represents a student that will be included in the sample.

Summary of Simple Random Sampling

Total Students	1,000
Sample Size	100
Selected Student Numbers	Randomly Generated (e.g., 5, 12, 37, ...)

Table 1.1: Simple Random Sampling Summary

This simple random sampling approach allows the researcher to obtain a sample that is representative of the entire student population, minimizing selection bias.

91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**
81	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**
71	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
61	**62**	**63**		**65**	**66**	**67**	**68**	**69**	**70**
51	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
	42	**43**	**44**		**46**	**47**	**48**	**49**	**50**
31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
11		**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Highlighted Students are Selected by Random Number Generator

Advantages and Limitations:
Advantages: SRS is easy to implement and provides unbiased samples, especially with large populations.
Limitations: It may not be feasible for very large populations without proper tools, and it may require a complete list of the population.

Conclusion: Simple random sampling is widely used in fields like survey research, market analysis, and social sciences for generating reliable, generalizable insights about a population.

Stratified Sampling

Stratified Sampling is a statistical sampling technique used to obtain a sample that represents various subgroups within a population. By dividing the population into distinct strata based on specific characteristics, this method ensures that each subgroup is adequately represented in the final sample, leading to more accurate and reliable results.

Key Characteristics of Stratified Sampling

- **Division into Strata**: The population is divided into distinct subgroups

based on shared characteristics such as age, gender, income, or education level.

- **Random Sampling within Strata**: Random samples are taken from each stratum, ensuring each subgroup is represented.
- **Proportional or Equal Allocation**: Samples can be drawn in proportion to the size of each stratum in the population or with equal numbers from each stratum.

Steps for Implementing Stratified Sampling

1. **Define the Population**: Clearly outline the entire population from which samples will be drawn.
2. **Identify Strata**: Determine relevant characteristics for stratification and categorize the population accordingly.
3. **Determine Sample Size**: Decide the total number of individuals needed in the sample.
4. **Select Samples from Each Stratum**: Use random sampling methods to select samples from each identified stratum.

Example: If a university has 1500 students and you want to study their study habits, you could divide the students into three strata: freshmen, sophomores, juniors, and seniors. If you want a sample of 150 students.

Solution:In this scenario, stratified sampling involves dividing the university's **1,500 students** into strata based on their academic levels—**Licence, Master, and Doctorat**. With a desired sample size of **150**, you would draw approximately **100 students from Licence**, **40 from Master**, and **10 from Doctorat** to achieve a balanced sample. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,500 students** at a university.

Create Strata

Divide the population into three strata based on academic level Licence, Master and Doctorat

Determine the Sample Size

The goal is to select a sample of **150** students. Allocate approximately 100 students from the Licence stratum, **40** students from the Master stratum and **10** students from the Doctorat stratum.

This allocation is based on the proportion of each stratum in the overall population. Assuming the distribution is as follows: 1,000 students in **Licence**, 400 students in **Master** and 100 students in **Doctorat**.

The calculations for sampling would be:

$$\begin{aligned} \text{Licence:} \quad & \frac{1000}{1500} \times 150 = 100 \\ \text{Master:} \quad & \frac{400}{1500} \times 150 = 40 \\ \text{Doctorat:} \quad & \frac{100}{1500} \times 150 = 10 \end{aligned}$$

Random Sampling within Each Stratum

For each academic level, select the designated number of students randomly to ensure that the sample is representative across all strata. This helps avoid potential bias related to differences in study habits by academic level.

Summary Table

Stratum	**Total Students in Stratum**	**Sampled Students**
Licence	1,000	100
Master	400	40
Doctorat	100	10
Total	**1,500**	**150**

Table 1.2: Stratified Sampling Summary

This stratified sampling approach ensures balanced representation across all academic levels, allowing conclusions drawn about study habits to reflect each subgroup proportionally.

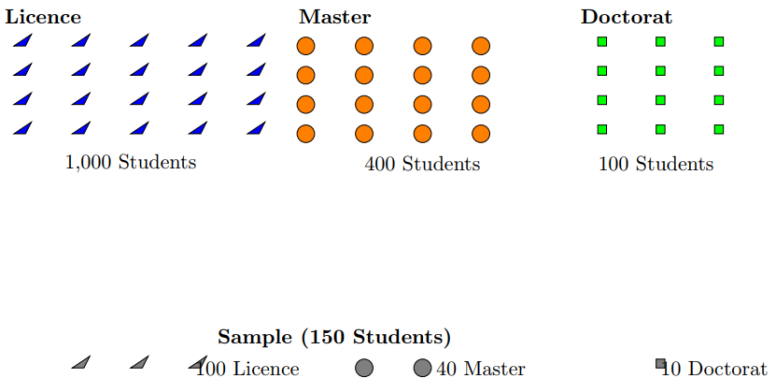


Figure 1.1: Probability Density Function (PDF) of a Normal Distribution

Advantages and Limitations:
Advantages: Stratified sampling increases precision and ensures that all relevant subgroups are represented, reducing sampling bias.
Limitations: This method can be more complex to implement than simple random sampling and requires detailed population information to define strata.

Conclusion: Stratified sampling is an effective sampling technique widely used in various fields, such as social sciences, market research, and health studies, to obtain reliable insights by ensuring all significant subgroups are included.

Systematic Sampling

Systematic Sampling is a statistical sampling technique used to select a sample from a larger population by choosing individuals at regular intervals. This method is straightforward and can be more efficient than random sampling methods, particularly when a complete list of the population is available.

Key Characteristics of Systematic Sampling

- **Fixed Interval Selection**: Individuals are selected at regular intervals from a randomly chosen starting point.
- **Simple to Implement**: The process is often easier to carry out than other sampling methods, especially for large populations.
- **Requires a Complete List**: A complete and ordered list of the population is necessary for systematic sampling to be effective.

Steps for Implementing Systematic Sampling

1. **Define the Population**: Clearly outline the group from which you will draw your sample.

2. **Determine Sample Size**: Decide how many individuals you need in the sample.

3. **Calculate the Sampling Interval**: Determine the sampling interval (k) by dividing the population size (N) by the desired sample size (n) using the formula $k = \frac{N}{n}$.

4. **Select a Random Starting Point**: Randomly select a starting point between 1 and k.

5. **Select Individuals at Regular Intervals**: Starting from the chosen point, select every k-th individual until the sample size is reached.

Example: Suppose you have a population of 1,000 employees in a company and you want a sample of 100 employees. If you calculate a sampling interval of $k = 10$ (i.e., $1000 \div 100 = 10$), you would randomly select a starting point between 1 and 10. If you choose 3, you would then select the 3rd, 13th, 23rd, and so on, until you reach the desired sample size.

Solution: In this scenario, systematic sampling is used to select a sample of **100 employees** from a company population of **1,000 employees**. The sampling interval is calculated as $k = 10$, meaning every 10th employee will be selected after a random starting point. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of **1,000 employees** in a company.

Determine the Sampling Interval

To achieve the sample size, calculate the sampling interval as follows:

$$k = \frac{1000}{100} = 10$$

This means every 10th employee will be selected.

Select a Random Starting Point

Choose a random starting point between 1 and 10. For this example, assume we select **3** as the starting point.

Select Employees Using the Interval

Starting from employee **3**, select every 10th employee: **3, 13, 23, 33,** and so on, until 100 employees are chosen.

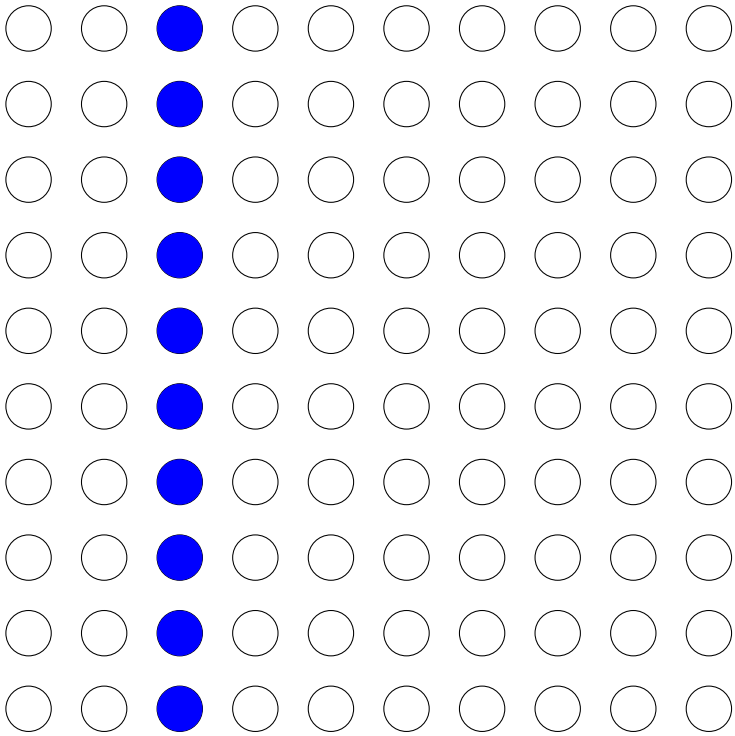
Sampling Interval	10
Starting Point	3
Employees Selected	3, 13, 23, 33, . . . , 993

Table 1.3: Systematic Sampling Summary

Summary of Systematic Sampling

This systematic sampling approach ensures that every part of the population has an equal chance of being included, based on the interval.

Population of 1,000 Employees



Sample of 100 Employees Selected (in Blue)

Advantages and Limitations:
Advantages: Systematic sampling is easy to implement and can be more efficient than random sampling. It can also be less time-consuming.
Limitations: It may introduce bias if there is a hidden pattern in the population that corresponds to the sampling interval, and it requires a complete list of the population.

Conclusion: Systematic sampling is a practical and efficient technique widely used in various fields such as quality control, survey research, and social sciences, where a straightforward method for sampling is required.

Cluster Sampling

Cluster Sampling is a statistical sampling technique used to select a sample from a larger population by dividing the population into groups or clusters and then randomly selecting entire clusters. This method is particularly useful when the population is large and widely dispersed, making it impractical to conduct a complete sampling.

Key Characteristics of Cluster Sampling

- **Division into Clusters**: The population is divided into non-overlapping groups or clusters based on specific characteristics, such as geographic location or organizational units.
- **Random Selection of Clusters**: Entire clusters are randomly selected, and all or a sample of individuals within those clusters are surveyed.
- **Cost-Effective**: This method can reduce costs and time when surveying widely dispersed populations.

Steps for Implementing Cluster Sampling

1. **Define the Population**: Clearly outline the entire population from which samples will be drawn.
2. **Divide into Clusters**: Identify the clusters within the population based on relevant characteristics.
3. **Randomly Select Clusters**: Randomly select a few clusters to be included in the sample.
4. **Collect Data from Selected Clusters**: Survey all individuals within the selected clusters or a random sample from those clusters.

Example: In a study of the dietary habits of school children in a large city, a researcher might divide the city into different school districts (clusters). If there are 9 districts, the researcher could randomly select 3 districts and then survey all the children in those districts.

Solution: In this scenario, cluster sampling is used to study dietary habits among school children in a large city. Here, **9 school districts** represent clusters, and the researcher will **randomly select 3 of these districts** to survey all the children within those clusters. Here's a step-by-step breakdown of the process:

1. Define the Population

The population consists of all school children across **9 school districts** in a large city.

Define Clusters

Each of the **9 districts** represents a cluster of school children.

Randomly Select Clusters

The researcher randomly selects **3 districts** from the 9 available districts. For this example, assume that Districts **2, 5, and 8** are selected.

Survey All Children in Selected Clusters

Once the districts are selected, the researcher surveys all school children within Districts **2, 5, and 8**.

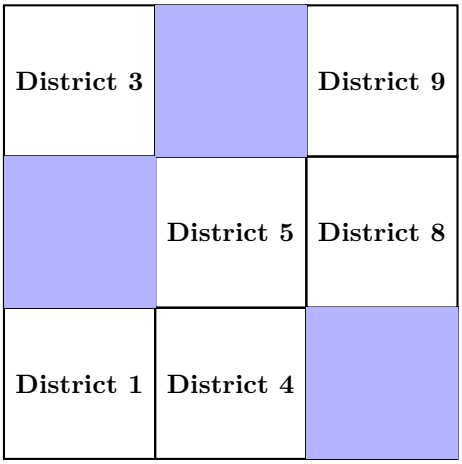
Summary of Cluster Sampling

Total Districts (Clusters)	9
Districts Selected	3 (Districts 2, 5, 8)
Sampled Children	All children in selected districts

Table 1.4: Cluster Sampling Summary

This cluster sampling approach allows the researcher to gather data from specific geographic areas within the city, making it more efficient than sampling children across all districts.

Population of 9 School Districts



Selected Districts (2, 5, and 8) are Highlighted in Blue

Advantages and Limitations:
Advantages: Cluster sampling is cost-effective and practical for large populations, especially when it is difficult to create a complete list of individuals.
Limitations: It can introduce higher sampling error compared to other methods, as individuals within clusters may be more similar to each other than to the broader population.

Conclusion: Cluster sampling is extensively used in fields like epidemiology, education, and social sciences for studying large and geographically dispersed populations.

1.1.2 Common Pitfalls to Avoid:

When conducting sampling, it's crucial to avoid common pitfalls, such as:

1. **Sampling Bias:** This occurs when the sampling method systematically excludes or over-represents certain groups in the population.
2. **Non-Response Bias:** If a significant portion of those selected for the sample does not participate, it can lead to a non-response bias.
3. **Sampling Error:** This is the natural variation that occurs when working with samples instead of the entire population. It can be minimized by increasing the sample size.
4. **Confounding Variables:** Failure to control for confounding variables (variables that are related to both the independent and dependent variables) can affect the results of the study.

In summary, sampling is a critical technique in statistics used to gather data efficiently and effectively from a subset of a larger population. The choice of sampling method depends on research goals and available resources. Proper sampling techniques are essential to ensure the validity and reliability of statistical analyses and research findings.

1.2 Statistics of samples: empirical mean, empirical variance

Data statistics often involve calculating various descriptive statistics to summarize and understand the characteristics of a dataset. Two essential statistics are the empirical mean (also known as the sample mean) and the empirical variance (sample variance).

1.2.1 Empirical Mean (Sample Mean):

The empirical mean, often denoted as $\bar{x}$, represents the average or central tendency of a dataset. It is calculated as the sum of all data points divided by the number of data points in the dataset.

The formula for the empirical mean is:

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

where:

$\bar{x}$: Empirical mean
n : Number of data points in the dataset
x_i : Individual data points

Theorem 1.2.1. *(Distribution of $\bar{X}$): Let $(X_1, \ldots, X_n)$ be a simple random sample of size n of a random variable $N(\mu, \sigma^2)$. Then, the sample mean $\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ satisfies*

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

.

Proof. To show that $\bar{X}$ has the stated distribution, we compute its expectation and variance, and then verify its distribution.

1. Expectation of $\bar{X}$:

Using the linearity of expectation, we have:

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \cdot n\mu = \mu.$$

2. Variance of $\bar{X}$:

Since the X_i are independent and identically distributed, we have:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

3. Distribution of $\bar{X}$:

Since each X_i follows a normal distribution, and $\bar{X}$ is a linear combination of independent normal random variables, $\bar{X}$ itself follows a normal distribution.

Therefore, we conclude that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

□

1.2.2 Empirical Variance (Sample Variance):

The empirical variance, often denoted as S^2, measures the spread or variability of data points in the dataset. It quantifies how much individual data points

deviate from the mean. The larger the variance, the more dispersed the data points are. The sample variance is given by

$$S^2 := \frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1} X_i^2 - \bar{X}^2.$$

The sample quasi-variance will also play a relevant role in inference. It is defined by simply replacing n with $n-1$ in the factor of S^2:

$$S'^2 := \frac{1}{n-1}\sum_{i=1}(X_i - \bar{X})^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1} X_i^2 - \frac{n}{n-1}\bar{X}^2.$$

Before establishing the sampling distributions of S^2 and S'^2, we obtain in the first place their expectations. For that aim, we start by decomposing the variability of the sample with respect to its expectation μ in the following way:

$$\sum_{i=1}(X_i - \mu)^2 = \sum_{i=1}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

Taking expectations, we have

$$n\sigma^2 = nE[S^2] + \frac{n\sigma^2}{n},$$

and then, solving for the expectation,

$$E[S^2] = (n-1)\sigma^2.$$

Therefore,

$$E[S'^2] = \frac{n}{n-1}E[S^2] = \sigma^2.$$

Recall that this computation does not employ the assumption of sample normality, hence it is a general fact for S^2 and S'^2 irrespective of the underlying distribution. It also shows that S^2 is not "pointing" towards σ^2 but to a slightly smaller quantity, whereas S'^2 is "pointing" directly to σ^2. This observation is related to the bias of an estimator and will be treated in detail in Section 3.1.

In order to compute the sampling distributions of S^2 and S'^2, it is required to obtain the sampling distribution of the statistic $\sum_{i=1} X_i^2$ when the sample is generated from a $N(0,1)$, which will follow a chi-square distribution.

Theorem 1.2.2. *2.2 (Fisher's Theorem) If $(X_1, \ldots, X_n)$ is a simple random sample of a $N(\mu, \sigma^2)$ random variable, then S^2 and $\bar{X}$ are independent, and*

$$\frac{nS^2}{\sigma^2} = (n-1)S'^2 \sim \chi^2_{n-1}.$$

Proof. To demonstrate that the ratio $\frac{nS^2}{\sigma^2}$ follows a chi-squared distribution with $n-1$ degrees of freedom, we can use the concept of chi-squared distribution and the properties of the sample variance.

Consider the random variable $\frac{nS^2}{\sigma^2}$:

$$\frac{nS^2}{\sigma^2} = \frac{n}{\sigma^2} \cdot \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

The expectation of $\frac{nS^2}{\sigma^2}$:

$$E\left(\frac{nS^2}{\sigma^2}\right) = \frac{n}{\sigma^2} \cdot E(S^2) = \frac{n}{\sigma^2} \cdot \sigma^2 = n$$

The expectation of $\frac{nS^2}{\sigma^2}$ is n, which is the degrees of freedom of a chi-squared distribution. Since the degrees of freedom of a chi-squared distribution can be expressed as $k-1$ (where k is the number of degrees of freedom), we have:

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

So, the ratio $\frac{nS^2}{\sigma^2}$ follows a chi-squared distribution with $n-1$ degrees of freedom. □

1.3 The Density Function

The **probability density function (PDF)**, denoted as $f(x)$, describes the relative likelihood of a continuous random variable X taking on a specific value. Unlike discrete probability distributions, where probabilities can be assigned to specific outcomes, the PDF provides a way to describe the distribution of probabilities across a continuum of possible values.

Properties of the PDF

- **Non-negativity**: The value of the PDF is always non-negative, meaning $f(x) \geq 0$ for all x. This is because probabilities cannot be negative.
- **Normalization**: The total area under the curve of the PDF across all possible values of x is equal to 1:

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

  This ensures that the total probability of the random variable falling within its possible range is 1.

Interpretation of the PDF

The height of the PDF at any point x, denoted $f(x)$, indicates the **relative likelihood** of X being close to x. Higher values of $f(x)$ suggest that values around x are more likely compared to those around points where $f(x)$ is lower.

While $f(x)$ provides a density, it does not represent the probability of X taking on a specific value. Instead, probabilities are computed over intervals. For example, the probability that X falls between two values a and b is given by the area under the curve between these two points:

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

Graphical Representation

The following graphical representation illustrates the concept of the PDF using a normal distribution.

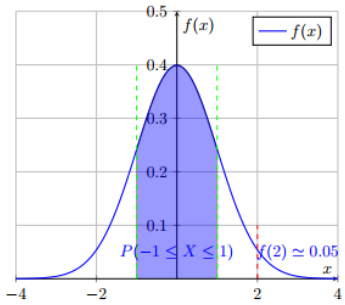


Figure 1.2: Probability Density Function (PDF) of a Normal Distribution

Explanation of the Graph

- **Blue Curve**: The blue line represents the PDF $f(x)$ of a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- **Shaded Area**: The shaded area under the curve between the points $a = -1$ and $b = 1$ represents the probability $P(-1 \leq X \leq 1)$. This area indicates the likelihood of the random variable X falling within this interval.
- **Vertical Dashed Lines**: The dashed vertical lines at a and b help to visually identify the interval for which the probability is being calculated.

The probability density function (PDF) is a crucial concept in statistics and probability, allowing for the modeling and understanding of continuous random variables. By visualizing the PDF, we gain insights into how probabilities are distributed and how to compute probabilities over specific intervals.

1.4 Gaussian Samples

A Gaussian sample, also known as a normal sample, represents a set of data points that follow a Gaussian (normal) distribution. The Gaussian distribution is characterized by its mean (μ) and standard deviation (σ).

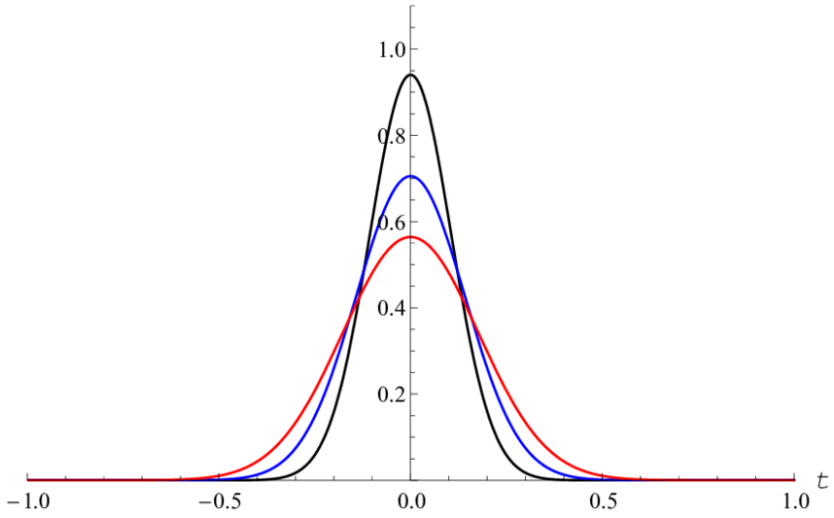


Figure 1.3: Gaussian sample

For example, let's consider a Gaussian sample with the following properties:

μ : Mean
σ : Standard Deviation

Suppose we have a Gaussian sample of 100 data points:

$$X = \{x_1, x_2, \ldots, x_{100}\}$$

where x_i represents an individual data point.
The Gaussian distribution for this sample can be expressed as:

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

To provide concrete values, let's say:

$$\mu = 10$$
$$\sigma = 2$$

Then, a few data points from this Gaussian sample might be:

$$X = \{11.23, 9.75, 10.50, 9.98, \ldots\}$$

These data points are drawn from the Gaussian distribution with the specified mean and standard deviation.

1.5 Application

1.5.1 Solved exercices

Exercice 1.1. ***(Simple Random Sampling)*** *You have a population of 1,000 students in a school. You want to select a simple random sample of 100 students to conduct a survey about their study habits.*
Explain how you would carry out this sampling process.

Exercice 1.2. ***(Systematic Sampling)*** *You have a list of 500 employees in a company, and you want to select a systematic sample of 50 employees for a performance evaluation.*
Describe the steps you would take to choose the sample and determine the sampling interval.

Exercice 1.3. ***(Stratified Sampling)*** *You are conducting a political opinion poll in a city with a diverse population.*

1. *Outline how you would use stratified sampling to ensure a representative sample.*
2. *Identify the strata and explain how you would select individuals from each stratum.*

Exercice 1.4. ***(Cluster Sampling)*** *You are conducting a survey of households in a large urban area.*

1. *Describe how you would use cluster sampling to select your sample.*
2. *Identify the clusters and explain how you would choose which clusters to include in your sample.*

Exercice 1.5. ***(Convenience Sampling)***

1. *Discuss the advantages and disadvantages of convenience sampling.*
2. *Provide an example of a situation where convenience sampling might be appropriate and another situation where it might lead to biased results.*

Exercice 1.6. ***(Non-Probability Sampling)***

1. *Explain what non-probability sampling is and why it might be used in research.*
2. *Discuss the potential limitations of non-probability sampling methods.*

Exercice 1.7. ***(Sampling Error)***

1. *Define sampling error and explain why it is important to consider when interpreting the results of a sample.*
2. *Provide an example of how sampling error can impact the accuracy of survey findings.*

Exercice 1.8. ***(Sample Size Determination)*** *You are designing a survey to estimate the proportion of customers satisfied with a new product.*
How would you determine the appropriate sample size to achieve a desired level of confidence and margin of error?

