Theory of probability and mathematical statistics

Estimations, Confidence Intervals, and Estimation Methods

BRAIS DE LA IGLESIA MARTÍNEZ

INGENIERÍA INFORMÁTICA (UVIGO)

1. Introduction

1.1 Project Objective

The objective of this project is to explore the fundamental concepts and methods of estimation in mathematical statistics. We will delve into the different types of estimations, the construction of confidence intervals, and the significance tests, along with practical implementations in MATLAB.

1.2 Importance of Mathematical Statistics Theory

Mathematical statistics is essential for analyzing and interpreting data, making it crucial for fields like science, engineering, economics, and more. Understanding estimation theory helps in making informed decisions based on data analysis.

1.3 Document Structure

This document is organized into several sections. We begin with fundamental concepts and types of estimations, followed by detailed discussions on confidence intervals and significance tests. We then explore the methods of moments and maximum likelihood estimators, culminating in practical MATLAB implementations.

2. Fundamental Concepts of Estimation Theory

2.1 Definition of Estimation

Estimation in statistics refers to the process of inferring the value of a population parameter based on a sample from that population. This involves using sample data to calculate a statistic that serves as an estimate of the parameter.

2.2 Properties of a Good Estimator

A good estimator should have the following properties:

- **Bias:** An estimator is unbiased if the expected value of the estimate equals the true parameter value. Formally,

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

- **Variance:** The estimator should have low variance, meaning it should produce estimates that are close to each other.

  $$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2]$$

- **Consistency:** As the sample size increases, the estimator should converge to the true parameter value.

  $$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$$

- **Efficiency:** Among unbiased estimators, the one with the smallest variance is considered efficient.

3. Types of Estimations

3.1 Point Estimation

Point estimation involves providing a single value as an estimate of the population parameter. Common