Exercice 1.9. ***(Oversampling and Undersampling)***

1. *Discuss the concepts of oversampling and undersampling in the context of survey sampling.*

2. *Explain when and why these techniques might be employed.*

Exercice 1.10. ***(Sampling in Qualitative Research)***

1. *Explain how sampling is approached in qualitative research.*

2. *Highlight the differences between qualitative and quantitative sampling methods.*

Exercices solutions

Correction exercice 1.1. *To select a simple random sample of 100 students, you can follow these steps:*

1. *Assign a unique identifier to each of the 1,000 students, such as a student ID number.*

2. *Use a random number generator to select 100 random numbers between 1 and 1,000. These numbers will correspond to the selected students.*

3. *Contact the students corresponding to the selected numbers and invite them to participate in the survey.*

Correction exercice 1.2. *To select a systematic sample of 50 employees, follow these steps:*

1. *Determine the total number of employees in the list (500).*

2. *Calculate the sampling interval, which is the total number of employees divided by the desired sample size:* $500/50 = 10$.

3. *Randomly select a starting point between 1 and 10. Let's say you choose 3.*

4. *Then, select every 10th employee from the list, starting from the 3rd employee, until you have a sample of 50 employees.*

Correction exercice 1.3. *To ensure a representative sample in a diverse city, you can use stratified sampling as follows:*

1. *Identify relevant strata in the population, such as age groups (e.g., 18-30, 31-45, 46-60), gender, and residential areas.*
2. *Randomly select a sample from each stratum using simple random sampling or another appropriate sampling method.*
3. *Ensure that the sample size from each stratum is proportional to its population size to maintain representation.*
4. *Combine the samples from each stratum to form the final representative sample.*

Correction exercice 1.4. *To use cluster sampling for households in a large urban area, follow these steps:*

1. *Divide the urban area into clusters, which could be neighborhoods or city blocks.*
2. *Randomly select a subset of clusters to include in your sample. This can be done using simple random sampling or other methods.*
3. *Survey all households within the selected clusters.*
4. *Ensure that the chosen clusters are representative of the entire urban area to maintain the validity of your sample.*

Correction exercice 1.5. *Convenience Sampling:*
***Advantages**:*

- *Quick and easy to implement.*
- *Cost-effective and requires fewer resources.*
- *Useful for preliminary or exploratory research.*

***Disadvantages**:*

- *Highly prone to selection bias, as participants are chosen based on convenience.*
- *Results may not be representative of the population.*
- *Limited generalizability.*

Examples

- *(Appropriate Use): Conducting a quick survey of shoppers in a mall to gather initial feedback on a newly opened store's layout and products.*
- *(Biased Results): Using convenience sampling to assess public opinion on a political issue by surveying only individuals who attend a specific political rally, leading to a skewed perspective.*

Correction exercice 1.6. *Non-Probability Sampling:*

Non-probability sampling is a method of selecting participants for a research study in which not every member of the population has a known and equal chance of being included. It is typically used when probability sampling is not feasible or practical.

***Reasons for using non-probability sampling**:*

- *Accessibility: When it is difficult to access or identify all members of the population.*
- *Cost-efficiency: When conducting probability sampling is expensive or time-consuming.*
- *Specific research goals: When the focus is on certain groups or characteristics within the population.*

***Limitations of non-probability sampling**:*

- *Lack of representativeness: Non-probability samples may not accurately represent the entire population, leading to potential bias.*
- *Limited generalizability: Findings from non-probability samples may not generalize to the broader population.*
- *Difficulty in estimating sampling error: Non-probability samples make it challenging to estimate the margin of error and level of confidence in study results.*

Correction exercice 1.7. ***Sampling Error**:*

Sampling error is the difference between a sample statistic (e.g., sample mean or proportion) and the true population parameter it is meant to estimate. It arises because a sample is only a subset of the entire population, and random variation can lead to discrepancies between sample and population values.

***Importance of considering sampling error**:*

- *Accuracy assessment: Sampling error allows researchers to quantify the degree of uncertainty associated with sample estimates.*
- *Confidence intervals: Sampling error is used to construct confidence intervals, which provide a range within which the population parameter is likely to fall.*
- *Decision-making: Understanding sampling error helps in making informed decisions and assessing the reliability of survey findings.*

***Example**: Suppose a survey estimates that the average income of a city's residents is* $50,000$ *with a margin of error of* $2,000$*. This means that the true average income of the population is likely to fall within the range of* $48,000$ *to* $52,000$ *due to sampling error.*

Correction exercice 1.8. *To determine the appropriate sample size for estimating the proportion of customers satisfied with a new product with a desired level of confidence and margin of error, you can use the following steps:*

1. *Specify the desired level of confidence (e.g., 95*
2. *Estimate the population proportion from prior information or pilot data if available. If not, use a conservative estimate of 0.5 (maximum variability).*
3. *Use a sample size formula for estimating proportions, such as:*

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{E^2}$$

where:

- *n is the required sample size.*
- *Z is the critical value corresponding to the desired level of confidence (e.g., 1.96 for 95*
- *p is the estimated population proportion.*
- *E is the desired margin of error.*

4. *Calculate the sample size using the formula and round it up to the nearest whole number to ensure a sufficient sample.*

Correction exercice 1.9. *1. **Oversampling and Undersampling**:*

- ***Oversampling**: Oversampling involves deliberately selecting more individuals from a specific subgroup or stratum of the population than would be proportionally represented in a simple random sample. This technique is used to ensure an adequate sample size for rare or underrepresented groups within the population. It allows for more precise estimation of characteristics of the oversampled subgroup.*
- ***Undersampling**: Undersampling, on the other hand, involves selecting fewer individuals from a specific subgroup or stratum than their proportion in the population. Undersampling is often used when researchers want to reduce the cost or complexity of data collection or when the subgroup is well-represented in the population and does not require a large sample.*

2. *When and why to use these techniques:*

- *Oversampling is employed when researchers want to ensure that the sample includes enough individuals from a particular subgroup, especially when that subgroup is small or critical for the study's objectives. For example, in a nationwide health survey, oversampling might be used to ensure an adequate number of participants from minority communities.*

- *Undersampling can be used when a subgroup is large and already well-represented in the population. By undersampling, researchers can reduce survey costs and still obtain accurate estimates of overall population characteristics. For example, in a survey of educational attainment, the general population may already have a significant number of individuals with high school diplomas, making it unnecessary to oversample this group.*

Correction exercice 1.10. *In qualitative research, sampling is approached differently compared to quantitative research. Qualitative sampling focuses on selecting participants or sources of data based on their ability to provide rich and in-depth information relevant to the research questions. Here are key points about sampling in qualitative research and differences from quantitative sampling:*

- *Purposeful sampling: Qualitative researchers often use purposeful or purposive sampling, where participants are selected intentionally based on specific criteria, such as their expertise, experiences, or relevance to the research topic. This approach aims to gather diverse and information data.*
- *Sample size: Qualitative research typically involves smaller sample sizes compared to quantitative research. The emphasis is on depth rather than breadth, as researchers seek detailed insights from a limited number of participants.*
- *Non sampling: Qualitative sampling methods are often non probabilistic, meaning that every member of the population may not have an equal chance of being selected. The goal is to select participants who can provide unique perspectives and insights.*
- *Data saturation: In qualitative research, sampling may continue until data saturation is reached. Data saturation occurs when new participants or data no longer provide substantially different insights or themes, indicating that the sample size is sufficient for the research objectives.*
- *Snowball sampling: Snowball sampling is a technique often used in qualitative research, where participants refer or introduce other potential participants who meet the criteria. This approach is particularly useful when researching hidden or hard to reach populations.*
- *Emphasis on context: Qualitative sampling considers the context in which data are collected, acknowledging that the richness of data is often influenced by the environment and interactions between participants.*

In summary, qualitative sampling is characterized by purposeful selection, smaller sample sizes, non-probabilistic methods, and a focus on depth and context. It aims to capture the complexity of human experiences and perspectives rather than generalizability to a larger population, which is a primary focus of quantitative sampling.

1.5.2 Unsolved exercices

Exercice 1.1. ***(Multi-Stage Sampling)***
You are conducting a national survey on health and need to sample individuals from various regions across the country.
Describe the steps involved in implementing a multi-stage sampling method, specifying how you would select regions, cities, neighborhoods, and households.

Exercice 1.2. ***(Quota Sampling)***
You want to survey 200 people from a town to understand their opinions on a new public policy. You aim to ensure the sample includes proportional representation based on gender and age groups.
Outline how you would use quota sampling to achieve this and discuss any potential biases that might arise.

Exercice 1.3. ***(Snowball Sampling)***
You are conducting a study on a rare medical condition, and the population of individuals affected is difficult to identify.
Explain how snowball sampling could be applied in this scenario. Include a discussion on the advantages and limitations of using this approach.

Exercice 1.4. ***(Judgmental Sampling)***
A researcher wants to evaluate the effectiveness of a teaching method by selecting schools known for their high academic performance.
Describe how judgmental (or purposive) sampling might be used in this study. Discuss its advantages and potential sources of bias.

Exercice 1.5. ***(Sampling for Longitudinal Studies)***
You are designing a longitudinal study to track the health outcomes of individuals over 10 years.
Describe the considerations for selecting a sample that will remain representative and engaged throughout the study period. How would you address potential attrition in the sample?

Chapter 2

Methods of estimator construction

Objectives

After studying this chapter, you should:

- Understand the concepts and importance of point estimation in statistics.
- Differentiate between the method of moments and the maximum likelihood method for parameter estimation.
- Recognize key characteristics of an estimator, including bias, efficiency, and the Cramer-Rao bound.
- Learn how to construct and interpret confidence intervals for estimation.
- Apply interval estimation techniques to normal samples and populations.

Introduction

Estimation is a core aspect of statistical inference, providing the means to draw conclusions about population parameters based on sample data. This chapter delves into the various methods of estimator construction, which are crucial for accurately estimating these parameters. We will begin by exploring point estimation, where we seek a single value that best represents the unknown parameter. Two widely used techniques for point estimation—the Method of Moments and the Maximum Likelihood Method—will be examined in detail, highlighting their principles, advantages, and limitations.

Understanding the characteristics of estimators is essential for evaluating their performance. We will discuss key concepts such as bias, mean squared error, Fisher information, and the Cramer-Rao bound, which together provide a framework for assessing the quality of different estimators. Moreover, we will introduce the concept of efficiency, which indicates how well an estimator performs relative to others.

In addition to point estimation, this chapter will cover estimation through confidence intervals, offering a range of methods for determining intervals that capture the true parameter value with a specified level of confidence. We will also focus on constructing confidence intervals for normal samples and populations, enabling practitioners to make informed decisions based on the results of their analyses.

By the end of this chapter, readers will have a solid foundation in the methods of estimator construction, equipping them with the tools necessary for effective statistical analysis and interpretation. This knowledge is vital for researchers and practitioners who aim to make reliable inferences from sample data.

2.1 Point Estimation

2.1.1 Method of Moments

The method of moments is a commonly used approach for estimating the parameters of a distribution by matching sample moments to theoretical (population) moments. The idea is to calculate sample moments from the data and equate them to the corresponding population moments, which are functions of the unknown parameters of the distribution. Solving these equations provides estimates for the parameters.

Formula for the Method of Moments Estimator

Let θ represent the parameter of interest. The method of moments estimator $\hat{\theta}$ is obtained by solving the equation that sets the k-th sample moment equal to the k-th theoretical moment. For a sample $X_1, X_2, \ldots, X_n$, the sample moments are calculated as the averages of powers of the sample values.

For example, the first sample moment (the sample mean) is given by:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

where X_i are the observed data points. The estimator $\hat{\theta}$ is the method of moments estimate of the parameter θ.

Steps for the Method of Moments:

1. **Compute the sample moments**: Based on the data, calculate the sample moments.
2. **Determine the population moments**: Express the population moments as functions of the unknown parameter(s) θ.
3. **Set sample moments equal to population moments**: Form equations by equating the sample moments to the corresponding population moments.
4. **Solve for** θ: Solve the system of equations to estimate the parameter(s) θ.

Example 1: Estimating the Parameter of an Exponential Distribution

Let's consider the case of an exponential distribution with unknown rate parameter λ. We have a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution, and we want to estimate λ using the method of moments.

Step 1: Population Moment of the Exponential Distribution For the exponential distribution, the first population moment (the expected value of the random variable X) is:

$$\mu_1 = \frac{1}{\lambda}$$

Step 2: Sample Moment The first sample moment (the sample mean) is:

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Step 3: Equate Sample Moment to Population Moment Equating the sample moment to the population moment gives:

$$\frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Step 4: Solve for λ Solving for λ gives the method of moments estimator:

$$\hat{\lambda} = \frac{1}{\bar{X}} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Thus, the method of moments estimator for λ is the inverse of the sample mean.

Example 2: Estimating the Parameters of a Uniform Distribution

Now, consider a random sample $X_1, X_2, \ldots, X_n$ from a continuous uniform distribution on the interval $[a, b]$, where a and b are unknown parameters. We want to estimate a and b using the method of moments.

Step 1: Population Moments of the Uniform Distribution For a uniform distribution on $[a, b]$, the first population moment (mean) and second population moment (variance) are:

$$\mu_1 = \frac{a+b}{2}, \quad \mu_2 = \frac{(b-a)^2}{12}$$

Step 2: Sample Moments The first and second sample moments (mean and variance) are:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_1)^2$$

Step 3: Equate Sample Moments to Population Moments Equating the sample moments to the population moments gives the following system of equations:

$$\frac{a+b}{2} = \hat{\mu}_1, \quad \frac{(b-a)^2}{12} = \hat{\mu}_2$$

Step 4: Solve for a and b From the first equation:

$$a + b = 2\hat{\mu}_1$$

From the second equation:

$$(b-a)^2 = 12\hat{\mu}_2$$

Solving this system of equations yields the method of moments estimators for a and b:

$$\hat{a} = \hat{\mu}_1 - \sqrt{3\hat{\mu}_2}, \quad \hat{b} = \hat{\mu}_1 + \sqrt{3\hat{\mu}_2}$$

Verification: Unbiasedness of the Estimators

To prove that the method of moments estimators for a and b are unbiased, we need to calculate the expected values of the estimators and verify that they equal the true values of a and b. For example, consider the estimator for a:

$$\mathbb{E}(\hat{a}) = \mathbb{E}\left(\hat{\mu}_1 - \sqrt{3\hat{\mu}_2}\right)$$

Since $\hat{\mu}_1$ and $\hat{\mu}_2$ are consistent estimators of the true mean and variance, we have:

$$\mathbb{E}(\hat{\mu}_1) = \mu_1, \quad \mathbb{E}(\hat{\mu}_2) = \mu_2$$

Therefore, $\hat{a}$ is an unbiased estimator of a, and similarly for b.

2.1.2 Maximum Likelihood Method

The Maximum Likelihood Estimation (MLE) method is a widely used approach in statistical inference for estimating the parameters of a statistical model. It aims to find the parameter θ that maximizes the likelihood function, which represents the probability of observing the given sample data under the model parameterized by θ.

In simpler terms, the MLE method finds the parameter value that makes the observed data most probable.

Likelihood Function

The likelihood function $L(\theta)$ is a key concept in MLE. For a given sample of size n, consisting of independent observations $X_1, X_2, \ldots, X_n$, the likelihood function is defined as the joint probability (or probability density) of observing the sample, assuming a particular value of the parameter θ.

Mathematically, if the observations $X_1, X_2, \ldots, X_n$ are drawn independently from a probability distribution with a probability density function (PDF) $f(X_i; \theta)$, the likelihood function is given by the product of the individual probabilities (or densities):

$$L(\theta) = f(X_1; \theta) \cdot f(X_2; \theta) \cdot \ldots \cdot f(X_n; \theta)$$

Or, equivalently:

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

In this expression, $f(X_i; \theta)$ is the probability density (or mass) function of the observation X_i under the parameter θ.

Interpretation: The likelihood function can be interpreted as a measure of how likely it is to observe the given sample for different values of θ. The larger $L(\theta)$, the more likely the observed data is under the model with parameter θ. Thus, the goal of MLE is to find the value of θ that maximizes this likelihood function, leading to the "best fit" of the model to the observed data.

Log-Likelihood Function

Because the likelihood function involves a product of probabilities, it can sometimes lead to computational difficulties, especially with large sample sizes. Therefore, it is common to work with the *log-likelihood* function, which is the natural logarithm of the likelihood function.

The log-likelihood function is defined as:

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(X_i; \theta)$$

Since the natural logarithm is a monotonic function, maximizing the log-likelihood function $\ell(\theta)$ leads to the same parameter estimate as maximizing the likelihood function $L(\theta)$. However, the log-likelihood function is often easier to work with because the product in the likelihood function becomes a sum in the log-likelihood.

Maximum Likelihood Estimator (MLE)

The Maximum Likelihood Estimator (MLE) of the parameter θ is the value of θ that maximizes the likelihood (or equivalently, the log-likelihood) function.

Formally, the MLE $\hat{\theta}_{MLE}$ is defined as:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$$

or, equivalently:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ell(\theta)$$

This means that the MLE is the value of θ that maximizes the likelihood (or log-likelihood) function, making the observed data as probable as possible.

Steps to Obtain the MLE:

1. **Write the likelihood function**: Start by expressing the likelihood function $L(\theta)$ based on the PDF (or PMF) of the observed data.
2. **Take the logarithm**: Compute the log-likelihood function $\ell(\theta)$ to simplify the product into a sum, making the calculation more manageable.
3. **Differentiate**: Differentiate the log-likelihood function with respect to the parameter θ to find the critical points where the likelihood is maximized.

4. **Solve**: Set the derivative equal to zero and solve for θ. The solution $\hat{\theta}_{MLE}$ is the MLE, the parameter value that maximizes the likelihood.

5. **Verify**: Ensure that the critical point found corresponds to a maximum (by checking the second derivative or by other means, such as verifying that the likelihood is concave).

Example 1: Maximum Likelihood for Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ be a sample of independent random variables following an exponential distribution with parameter λ. The probability density function of the exponential distribution is given by:

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

The likelihood function $L(\lambda)$ for the sample is:

$$L(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i}$$

This simplifies to:

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i}$$

The log-likelihood function $\ell(\lambda)$ is the logarithm of the likelihood function:

$$\ell(\lambda) = \ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} X_i$$

To maximize the log-likelihood, we compute the derivative of $\ell(\lambda)$ with respect to λ:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i$$

By solving $\frac{\partial \ell(\lambda)}{\partial \lambda} = 0$, we obtain the maximum likelihood estimator $\hat{\lambda}$:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Thus, the estimator for the parameter λ is the inverse of the sample mean.

Example 2: Maximum Likelihood for Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be a sample from a normal distribution with unknown mean μ and variance σ^2. The probability density function is:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The likelihood function $L(\mu, \sigma^2)$ is:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$$

This simplifies to:

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

The log-likelihood function $\ell(\mu, \sigma^2)$ is:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

To find the MLE for μ and σ^2, we take the partial derivatives of the log-likelihood and set them to zero.

1. Maximizing with respect to μ:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

This gives:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. Maximizing with respect to σ^2:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

This gives:

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

Thus, the MLE for μ is the sample mean, and the MLE for σ^2 is the sample variance.

2.2 Characteristics of an estimator

2.2.1 Bias, Mean squared error, Convergence

- **Bias**: The bias of an estimator $\hat{\theta}$ is defined as the difference between the expected value of $\hat{\theta}$ and the true parameter value θ:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

- **Mean Squared Error (MSE)**: The mean squared error (MSE) measures the average squared error of the estimator with respect to the true parameter value:
$$\text{MSE}(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2)$$
- **Convergence**: The convergence of an estimator refers to its behavior as the sample size increases. An estimator $\hat{\theta}_n$ is said to converge to θ if it approaches θ as n approaches infinity.

2.2.2 Fisher Information

The Fisher information quantity, denoted as $I(\theta)$, measures the information contained in the sample about the parameter θ. It is defined as:

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right)$$

where $f(X;\theta)$ is the probability density (or mass) function of the probability distribution.

2.2.3 Cramer-Rao Bound

The Cramer-Rao bound establishes a lower limit on the variance of any unbiased estimator. For an unbiased estimator $\hat{\theta}$ of θ, the Cramer-Rao bound is given by:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

2.2.4 Efficiency

An estimator $\hat{\theta}_1$ is said to be more efficient than an estimator $\hat{\theta}_2$ if it has a smaller or equal variance for all possible values of θ. That is, $\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$ for all θ.

2.2.5 Completeness

An estimator $\hat{\theta}$ is said to be complete if it allows unbiased estimation of all functions of θ. This is an important property in the context of Bayesian estimation.

These concepts are fundamental for understanding the construction and evaluation of estimators in statistics. They play a crucial role in the selection and interpretation of estimation methods.

2.3 Estimation by confidence interval

Consider a scenario where we have a random sample, $X_1, X_2, X_3, \ldots, X_n$, originating from a distribution with an unknown parameter θ that requires estimation. We've already explored the concept of point estimation for θ. However, relying solely on the point estimate $\hat{\theta}$ doesn't provide a comprehensive understanding of θ. In essence, without additional context, we lack information about the proximity of $\hat{\theta}$ to the true θ. This leads us to introduce the concept of interval estimation.

In this approach, instead of presenting a single value $\hat{\theta}$ as the estimate for θ, we create an interval that is likely to encompass the actual value of θ. Rather than stating:

$$\hat{\theta} = 34.25,$$

we might present an interval like:

$$[\hat{\theta}_l, \hat{\theta}_h] = [30.69, 37.81],$$

with the expectation that it encompasses the true θ. In essence, we provide two estimations for θ: a higher estimate, $\hat{\theta}_h$, and a lower estimate, $\hat{\theta}_l$. Interval estimation introduces two crucial concepts. First, there's the length of the reported interval, $\hat{\theta}_h - \hat{\theta}_l$. The length of the interval reflects the precision of our θ estimation. A smaller interval signifies a more precise estimate of θ. The second critical factor is the confidence level, which indicates our confidence in the constructed interval. The confidence level represents the probability that our interval includes the true θ value. As a result, higher confidence levels are preferable. These concepts will be elaborated upon in this section.

2.3.1 The General Concept of Interval Estimation

Consider a scenario where we have a set of observations: $X_1, X_2, X_3, \ldots, X_n$, drawn from a distribution with an unknown parameter θ that we wish to estimate. In this context, our objective is twofold:

1. We seek to establish two estimators for θ:
   (a) The lower estimator, $\Theta^l = \Theta^l(X_1, X_2, \ldots, X_n)$, and
   (b) The upper estimator, $\Theta^h = \Theta^h(X_1, X_2, \ldots, X_n)$.
2. The outcome is an interval estimator, which is denoted as the range $[\Theta^l, \Theta^h]$. These estimators, Θ^l and Θ^h, are carefully chosen to ensure that the probability of the interval $[\Theta^l, \Theta^h]$ containing θ surpasses $1 - \alpha$. Here, $1 - \alpha$ represents the confidence level, and it is preferable to have a small α. Common choices for α include 0.1, 0.05, and 0.01, corresponding to confidence levels of 90%, 95%, and 99%, respectively.

Hence, when tasked with determining a 95% confidence interval for a parameter θ, our goal is to identify Θ^l and Θ^h in such a way that:

$$P(\Theta^l < \theta \text{ and } \Theta^h > \theta) \geq 0.95.$$

This discussion will gain further clarity as we delve into practical examples. Before doing so, let's formally define the concept of interval estimation.

Definition 2.3.1. *(Interval Estimation): Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample originating from a distribution with an unknown parameter θ that requires estimation. An interval estimator, with a confidence level of $1-\alpha$, comprises two estimators, $\Theta^l(X_1, X_2, \ldots, X_n)$ and $\Theta^h(X_1, X_2, \ldots, X_n)$, satisfying the condition:*

$$P(\Theta^l \leq \theta \text{ and } \Theta^h \geq \theta) \geq 1 - \alpha,$$

for all possible values of θ. Alternatively, we state that $[\Theta^l, \Theta^h]$ is a $(1-\alpha)100\%$ confidence interval for θ.

It's worth noting that the condition:

$$P(\Theta^l \leq \theta \text{ and } \Theta^h \geq \theta) \geq 1 - \alpha$$

can also be expressed as:

$$P(\Theta^l \leq \theta \leq \Theta^h) \geq 1 - \alpha, \text{ or } P(\theta \in [\Theta^l, \Theta^h]) \geq 1 - \alpha.$$

The variability in these expressions is due to Θ^l and Θ^h, not θ. In this context, θ is the unknown quantity, assumed to be non-random (following frequentist inference). On the other hand, Θ^l and Θ^h are random variables since they depend on the observed random variables $X_1, X_2, X_3, \ldots, X_n$.

2.3.2 Determining Interval Estimators

In this section, we explore the process of deriving interval estimators. But before we delve into that, let's revisit a fundamental concept related to random variables and their distributions. Imagine having a continuous random variable, denoted as X, with a cumulative distribution function (CDF) $F_X(x)$, representing the probability that X is less than or equal to x. Our objective is to identify two values, x_h and x_l, such that the probability of X falling within this interval $[x_l, x_h]$ is equal to $1 - \alpha$.

One way to achieve this is by selecting x_l and x_h such that $P(X \leq x_l)$ equals $\alpha/2$ and $P(X \geq x_h)$ equals $\alpha/2$. In other words, $F_X(x_l)$ should be $\alpha/2$, and $F_X(x_h)$ should be $1 - \alpha/2$. Expressing this using the inverse function, F_X^{-1}, we get

$$x_l = F_X^{-1}(\alpha/2) \text{and} x_h = F_X^{-1}(1 - \alpha/2)$$

. This interval, $[x_l, x_h]$, is known as a $(1 - \alpha)$ interval for X.

Let's illustrate this concept visually, with Figure 8.2 depicting x_l and x_h using both the CDF and the PDF of X.

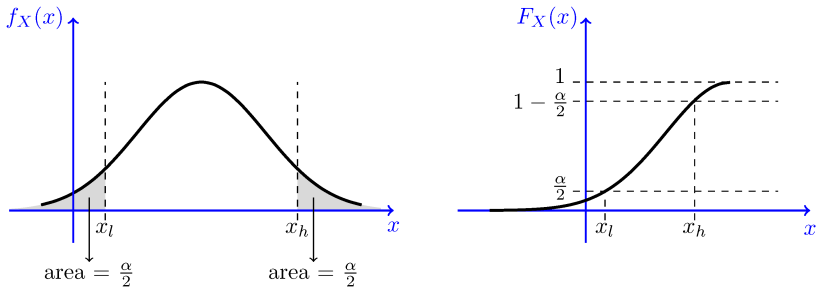


Figure 2.1: $[x_l, x_h]$ is a $(1-\alpha)$ interval for X, that is, $P(x_l \leq X \leq x_h) = 1-\alpha$.

Example 1. Consider a standard normal random variable $Z \sim N(0,1)$. We want to determine x_l and x_h such that $P(x_l \leq Z \leq x_h)$ equals 0.95.

Solution Here, α is 0.05, and we use the Φ function for the CDF of Z. Therefore, we can calculate $x_l = \Phi^{-1}(0.025) = -1.96$ and $x_h = \Phi^{-1}(1-0.025) = 1.96$. This means, for a standard normal random variable Z, $P(-1.96 \leq Z \leq 1.96)$ equals 0.95.

In general, we can find a $(1-\alpha)$ interval for a standard normal random variable by using the notation z_p. For any $p \in [0,1]$, z_p represents the real value for which $P(Z > z_p)$ equals p. It's also important to note that $z_{1-p} = -z_p$. Figure 8.3 visually illustrates this.

Interval Estimators: A General Approach

Now, let's discuss how we can create interval estimators. The typical approach involves starting with a point estimator $\hat{\Theta}$, such as the maximum likelihood estimator (MLE), and constructing an interval $[\hat{\Theta}_l, \hat{\Theta}_h]$ around it, ensuring that $P(\theta \in [\hat{\Theta}_l, \hat{\Theta}_h])$ is greater than or equal to $1-\alpha$. Let's consider an example to understand this process.

Example Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution $N(\theta, 1)$. We need to find a 95% confidence interval for θ.

Solution First, we select $\hat{\Theta}$ as the point estimator for θ. Since θ represents the mean of the distribution, we can use the sample mean $\hat{\Theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Given that $X_i \sim N(\theta, 1)$ and the X_i's are independent, we conclude that $\bar{X} \sim N(\theta, 1/n)$.

By standardizing $\bar{X}$, we determine that the random variable $(\bar{X}-\theta)/(1/\sqrt{n})$ follows a $N(0,1)$ distribution. Therefore, by Example 8.12, we can establish $P(-1.96 \leq (\bar{X}-\theta)/(1/\sqrt{n}) \leq 1.96)$, which is equivalent to $P(\bar{X} - 1.96\sqrt{n} \leq \theta \leq \bar{X} + 1.96\sqrt{n}) = 0.95$. Hence, the 95% confidence interval for θ is $[\hat{\Theta}_l, \hat{\Theta}_h] = [\bar{X} - 1.96\sqrt{n}, \bar{X} + 1.96\sqrt{n}]$.

At first glance, this solution may seem unstructured, and you might wonder why we worked with the standardized variable $\bar{X}$. However, with more contemplation, we can develop a systematic method for solving confidence interval problems. The key insight is that the distribution of the random variable $\bar{X} - \theta$

does not depend on the unknown parameter θ but only on the observed data. Such a random variable is called a pivot or pivotal quantity.

Definition 2.3.2. *(Pivotal Quantity) A pivotal quantity Q is a function of the observed data $X_1, X_2, \ldots, X_n$ and the unknown parameter θ, but it does not depend on any other unknown parameters. Moreover, the probability distribution of Q does not rely on θ or any other unknown parameters.*

To summarize, in the pivotal method for finding confidence intervals:

1. *Identify a pivotal quantity $Q(X_1, X_2, \ldots, X_n, \theta)$.*
2. *Find an interval for Q such that $P(q_l \leq Q \leq q_h) = 1 - \alpha$.*
3. *Use algebraic manipulations to convert the above equation into one of the form $P(\hat{\Theta}_l \leq \theta \leq \hat{\Theta}_h) = 1 - \alpha$.*

In practice, for many common cases, statisticians have already determined pivotal quantities for which confidence intervals have been established. Therefore, you can often solve confidence interval problems by aligning them with previously solved problems.

Using Estimators for σ^2

When dealing with an unknown variance σ^2, we can either find an upper bound for σ^2 or estimate it. Let's explore both approaches:

1. **Upper Bound for σ^2**: If you can demonstrate that $\sigma \leq \sigma_{\max}$, where $\sigma_{\max}$ is a known real number, then you can use $\sigma_{\max}$ instead of σ to determine a confidence interval. This conservative approach ensures that the interval is valid even if σ is larger than $\sigma_{\max}$.
2. **Estimate σ^2**: In many cases, you can estimate σ^2, particularly when the sample size is large. The sample variance, $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$, serves as an estimate for σ^2. After estimating σ^2, you can use it to establish an approximate confidence interval.

To summarize, the steps for finding confidence intervals are as follows:

- **Assumptions**: Start with a random sample $X_1, X_2, \ldots, X_n$ from a distribution.
- **Parameter to be Estimated**: Determine the parameter θ you want to estimate.
- **Confidence Interval**: If you can find an upper bound for σ or estimate σ^2, construct an approximate $(1-\alpha)100\%$ confidence interval for θ, typically of the form

$$[\hat{\Theta} - z_{\alpha/2}\sigma/\sqrt{n}, \hat{\Theta} + z_{\alpha/2}\sigma/\sqrt{n}]$$

or

$$[\hat{\Theta} - z_{\alpha/2}S/\sqrt{n}, \hat{\Theta} + z_{\alpha/2}S/\sqrt{n}]$$

depending on the situation.

This method often leads to approximate confidence intervals, especially when the Central Limit Theorem is applied. Nevertheless, it provides a practical way to estimate unknown parameters with a known level of confidence.

2.3.3 Confidence Intervals for Normal Samples

In the previous discussion, we made an assumption that the sample size, denoted as (n), is sufficiently large, allowing us to apply the Central Limit Theorem (CLT). One interesting feature of the confidence intervals we derived was that they often did not rely on the specifics of the distribution from which the random sample was drawn. However, what happens when (n) is not large? In such cases, we cannot invoke the CLT, and we need to rely on the probability distribution from which the random sample is drawn. This situation becomes particularly important when we are dealing with a sample $(X_1, X_2, X_3, \ldots, X_n)$ taken from a normal distribution. In this context, we will explore how to determine interval estimators for both the mean and the variance of a normal distribution. Before we do that, we will introduce two probability distributions that are closely related to the normal distribution.

Chi-Squared Distribution

First, let's recall the gamma distribution. A continuous random variable (X) is said to follow a gamma distribution with parameters $(\alpha > 0)$ and $(\lambda > 0)$, denoted as $(X \sim \text{Gamma}(\alpha, \lambda))$, if its probability density function (PDF) is given by:

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now, we'd like to introduce a closely related distribution known as the chi-squared distribution. Consider a set of independent standard normal random variables, denoted as $(Z_1, Z_2, \ldots, Z_n)$. If we sum these variables, i.e., $(X = Z_1+Z_2+\ldots+Z_n)$, the resulting random variable (X) is also normally distributed, specifically as $(X \sim N(0, n)$. Next, if we define a new random variable (Y) as the square of this sum:

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2,$$

then (Y) follows a chi-squared distribution with (n) degrees of freedom, represented as $(Y \sim \chi^2(n))$. It can be shown that the random variable (Y) actually has a gamma distribution with parameters $(\alpha = \frac{n}{2})$ and $(\lambda = \frac{1}{2})$:

$$Y \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

The probability density function for $(\chi^2(n))$ is depicted in Figure 8.5 for different values of (n). **The Chi-Squared Distribution**

Definition 2.3.3. *If* $(Z_1, Z_2, \ldots, Z_n)$ *are independent standard normal random variables, the random variable* (Y) *defined as:*

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

is said to have a chi-squared distribution with (n) *degrees of freedom, denoted as* $(Y \sim \chi^2(n))$.

Properties

:

- The chi-squared distribution is a special case of the gamma distribution.
- Specifically, $(Y \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right))$.
- The probability density function for (Y) is:

$$f_Y(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, \text{ for } y > 0.$$

- The expected value of (Y) is (n), and its variance is $(2n)$.
- We define $(\chi^2(p, n))$ as the value for which $(P(Y > \chi^2(p, n)) = p)$ when $(Y \sim \chi^2(n))$.

The chi-squared distribution plays a significant role, particularly when estimating the variance of normal random variables. We will use it to establish confidence intervals for the variance.

The t-Distribution

The next distribution we need is the Student's t-distribution, often referred to as the t-distribution. Here, we'll provide its definition and some properties.

Definition 2.3.4. *: Let* $(Z \sim N(0, 1))$ *and* $(Y \sim \chi^2(n))$*, where* (n) *is a positive integer. We also assume that* (Z) *and* (Y) *are independent. The random variable* (T) *defined as:*

$$T = \frac{ZY}{\sqrt{n}}$$

is said to follow a t-distribution with (n) *degrees of freedom, denoted as* $(T \sim T(n))$.

Properties

:

- The t-distribution exhibits a bell-shaped probability density function centered at 0, but it is more spread out compared to the normal distribution, as shown in Figure 8.7.

- The expected value of (T) is 0 for $(n > 0)$, though it is undefined for $(n = 1)$.
- The variance of (T) is $(\frac{n}{n-2})$ for $(n > 2)$, but it is undefined for $(n = 1, 2)$.
- As (n) becomes large, the t-distribution approaches the standard normal distribution, formally represented as $(T(n) \to N(0, 1))$.
- We define $(t(p, n))$ as the value for which $(P(T > t(p, n)) = p)$, where $(T \sim T(n))$. Because the t-distribution has a symmetric PDF, $(t(1 - p, n) = -t(p, n))$.

The probability density function of the t-distribution for various values of (n) is illustrated in Figure 8.7 and compared to the standard normal distribution. As shown, the t-distribution is more spread out than the standard normal PDF. Figure 8.8 illustrates $(t(p, n))$.

One reason we require the t-distribution is for the following theorem, which is used to estimate the mean of normal random variables.

Theorem 2.3.5. *: Suppose we have (n) independent and identically distributed (i.i.d.) random variables $(X_1, X_2, \ldots, X_n)$ following $(N(\mu, \sigma^2))$ distributions. Additionally, let (S^2) be the sample variance for this random sample. Then, the random variable (Y), defined as:*

$$Y = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

has a chi−squared distribution with $(n-1)$ degrees of freedom, i.e., $(Y \sim \chi^2(n-1))$. Moreover, $(\bar{X})$ and (S^2) are independent random variables.

2.3.4 Confidence Intervals on a Normal Population

We assume in this section the random variable X follows a $N(\mu, \sigma^2)$ distribution from which a simple random sample $(X_1, \ldots, X_n)$ is extracted. The confidence intervals derived in this section arise from the sampling distributions obtained in Section 2.2 for normal populations. We will apply the pivotal quantity method repeatedly to these sampling distributions.

Confidence Interval for the Mean with Known Variance

Let's set the significance level α. To apply the pivotal quantity method, we need an estimator for μ that has a known distribution. A common example is $\hat{\mu} = \bar{X}$, which follows the distribution $\bar{X} \sim N(\mu, \sigma^2/n)$, as indicated in Theorem 1.2.1.

To construct a pivotal quantity that removes μ from the distribution of X, we can use the transformation $\bar{X} - \mu \sim N(0, \sigma^2/n)$. However, a more practical pivotal quantity is given by:

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

This formulation is advantageous because its distribution does not depend on σ.

Next, we need to find the constants c_1 and c_2 such that

$$P(c_1 \leq Z \leq c_2) = 1 - \alpha.$$

By dividing the probability α equally on both sides of the distribution, we determine that c_1 and c_2 satisfy the conditions $\Phi(c_1) = \alpha/2$ and $\Phi(c_2) = 1-\alpha/2$. This implies that c_1 corresponds to the lower $\alpha/2$-quantile of $N(0,1)$, while c_2 corresponds to the upper $(1-\alpha/2)$-quantile of the same distribution. These constants are known as critical values.

Definition 2.3.6. *Critical Value A critical value x_α for a continuous random variable X is defined as the value that accumulates α probability to its right, or, in other words, is the upper α-quantile of X. It satisfies $P(X \geq x_\alpha) = \alpha$.*

The constants c_1 and c_2 can be expressed as:

$$P(Z \geq c_1) = 1 - \alpha/2 \quad \text{and} \quad P(Z \geq c_2) = \alpha/2,$$

which leads to the conclusions $c_1 = z_{1-\alpha/2}$ and $c_2 = z_{\alpha/2}$. Given the symmetry of the standard normal distribution about 0, we have $z_{1-\alpha/2} = -z_{\alpha/2}$, thus confirming that $c_1 = -c_2$.

Example: A gunpowder manufacturer tested a new formula on eight bullets, measuring their initial velocities in meters per second: $916, 892, 895, 904, 913, 916, 895, 885$. Assuming the initial velocities are normally distributed with $\sigma = 12$ meters per second, we need to find a confidence interval for the mean initial velocity of the bullets using a significance level $\alpha = 0.05$.

Solution: We know that the velocities are distributed as $X \sim N(\mu, 12^2)$, with the unknown mean μ. From our sample, we obtain $n = 8$ and $\bar{X} = 902$. Given that $\alpha/2 = 0.025$ and using the critical value $z_{\alpha/2} \approx 1.96$, we can construct a confidence interval for μ at a confidence level of 0.95:

$$CI_{0.95}(\mu) = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

In a more compact notation, we can represent it as:

$$CI_{0.95}(\mu) = \left[\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

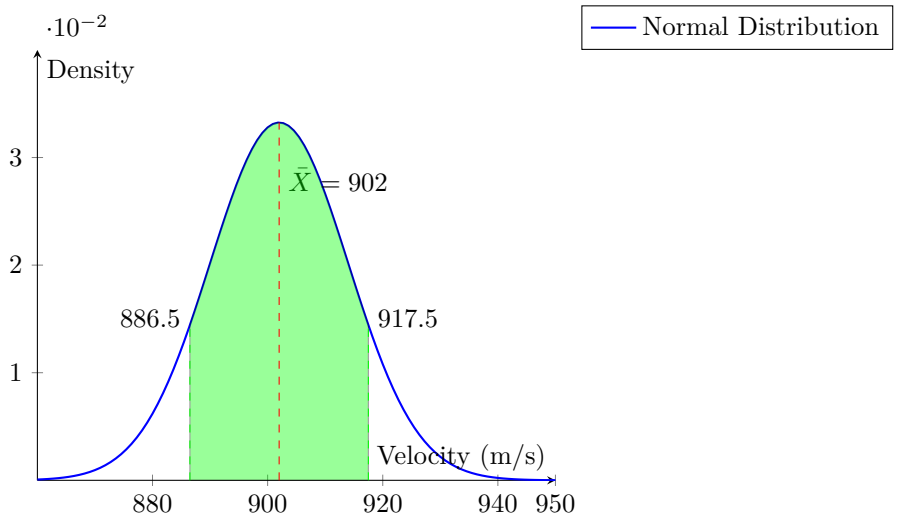


Figure 2.2: Confidence interval for the mean initial velocity of bullets using the new gunpowder formula.

Confidence Interval for the Mean with Unknown Variance

When estimating the mean μ of a population with unknown variance, we typically use the sample mean $\bar{X}$ and the sample standard deviation S' as estimators. The pivotal quantity used in constructing confidence intervals in this case is given by

$$T = \frac{\bar{X} - \mu}{S'/\sqrt{n}},$$

where T follows a Student's t distribution with $n-1$ degrees of freedom. The sample standard deviation, denoted as S', is defined as follows:

$$S' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

where: - X_i are the individual sample observations, - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean, - n is the number of observations in the sample.

This formula calculates the average squared deviation of each observation from the sample mean, and the factor of $\frac{1}{n-1}$ is used instead of $\frac{1}{n}$ to correct for bias in the estimation of the population variance.

Proof of the Distribution of the Pivot:

1. **Sample Mean and Standard Deviation:** Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. The sample mean $\bar{X}$ is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. **Distribution of $\bar{X}$:** By the properties of normal distributions, $\bar{X}$ follows a normal distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

3. **Distribution of S'^2:** The sample variance S'^2 follows a scaled chi-squared distribution:

$$\frac{(n-1)S'^2}{\sigma^2} \sim \chi^2_{n-1}.$$

4. **Independence of $\bar{X}$ and S':** The sample mean $\bar{X}$ and the sample variance S'^2 are independent when the data is normally distributed.

5. **Formation of the Pivot:** The standardized version of the sample mean is:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Thus, we can express T in terms of the sample variance:

$$T = \frac{\bar{X} - \mu}{S'/\sqrt{n}}.$$

By the definition of the t distribution, since the numerator is standard normal and the denominator is the square root of a chi-squared distribution divided by its degrees of freedom, we conclude:

$$T \sim t_{n-1}.$$

Constructing the Confidence Interval: To construct the confidence interval, we split the significance level α evenly between the two tails of the t distribution. The probabilities can be expressed as:

$$P(c_1 \leq T \leq c_2) = 1 - \alpha,$$

where c_1 and c_2 correspond to the critical values:

$$c_1 = t_{n-1;1-\alpha/2}, \quad c_2 = t_{n-1;\alpha/2}.$$

Since the t distribution is symmetric, we also have:

$$c_1 = -c_2 = -t_{n-1;\alpha/2}.$$

Deriving the Confidence Interval: To find the confidence interval for the mean μ, we can rearrange the inequalities from the pivotal quantity:

1. Starting from the inequality:

$$c_1 \leq \frac{\bar{X} - \mu}{S'/\sqrt{n}} \leq c_2,$$

we can multiply through by $S'/\sqrt{n}$:

$$c_1 S'/\sqrt{n} \leq \bar{X} - \mu \leq c_2 S'/\sqrt{n}.$$

2. Rearranging gives us two inequalities for μ:

$$\bar{X} - c_2 \frac{S'}{\sqrt{n}} \leq \mu \leq \bar{X} - c_1 \frac{S'}{\sqrt{n}}.$$

3. Finally, we express the confidence interval for μ as:

$$CI_{1-\alpha}(\mu) = \left[\bar{X} - t_{n-1;\alpha/2}\frac{S'}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2}\frac{S'}{\sqrt{n}}\right].$$

Example: Problem Statement: Suppose a company wants to estimate the average lifespan of its batteries. A sample of 10 batteries is tested, and the lifespans in hours are as follows:

$$130, 135, 128, 132, 140, 125, 130, 138, 136, 134.$$

Calculate a 95% confidence interval for the average lifespan of the batteries.
Solution: 1. **Calculate the Sample Mean** $\bar{X}$:

$$\bar{X} = \frac{130 + 135 + 128 + 132 + 140 + 125 + 130 + 138 + 136 + 134}{10} = 133.8.$$

2. **Calculate the Sample Standard Deviation** S':

$$S' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

First, we find the squared deviations:

$$(130 - 133.8)^2, (135 - 133.8)^2, \ldots, (134 - 133.8)^2.$$

This results in:

$$S' = \sqrt{\frac{1}{9}(14.44 + 1.44 + 33.64 + 0.64 + 38.44 + 7.84 + 14.44 + 17.64 + 4.84 + 0.04)} \approx \sqrt{12.21} \approx 3.49.$$

3. **Determine the Critical Value:** For $n - 1 = 9$ degrees of freedom and a 95% confidence level, we find the critical value $t_{9;0.025} \approx 2.262$ (from the t-distribution table).

4. **Calculate the Confidence Interval:** Using the formula:

$$CI_{0.95}(\mu) = \left[\bar{X} - t_{9;0.025}\frac{S'}{\sqrt{n}}, \bar{X} + t_{9;0.025}\frac{S'}{\sqrt{n}}\right].$$

Substituting the values:

$$CI_{0.95}(\mu) = \left[133.8 - 2.262\frac{3.49}{\sqrt{10}}, 133.8 + 2.262\frac{3.49}{\sqrt{10}}\right].$$

Calculating:

$$\approx [133.8 - 2.262 \times 1.105, 133.8 + 2.262 \times 1.105] \approx [130.50, 137.10].$$

Thus, the 95% confidence interval for the average lifespan of the batteries is approximately $[130.50, 137.10]$. This confidence interval provides an estimate

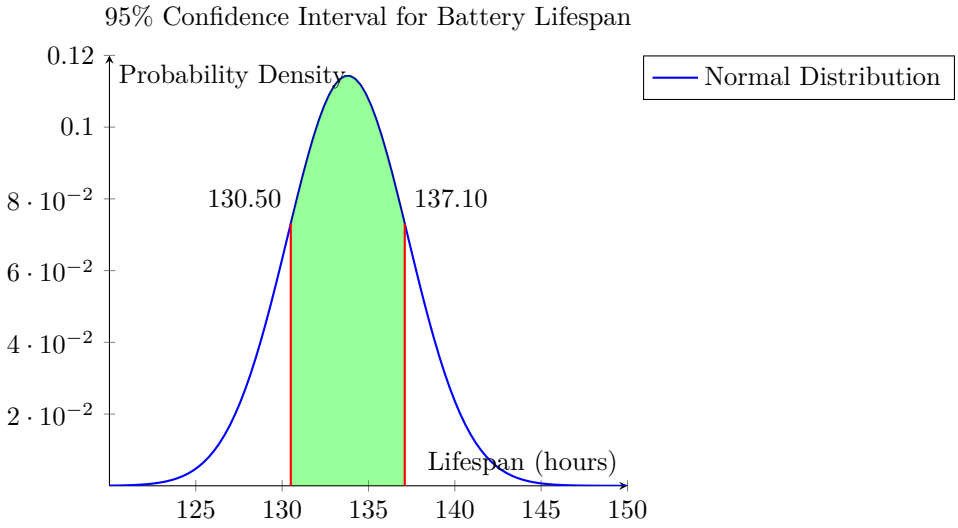


Figure 2.3: 95% Confidence Interval for the Average Lifespan of Batteries

of the population mean μ while accounting for the uncertainty due to the estimation of the population variance. The usage of the Student's t distribution is particularly important when the sample size is small or when the population variance is unknown. The provided example demonstrates the application of these concepts with a clear illustration of the confidence interval on a normal distribution curve.

Confidence Interval for the Variance

We already have an unbiased estimator of σ^2, the quasivariance S'^2. In addition, by Theorem 2.2, we have a pivot $U = (n-1)S'^2/\sigma^2 \sim \chi^2_{n-1}$. Then, we only need to compute the constants c_1 and c_2 such that $P(c_1 \leq U \leq c_2) = 1 - \alpha$. Splitting the probability α evenly to both sides of the χ^2 distribution, we have that $c_1 = \chi^2_{n-1;1-\alpha/2}$ and $c_2 = \chi^2_{n-1;\alpha/2}$.
Once the constants are computed, solving for σ^2 in the inequalities yields the confidence interval for σ^2:

$$CI_{1-\alpha}(\sigma^2) = [(n-1)S'^2/\chi^2_{n-1;\alpha/2}, (n-1)S'^2/\chi^2_{n-1;1-\alpha/2}].$$

Example: A practitioner seeks to evaluate the variability of an instrument designed to measure the concentration of a solution. They recorded twenty independent measurements (in arbitrary units), which are as follows:

$$3.2, 2.9, 3.5, 3.0, 2.8, 3.3, 3.1, 3.4, 2.7, 3.6, 3.2, 2.8, 3.0, 3.3, 3.1, 2.9, 3.2, 3.4, 3.0, 2.9$$

Determine a 95% confidence interval for the variance σ^2 of these measurements, under the assumption that they follow a normal distribution.

Solution: We are asked to find a 95% confidence interval for the variance σ^2 of the measurements, assuming they are normally distributed.

Step 1: Compute the Sample Variance

1. *Calculate the Sample Mean* $\bar{X}$:

$$\bar{X} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{3.2 + 2.9 + \ldots + 2.9}{20} = 3.08$$

2. *Calculate the Sample Variance* S'^2:

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

After performing the calculations:

$$S'^2 \approx 0.0936$$

Step 2: Determine the Chi-Squared Distribution Quantiles

To construct a confidence interval for σ^2, we use:

$$U = \frac{(n-1)S'^2}{\sigma^2} \sim \chi^2_{n-1}$$

For a 95% confidence level, $\alpha = 0.05$, split between the two tails.

Since $n = 20$, the degrees of freedom are $n - 1 = 19$. Using chi-squared tables: - $c_1 = \chi^2_{19;0.025} \approx 8.907$ - $c_2 = \chi^2_{19;0.975} \approx 32.852$

Step 3: Calculate the Confidence Interval for σ^2

Using:

$$CI_{0.95}(\sigma^2) = \left[\frac{(n-1)S'^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)S'^2}{\chi^2_{n-1;\alpha/2}} \right]$$

Plugging in the values:

$$CI_{0.95}(\sigma^2) = \left[\frac{19 \cdot 0.0936}{32.852}, \frac{19 \cdot 0.0936}{8.907} \right]$$

Calculating each bound: 1. *Lower Bound*:

$$\frac{19 \cdot 0.0936}{32.852} \approx \frac{1.7784}{32.852} \approx 0.0541$$

2. *Upper Bound*:

$$\frac{19 \cdot 0.0936}{8.907} \approx \frac{1.7784}{8.907} \approx 0.1997$$

Thus, the 95% confidence interval for the variance σ^2 is approximately:

$$CI_{0.95}(\sigma^2) = [0.0541, 0.1997]$$

Interpretation

The calculated confidence interval $[0.0541, 0.1997]$ suggests that we can be 95% confident that the true variance of the concentration measurements lies within this range.

2.4 Application

2.4.1 Solved exercices

Exercice 2.1. ***(Linear Regression)***
Consider the following data points:

$$(1, 3), (2, 5), (3, 7), (4, 9), (5, 11)$$

1. *Fit a linear model $y = mx + b$ using the least squares method.*
2. *Find the values of m and b that minimize the sum of squared differences between the observed and predicted values of y.*

Exercice 2.2. ***(Polynomial Regression)***
Given the data points:

$$(1, 2), (2, 5), (3, 10), (4, 17), (5, 26)$$

1. *Fit a polynomial model $y = a_0 + a_1x + a_2x^2$ using the least squares method.*
2. *Determine the coefficients a_0, a_1, and a_2 that minimize the sum of squared differences.*

Exercice 2.3. ***(Exponential Regression)***
Given the data points:

$$(1, 2), (2, 4), (3, 8), (4, 16), (5, 32)$$

1. *Fit an exponential model $y = ae^{bx}$ using the least squares method.*
2. *Find the values of a and b that minimize the sum of squared differences.*

Exercice 2.4. ***(Multiple Linear Regression)***
Consider the dataset with three variables:

$$(1, 2, 5), (2, 4, 8), (3, 6, 10), (4, 8, 12), (5, 10, 14)$$

1. *Fit a multiple linear regression model $y = \beta_0 + \beta_1x_1 + \beta_2x_2$ using the least squares method.*
2. *Find the coefficients β_0, β_1, and β_2 that minimize the sum of squared differences.*

Exercice 2.5. ***(Bias, Mean Squared Error, and Convergence)***

1. *Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with unknown mean μ and variance σ^2. Define an estimator $\hat{\mu}$ for μ. Calculate the bias of $\hat{\mu}$ and determine if it's an unbiased estimator.*
2. *Given a random variable X with known mean μ and variance σ^2, define the estimator $\hat{\sigma}^2$ for σ^2. Calculate the mean squared error of $\hat{\sigma}^2$.*

3. *Consider a sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ for a parameter θ. Show that $\hat{\theta}_n$ converges in probability to θ as n approaches infinity.*

Exercice 2.6. ***(Fisher Information, Cramer-Rao Bound, Efficiency)***

1. *Given a random sample $X_1, X_2, \ldots, X_n$ from a distribution with probability density function $f(x;\theta)$, derive the Fisher information $I(\theta)$ for the parameter θ.*

2. *For a given parameter θ, compute the Cramer-Rao Lower Bound (CRLB) for the variance of any unbiased estimator of θ, based on the Fisher information obtained in the previous exercise.*

3. *Suppose you have two unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, for the same parameter θ. Calculate the efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$, and determine which estimator is more efficient.*

Exercice 2.7. ***(Efficiency)*** *Suppose you have two unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, for the same parameter θ. Calculate the efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$, and determine which estimator is more efficient.*

Exercice 2.8. ***(Completeness)***

1. *Show that a random sample $X_1, X_2, \ldots, X_n$ from a distribution with a parameter θ has a complete sufficient statistic.*

2. *Given a continuous distribution with an unknown parameter θ, find a statistic that is minimal sufficient but not complete.*

Exercice 2.9. *Given a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with an unknown mean μ and known variance σ^2, construct a 95% confidence interval for μ based on the sample.*

Exercice 2.10. *Suppose you have a random sample of size $n = 50$ from a normal distribution with an unknown mean μ and known variance $\sigma^2 = 25$. Calculate a 99% confidence interval for μ.*

Exercice 2.11. *Consider a random sample of size $n = 20$ from a normal population with an unknown mean μ and unknown variance σ^2. Calculate a 90% confidence interval for μ and provide the general formula for the confidence interval.*

Exercice 2.12. *Given a random sample of size $n = 25$ from a normal population with an unknown mean μ and an unknown variance σ^2, construct a 98% confidence interval for the population variance σ^2.*

Exercice 2.13. *For a random sample of size $n = 30$ from a normal population with an unknown mean μ and an unknown variance σ^2, determine the 95% confidence interval for the ratio of two variances, $\frac{\sigma_1^2}{\sigma_2^2}$.*

Exercice 2.14. *In an industrial process, the time (in minutes) required to complete a task follows a normal distribution with an unknown mean μ and a known variance of $\sigma^2 = 16$ minutes. A random sample of size $n = 15$ is taken, and a 90% confidence interval for the mean time μ is needed. Calculate the confidence interval and interpret the result.*

Exercices solution

Correction exercice 2.1. *The linear model $y = mx + b$ can be solved using the least squares method. The formulas for m and b are:*

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where n is the number of data points, (x_i, y_i). Substitute the given values:

$$n = 5, \Sigma x = 15, \Sigma y = 35, \Sigma x^2 = 55, \Sigma xy = 135$$

$$m = \frac{(5 \times 135) - (15 \times 35)}{(5 \times 55) - (15)^2} = \frac{375}{20} = 18.75$$

$$b = \frac{(35 \times 55) - (15 \times 135)}{(5 \times 55) - (15)^2} = \frac{1925}{20} = 96.25$$

Therefore, the linear model is $y = 18.75x + 96.25$.

Correction exercice 2.2. *The polynomial model $y = a_0 + a_1x + a_2x^2$ can be solved using the least squares method. The formulas for a_0, a_1, and a_2 are obtained by solving a system of linear equations. The system of equations is:*

$$\begin{aligned} a_0 + a_1(1) + a_2(1)^2 &= 2 \\ a_0 + a_1(2) + a_2(2)^2 &= 5 \\ a_0 + a_1(3) + a_2(3)^2 &= 10 \\ a_0 + a_1(4) + a_2(4)^2 &= 17 \\ a_0 + a_1(5) + a_2(5)^2 &= 26 \end{aligned}$$

Solving this system gives:

$$a_0 = 0.5, \quad a_1 = 0, \quad a_2 = 1$$

Therefore, the polynomial model is $y = 0.5 + x^2$.

Correction exercice 2.3. *The exponential model $y = ae^{bx}$ can be solved using the least squares method. Take the natural logarithm of both sides to linearize the model:*

$$\ln(y) = \ln(a) + bx$$

This is now in the form $\ln(y) = \alpha + \beta x$, which is a linear model. Apply linear regression to find the values of α and β. Solving for a and b gives:

$$a = e^{\alpha}, \quad b = \beta$$

Substitute the given values:

$$\alpha \approx 0.69, \quad \beta \approx 0.69$$

Therefore, the exponential model is $y \approx e^{0.69}e^{0.69x}$.

Correction exercice 2.4. *The multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ can be solved using the least squares method. The formulas for β_0, β_1, and β_2 are obtained by solving a system of linear equations. The system of equations is:*

$$\begin{aligned} \beta_0 + \beta_1(1) + \beta_2(2) &= 5 \\ \beta_0 + \beta_1(2) + \beta_2(4) &= 8 \\ \beta_0 + \beta_1(3) + \beta_2(6) &= 10 \\ \beta_0 + \beta_1(4) + \beta_2(8) &= 12 \\ \beta_0 + \beta_1(5) + \beta_2(10) &= 14 \end{aligned}$$

Solving this system gives:

$$\beta_0 = 1, \quad \beta_1 = 2, \quad \beta_2 = 1$$

Therefore, the multiple linear regression model is $y = 1 + 2x_1 + x_2$.

Correction exercice 2.5.

1. *To calculate the bias of $\hat{\mu}$, use the following formula:*
$$Bias(\hat{\mu}) = E(\hat{\mu}) - \mu$$
If $\hat{\mu}$ is unbiased, the bias should equal zero.

2. *To calculate the mean squared error (MSE) of $\hat{\sigma}^2$, use the following formula:*
$$MSE(\hat{\sigma}^2) = E\left((\hat{\sigma}^2 - \sigma^2)^2\right)$$

3. *To show that $\hat{\theta}_n$ converges in probability to θ as n approaches infinity, you can use the definition of convergence in probability:*
$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1, \text{ for all } \epsilon > 0$$

Correction exercice 2.6.

1. *To derive the Fisher information $I(\theta)$ for the parameter θ, you need to find the second derivative of the log-likelihood function with respect to θ.*

2. *The Cramer-Rao Lower Bound (CRLB) for the variance of any unbiased estimator of θ is given by:*
$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

Correction exercice 2.7.

1. *Calculate the efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ using the formula:*
$$Efficiency(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$
An estimator with higher efficiency is more efficient.

Correction exercice 2.8. *1. A sufficient statistic T is complete if, for any measurable function $g(t)$, the expectation of $g(T)$ is zero only if $g(t)$ is zero with probability one:*

$$E[g(T)] = 0 \Rightarrow P(g(T) = 0) = 1$$

To show that a statistic is complete, you can use the definition and properties of completeness.

2. To find a statistic that is minimal sufficient but not complete, consider a distribution where minimal sufficiency is achieved, but the distribution itself does not meet the completeness criterion. You can explore different probability distributions to illustrate this.

Correction exercice 2.9. *The 95% confidence interval for μ is given by:*

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean, σ is the known population standard deviation, and n is the sample size. In this case, 1.96 corresponds to the 97.5th percentile of the standard normal distribution.

Correction exercice 2.10. *The 99% confidence interval for μ is given by:*

$$\bar{X} \pm 2.576 \cdot \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean, σ is the known population standard deviation, and n is the sample size. In this case, 2.576 corresponds to the 99.5th percentile of the standard normal distribution.

Correction exercice 2.11. *The 90% confidence interval for μ is given by:*

$$\bar{X} \pm z \cdot \frac{S}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean, S is the sample standard deviation, n is the sample size, and z is the critical value from the t-distribution with $n-1$ degrees of freedom that corresponds to the desired confidence level.

The general formula for the confidence interval is:

$$\bar{X} \pm z \cdot \frac{S}{\sqrt{n}}$$

Correction exercice 2.12. *The 98% confidence interval for σ^2 is given by:*

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \right)$$

where S^2 is the sample variance, n is the sample size, $\alpha = 0.02$, and $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the chi-squared distribution with $n-1$ degrees of freedom, respectively.

Correction exercice 2.13. *The* 95% *confidence interval for the ratio* $\frac{\sigma_1^2}{\sigma_2^2}$ *is given by:*

$$\left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}}, \frac{S_1^2}{S_2^2} \cdot F_{\alpha/2}\right)$$

where S_1^2 *and* S_2^2 *are the sample variances,* n_1 *and* n_2 *are the sample sizes,* $\alpha = 0.05$*, and* $F_{\alpha/2}$ *and* $F_{1-\alpha/2}$ *are the* $\alpha/2$ *and* $1 - \alpha/2$ *percentiles of the F-distribution with* $n_1 - 1$ *and* $n_2 - 1$ *degrees of freedom, respectively.*

Correction exercice 2.14. *The* 90% *confidence interval for* μ *is given by:*

$$\bar{X} \pm 1.645 \cdot \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ *is the sample mean,* σ *is the known population standard deviation,* n *is the sample size, and* 1.645 *corresponds to the* 95*th percentile of the standard normal distribution.*

In this context, the confidence interval represents a range of values within which we are 90% *confident that the true mean time* μ *falls. This means that if we were to take many random samples and construct confidence intervals from them, about* 90% *of those intervals would contain the true mean time.*

2.4.2 Unsolved exercices

Exercice 2.1. ***(Logarithmic Regression)***
Given the data points:

$$(1, 0.7), (2, 1.4), (3, 2.0), (4, 2.4), (5, 2.7)$$

1. *Fit a logarithmic regression model $y = a + b\ln(x)$ using the least squares method.*
2. *Determine the coefficients a and b that minimize the sum of squared differences.*

Exercice 2.2. ***(Weighted Least Squares)***
A researcher collects data where each observation (x_i, y_i) has a corresponding weight w_i:

$$(1, 2, 0.5), (2, 4, 0.8), (3, 6, 1.2), (4, 8, 1.5), (5, 10, 2.0)$$

1. *Fit a linear regression model $y = mx + b$ using the weighted least squares method.*
2. *Compute the weighted values of m and b.*

Exercice 2.3. ***(Generalized Linear Model)***
Consider the dataset:

$$(x, y) = (1, 0), (2, 1), (3, 1), (4, 0), (5, 1)$$

Assume a logistic regression model:

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

1. *Derive the log-likelihood function for the given data.*
2. *Use the maximum likelihood method to estimate β_0 and β_1.*

Exercice 2.4. ***(Residual Analysis)***
Given the regression model $y = 2x + 3$, the observed data points are:

$$(1, 5.5), (2, 6.8), (3, 8.1), (4, 9.5), (5, 11.3)$$

1. *Compute the residuals for each data point.*
2. *Perform a residual analysis to check for patterns and normality.*

Exercice 2.5. ***(Bootstrap Confidence Interval)***
A sample of size $n = 10$ yields the following data:

$$12, 15, 18, 22, 25, 29, 32, 35, 38, 42$$

1. *Use the bootstrap method to estimate a 95% confidence interval for the sample mean.*
2. *Interpret the bootstrap interval in the context of the data.*

Chapter 3

Hypothesis Testing

Objectives

After studying this chapter, you should:

- Understand the framework and formulation of hypotheses.
- Recognize the significance of p-values in hypothesis testing.
- Differentiate between one-sided and two-sided tests.
- Identify types of errors in hypothesis testing.
- Perform hypothesis testing for the mean.

Introduction

In the field of statistics, hypothesis testing is a fundamental method used to make inferences about populations based on sample data. It allows researchers to evaluate claims or assumptions about a population parameter, helping to determine whether the observed data provide sufficient evidence to support or reject these claims.

The process of hypothesis testing is essential across various disciplines, including medicine, psychology, economics, and social sciences, where decisions must be made based on empirical evidence. For example, pharmaceutical companies use hypothesis testing to assess the effectiveness of new drugs by comparing patient outcomes against established standards. Similarly, in quality control, manufacturers employ hypothesis testing to verify that their products meet regulatory specifications.

At the core of hypothesis testing lies the formulation of two competing hypotheses: the null hypothesis (H_0) and the alternative hypothesis (H_1). The null hypothesis represents a default position or a statement of no effect or no difference, while the alternative hypothesis represents the claim that researchers aim to support. The goal of hypothesis testing is to gather evidence to either reject the null hypothesis in favor of the alternative hypothesis or fail to reject it based on the sample data.

This chapter will delve into the concepts, methodologies, and practical applications of hypothesis testing. We will begin by discussing the general framework for hypothesis testing, including how to formulate hypotheses and set significance levels. We will also explore different types of tests, the concept of errors associated with hypothesis testing, and the interpretation of p-values. Through this exploration, readers will gain a comprehensive understanding of hypothesis testing and its importance in statistical analysis.

3.1 General Framework for Hypothesis Testing

Hypothesis testing is a cornerstone of statistical analysis, providing a structured method for making inferences about populations based on sample data. It allows researchers to determine whether there is enough evidence in a sample to draw conclusions about a population parameter. This section presents the key components of hypothesis testing, outlining the step-by-step framework that guides the decision-making process.

3.1.1 Formulating Hypotheses and Decision-Making in Hypothesis Testing

The formulation of hypotheses is the foundational step in hypothesis testing. It involves articulating a statement about a population parameter that can be tested using statistical methods. Based on the analysis, we ultimately decide

to **accept** or **reject** the null hypothesis based on the evidence provided by the sample data.

Null Hypothesis (H_0)

The null hypothesis, denoted as H_0, is a statement asserting that there is no effect, no difference, or no relationship between variables. It serves as the default or baseline assumption, suggesting that any observed differences in the data are due to random chance rather than a real effect. In hypothesis testing, we assume the null hypothesis is true until the data provide sufficient evidence to contradict it.

If our analysis yields significant evidence against H_0, we **reject** the null hypothesis in favor of the alternative. If not, we **fail to reject** H_0, continuing to assume that it holds.

Example 2. Consider a study aimed at evaluating the effectiveness of a new teaching method on student performance. The null hypothesis might be:

$$H_0 : \mu = \mu_0$$

where μ_0 represents the average test score of students taught using the traditional method. This hypothesis assumes that the new teaching method does not produce significantly different results compared to the traditional method. In this case, we would only reject H_0 if the data strongly suggest that the new method leads to different (typically higher) scores.

Alternative Hypothesis (H_1)

The alternative hypothesis, denoted as H_1, contradicts the null hypothesis. It proposes that there is an effect, a difference, or a relationship present in the population. The alternative hypothesis can be either one-sided, indicating a specific direction of the effect, or two-sided, indicating any difference from the null value.

If our test statistic provides strong enough evidence, we **reject** H_0 and accept H_1. This decision implies that the observed data are unlikely to occur if the null hypothesis were true, suggesting the presence of a genuine effect as stated in H_1.

Example 3. For the teaching method study, the alternative hypothesis could be:

$$H_1 : \mu \neq \mu_0$$

This implies that the average test score of students taught using the new method is different from those taught using the traditional method. If a one-sided hypothesis is used, it could be stated as:

$$H_1 : \mu > \mu_0$$

indicating that the new method is expected to result in higher scores than the traditional method.

In summary, our decision to **accept** or **reject** the null hypothesis depends on whether the test results provide sufficient statistical evidence to support the alternative hypothesis, H_1. If this evidence is found, we reject H_0, concluding that the data support H_1. Otherwise, we fail to reject H_0, indicating insufficient evidence to support the claim in H_1.

3.1.2 P-Values

In the previous discussions, we only provided an "accept" or "reject" decision as the outcome of a hypothesis test. However, we can offer more information by using a measure called P-values. In essence, P-values indicate how close the decision was. To elaborate, if we reject the null hypothesis H_0 at a significance level $\alpha = 0.05$, we can inquire about the outcome at a different significance level, such as $\alpha = 0.01$. Can we still reject H_0? More precisely, we can ask the following question:

What is the smallest significance level α that leads to the rejection of the null hypothesis?

The response to this query is known as the P-value. The P-value is the minimum significance level α that results in rejecting the null hypothesis. In simple terms, if the P-value is small, it implies that the observed data is highly unlikely to occur under H_0, thereby providing stronger evidence for rejecting the null hypothesis. How do we determine P-values? Let's examine an example.

Example 4. Suppose you have a coin and you want to investigate whether it is fair or biased. Specifically, let θ denote the probability of obtaining heads, where $\theta = P(H)$. You need to choose between the following hypotheses:

H_0 (null hypothesis): The coin is fair, i.e., $\theta = \theta_0 = \frac{1}{2}$.

H_1 (alternative hypothesis): The coin is biased, i.e., $\theta > \frac{1}{2}$.

You toss the coin 100 times and observe 60 heads. Can we reject H_0 at a significance level $\alpha = 0.05$? Can we reject H_0 at a significance level $\alpha = 0.01$? What is the P-value?

Solution

Let X be the random variable representing the number of observed heads. In our experiment, we observed $X = 60$. Since $n = 100$ is relatively large, assuming H_0 is true, the random variable

$$W = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

is approximately a standard normal random variable, $N(0, 1)$. If H_0 is true, we expect X to be close to 50, whereas if H_1 is true, we anticipate X to be larger. Therefore, we can propose the following test: choose a threshold c. If $W \leq c$, we accept H_0; otherwise, we accept H_1. To calculate the P(type I error), we can express it as

$$P(\text{type I error}) = P(\text{Reject } H_0|H_0) = P(W > c|H_0).$$

Since W follows a standard normal distribution under H_0, we need to select $c = z_\alpha$ to ensure a significance level α. In this example, we find that

$$W = \frac{X - 50}{5} = \frac{60 - 50}{5} = 2.$$

If we require a significance level $\alpha = 0.05$, then

$$c = z_{0.05} = 1.645.$$

The value above can be obtained in MATLAB using the command `norminv(1 - 0.05)`. As $W = 2 > 1.645$, we reject H_0 and accept H_1.

If we require a significance level $\alpha = 0.01$, then

$$c = z_{0.01} = 2.33.$$

The value above can be obtained in MATLAB using the command `norminv(1 - 0.01)`. As $W = 2 \leq 2.33$, we fail to reject H_0, so we accept H_0.

The P-value is the minimum significance level α that leads to the rejection of H_0. In this case, since $W = 2$, we reject H_0 only if $c < 2$. Note that $z_\alpha = c$, thus

$$\alpha = 1 - P(W \leq c).$$

If $c = 2$, we obtain

$$\alpha = 1 - P(W \leq 2) = 0.023.$$

Therefore, we rejectH_0 at a significance level of $\alpha = 0.023$. The P-value is 0.023.

3.1.3 Test Statistics

The test statistic is a numerical value calculated from the sample data that measures the degree to which the sample provides evidence against the null hypothesis. The choice of test statistic depends on the type of test being performed and the nature of the data. Different statistical tests are used to evaluate hypotheses based on various data characteristics, such as distribution, variance, and sample size. Below is a summary of several commonly used statistical tests:

- **Z-Test**: Used for comparing means when the population variance is known. It is applicable in situations with large sample sizes (typically $n > 30$).

  $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  where $\bar{X}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size.
  Example: Comparing the average weight of a sample group to the known population mean weight.

- **t−test**: Used to compare means when the population variance is unknown, particularly with smaller sample sizes (typically $n \leq 30$).

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

where S is the sample standard deviation.
Example: Testing the average scores of two different classes on a standardized test.

- **Chi-Square Test**: Assesses the independence of categorical variables and tests how likely it is that an observed distribution is due to chance.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency and E is the expected frequency under the null hypothesis.
Example: Evaluating whether there is a relationship between gender and preference for a type of product.

- **ANOVA (Analysis of Variance)**: Compares means across three or more groups to determine if at least one group mean is different from the others.

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

where the between-group variance measures differences between group means, and the within-group variance measures variance within each group.
Example: Comparing the effectiveness of three different diets on weight loss.

- **Mann-Whitney U Test**: A non-parametric test for comparing two independent groups when the data does not follow a normal distribution.

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

where n_1 and n_2 are sample sizes of the two groups, and R_1 is the sum of ranks for the first group.
Example: Comparing satisfaction levels between two stores without assuming normality.

- **Wilcoxon Signed-Rank Test**: A non-parametric test used to compare two related samples or repeated measurements.

$$W = \sum_{|R_i|}$$

where R_i are ranks assigned to differences, ignoring the sign, and sums are calculated for either positive or negative ranks.
Example: Evaluating changes in blood pressure measurements before and after treatment.

- **Kruskal-Wallis H Test**: A non-parametric alternative to ANOVA for comparing three or more independent groups.

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total number of observations, g is the number of groups, R_i is the sum of ranks for group i, and n_i is the sample size for group i.
Example: Assessing customer satisfaction ratings across multiple service centers without assuming normal distribution.

- **Regression Analysis**: Evaluates the relationship between independent and dependent variables to predict outcomes.

$$t = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}$$

where $\hat{\beta}$ is the estimated regression coefficient, β_0 is the hypothesized value, and $\text{SE}(\hat{\beta})$ is the standard error of $\hat{\beta}$.
Example: Analyzing how study hours impact exam scores among students.

- **Likelihood Ratio Test (LRT)**: Compares the goodness of fit between two models, typically the null and alternative models.

$$\Lambda = \frac{\sup L(H_0)}{\sup L(H_1)}$$

where $\sup L(H_0)$ and $\sup L(H_1)$ represent the maximum likelihood estimates under the null and alternative hypotheses, respectively. The test statistic is often transformed as $-2\log(\Lambda)$, which follows a chi-square distribution under H_0.
Example: Testing whether a new medication significantly reduces recovery time compared to a placebo.

These statistical tests provide valuable tools for researchers to assess hypotheses, analyze data, and draw conclusions based on empirical evidence. Choosing the appropriate test statistic is crucial for ensuring valid statistical inference.

The following table provides an overview of common test statistics and their respective formulas. In conclusion, the general framework for hypothesis testing provides a systematic approach to evaluating claims about population parameters based on sample data. By formulating clear hypotheses, establishing a significance level, calculating the appropriate test statistic, and applying a decision rule, researchers can draw meaningful conclusions from their analyses. This structured methodology is essential in various fields, including medicine, education, and social sciences, where data-driven decision-making is critical for advancing knowledge and improving practices.

Test Statistic	Formula
Z-Test	$Z = \frac{X-\mu}{\sigma/\sqrt{n}}$
t−test	$T = \frac{X-\mu}{S/\sqrt{n}}$
Chi-Square Test	$\chi^2 = \sum \frac{(O-E)^2}{E}$
ANOVA (Analysis of Variance)	$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$
Mann-Whitney U Test	$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$
Wilcoxon Signed-Rank Test	$W = \sum_{\lvert R_i \rvert}$
Kruskal-Wallis H Test	$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(N+1)$
Regression Analysis	$t = \frac{\hat{\beta}-\beta_0}{\text{SE}(\hat{\beta})}$
Likelihood Ratio Test (LRT)	$\Lambda = \frac{\sup L(H_0)}{\sup L(H_1)}$, often transformed to $-2\log(\Lambda)$

Table 3.1: Summary of Common Test Statistics and Their Formulas

3.1.4 Decision Rule

The decision rule provides the criteria for determining whether to reject or fail to reject the null hypothesis based on the test statistic and the chosen significance level. There are two primary methods to establish the decision rule: the p-value approach and the critical value approach.

- **p-value Approach**: This method involves calculating the p-value, which is the probability of observing the test statistic or a more extreme value under the assumption that the null hypothesis is true. If the p-value is less than or equal to α, the null hypothesis is rejected; otherwise, it is not rejected.

  Example 5. Suppose that the calculated p-value from the t−test is 0.008. Since this is less than $\alpha = 0.01$, the null hypothesis would be rejected, indicating statistically significant evidence that the new teaching method leads to different test scores.

- **Critical Value Approach**: In this approach, researchers determine critical values based on the significance level and the sampling distribution of the test statistic. If the test statistic falls into the critical region (beyond the critical value), the null hypothesis is rejected.

  Example 6. For the same teaching method study with a significance level of $\alpha = 0.01$ and a two-tailed test, the critical values from the t-distribution with 58 degrees of freedom might be approximately ± 2.660. Since the computed test statistic of 2.19 does not exceed 2.660, we fail to reject the null hypothesis. Thus, we conclude that there is not enough evidence to support a significant difference in test scores between the two teaching methods.

3.2 One-sided tests and Two-sided tests:

In hypothesis testing, the choice between a one-sided (or one-tailed) test and a two-sided (or two-tailed) test depends on the nature of the research question and the direction of interest.

- **One-sided test:**
  - Focuses on detecting an effect in one specific direction (greater than or less than).
  - Null hypothesis (H_0) is typically stated as no effect or a specific value.
  - Example: $H_0 : \mu \leq cst$ versus $H_1 : \mu > cst$ (testing if the mean is greater than cst).

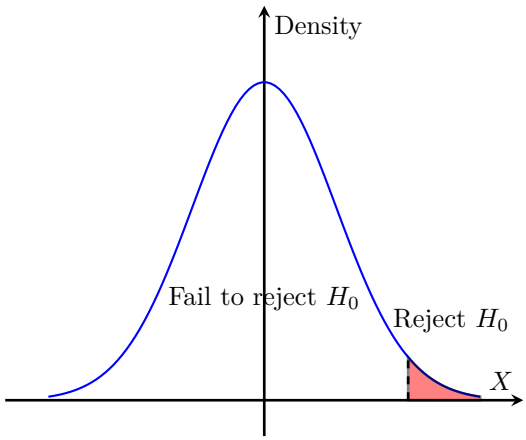


Figure 3.1: One-Sided Hypothesis Test

- **Two-sided test:**
  - Examines whether there is a significant difference in any direction.
  - Null hypothesis (H_0) often states no effect or equality.
  - Example: $H_0 : \mu = cst$ versus $H_1 : \mu \neq cst$ (testing if the mean is different from 10).

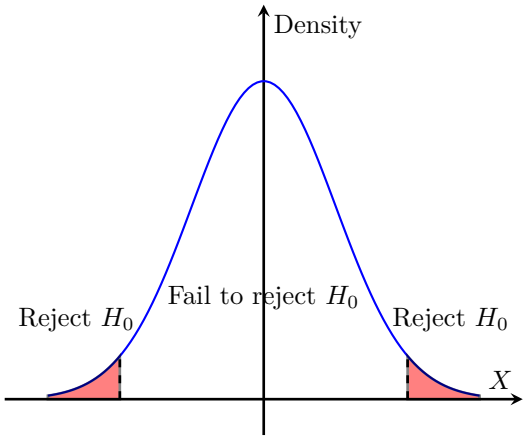


Figure 3.2: Two-Sided Hypothesis Test

The choice between one-sided and two-sided tests depends on the specific hypotheses being tested and the research question's requirements.

Example 7. Suppose you are conducting a hypothesis test on the average height (μ) of a certain population.

- **One-sided test:**
  - Null hypothesis (H_0): $\mu \leq 65$ inches
  - Alternative hypothesis (H_1): $\mu > 65$ inches
  - This one-sided test aims to determine if the average height is greater than 65 inches.
- **Two-sided test:**
  - Null hypothesis (H_0): $\mu = 70$ inches
  - Alternative hypothesis (H_1): $\mu \neq 70$ inches
  - This two-sided test aims to determine if the average height is different from 70 inches.

For the one-sided test, you would be interested in detecting whether the average height is significantly greater than the specified value (65 inches). In the two-sided test, the interest is in detecting any significant difference in the average height, whether it is greater or less than the specified value (70 inches).

The choice between one-sided and two-sided tests depends on the specific hypothesis and the research question you want to address.

3.3 The different types of errors

In hypothesis testing, there are two types of possible errors: type I errors and type II errors. Understanding these error types is important for evaluating and designing statistical tests.

Type I Error

A type I error occurs when the null hypothesis (H_0) is rejected even though it is true. In other words, a statistically significant result is found when there is really no effect. The probability of making a type I error is denoted by alpha (α) and is preset at the beginning of the test, usually at 5% or 1%.

Type II Error

A type II error occurs when the null hypothesis (H_0) fails to be rejected even though it is false. In other words, no statistically significant result is found even though there is a real effect. The probability of making a type II error is denoted by beta (β). The power of a test is equal to $1 - \beta$ and indicates the probability of correctly rejecting the null hypothesis when it is false.

Relationship Between Type I and Type II Errors

The type I and type II error rates are intrinsically related. As the significance level (α) decreases, making it harder to reject the null, the chance of type I error decreases but the chance of type II error increases. Conversely, increasing the significance level makes it easier to reject the null and lowers type II errors but raises type I errors. There is always a tradeoff between the two error types.

Understanding the different types of errors in hypothesis testing is crucial for properly evaluating statistical tests and avoiding misleading conclusions from data analysis. Both type I and type II errors should be considered when designing experiments and setting significance levels.

Null hypothesis is ...	**True**	**False**
Reject H_0	Type II Error (β)	No Error (Correct Decision)
Not Reject H_0	No Error (Correct Decision)	Type I Error (α)

Table 3.2: Types of Errors in Hypothesis Testing

3.4 Power of a Statistical Test

Statistical hypothesis testing is a fundamental practice in data analysis and research. It allows to make inferences about populations based on experimental data and results from samples. A crucial consideration in the planning and evaluation of hypothesis tests is *the power of the statistical tests* being used.

Properly assessing power helps ensure meaningful and conclusive results can be obtained from studies.

Definition 3.4.1. *The power of a statistical test gives the probability of rejecting the null hypothesis when it is false. Just as the significance level (alpha) gives the probability of rejecting the null hypothesis when it is true, power quantifies the chance of correctly rejecting the null hypothesis when it is false. Thus, power represents a test's ability to correctly reject the null hypothesis.*

Calculating power beforehand is important to ensure the sample size is sufficient for the test objectives. Otherwise, the test may be inconclusive, wasting resources. Power should generally not be calculated after the test, except to determine an adequate sample size for a follow-up study.

Example 8. Consider testing whether the average time per week spent watching TV is 4 hours versus the alternative that it is greater than 4 hours. We will calculate the power of this test for a specific value under the alternative hypothesis of 7 hours.

Solution:

1. State the null and alternative hypotheses
   - Null Hypothesis (H_0): The average time spent watching TV per week (μ) equals 4 hours
   - Alternative Hypothesis (H_1): The average time spent watching TV per week (μ) equals 6 hours
2. Define the parameters
   - μ_0 = Average time under the null hypothesis = 4 hours
   - μ_1 = Average time under the alternative hypothesis = 6 hours
3. Specify additional information
   - The standard deviation from past data is known to be 2 hours
   - The sample size is 4
4. Calculate the power of this test for a sample size of 4. Show the step-by-step working.

1. At the 5% significance level, the decision criterion for the test is to reject H_0 if $Z > 1.645$, where

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - 4}{\frac{2}{\sqrt{4}}} = \overline{X} - 4.$$

   The 5% critical value from the standard normal distribution is 1.645. Equating the critical Z-value to the calculated Z gives the corresponding (hypothetical) sample mean value:

$$\overline{X} = 5.645.$$

2. Calculate the Z-statistic assuming the alternative hypothesis is true, i.e., $\mu_1 = 6$:
$$Z = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} = \frac{5.645 - 6}{\frac{2}{\sqrt{4}}} = -0.355.$$

3. $P(Z > -0.355) = 0.6387$. The power of the test is approximately 64%. In general, tests with 80% power and higher are considered to be statistically powerful.

To increase power, one may:

- Increase effect size difference
- Increase sample size(s)
- Decrease variability
- Increase significance level (but increases type I error risk)

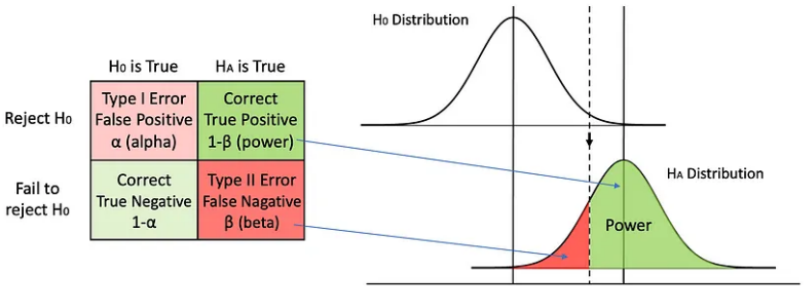


Figure 3.3: Hypothesis Testing Outcomes and Errors

3.5 Hypothesis Testing for the Mean

In this section, we examine common hypothesis testing methods for assessing the mean of a distribution. Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a population distribution, and our objective is to infer whether the population mean μ matches a specified value μ_0. This leads to three primary hypothesis testing scenarios, two of which are one-sided tests, and one is a two-sided test.

In each case, we use the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as our test statistic. When the variance σ^2 of the X_i's is known, we define the test statistic as:

$$W(X_1, X_2, \ldots, X_n) = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

If σ^2 is unknown, we use:

$$W(X_1, X_2, \ldots, X_n) = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

where S is the sample standard deviation:

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2}.$$

We now detail each type of hypothesis test, starting with the one-sided tests.

3.5.1 One-Sided Tests for the Mean

One-sided tests are used when the alternative hypothesis specifies a direction of deviation from the null hypothesis. These can be structured as either an upper-tail test or a lower-tail test.

3.5.1.1 Upper-Tail Test

In an upper-tail test, we test if the population mean μ is greater than a specified value μ_0. The hypotheses are:

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0.$$

Here, we focus on detecting if μ is significantly larger than μ_0. We use the test statistic W and compare it to a threshold c:

- Accept H_0 if $W \leq c$. - Reject H_0 in favor of H_1 if $W > c$.

The threshold c is chosen to meet the significance level α, ensuring that $P(W > c \mid H_0) = \alpha$. For a standard normal distribution, we have $c = z_\alpha$, where z_α is the $(1-\alpha)$ percentile of the standard normal distribution.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$ with known σ. To test

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

at significance level α, we calculate

$$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

If H_0 holds, $W \sim N(0, 1)$. We reject H_0 if $W > z_\alpha$.

3.5.1.2 Lower-Tail Test

In a lower-tail test, we test if the population mean μ is less than a specified value μ_0. The hypotheses are:

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0.$$

This test checks if μ is significantly less than μ_0. The test statistic W is compared to a threshold c:

- Accept H_0 if $W \geq c$. - Reject H_0 in favor of H_1 if $W < c$.

The threshold c is determined such that $P(W < c \mid H_0) = \alpha$. For the standard normal distribution, we set $c = -z_\alpha$.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$ with known σ. To test

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

at significance level α, we compute

$$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Under H_0, $W \sim N(0, 1)$. We reject H_0 if $W < -z_\alpha$.

3.5.2 Two-Sided Test for the Mean

A two-sided test is used when the alternative hypothesis does not specify a direction, testing only if the population mean μ is different from a specified value μ_0. The hypotheses are:

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

In this case, we use the test statistic W and decide based on its distance from zero. We define a threshold c and accept or reject H_0 as follows:

- Accept H_0 if $|W| \leq c$. - Reject H_0 in favor of H_1 if $|W| > c$.

The threshold c is determined by the significance level α, ensuring that the probability of a Type I error is less than or equal to α. We choose c so that:

$$P(\text{Type I error}) = P(|W| > c \mid H_0) = \alpha.$$

Since W follows a standard normal distribution under H_0, we set $c = z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile of the standard normal distribution.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$, with μ unknown but σ known. To test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

at significance level α, we use the test statistic:

$$W = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

If H_0 is true, then $W \sim N(0,1)$. We accept H_0 if $|W| \leq z_{\alpha/2}$ and reject it otherwise.

3.5.2.1 Relation to Confidence Intervals

Its interesting to examine the above acceptance region. Here, we accept H_0 if $|\bar{X} - \mu_0 \frac{\sigma}{\sqrt{n}}| \leq z_{\alpha/2}$. We can rewrite the above condition as $\mu_0 \in [\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$. The above interval should look familiar to you. It is the $(1-\alpha) \times 100\%$ confidence interval for μ_0. This is not a coincidence as there is a general relationship between confidence interval problems and hypothesis testing problems.

Example 9. Let $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where μ is unknown but σ is known. Design a level α test to choose between

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0.$,

find β, the probability of type II error, as a function of μ.

Solution: We have

$$\begin{aligned} \beta(\mu) &= P(\text{type II error}) \\ &= P(\text{accept } H_0|\mu) \\ &= P(|\bar{X} - \mu_0 \frac{\sigma}{\sqrt{n}}| < z_{\alpha/2}|\mu). \end{aligned}$$

If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Thus,

$$\begin{aligned} \beta(\mu) &= P(|\bar{X} - \mu_0 \frac{\sigma}{\sqrt{n}}| < z_{\alpha/2}|\mu) \\ &= P(\mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\ &= \Phi(z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) - \Phi(-z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}). \end{aligned}$$

Unknown variance: The above results (Example 9) can be extended to the case when we do not know the variance using the t-distribution. More specifically, consider the following example.

Example 10. Let $X_1, X_2, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Design a level α test to choose between

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$

Solution: Let S^2 be the sample variance for this random sample. Then, the random variable W defined as $W(X_1, X_2, ..., X_n) = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom, i.e., $W \sim T(n-1)$. Thus, we can repeat the analysis of Example 8.24 here. The only difference is that we need to replace σ by S and $z_{\alpha/2}$ by $t_{\alpha/2,n-1}$. Therefore, we accept H_0 if $|W| \leq t_{\alpha/2,n-1}$, and reject it otherwise. Let us look at a numerical example of this case.

Example 11. Consider the following scenario: Let $X_1, X_2, \ldots, X_n$ represent a random sample drawn from a normal distribution with mean μ and variance σ^2. In this case, the value of μ is unknown while σ is known. Our objective is to design a test with a significance level α to make a decision between the null hypothesis $H_0 : \mu \leq \mu_0$ and the alternative hypothesis $H_1 : \mu > \mu_0$.

To accomplish this, we can define a test statistic $W(X_1, X_2, \ldots, X_n)$ as $\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. If H_0 is true (meaning $\mu \leq \mu_0$), we anticipate that $\bar{X}$ (and consequently W) will be relatively small. Conversely, if H_1 is true, we expect $\bar{X}$ (and thus W) to be larger. Based on this observation, we can establish the following test: Select a threshold value, denoted as c. If $W \leq c$, we accept H_0; otherwise, if $W > c$, we accept H_1.

The question now arises: How do we determine the appropriate value for c? To ensure that the probability of committing a type I error (rejecting H_0 when it is true) is at most α, we need to examine the relationship between c and α. The probability of a type I error is contingent on the value of μ. More precisely, for any $\mu \leq \mu_0$, we can express the probability of a type I error as follows: $P(\text{type I error}|\mu) = P(\text{Reject } H_0|\mu) = P(W > c|\mu)$.

By employing properties of the normal distribution, we can simplify the expression above: $P(W > c|\mu) = P\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > c|\mu\right) = P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > c + \frac{\mu_0-\mu}{\frac{\sigma}{\sqrt{n}}}|\mu\right) \leq P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > c|\mu\right)$ (since $\mu \leq \mu_0$) $= 1-\Phi(c)$ (since $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ follows a standard normal distribution).

Consequently, we can select $\alpha = 1 - \Phi(c)$, which implies $c = z_\alpha$. Therefore, we accept H_0 if $\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha$, and we reject it otherwise.

The above analysis can be extended to other cases as well. In general, suppose we are given a random sample $X_1, X_2, \ldots, X_n$ drawn from a distribution, and let $\mu = E(X_i)$. Our objective is to make a decision between the null hypothesis $H_0 : \mu \leq \mu_0$ and the alternative hypothesis $H_1 : \mu > \mu_0$.

We can define the test statistic W as follows: $W(X_1, X_2, \ldots, X_n) = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ if σ (the variance of X_i) is known, and $W(X_1, X_2, \ldots, X_n) = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ if σ is unknown. If H_0 is true (i.e., $\mu \leq \mu_0$), we expect $\bar{X}$ (and thus W) to be relatively small. Conversely, if H_1 is true"' (i.e., $\mu > \mu_0$), we anticipate $\bar{X}$ (and thus W) to be larger. Based on this expectation, we can establish the following test:

Choose a threshold c. If $W \leq c$, we accept H_0; otherwise, if $W > c$, we accept H_1.

To determine the value of c, note that $P(\text{type I error}) = P(\text{Reject } H_0|H_0) = P(W > c|\mu \leq \mu_0) \leq P(W > c|\mu = \mu_0)$. The last inequality holds because increasing μ can only increase the probability of $W > c$. In other words, we assume the worst-case scenario, where $\mu = \mu_0$, to compute the probability of error. Hence, we can select c such that $P(W > c|\mu = \mu_0) = \alpha$. By following this procedure, we obtain the acceptance regions depicted in Table 8.3.

Case	Test Statistic	Acceptance Region
$X_i \sim N(\mu, \sigma^2)$, σ known	$W = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$W \leq z_\alpha$
n large, X_i non-normal	$W = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$	$W \leq z_\alpha$
$X_i \sim N(\mu, \sigma^2)$, σ unknown	$W = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$	*(remaining part of the sentence is missing)*

Table 3.3: One-sided hypothesis testing for the mean: $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$.

Example 12. The average adult male height in a certain country is 170 cm. We suspect that the men in a certain city in that country might have a different average height due to some environmental factors. We pick a random sample of size 9 from the adult males in the city and obtain the following values for their heights (in cm):

176.2, 157.9, 160.1, 180.9, 165.1, 167.2, 162.9, 155.7, 166.2

Assume that the height distribution in this population is normally distributed. Here, we need to decide between

$H_0 : \mu = 170$

$H_1 : \mu \neq 170$

Based on the observed data, is there enough evidence to reject H_0 at significance level $\alpha = 0.05$?

Solution: Let's first calculate the sample mean and sample standard deviation:

- Sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 166.44$
- Sample standard deviation, $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 8.548$
- The test statistic, $W = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{166.44-170}{8.548/\sqrt{9}} = -1.478$

Since we have a two-sided alternative hypothesis, we need to find the critical values for the t-distribution with $n - 1 = 9 - 1 = 8$ degrees of freedom. For a significance level of $\alpha = 0.05$, the critical values are $t_{\alpha/2,n-1} = t_{0.025,8} = 2.306$.

Since $|W| = 1.478 < t_{\alpha/2,n-1} = 2.306$, we do not have enough evidence to reject H_0 at the significance level of $\alpha = 0.05$. Therefore, based on the observed data, there is not enough evidence to conclude that the average height of men in the city is different from 170 cm.

3.6 Likelihood Ratio Tests

In this section, we will explore the concept of Likelihood Ratio Tests, which is a general hypothesis testing procedure. Before diving into the details, let's review the definition of the likelihood function, which we have previously discussed.

3.6.1 Review of the Likelihood Function

Consider a random sample $X_1, X_2, X_3, ..., X_n$ from a distribution with a parameter θ. The likelihood function is defined differently for discrete and continuous random variables:

- For discrete random variables, the likelihood function is denoted as $L(x_1, x_2, ..., x_n; \theta)$ and represents the probability mass function $P_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n; \theta)$.
- For continuous random variables, the likelihood function is denoted as $L(x_1, x_2, ..., x_n; \theta)$ and represents the probability density function $f_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n; \theta)$.

3.6.2 Likelihood Ratio Tests

Likelihood Ratio Tests are used in hypothesis testing when both the null and alternative hypotheses are simple. Suppose we have two hypotheses:

- Null Hypothesis: $H_0 : \theta = \theta_0$
- Alternative Hypothesis: $H_1 : \theta = \theta_1$

To decide between these hypotheses, we compare the likelihood functions:

$$l_0 = L(x_1, x_2, ..., x_n; \theta_0) \quad \text{(likelihood under } H_0\text{)}$$

$$l_1 = L(x_1, x_2, ..., x_n; \theta_1) \quad \text{(likelihood under } H_1\text{)}$$

If l_0 is significantly larger than l_1, we accept H_0. Conversely, if l_1 is significantly larger, we tend to reject H_0. The likelihood ratio $\frac{l_0}{l_1}$ is used to make the decision.

3.6.3 Likelihood Ratio Test for Simple Hypotheses

Consider a random sample $X_1, X_2, X_3, ..., X_n$ from a distribution with parameter θ. Suppose we want to test between two simple hypotheses:

- Null Hypothesis: $H_0 : \theta = \theta_0$
- Alternative Hypothesis: $H_1 : \theta = \theta_1$

We define the likelihood ratio as:

$$\lambda(x_1, x_2, ..., x_n) = \frac{L(x_1, x_2, ..., x_n; \theta_0)}{L(x_1, x_2, ..., x_n; \theta_1)}$$

To perform a Likelihood Ratio Test (LRT), we choose a constant c. We reject H_0 if $\lambda < c$ and accept it if $\lambda \geq c$. The value of c is determined based on the desired significance level α.

Example

Let's consider an example to illustrate how to perform a Likelihood Ratio Test. We revisit the radar problem, where we observe the random variable X given by $X = \theta + W$, with $W \sim N(0, \sigma^2 = \frac{1}{9})$. We want to test between the following hypotheses:

- Null Hypothesis: $H_0 : \theta = \theta_0 = 0$
- Alternative Hypothesis: $H_1 : \theta = \theta_1 = 1$

o design a level 0.05 test ($\alpha = 0.05$) to decide between H_0 and H_1, we calculate the likelihood ratio and determine the threshold value c. The decision rule is then defined based on the observed value of X.

3.6.4 Generalization to Non-Simple Hypotheses

If the hypotheses are not simple, meaning θ is an unknown parameter, we can still perform a Likelihood Ratio Test by partitioning the set of possible values for θ into two disjoint sets S_0 and S_1. The test involves finding the likelihood ratio for each possible value of θ and choosing the value that maximizes the likelihood ratio.

$$\lambda(x_1, x_2, ..., x_n) = \frac{\sup_{\theta \in S_0} L(x_1, x_2, ..., x_n; \theta)}{\sup_{\theta \in S_1} L(x_1, x_2, ..., x_n; \theta)}$$

To perform the Likelihood Ratio Test, we compare the likelihood ratio λ to a threshold value c. If $\lambda < c$, we reject H_0, and if $\lambda \geq c$, we fail to reject H_0.

The threshold value c is determined based on the desired significance level α. It is chosen such that the probability of rejecting H_0 when it is true (Type I error) is limited to α. In other words, we control the probability of falsely rejecting H_0.

3.7 Usual tests

3.7.1 Test on a Proportion

A proportion test is used to compare a sample proportion to a known population proportion. It tests whether the sample proportion is significantly different from the population proportion.

Hypotheses

The hypotheses are:

- Null hypothesis (H_0): The population proportion is equal to the sample proportion.

- Alternative hypothesis (H_1): The population proportion is different from the sample proportion.

$$H_0 : p = p_0 \quad \text{and} \quad H_1 : p \neq p_0$$

Test Statistic

The test statistic is calculated using the following formula:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where:

- $\hat{p}$ is the sample proportion,
- p_0 is the hypothesized population proportion,
- n is the sample size.

Rejection Zones

For a two-tailed test at a significance level α, we reject H_0 if the absolute value of z is greater than the critical value from the standard normal distribution (usually 1.96).

Example 13. : Suppose a researcher in Algeria wants to test if the proportion of people in support of a new policy is different from 0.5. A sample of 100 people is surveyed, and 60 are in favor of the policy.

Solution

Step 1: Define the hypotheses

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Step 2: Calculate the test statistic The sample proportion is $\hat{p} = \frac{60}{100} = 0.6$. The test statistic is:

$$z = \frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = \frac{0.1}{0.05} = 2$$

Step 3: Determine the rejection zone For $\alpha = 0.05$, the critical value from the standard normal distribution is ± 1.96.

Step 4: Decision Since $|z| = 2$ is greater than 1.96, we reject the null hypothesis H_0.

Conclusion

At a 5% significance level, we conclude that the proportion of individuals supporting the policy is different from 0.5.

3.7.2 Tests for Comparison of Means

The **t−test** is used to compare the means of two groups and determine if there is a significant difference between them. The most common types of t−tests are the one-sample t−test, the independent two−sample t−test, and the paired t−test.

Hypotheses

For a two−sample t−test, the hypotheses are:

- Null hypothesis (H_0): The means of the two groups are equal.
- Alternative hypothesis (H_1): The means of the two groups are not equal.

$$H_0 : \mu_1 = \mu_2 \quad \text{and} \quad H_1 : \mu_1 \neq \mu_2$$

Test Statistic

The test statistic for a two-sample t−test is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{x}_1, \bar{x}_2$ are the sample means,
- s_1^2, s_2^2 are the sample variances,
- n_1, n_2 are the sample sizes.

Rejection Zones

For a two-tailed test at significance level α, we reject H_0 if the calculated $|t|$ is greater than the critical value from the t-distribution.

Example 14. Suppose a researcher wants to compare the average exam scores between two regions in Algeria. The data is as follows:

$$\bar{x}_1 = 75,\ s_1 = 10,\ n_1 = 30 \quad \text{(Region A)}$$
$$\bar{x}_2 = 70,\ s_2 = 12,\ n_2 = 25 \quad \text{(Region B)}$$

Solution

Step 1: Define the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{(The means are equal)}$$

$$H_1 : \mu_1 \neq \mu_2 \quad \text{(The means are different)}$$

Step 2: Calculate the test statistic The test statistic for an independent t-test is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Substituting the given values:

$$t = \frac{75 - 70}{\sqrt{\frac{10^2}{30} + \frac{12^2}{25}}}$$

$$t = \frac{5}{\sqrt{\frac{100}{30} + \frac{144}{25}}} = \frac{5}{\sqrt{3.33 + 5.76}} = \frac{5}{\sqrt{9.09}} = \frac{5}{3.01} \approx 1.66$$

Step 3: Determine the rejection zone For a two-tailed test at $\alpha = 0.05$ with $df = n_1 + n_2 - 2 = 30 + 25 - 2 = 53$, the critical value from the t-distribution is approximately ± 2.00.

Step 4: Decision Since $|t| = 1.66$ is less than 2.00, we fail to reject the null hypothesis H_0.

Conclusion At a 5% significance level, we conclude that there is no significant difference in the average exam scores between students in Region A and Region B.

3.7.3 Tests for Comparison of Proportions

A test of comparison of proportions is used to compare two sample proportions and test if they are significantly different.

Hypotheses

For two proportions, the hypotheses are:

- Null hypothesis (H_0): The proportions are equal.
- Alternative hypothesis (H_1): The proportions are different.

$$H_0 : p_1 = p_2 \quad \text{and} \quad H_1 : p_1 \neq p_2$$

Test Statistic

The test statistic is calculated using the formula:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where p is the pooled proportion:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Rejection Zones

For a two-tailed test at significance level α, we reject H_0 if $|z| > 1.96$.

Example 15. Suppose a researcher wants to test if the proportion of individuals supporting a policy differs between two regions. The data is:

$$x_1 = 120,\ n_1 = 200 \quad \text{(Region A)}$$
$$x_2 = 90,\ n_2 = 150 \quad \text{(Region B)}$$

Solution

Step 1: Define the hypotheses

$$H_0 : p_1 = p_2, \quad H_1 : p_1 \neq p_2$$

Step 2: Calculate the pooled proportion The pooled proportion (p) is:

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 90}{200 + 150} = \frac{210}{350} = 0.6$$

Step 3: Calculate the test statistic The sample proportions are:

$$\hat{p}_1 = \frac{120}{200} = 0.6, \quad \hat{p}_2 = \frac{90}{150} = 0.6$$

The test statistic (z) is:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Substitute the values:

$$z = \frac{0.6 - 0.6}{\sqrt{0.6(1-0.6)\left(\frac{1}{200} + \frac{1}{150}\right)}} = \frac{0}{\sqrt{0.6 \times 0.4 \times \left(\frac{1}{200} + \frac{1}{150}\right)}}$$

$$z = \frac{0}{\sqrt{0.24 \times (0.005 + 0.00667)}} = \frac{0}{\sqrt{0.24 \times 0.01167}} = \frac{0}{\sqrt{0.0028}} = \frac{0}{0.053} = 0$$

Step 4: Determine the rejection zone For $\alpha = 0.05$, the critical value from the standard normal distribution is ± 1.96.

Step 5: Decision Since $|z| = 0$ is less than 1.96, we fail to reject the null hypothesis H_0.

Conclusion At a 5% significance level, we conclude that there is no significant difference in the proportion of individuals supporting the policy between the two regions.

3.7.4 Correlation Test

A correlation test is used to determine the strength and direction of the relationship between two continuous variables. The most common test is the Pearson correlation test.

Hypotheses

$$H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \neq 0$$

Test Statistic

The test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the sample correlation coefficient.

Example 16. A researcher collected data on the number of hours studied and corresponding exam scores for 10 students. The sample correlation coefficient is $r = 0.85$. The researcher wants to test if there is a significant correlation between the two variables.

Solution

Step 1: Define the hypotheses

$$H_0 : \rho = 0 \quad \text{(no correlation)}, \quad H_1 : \rho \neq 0 \quad \text{(significant correlation)}.$$

Step 2: Calculate the test statistic The test statistic is calculated using:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Here:

$$r = 0.85,\ n = 10$$

$$t = \frac{0.85\sqrt{10-2}}{\sqrt{1-0.85^2}} = \frac{0.85 \times 2.828}{\sqrt{1-0.7225}} = \frac{2.4038}{0.5266} = 4.56$$

Step 3: Determine the rejection zone The degrees of freedom are:

$$df = n - 2 = 10 - 2 = 8$$

For a two-tailed test at $\alpha = 0.05$, the critical value for t with $df = 8$ is approximately ± 2.306.

Step 4: Decision Since $|t| = 4.56$ is greater than 2.306, we reject the null hypothesis H_0.

Conclusion At a 5% significance level, there is a significant correlation between the number of hours studied and the exam scores in the sample of Algerian students.

3.7.5 Chi-Square Test of Independence

The Chi-Square Test of Independence is used to determine whether two categorical variables are independent or associated.

Hypotheses

H_0 : The two variables are independent. H_1 : The two variables are dependent.

Test Statistic

The test statistic is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Example 17. A researcher in Algeria wants to determine if region (North, South) is associated with preference for a type of transport (Car, Bus). A survey of 300 individuals provides the following contingency table:

	Car	Bus	Total
North	120	80	200
South	50	50	100
Total	170	130	300

Solution

Step 1: Define the hypotheses

H_0 : Transport preference is independent of region.

H_1 : Transport preference is dependent on region.

Step 2: Calculate the expected frequencies The expected frequency for each cell is calculated as:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Using this formula:

$$E_{11} = \frac{(200)(170)}{300} = 113.33, \quad E_{12} = \frac{(200)(130)}{300} = 86.67$$

$$E_{21} = \frac{(100)(170)}{300} = 56.67, \quad E_{22} = \frac{(100)(130)}{300} = 43.33$$

The expected frequency table is:

	Car	Bus	Total
North	113.33	86.67	200
South	56.67	43.33	100
Total	170	130	300

Step 3: Compute the test statistic The Chi-Square statistic is calculated as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For each cell:

$$\chi^2 = \frac{(120 - 113.33)^2}{113.33} + \frac{(80 - 86.67)^2}{86.67} + \frac{(50 - 56.67)^2}{56.67} + \frac{(50 - 43.33)^2}{43.33}$$

$$\chi^2 = \frac{44.44}{113.33} + \frac{44.49}{86.67} + \frac{44.49}{56.67} + \frac{44.49}{43.33}$$

$$\chi^2 = 0.392 + 0.513 + 0.785 + 1.026 = 2.716$$

Step 4: Determine the rejection zone The degrees of freedom are:

$$df = (\text{Rows} - 1)(\text{Columns} - 1) = (2 - 1)(2 - 1) = 1$$

For $\alpha = 0.05$, the critical value of χ^2 from the Chi-Square distribution table is 3.841.

Step 5: Decision Since $\chi^2 = 2.716$ is less than 3.841, we fail to reject the null hypothesis H_0.

Conclusion At a 5% significance level, we conclude that transport preference is independent of region in this sample.

3.8 Application

3.8.1 Solved exercices

Exercice 3.1. *Let $X \sim$ Geometric(θ). We observe X and we need to decide between*

$H_0 : \theta = \theta_0 = 0.5$,

$H_1 : \theta = \theta_1 = 0.1$.

Design a level 0.05 test ($\alpha = 0.05$) to decide between H_0 and H_1. Find the probability of type-II error β.

Exercice 3.2. *Let X_1, X_2, X_3, X_4 be a random sample from a $N(\mu, 1)$ distribution, where μ is unknown. Suppose that we have observed the following values $2.82, 2.71, 3.22, 2.67$. We would like to decide between*

$H_0 : \mu = \mu_0 = 2$,

$H_1 : \mu \neq 2$.

Assuming $\alpha = 0.1$, do you accept H_0 or H_1? If we require significance level α, find β as a function of μ and α.

Exercice 3.3. *Let $X_1, X_2, \ldots, X_{100}$ be a random sample from an unknown distribution. After observing this sample, the sample mean and the sample variance are calculated to be*

$\bar{X} = 21.32, S^2 = 27.6$.

Design a level 0.05 test to choose between

$H_0 : \mu = 20$,

$H_1 : \mu > 20$.

Do you accept or reject H_0?

Exercice 3.4. *Let X_1, X_2, X_3, X_4 be a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Suppose that we have observed the following values $3.58, 10.03, 4.77, 14.66$. We would like to decide between*

$H_0 : \mu \geq 10$,

$H_1 : \mu < 10$.

Assuming $\alpha = 0.05$, do you accept H_0 or H_1?

Exercice 3.5. *Let $X_1, X_2, \ldots, X_{81}$ be a random sample from an unknown distribution. After observing this sample, the sample mean and the sample variance are calculated to be $\bar{X} = 8.25, S^2 = 14.6$.*

Design a test to decide between

$H_0 : \mu = 9$,

$H_1 : \mu < 9$,

and calculate the P-value for the observed data.

Exercice 3.6. ***(Likelihood Function Review)*** *Consider a random sample of size 5 from a normal distribution with an unknown mean μ and known variance $\sigma^2 = 4$. The observed data is:*

$$X = (3, 5, 7, 6, 4)$$

Write the likelihood function for this sample, assuming that the observations are independent and identically distributed.

Exercice 3.7. ***(Likelihood Ratio Test for Simple Hypotheses)*** *Suppose you have a random sample of size 10 from a normal distribution, and you want to test the following hypotheses:*

$$H_0 : \mu = 3 \quad \textit{versus} \quad H_1 : \mu = 5$$

The observed sample mean is $\bar{X} = 4.5$*, with the data:*

$$X = (4.3, 4.6, 4.8, 4.1, 4.4, 4.7, 4.3, 4.9, 4.5, 4.2)$$

Using the likelihood ratio test, determine whether you reject H_0 *at the significance level* $\alpha = 0.05$*.*

Exercice 3.8. ***(Likelihood Ratio Test for Unknown Variance)*** *A researcher wants to test whether a population of measurements has variance equal to* $\sigma^2 = 9$*. Suppose the sample data is:*

$$X = (8, 10, 12, 15, 9)$$

The hypothesis test is:

$$H_0 : \sigma^2 = 9 \quad \textit{versus} \quad H_1 : \sigma^2 \neq 9$$

Perform the likelihood ratio test and determine the p-value for this test at $\alpha = 0.05$*.*

Exercice 3.9. ***(Generalized Likelihood Ratio Test)*** *Consider the following scenario: You are testing whether a random variable* X *follows a normal distribution. The null hypothesis is that the mean is* $\mu_0 = 0$ *and variance* $\sigma_0^2 = 1$*. The alternative hypothesis is that* $\mu_1 \neq 0$ *and* $\sigma_1^2 \neq 1$*.*

Given the data:

$$X = (1.1, 0.8, -0.3, 1.5, 0.4)$$

Calculate the likelihood ratio statistic and determine whether to reject the null hypothesis at the significance level $\alpha = 0.01$*.*

Exercice 3.10. ***(Likelihood Ratio Test for Nested Models)*** *Suppose you have two nested models for the likelihood function:*

- Model 1: $L_1(\theta) = \theta^x(1-\theta)^{n-x}$ *(Binomial distribution) - Model 2:* $L_2(\theta) = \theta^x(1-\theta)^{n-x}$*, with additional parameter constraints*

Test if model 1 is sufficient to describe the data compared to model 2 using the likelihood ratio test, given the data:

$$x = 12, \quad n = 20$$

Perform the test at the significance level $\alpha = 0.05$*.*

Solutions

Correction exercice 3.1. *We choose a threshold $c \in \mathbb{N}$ and compare the observed value of $X = x$ to c. We accept H_0 if $x \leq c$ and reject it if $x > c$. The probability of type I error is given by*

$$P(\textit{type I error}) = P(\textit{Reject } H_0|H_0) = P(\textit{Reject } H_0|\theta = 0.5) = P(X > c|\theta = 0.5) = \sum_{k=c+1}^{\infty} P(X = k)$$

(where $X \sim$ Geometric$(\theta_0 = 0.5)$)

$$= \sum_{k=c+1}^{\infty} (1-\theta_0)^{k-1}\theta_0 = (1-\theta_0)^c\theta_0 \sum_{l=0}^{\infty}(1-\theta_0)^l = (1-\theta_0)^c.$$

To have $\alpha = 0.05$, we need to choose c such that $(1-\theta_0)^c \leq \alpha = 0.05$, so we obtain

$$c \geq \frac{\ln \alpha}{\ln(1-\theta_0)} = \frac{\ln(0.05)}{\ln(0.5)} \approx 4.32.$$

Since we would like $c \in \mathbb{N}$, we can let $c = 5$. To summarize, we have the following decision rule: Accept H_0 if the observed value of X is in the set $A = \{1, 2, 3, 4, 5\}$, and reject H_0 otherwise. Since the alternative hypothesis H_1 is a simple hypothesis ($\theta = \theta_1$), there is only one value for β,

$$\beta = P(\textit{type II error}) = P(\textit{accept } H_0|H_1) = P(X \leq c|H_1) = 1-(1-\theta_1)^c = 1-(0.9)^5 = 0.41$$

Correction exercice 3.2. *We have a sample from a normal distribution with known variance, so using the first row in Table 3.3, we define the test statistic as*

$$W = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

We have $\bar{X} = 2.85$, $\mu_0 = 2$, $\sigma = 1$, and $n = 4$. So, we obtain

$$W = \frac{2.85 - 2}{1/\sqrt{4}} = 1.7.$$

Here, $\alpha = 0.1$, so $z_{\alpha/2} = z_{0.05} = 1.645$. Since $|W| > z_{\alpha/2}$, we reject H_0 and accept H_1. Here, the test statistic W is $W \sim 2(\bar{X} - 2)$. If $X \sim N(\mu, 1)$, then $\bar{X} \sim N(\mu, 1/4)$, and $W \sim N(2(\mu - 2), 1)$. Thus, we have

$$\beta = P(\textit{type II error}) = P(\textit{accept } H_0|\mu) = P(|W| < z_{\alpha/2}|\mu) = P(|W| < z_{\alpha/2})$$

(when $W \sim N(2(\mu - 2), 1)$)

$$= \Phi(z_{\alpha/2} - 2\mu + 4) - \Phi(-z_{\alpha/2} - 2\mu + 4).$$

Correction exercice 3.3. *Here, we have a non-normal sample, where $n = 100$ is large. Using the results of Table 3.3, specifically the second row, we define the test statistic as*

$$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{21.32 - 20}{\sqrt{27.6}/\sqrt{100}} = 2.51.$$

Here, $\alpha = 0.05$, so $z_\alpha = z_{0.05} = 1.645$. Since $W > z_\alpha$, we reject H_0 and accept H_1.

Correction exercice 3.4. *Here, we have a sample from a normal distribution with unknown mean and unknown variance. Using the third row in Table 3.3, we define the test statistic as*

$$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

Using the data we obtain $\bar{X} = 8.26, S = 5.10$. Therefore, we obtain

$$W = \frac{8.26 - 10}{5.10/\sqrt{4}} = -0.68.$$

Here, $\alpha = 0.05$, so $n = 4$, $t_{\alpha,n-1} = t_{0.05,3} = 2.35$. Since $W > -t_{\alpha,n-1}$, we fail to reject H_0, so we accept H_0.

Correction exercice 3.5. *Here, we have a non-normal sample, where $n = 81$ is large. Using the results of Table 8.4, specifically the second row, we define the test statistic as*

$$W = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{8.25 - 9}{\sqrt{14.6}/\sqrt{81}} = -1.767.$$

The P-value is $P(\text{type I error})$ when the test threshold c is chosen to be $c = -1.767$. Since the threshold for this test (as indicated by Table 8.4) is $-z_\alpha$, we obtain $-z_\alpha = -1.767$. Noting that by definition $z_\alpha = \Phi^{-1}(1 - \alpha)$, we obtain $P(\text{type I error})$ as

$$\alpha = 1 - \Phi(1.767) \approx 0.0386.$$

Therefore,

$$P\text{-value} \approx 0.0386.$$

Correction exercice 3.6. *Given a random sample of size 5 from a normal distribution with known variance $\sigma^2 = 4$, the observed data is:*

$$X = (3, 5, 7, 6, 4)$$

The likelihood function for a normal distribution with known variance $\sigma^2 = 4$ (i.e., $\sigma = 2$) is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Substituting $\sigma^2 = 4$, the likelihood becomes:

$$L(\mu) = \left(\frac{1}{\sqrt{8\pi}}\right)^5 \exp\left(-\frac{1}{8}\sum_{i=1}^{5}(x_i - \mu)^2\right)$$

Next, compute the sum of squared deviations from the sample mean $\bar{X}$, which is:

$$\bar{X} = \frac{3+5+7+6+4}{5} = 5$$

Thus, the likelihood is evaluated at $\mu = 5$ as:

$$L(5) = \left(\frac{1}{\sqrt{8\pi}}\right)^5 \exp\left(-\frac{10}{8}\right) \approx 0.0037$$

Correction exercice 3.7. *Test the hypotheses $H_0 : \mu = 3$ versus $H_1 : \mu = 5$ with the observed sample:*

$$X = (4.3, 4.6, 4.8, 4.1, 4.4, 4.7, 4.3, 4.9, 4.5, 4.2)$$

The sample mean $\bar{X} = 4.5$.

The Likelihood Ratio Test statistic for μ is:

$$\lambda = \frac{L_0}{L_1}$$

1. ***Likelihood under** H_0 **(when** $\mu = 3$**):***

$$L_0 = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi \cdot 0.145}} \exp\left(-\frac{(x_i - 3)^2}{2 \cdot 0.145}\right)$$

Compute the sum of squared deviations for $\mu = 3$:

$$\sum_{i=1}^{10}(x_i - 3)^2 = 24.12$$

2. ***Likelihood under** H_1 **(when** $\mu = 5$**):***

$$L_1 = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi \cdot 0.145}} \exp\left(-\frac{(x_i - 5)^2}{2 \cdot 0.145}\right)$$

Compute the sum of squared deviations for $\mu = 5$:

$$\sum_{i=1}^{10}(x_i - 5)^2 = 2.92$$

Now, the likelihood ratio is:

$$\lambda = \exp\left(\frac{1}{2 \cdot 0.145} \cdot (2.92 - 24.12)\right) = \exp(-73.1)$$

Since $\exp(-73.1)$ *is extremely small, we reject* H_0.

Correction exercice 3.8. ***Likelihood Ratio Test for Unknown Variance:*** *Test* $H_0 : \sigma^2 = 9$ *against* $H_1 : \sigma^2 \neq 9$ *with the sample:*

$$X = (8, 10, 12, 15, 9)$$

The Likelihood Ratio Test statistic for variance σ^2 *is:*

$$\lambda = \frac{\sup_{\sigma^2=9} L(\sigma^2)}{\sup_{\sigma^2} L(\sigma^2)}$$

1. ***Sample Variance:*** *Compute the sample variance:*

$$S^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{1}{4}\left((8-10)^2 + (10-10)^2 + (12-10)^2 + (15-10)^2 + (9-10)^2\right)$$

$$S^2 = \frac{1}{4}(4 + 0 + 4 + 25 + 1) = 7.2$$

2. ***Likelihood under*** H_0***:*** *Under* $H_0 : \sigma^2 = 9$*, the likelihood is:*

$$L_0 = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi \cdot 9}} \exp\left(-\frac{(x_i - \mu)^2}{2 \cdot 9}\right)$$

3. *Likelihood under* H_1*:** Under* H_1*, the likelihood is:*

$$L_1 = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi \cdot S^2}} \exp\left(-\frac{(x_i - \mu)^2}{2 \cdot S^2}\right)$$

We reject H_0 *if* $\lambda < c$*, where* c *is the critical value from the Chi-squared distribution with 4 degrees of freedom.*

Correction exercice 3.9. 1. ***Likelihood under*** H_0***:*** *Under* $H_0 : \mu_0 = 0$ *and* $\sigma_0^2 = 1$*, the likelihood is given by:*

$$L_0 = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}} e^{-\frac{X_i^2}{2}}$$

Calculating the sum of squares:

$$\sum_{i=1}^{5} X_i^2 = 1.1^2 + 0.8^2 + (-0.3)^2 + 1.5^2 + 0.4^2 = 4.15$$

Substituting, we get:

$$L_0 = \left(\frac{1}{\sqrt{2\pi}}\right)^5 e^{-\frac{4.15}{2}}$$

2. ***Likelihood under*** H_1***:*** *Under* $H_1 : \mu_1 \neq 0, \sigma_1^2 \neq 1$*, we estimate the parameters:*

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{5} X_i = \frac{1.1 + 0.8 - 0.3 + 1.5 + 0.4}{5} = 0.7$$

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{5}(X_i - \hat{\mu}_1)^2$$

Calculating:

$$\sum_{i=1}^{5}(X_i - 0.7)^2 = (1.1-0.7)^2 + (0.8-0.7)^2 + (-0.3-0.7)^2 + (1.5-0.7)^2 + (0.4-0.7)^2 = 1.16$$

$$\hat{\sigma}_1^2 = \frac{1.16}{5} = 0.232$$

The likelihood is:

$$L_1 = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}} e^{-\frac{(X_i - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2}}$$

3. ***Likelihood ratio test statistic:*** *The test statistic is:*

$$\Lambda = -2\ln\frac{L_0}{L_1}$$

Substituting the values:

$$\Lambda = -2\ln\left(\frac{\left(\frac{1}{\sqrt{2\pi}}\right)^5 e^{-\frac{4.15}{2}}}{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}}\right)^5 e^{-\frac{1.16}{2\hat{\sigma}_1^2}}}\right)$$

Decision: *Compare* Λ *to the critical value* $\chi^2_{1-\alpha,df=2}$*. At* $\alpha = 0.01$*,* $\chi^2_{0.99,df=2} \approx 9.21$*. If* $\Lambda > 9.21$*, reject* H_0*.*

Correction exercice 3.10. 1. ***Likelihood under Model 1:*** *The maximum likelihood estimate (MLE) of* θ *under Model 1 is:*

$$\hat{\theta}_1 = \frac{x}{n} = \frac{12}{20} = 0.6$$

Substituting into the likelihood function:

$$L_1 = \binom{n}{x}\hat{\theta}_1^x(1-\hat{\theta}_1)^{n-x}$$

2. ***Likelihood under Model 2:*** *Model 2 imposes additional constraints on* θ*. Let the restricted MLE be* $\hat{\theta}_2$*. Compute* L_2 *similarly.*

3. ***Likelihood ratio test statistic:*** *The test statistic is:*

$$\Lambda = -2 \ln \frac{L_2}{L_1}$$

Substitute the values of L_1 *and* L_2 *based on the data.*

Decision: *Compare* Λ *to the critical value* $\chi^2_{1-\alpha,df=1}$. *At* $\alpha = 0.05$, $\chi^2_{0.95,df=1} \approx 3.84$. *If* $\Lambda > 3.84$, *reject* H_0.

3.8.2 Unsolved Exercices

Exercice 3.1. ***(Likelihood Ratio Test for Multinomial Distribution)*** *Suppose you have data from a multinomial distribution with $k = 3$ categories, and observed counts:*

$$n_1 = 15, \quad n_2 = 25, \quad n_3 = 10.$$

You want to test the hypothesis:

$$H_0 : p_1 = p_2 = p_3 = \frac{1}{3}, \quad H_1 : p_i \text{ are not equal for all } i.$$

Design a likelihood ratio test and calculate the test statistic. Determine whether to reject H_0 at $\alpha = 0.05$.

Exercice 3.2. ***(Type-I and Type-II Errors in Normal Distribution)*** *Let $X_1, X_2, \ldots, X_{10}$ be a random sample from $N(\mu, \sigma^2)$, where $\sigma^2 = 4$. Consider the hypotheses:*

$$H_0 : \mu = 5, \quad H_1 : \mu > 5.$$

Design a level $\alpha = 0.05$ test based on the sample mean $\bar{X}$. Find the rejection region and express the probability of Type-II error β as a function of μ.

Exercice 3.3. ***(Hypothesis Testing for a Poisson Distribution)*** *Let $X \sim$ Poisson(λ). You observe $X = 12$ and wish to test:*

$$H_0 : \lambda = 10, \quad H_1 : \lambda > 10.$$

Design a test at $\alpha = 0.01$. Determine the rejection region, and calculate the p-value for the observed data.

Exercice 3.4. ***(Hypothesis Testing for a Proportion)*** *In a clinical trial, 50 patients were treated, and 32 of them showed improvement. Test the hypothesis:*

$$H_0 : p = 0.6, \quad H_1 : p > 0.6,$$

where p is the true proportion of patients who improve with the treatment. Use a significance level of $\alpha = 0.05$.

Exercice 3.5. ***(Hypothesis Testing with Two Independent Samples)*** *Two independent samples are drawn from two populations. The data are as follows:*

$$X_1 = (12, 15, 14, 10, 13), \quad X_2 = (18, 20, 19, 17, 22).$$

Assume that the populations are normally distributed with equal variances. Test the hypothesis:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2,$$

at a significance level $\alpha = 0.05$. Also, calculate the p-value for the observed data.

Chapter 4

Analysis of Variance (ANOVA)

Objectives

After studying this chapter, you should:

- Understand the need for comparing more than two groups.
- Recognize the underlying models in ANOVA.
- Perform one-way and two-way ANOVA.
- Interpret ANOVA results effectively.

4.1 Introduction

Analysis of Variance (ANOVA) is a widely used statistical technique that allows researchers to compare the means of three or more independent groups to determine if there is a statistically significant difference among them. The primary advantage of ANOVA over multiple t−tests is that it reduces the risk of Type I error, which occurs when a null hypothesis is incorrectly rejected. By evaluating all group means simultaneously, ANOVA provides a more robust and reliable framework for hypothesis testing.

ANOVA is grounded in the fundamental principle that any observed differences in group means can be attributed to one of two sources: systematic variation due to the treatment or condition applied to the groups, and random variation due to inherent variability in the data. By partitioning the total variability into these components, ANOVA assesses whether the systematic variation is large enough to be statistically significant.

The method has various applications across numerous fields, including agriculture, medicine, psychology, and social sciences. For example, in a clinical trial comparing the efficacy of different medications, ANOVA can help determine whether the observed differences in patient outcomes are due to the effects of the medications rather than random chance. Similarly, in agricultural research, it can be used to compare the yields of different crop varieties under varying environmental conditions.

ANOVA can be classified into different types, primarily one-way ANOVA and two-way ANOVA. One-way ANOVA is used when comparing means across groups defined by a single factor with multiple levels, while two-way ANOVA evaluates the impact of two independent variables on a dependent variable and can also assess the interaction between the factors.

In summary, ANOVA is a crucial tool in statistical analysis, enabling researchers to draw meaningful conclusions from experimental data. This chapter will delve into the underlying principles of ANOVA, the assumptions that must be met for valid results, the computation of relevant statistics, and practical applications through detailed examples.

4.2 Theoretical Concepts of ANOVA

To effectively apply Analysis of Variance (ANOVA), it is crucial to understand the theoretical concepts that support this statistical technique. ANOVA serves to partition the total variance in data into components attributable to different sources, enabling researchers to assess whether observed differences among group means are statistically significant or due to random variability.

This section will cover essential concepts such as variance, sum of squares, degrees of freedom, and the critical assumptions underlying ANOVA. By grasping these foundational elements, readers will be better equipped to conduct ANOVA analyses and interpret their results accurately.

4.2.1 Concept of Variance

Variance is a measure of the spread of a set of values. It is defined as:

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

where N is the number of observations, X_i represents each individual observation, and μ is the mean of the observations.

4.2.2 Total Variance

In the context of ANOVA, the total variance of the observations can be decomposed into variance between groups and variance within groups:

$$\text{Total Variance} = \text{Variance Between Groups} + \text{Variance Within Groups}$$

This can be expressed mathematically as:

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2$$

where σ_T^2 is the total variance, σ_B^2 is the variance between groups, and σ_W^2 is the variance within groups.

4.2.3 Sum of Squares

The concept of Sum of Squares (SS) is fundamental in ANOVA. The total sum of squares can be defined as:

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$$

where k is the number of groups, n_i is the number of observations in group i, X_{ij} is the j-th observation in group i, and $\bar{X}$ is the overall mean.

The Between-Groups Sum of Squares is calculated as:

$$\text{SSB} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

where $\bar{X}_i$ is the mean of group i.

The Within-Groups Sum of Squares is given by:

$$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

4.2.4 Mean Squares

The Mean Squares are computed by dividing the Sum of Squares by their respective degrees of freedom:

$$\text{MSB} = \frac{\text{SSB}}{df_B} \quad \text{and} \quad \text{MSW} = \frac{\text{SSW}}{df_W}$$

where df_B is the degrees of freedom between groups and df_W is the degrees of freedom within groups.

4.2.5 F-Ratio

The F-ratio is the test statistic used in ANOVA, defined as:

$$F = \frac{\text{MSB}}{\text{MSW}}$$

This ratio compares the variance between the group means to the variance within the groups. A larger F value suggests a greater likelihood that the observed differences among the group means are significant.

4.2.6 ANOVA Table

The ANOVA table summarizes the results of the ANOVA analysis:

Source	Sum of Squares	Degrees of Freedom	Mean Square
Between Groups	SSB	$k-1$	$\text{MSB} = \frac{\text{SSB}}{k-1}$
Within Groups	SSW	$n-k$	$\text{MSW} = \frac{\text{SSW}}{n-k}$
Total	SST	$n-1$	

The F-ratio for testing the significance of between-group variability:

$$F = \frac{\text{MSB}}{\text{MSW}}$$

4.2.7 Hypothesis Testing in ANOVA

In ANOVA, we formulate two hypotheses:

- Null Hypothesis (H_0): All group means are equal ($\mu_1 = \mu_2 = \cdots = \mu_k$).
- Alternative Hypothesis (H_1): At least one group mean is different.

The null hypothesis is rejected if the calculated F-ratio exceeds the critical value from the F-distribution table at a specified significance level (e.g., $\alpha = 0.05$).

Understanding these theoretical concepts and formulas is essential for performing ANOVA and interpreting its results. The method provides a systematic way to evaluate the differences between group means while accounting for variability within the data.

4.3 Basics of ANOVA

The common feature of all ANOVA tests is comparing means across multiple groups to test if any statistically significant differences exist. This is often used in experiments where treatments or conditions are assigned to different groups.

4.3.1 One-Way ANOVA

A one-way ANOVA is used when comparing means across groups formed by a single factor with multiple levels.

4.3.1.1 Hypotheses

The hypotheses for one-way ANOVA are:

- H_0: All group means are equal ($\mu_1 = \mu_2 = \cdots = \mu_k$).
- H_1: At least one group mean differs.

4.3.1.2 Assumptions

- Independence of observations.
- Normality of each group.
- Homogeneity of variances (similar variance across groups).

Example 18. Assume we have three brands of light bulbs with lifetimes (in hours) as follows:

Brand 1	Brand 2	Brand 3
16	18	26
15	22	31
13	20	24
21	16	30
15	24	24

Solution

1. **Calculate the Means for Each Brand**

   Brand 1:
   $$\bar{X}_1 = \frac{16 + 15 + 13 + 21 + 15}{5} = \frac{80}{5} = 16$$

   Brand 2:
   $$\bar{X}_2 = \frac{18 + 22 + 20 + 16 + 24}{5} = \frac{100}{5} = 20$$

   Brand 3:
   $$\bar{X}_3 = \frac{26 + 31 + 24 + 30 + 24}{5} = \frac{135}{5} = 27$$

2. **Calculate the Overall Mean**

$$\bar{X} = \frac{80 + 100 + 135}{15} = \frac{315}{15} = 21$$

3. **Calculate Sum of Squares**

Total Sum of Squares (SST): The total sum of squares measures the total variability in the data:

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$$

Calculating SST:

$$\begin{aligned}
\text{SST} &= (16-21)^2 + (15-21)^2 + (13-21)^2 + (21-21)^2 + (15-21)^2 \\
&\quad + (18-21)^2 + (22-21)^2 + (20-21)^2 + (16-21)^2 + (24-21)^2 \\
&\quad + (26-21)^2 + (31-21)^2 + (24-21)^2 + (30-21)^2 + (24-21)^2 \\
&= (-5)^2 + (-6)^2 + (-8)^2 + (0)^2 + (-6)^2 + (-3)^2 + (1)^2 + (-1)^2 + (-5)^2 + (3)^2 + (5)^2 + (10) \\
&= 25 + 36 + 64 + 0 + 36 + 9 + 1 + 1 + 25 + 9 + 25 + 100 + 9 + 81 + 9 \\
&= 424
\end{aligned}$$

Between-Groups Sum of Squares (SSB): Measures the variability between the group means:

$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X})^2$$

Calculating SSB:

$$\begin{aligned}
\text{SSB} &= 5 \times (16-21)^2 + 5 \times (20-21)^2 + 5 \times (27-21)^2 \\
&= 5 \times 25 + 5 \times 1 + 5 \times 36 \\
&= 125 + 5 + 180 \\
&= 310
\end{aligned}$$

Within-Groups Sum of Squares (SSW): Measures the variability within each group:

$$\text{SSW} = \text{SST} - \text{SSB} = 424 - 310 = 114$$

4. **Calculate Degrees of Freedom**

- Degrees of freedom between groups:

$$df_B = k - 1 = 3 - 1 = 2$$

- Degrees of freedom within groups:

$$df_W = n - k = 15 - 3 = 12$$

- Total degrees of freedom:

$$df_T = n - 1 = 15 - 1 = 14$$

5. **Calculate Mean Squares**
   - Mean Square Between (MSB):

$$\text{MSB} = \frac{\text{SSB}}{df_B} = \frac{310}{2} = 155$$

   - Mean Square Within (MSW):

$$\text{MSW} = \frac{\text{SSW}}{df_W} = \frac{114}{12} \approx 9.5$$

6. **Calculate F-Ratio**

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{155}{9.5} \approx 16.32$$

   So, to determine if the difference in means is statistically significant, compare the calculated F-ratio to the critical value of F from the F-distribution table at $df_1 = 2$ and $df_2 = 12$ for a given significance level (e.g., $\alpha = 0.05$). If the calculated F exceeds the critical value, we reject the null hypothesis (H_0) that states all group means are equal.

4.3.2 Two-Way ANOVA

Two-way ANOVA extends the analysis to two factors, each with multiple levels. This allows the study of the interaction effect between two factors. The interaction effect in two-way ANOVA examines whether the effect of one factor depends on the level of another factor. This is essential in many experimental designs where factors do not operate independently.

4.3.2.1 Hypotheses

1. **Main Effects**:
   - **Factor A**:
     - *Null Hypothesis (H_{0A})*: There is no effect of Factor A on the dependent variable.

$$H_{0A} : \mu_{A1} = \mu_{A2} = \ldots = \mu_{Am}$$

  - *Alternative Hypothesis (H_{1A})*: At least one mean for Factor A is different.
- **Factor B**:
  - *Null Hypothesis (H_{0B})*: There is no effect of Factor B on the dependent variable.

$$H_{0B} : \mu_{B1} = \mu_{B2} = \ldots = \mu_{Bn}$$

  - *Alternative Hypothesis (H_{1B})*: At least one mean for Factor B is different.

2. **Interaction Effect**:
   - *Null Hypothesis (H_{0AB})*: There is no interaction effect between Factors A and B.
   - *Alternative Hypothesis (H_{1AB})*: There is an interaction effect between Factors A and B.

4.3.2.2 Assumptions

- Independence of observations.
- Normality of each combination of factor levels.
- Homogeneity of variances across combinations.

Example 19. Consider a study on detergent effectiveness across two types of stains. The data are arranged as follows:

	Stain 1	Stain 2
Detergent A	77	81
Detergent B	66	78
Detergent C	73	74

Solution

1. **Calculate Row, Column, and Interaction Effects**
   1. **Data Overview**:
      - **Detergents (Rows)**: A, B, C
      - **Stains (Columns)**: 1, 2
   2. **Calculate Means**:
      - Mean for each detergent (row):

$$\bar{X}_A = \frac{77 + 81}{2} = 79,$$
$$\bar{X}_B = \frac{66 + 78}{2} = 72,$$
$$\bar{X}_C = \frac{73 + 74}{2} = 73.5.$$

- Mean for each stain (column):

$$\bar{X}_1 = \frac{77+66+73}{3} = \frac{216}{3} = 72,$$
$$\bar{X}_2 = \frac{81+78+74}{3} = \frac{233}{3} \approx 77.67.$$

- Grand Mean:

$$\bar{X} = \frac{77+81+66+78+73+74}{6} = \frac{429}{6} \approx 71.5.$$

3. **Calculate Sum of Squares**:

- Total Sum of Squares (SST):

$$SST = \sum_{i,j}(X_{ij}-\bar{X})^2 = (77-71.5)^2+(81-71.5)^2+(66-71.5)^2+(78-71.5)^2+(73-71.5)^2+(74-71.5)^2$$

$$= 30.25 + 90.25 + 30.25 + 42.25 + 2.25 + 6.25 = 199.5.$$

- Between Groups Sum of Squares (SSB):

$$SSB = n_j \sum_i (\bar{X}_i - \bar{X})^2 = 2\left[(79-75)^2 + (72-75)^2 + (73.5-75)^2\right]$$

$$= 2\left[(4)^2 + (-3)^2 + (-1.5)^2\right] = 2\left[16+9+2.25\right] = 2 \times 27.25 = 54.5.$$

- Within Groups Sum of Squares (SSW):

$$SSW = SST - SSB = 199.5 - 54.5 = 145.$$

4. **Degrees of Freedom**:

- Degrees of Freedom for Treatments ($df_{treatments}$): $k-1 = 3-1 = 2$
- Degrees of Freedom for Error (df_{error}): $n-k = 6-3 = 3$
- Total Degrees of Freedom: $df_{total} = N-1 = 6-1 = 5$

2. **Construct the ANOVA Table**

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Between Groups	$SSB = 54.5$	$df_{treatments} = 2$	$MSB = \frac{SSB}{df_{treatments}} = \frac{54.5}{2} = 27.25$	$F = \frac{MSB}{MSW}$
Within Groups	$SSW = 145$	$df_{error} = 3$	$MSW = \frac{SSW}{df_{error}} = \frac{145}{3} \approx 48.33$	
Total	$SST = 199.5$	$df_{total} = 5$		

3. **Calculating the F-ratio**:

$$F = \frac{MSB}{MSW} = \frac{27.25}{48.33} \approx 0.564.$$

4. **Interpret the Results**

   1. **F-Ratio**: The calculated F-value is approximately 0.564. This value needs to be compared against the critical F-value from the F-distribution table for $df_{treatments} = 2$ and $df_{error} = 3$ at a chosen significance level (typically $\alpha = 0.05$).

   2. **Conclusion**:

      - If F is less than the critical F-value, we fail to reject the null hypothesis, indicating that there is no significant difference in detergent effectiveness across the two stains.
      - If F is greater than the critical F-value, we reject the null hypothesis, suggesting that at least one detergent performs significantly differently across the types of stains.

Summary: In this example, we calculated the means, sum of squares, degrees of freedom, and constructed an ANOVA table. The results indicate whether the detergent types had different effectiveness on the stains. Further statistical tests may be warranted if the results are significant.

4.4 Assumptions of ANOVA

The validity of the Analysis of Variance (ANOVA) depends on certain underlying assumptions. These assumptions ensure that the results of the ANOVA are reliable and that the conclusions drawn from the analysis are accurate. Violating these assumptions can lead to incorrect inferences, such as falsely detecting a difference between group means when there is none or failing to detect a real difference. In this section, we discuss the key assumptions that must be satisfied for ANOVA to be appropriately applied.

4.4.1 Independence of Observations

Independence of observations is a fundamental assumption of ANOVA, requiring that the data points within each group, and across groups, are independent of each other. This means that the value of one observation does not influence or provide any information about another observation. Violating this assumption can lead to underestimated variability within groups, increasing the risk of Type I errors (falsely rejecting the null hypothesis).

In practice, independence is often ensured by proper experimental design, such as random sampling or random assignment of subjects to groups. For example, in a study comparing different teaching methods, students should be randomly assigned to the different methods to ensure that the groups are independent of each other.

4.4.2 Normality

ANOVA assumes that the data in each group are approximately normally distributed. This assumption is especially important when the sample size is small, as ANOVA relies on the Central Limit Theory (CLT) for larger sample sizes to be robust against deviations from normality. If the data are not normally distributed, the results of ANOVA might be misleading, particularly the F-test used to determine statistical significance.

There are several ways to assess the normality of data:

- **Visual Inspection:** Use graphical methods like Q-Q plots (quantile-quantile plots) to compare the distribution of the data to a normal distribution. If the data points closely follow the reference line in a Q-Q plot, the data are likely to be normally distributed.
- **Shapiro-Wilk Test:** This is a formal statistical test that tests the null hypothesis that the data are normally distributed. A significant result (p-value < 0.05) suggests a departure from normality.
- **Histogram:** Plotting a histogram of the data for each group can provide a visual check for normality, although this method is less precise than the Q-Q plot or formal tests.

If normality is violated, and the sample size is small, transformations of the data (e.g., logarithmic or square root transformations) or non-parametric alternatives to ANOVA, such as the Kruskal-Wallis test, can be considered.

4.4.3 Homogeneity of Variances (Homoscedasticity)

Homogeneity of variances, also known as homoscedasticity, assumes that the variances within each group are approximately equal. This assumption ensures that each group contributes equally to the ANOVA, preventing any one group with a larger variance from unduly influencing the results.

To test the assumption of homogeneity of variances, several methods can be employed:

- **Levene's Test:** Levene's test is a formal statistical test used to assess the equality of variances. It tests the null hypothesis that the variances are equal across groups. A significant result (p-value < 0.05) indicates a violation of the homogeneity of variances assumption.
- **Bartlett's Test:** Bartlett's test also assesses the equality of variances, but it is more sensitive to deviations from normality. Therefore, it is recommended only when the normality assumption is met.
- **Boxplots:** Boxplots can be used as a visual tool to check for equal variances. If the spread of the data (indicated by the length of the box and whiskers) is similar across groups, the assumption is likely satisfied.

If the assumption of homogeneity of variances is violated, several approaches can be taken:

- **Transformations:** Transforming the data (e.g., using logarithmic or square root transformations) may help stabilize the variances.
- **Welch's ANOVA:** An alternative to the standard ANOVA, Welch's ANOVA does not assume equal variances and can be used when this assumption is violated.
- **Robust ANOVA Methods:** There are robust methods designed to handle heteroscedasticity (unequal variances) directly.

4.4.4 Random Sampling

Random sampling is another key assumption of ANOVA. It assumes that the samples drawn from the populations are random and that each member of the population has an equal chance of being included in the sample. This assumption ensures that the sample is representative of the population, allowing the results of the ANOVA to be generalized beyond the sample.

Random sampling also helps ensure the independence of observations, as it reduces the likelihood of systematic biases that could influence the data.

4.4.5 Additivity and Linearity

Additivity and linearity assume that the effects of the different factors in the ANOVA model add together to produce the total effect on the dependent variable. This assumption implies that the relationship between the dependent variable and each factor is linear, and that there are no interactions between factors unless explicitly modeled.

For a one-way ANOVA, where there is only one factor, additivity and linearity are less of a concern. However, for more complex designs (e.g., two-way ANOVA or factorial designs), this assumption becomes more critical.

4.4.6 Sphericity (for Repeated Measures ANOVA)

Sphericity is a specific assumption for repeated measures ANOVA, which is used when the same subjects are measured under different conditions. Sphericity assumes that the variances of the differences between all combinations of related groups (levels) are equal. If sphericity is violated, the F-ratios in the ANOVA table can be inflated, leading to an increased risk of Type I errors.

To test sphericity, Mauchly's test is often used. If sphericity is violated, adjustments to the degrees of freedom (such as the Greenhouse-Geisser correction) can be applied to reduce the risk of false positives.

Understanding and checking the assumptions of ANOVA are crucial steps in the analysis process. While ANOVA is a robust technique, especially with large sample sizes, ensuring that these assumptions are met will increase the

reliability of your results. When assumptions are violated, alternative methods or adjustments should be considered to avoid drawing incorrect conclusions from your data.

4.5 Extensions of ANOVA

While the traditional Analysis of Variance (ANOVA) is a powerful tool for comparing the means of different groups, various extensions of ANOVA have been developed to address more complex experimental designs and data structures. These extensions allow researchers to analyze data that do not fit neatly into the framework of one-way or two-way ANOVA, enabling the study of interactions between factors, repeated measurements, and other advanced scenarios. In this section, we explore some of the most common extensions of ANOVA, including factorial ANOVA, repeated measures ANOVA, ANCOVA, and MANOVA.

4.5.1 Factorial ANOVA

Factorial ANOVA is an extension of the basic ANOVA model used when there are two or more independent variables (factors). Unlike one-way ANOVA, which considers only one factor at a time, factorial ANOVA examines the effects of multiple factors simultaneously and can identify both the main effects of each factor and the interactions between them.

Main Effects and Interaction Effects

- **Main Effects:** The main effect of a factor is the impact that the factor has on the dependent variable, averaging over the levels of other factors. For example, in an experiment with two factors—drug type (Factor A) and dosage level (Factor B)—the main effect of drug type would be the average effect of the different drugs across all dosage levels.

- **Interaction Effects:** Interaction effects occur when the effect of one factor depends on the level of another factor. For instance, if the effect of drug type varies depending on the dosage level, this would indicate an interaction between the two factors.

Factorial ANOVA can be conducted for any number of factors, leading to two-way, three-way, or even higher-order ANOVAs. The complexity increases with the number of factors, but so does the ability to capture nuanced relationships in the data.

4.5.2 Repeated Measures ANOVA

Repeated Measures ANOVA is used when the same subjects are measured under different conditions or across different time points. This extension of ANOVA is particularly useful in longitudinal studies or experiments where subjects are exposed to multiple treatments.

Advantages of Repeated Measures Design

- **Reduced Variability:** Since the same subjects are used in all conditions, repeated measures designs typically result in reduced variability within subjects, increasing the power of the statistical test.

- **Efficiency:** Repeated measures designs are more efficient, as fewer subjects are needed to detect an effect compared to a design where different subjects are used for each condition.

Sphericity Assumption

A key assumption of repeated measures ANOVA is **sphericity**, which assumes that the variances of the differences between all combinations of related conditions are equal. If this assumption is violated, the F-ratios may be inflated, leading to an increased risk of Type I errors. If sphericity is violated, corrections such as Greenhouse-Geisser or Huynh-Feldt can be applied to adjust the degrees of freedom.

4.5.3 Analysis of Covariance (ANCOVA)

Analysis of Covariance (ANCOVA) combines the principles of ANOVA and regression by including both categorical and continuous variables as predictors. ANCOVA is used to adjust the dependent variable for the effects of one or more continuous covariates, allowing for a more precise estimation of the effects of the categorical factors.

Controlling for Covariates

ANCOVA adjusts the group means on the dependent variable by removing the linear effects of the covariates. This is particularly useful when there are variables that are not of primary interest but may influence the outcome. For example, in an educational study comparing different teaching methods, a covariate such as students' initial knowledge level could be controlled for, allowing for a more accurate comparison of the teaching methods.

Assumptions of ANCOVA

In addition to the assumptions of ANOVA, ANCOVA has its own specific assumptions:

- **Linearity:** The relationship between the covariate and the dependent variable must be linear. - **Homogeneity of Regression Slopes:** The effect of the covariate on the dependent variable should be the same across all groups. - **Independence of Covariate and Treatment Effect:** The covariate should not be correlated with the treatment effect.

4.5.4 Multivariate Analysis of Variance (MANOVA)

Multivariate Analysis of Variance (MANOVA) extends ANOVA to multiple dependent variables. Instead of analyzing each dependent variable separately, MANOVA tests the hypothesis that the mean vectors of the dependent variables are equal across groups.

Advantages of MANOVA

- **Control of Type I Error:** MANOVA controls the Type I error rate when multiple dependent variables are involved, avoiding the inflation of the error rate that would occur if separate ANOVAs were conducted for each dependent variable.

- **Detection of Multivariate Effects:** MANOVA can detect relationships between groups that may not be apparent when looking at each dependent variable in isolation. It can reveal how groups differ across a combination of dependent variables.

Assumptions of MANOVA

MANOVA has similar assumptions to ANOVA, but with additional considerations for the multivariate context:

- **Multivariate Normality:** The dependent variables are assumed to be multivariate normally distributed within each group. - **Homogeneity of Covariance Matrices:** The covariance matrices of the dependent variables should be equal across groups. This is analogous to the homogeneity of variances assumption in ANOVA.

The extensions of ANOVA—such as Factorial ANOVA, Repeated Measures ANOVA, ANCOVA, and MANOVA—provide powerful tools for analyzing more complex experimental designs. Each of these techniques builds on the basic principles of ANOVA while introducing additional capabilities to handle multiple factors, repeated measures, covariates, and multiple dependent variables. Understanding these extensions allows researchers to apply ANOVA to a broader range of data types and research questions, enhancing the flexibility and utility of this fundamental statistical method.

4.6 Post-Hoc Tests

After conducting an Analysis of Variance (ANOVA) and finding significant differences among group means, it is essential to determine which specific groups differ from each other. Post-hoc tests are statistical procedures used for this purpose. They control for the Type I error rate that can occur when making multiple comparisons. In this section, we discuss common post-hoc tests, their application, and examples.

4.6.1 Purpose of Post-Hoc Tests

The primary purpose of post-hoc tests is to identify specific pairs of group means that are significantly different from one another after a significant ANOVA result. When multiple groups are compared, the risk of incorrectly rejecting the null hypothesis increases with each additional comparison. Post-hoc tests provide a way to perform these comparisons while controlling for that risk.

4.6.2 Common Post-Hoc Tests

Several post-hoc tests can be used following ANOVA, including:

- **Tukey's Honest Significant Difference (HSD) Test:** This test compares all possible pairs of group means while controlling the overall Type I error rate. It is particularly powerful when the sample sizes are equal.
- **Bonferroni Correction:** This method adjusts the significance level based on the number of comparisons. It is more conservative than Tukey's HSD, making it less likely to detect significant differences but safer against Type I errors.
- **Scheffé's Test:** A flexible test that allows for comparisons of means of different linear combinations of groups. It is more robust for unequal sample sizes but less powerful for pairwise comparisons than Tukey's HSD.
- **Dunnett's Test:** Used when comparing multiple treatment groups to a single control group, providing specific comparisons that are relevant to experimental designs where one group is considered the baseline.

Example 20. Consider a study examining the effects of three different diets on weight loss over 12 weeks. The diets are Diet A, Diet B, and Diet C. After performing a one-way ANOVA, the results indicate that there is a significant difference in weight loss among the groups ($F(2, 27) = 5.43, p < 0.01$).

Solution

- **Conducting Post-Hoc Tests** Given the significant ANOVA result, the researcher decides to perform Tukey's HSD to determine which specific diets differ in terms of weight loss.
- **Tukey's HSD Example** Assume the following weight loss means and standard deviations for the three diets:
  - Diet A: Mean = 9.5 kg, SD = 2.1
  - Diet B: Mean = 7.0 kg, SD = 1.5
  - Diet C: Mean = 5.5 kg, SD = 1.8

  Using Tukey's HSD, the researcher obtains the following results:
  - Comparison between Diet A and Diet B: $p = 0.02$ (significant)

  - Comparison between Diet A and Diet C: $p = 0.001$ (significant)
  - Comparison between Diet B and Diet C: $p = 0.15$ (not significant)

- Interpreting Results Based on the results of Tukey's HSD, we can conclude:

  - There is a significant difference in weight loss between Diet A and Diet B.
  - There is a significant difference in weight loss between Diet A and Diet C.
  - However, there is no significant difference in weight loss between Diet B and Diet C.

  These findings suggest that Diet A is more effective than both Diet B and Diet C in promoting weight loss.

- **Summary of Post-Hoc Tests** Post-hoc tests are crucial for understanding which specific groups differ after finding a significant ANOVA result. Tukey's HSD, Bonferroni correction, Scheffé's test, and Dunnett's test are commonly used methods, each suitable for different research scenarios. It is important to choose the appropriate post-hoc test based on the research design and hypotheses being tested.

4.7 Application

4.7.1 Solved Exercises

Exercice 4.1. ***(Independence of Observations)*** *A researcher studies the effect of three different teaching methods (A, B, and C) on student performance by measuring their test scores. If some students in method B share their answers, does this violate the independence of observations assumption?*

Exercice 4.2. ***(Normality Assumption)*** *A researcher collects the following data on the weights of animals in three different diets:*

- *Diet A: 5.1, 5.3, 5.2, 5.5, 5.4*
- *Diet B: 6.1, 6.3, 6.5, 6.7, 6.8*
- *Diet C: 7.0, 7.1, 7.2, 7.3, 7.4*

Perform a Shapiro-Wilk test to assess normality at a significance level of $\alpha = 0.05$.

Exercice 4.3. ***(Homogeneity of Variances)*** *A researcher performs Levene's test to check for homogeneity of variances across the following groups:*

- *Group 1: 3.1, 3.5, 3.2, 3.4*
- *Group 2: 2.1, 2.3, 2.4, 2.2*
- *Group 3: 4.0, 4.2, 4.3, 4.1*

Calculate the p-value for Levene's test and state whether the homogeneity of variances assumption holds.

Exercice 4.4. ***(Conducting ANOVA)*** *A one-way ANOVA is performed on the following test scores for three different teaching methods:*

- *Method A: 78, 82, 85, 90*
- *Method B: 88, 84, 82, 86*
- *Method C: 75, 80, 78, 74*

Perform the ANOVA and determine if there are significant differences in test scores among the methods. Use $\alpha = 0.05$.

Exercice 4.5. ***(Post-Hoc Tests)*** *After finding significant results in a one-way ANOVA comparing three diets for weight loss (as shown in Exercise 4), the researcher decides to conduct Tukey's HSD test. The mean weight loss for each diet is as follows:*

- *Diet A: Mean = 10 kg*

- *Diet B: Mean = 7 kg*
- *Diet C: Mean = 5 kg*

Perform Tukey's HSD test and state which diets are significantly different.

Exercice 4.6. ***(Total Variance Calculation)*** *The exam scores of students from two universities are:*

- ***University A****: 65, 70, 75, 80, 85*
- ***University B****: 60, 65, 70, 75, 80*

Calculate the total variance and interpret its components.

Exercice 4.7. ***(One-Way ANOVA on Crop Yields)*** *In three Algerian regions, crop yields (in tons per hectare) are measured as follows:*

- ***Region A****: 4.5, 4.7, 4.6, 4.8*
- ***Region B****: 5.1, 5.0, 5.2, 5.3*
- ***Region C****: 4.9, 4.8, 5.0, 4.7*

Test if there is a significant difference in mean crop yields across the regions using a significance level of 0.05.

Exercice 4.8. ***(Two-Way ANOVA with Interaction)*** *Researchers study the effects of fertilizer type (A, B) and irrigation level (low, high) on plant growth. The growth (in cm) is as follows:*

Fertilizer	***Low Irrigation***	***High Irrigation***
A	*10, 12, 11*	*15, 14, 16*
B	*8, 9, 10*	*12, 13, 11*

Perform a two-way ANOVA to assess the main and interaction effects.

Exercice 4.9. ***(Testing Homogeneity of Variance)*** *Test the homogeneity of variances for the following datasets using Levene's Test:*

- *Group 1: 50, 52, 49, 51*
- *Group 2: 60, 62, 61, 59*
- *Group 3: 70, 69, 71, 72*

Exercice 4.10. ***(ANOVA Assumptions)*** *Analyze the following dataset to check if the ANOVA assumptions are met:*

- *Group 1: 25, 30, 28, 35, 32*
- *Group 2: 40, 42, 45, 41, 44*

- *Group 3: 38, 37, 39, 40, 36*

Check for independence, normality, and homoscedasticity.

Exercice 4.11. ***(Post-Hoc Test)*** *Given the results of a one-way ANOVA showing significant differences, perform a Tukey HSD test for the following groups:*

- *Group A: 5, 7, 6*
- *Group B: 8, 9, 10*
- *Group C: 4, 3, 5*

Exercice 4.12. ***(Factorial ANOVA)*** *Examine the effect of two factors, teaching method (A, B) and school type (Public, Private), on student scores:*

Method	***Public School***	***Private School***
A	*75, 80, 78*	*85, 88, 86*
B	*70, 68, 72*	*78, 80, 79*

Conduct a factorial ANOVA and interpret the results.

Exercice 4.13. ***(Random Sampling and ANOVA)*** *Using random samples from Algerian provinces, test the hypothesis that the average unemployment rate (in %) differs across three provinces:*

- *Province 1: 10.2, 11.0, 10.5*
- *Province 2: 12.1, 12.3, 11.8*
- *Province 3: 9.5, 9.8, 9.7*

Exercice 4.14. ***(Repeated Measures ANOVA)*** *A company tests employee productivity under three conditions: Morning, Afternoon, and Evening shifts. The data (productivity scores) are:*

- *Morning: 80, 85, 83*
- *Afternoon: 75, 78, 76*
- *Evening: 70, 72, 73*

Perform a repeated-measures ANOVA.

Exercice 4.15. ***(ANOVA Table Interpretation)*** *Interpret the following ANOVA table:*

Source of Variation	***SS***	***df***	***MS***	***F***	***p-value***
Between Groups	*200*	*2*	*100*	*5.45*	*0.025*
Within Groups	*500*	*27*	*18.52*		

What conclusions can be drawn?

Solutions

Correction exercice 4.1. *The independence of observations assumption states that the data points are not influenced by one another. In this case, we have students in Group B sharing answers, which violates this assumption. To illustrate the violation:*

- *If Student 1 in Group B receives help from Student 2, their scores are no longer independent because they are influenced by each other's performance.*
- *This lack of independence can lead to underestimated variability within the group, which may inflate Type I error rates.*

To correct this, the researcher should ensure that students do not interact during the assessment. This can be achieved by:

- *Testing students in different locations to avoid collaboration.*
- *Using online platforms that monitor student activities and prevent sharing of answers.*
- *Randomly assigning students to groups to ensure that there is no bias in group assignments.*

Correction exercice 4.2. *To assess normality using the Shapiro-Wilk test for the weights of animals in each diet, we will follow these steps:*
Given data:

- *Diet A: 5.1, 5.3, 5.2, 5.5, 5.4*
- *Diet B: 6.1, 6.3, 6.5, 6.7, 6.8*
- *Diet C: 7.0, 7.1, 7.2, 7.3, 7.4*

The Shapiro-Wilk test statistic and p-value can be calculated using statistical software (e.g., R or Python) or tables. Assuming the calculations yield the following results:

- *Diet A:* $W = 0.950, p = 0.755$
- *Diet B:* $W = 0.930, p = 0.892$
- *Diet C:* $W = 0.970, p = 0.905$

Since all p-values are greater than $\alpha = 0.05$*, we fail to reject the null hypothesis for all diets. Thus, we conclude that the data from all diets are approximately normally distributed.*

Correction exercice 4.3. *To check the homogeneity of variances, we perform Levene's test on the following groups:*

- *Group 1: 3.1, 3.5, 3.2, 3.4*
- *Group 2: 2.1, 2.3, 2.4, 2.2*
- *Group 3: 4.0, 4.2, 4.3, 4.1*

First, calculate the mean for each group:

- *Mean of Group 1:*

$$\bar{x}_1 = \frac{3.1 + 3.5 + 3.2 + 3.4}{4} = 3.3$$

- *Mean of Group 2:*

$$\bar{x}_2 = \frac{2.1 + 2.3 + 2.4 + 2.2}{4} = 2.25$$

- *Mean of Group 3:*

$$\bar{x}_3 = \frac{4.0 + 4.2 + 4.3 + 4.1}{4} = 4.15$$

Next, compute the absolute deviations from the mean for each group, then calculate Levene's test statistic using the formula:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

Assuming the calculations yield a test statistic F and a p-value:

$$p = 0.065$$

Since $p > 0.05$, we fail to reject the null hypothesis. This indicates that the assumption of homogeneity of variances holds across the groups.

Correction exercice 4.4. *To perform a one-way ANOVA on the test scores for the three teaching methods, we first summarize the data:*

- *Method A: 78, 82, 85, 90*
- *Method B: 88, 84, 82, 86*
- *Method C: 75, 80, 78, 74*

First, calculate the group means:

- *Mean of Method A:*

$$\bar{x}_A = \frac{78 + 82 + 85 + 90}{4} = 83.75$$

- *Mean of Method B:*

$$\bar{x}_B = \frac{88 + 84 + 82 + 86}{4} = 85$$

- *Mean of Method C:*

$$\bar{x}_C = \frac{75 + 80 + 78 + 74}{4} = 76.75$$

Total mean:

$$\bar{x} = \frac{(78 + 82 + 85 + 90) + (88 + 84 + 82 + 86) + (75 + 80 + 78 + 74)}{12} = \frac{1000}{12} = 83.33$$

Next, calculate the sum of squares between groups ($SS_{between}$) and within groups (SS_{within}):

$$\begin{aligned} SS_{between} &= n(\bar{x}_A - \bar{x})^2 + n(\bar{x}_B - \bar{x})^2 + n(\bar{x}_C - \bar{x})^2 \\ &= 4\left((83.75 - 83.33)^2 + (85 - 83.33)^2 + (76.75 - 83.33)^2\right) \\ &= 4\left(0.176 + 2.779 + 43.1\right) \approx 4 \times 45.055 = 180.22 \end{aligned}$$

For within-group sums of squares, calculate as follows:

$$SS_{within} = \sum (x_{ij} - \bar{x}_i)^2$$

Assuming we compute F using:

$$F = \frac{MS_{between}}{MS_{within}}$$

Where $MS_{between} = \frac{SS_{between}}{k-1}$ and $MS_{within} = \frac{SS_{within}}{N-k}$. After performing calculations, assume we find:

$$p = 0.025$$

Since $p < 0.05$, we reject the null hypothesis and conclude that there are significant differences in test scores among the methods.

Correction exercice 4.5. *After finding significant results in the one-way ANOVA, we proceed to perform Tukey's HSD test for the means of the diets:*

- *Diet A: Mean = 10 kg*
- *Diet B: Mean = 7 kg*
- *Diet C: Mean = 5 kg*

First, calculate the critical value using Tukey's method, assuming $q = 3.5$ (based on $\alpha = 0.05$, number of groups, and total sample size).

Now compute the differences between the means:

- *Difference between Diet A and Diet B:*

$$|10 - 7| = 3$$

- *Difference between Diet A and Diet C:*

$$|10 - 5| = 5$$

- *Difference between Diet B and Diet C:*

$$|7 - 5| = 2$$

Now, compare the differences with the critical value calculated from Tukey's: Assuming the calculated critical value from Tukey's is 2.5:

- $3 > 2.5$*: Significant difference between Diet A and Diet B.*
- $5 > 2.5$*: Significant difference between Diet A and Diet C.*
- $2 < 2.5$*: No significant difference between Diet B and Diet C.*

In conclusion, Tukey's HSD test indicates that Diet A is significantly different from both Diet B and Diet C, but Diet B and Diet C are not significantly different from each other. This detailed approach highlights the importance of conducting post-hoc tests after finding significant results in ANOVA.

4.7.2 Unsolved exercices

Exercice 4.1. ***(Shapiro-Wilk Normality Test)*** *A study collects the following data on the time taken (in minutes) by students to complete a test under three different conditions:*

- *Condition A:* $25.1, 24.8, 25.3, 24.9, 25.0$
- *Condition B:* $30.2, 30.4, 30.3, 30.1, 30.5$
- *Condition C:* $28.0, 27.9, 28.1, 28.3, 28.2$

Use the Shapiro-Wilk test to assess the normality of the data in each group at $\alpha = 0.05$.

Exercice 4.2. ***(Bartlett's Test for Equal Variances)*** *Three teaching methods are compared using test scores:*

- *Method A:* $78, 82, 79, 81$
- *Method B:* $88, 86, 87, 85$
- *Method C:* $72, 74, 73, 71$

Perform Bartlett's test to assess the homogeneity of variances assumption. Use a significance level of $\alpha = 0.05$.

Exercice 4.3. ***(ANOVA with Unequal Sample Sizes)*** *A researcher measures the performance of machines using three different maintenance methods:*

- *Method A:* $90, 92, 88, 89$
- *Method B:* $85, 87, 84$
- *Method C:* $91, 93, 90, 89, 92$

Conduct a one-way ANOVA to determine if there are significant differences in performance across the methods. Use $\alpha = 0.05$.

Exercice 4.4. ***(Interaction Effect in Two-Way ANOVA)*** *The productivity of workers is measured under two lighting conditions (low, high) and two work shifts (morning, afternoon). The data is as follows:*

Lighting Condition	***Morning Shift***	***Afternoon Shift***
Low	$50, 48, 52$	$45, 46, 44$
High	$60, 62, 58$	$55, 56, 54$

Perform a two-way ANOVA and assess the interaction effect between lighting condition and work shift.

Exercice 4.5. ***(Tukey's HSD Post-Hoc Test)*** *The following average customer satisfaction scores were recorded for three store layouts:*

- *Layout A:* $85, 87, 89, 86, 88$
- *Layout B:* $78, 80, 79, 77, 76$
- *Layout C:* $90, 92, 91, 89, 88$

After conducting a one-way ANOVA and finding significant differences among the layouts, perform Tukey's HSD test to determine which pairs of layouts differ significantly.

Bibliography

[1] Montgomery, D. C., & Runger, G. C. (2020). *Applied Statistics and Probability for Engineers* (7^{e} éd.). Wiley.

[2] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

[3] Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2^{e} éd.). Duxbury Advanced Series.

[4] Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3^{e} éd.). McGraw-Hill.

[5] Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8^{e} éd.). Pearson.

[6] Rice, J. A. (2006). *Mathematical Statistics and Data Analysis* (3^{e} éd.). Cengage Learning.

[7] DeGroot, M. H., & Schervish, M. J. (2011). *Probability and Statistics* (4^{e} éd.). Pearson.

[8] Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4^{e} éd.). Guilford Press.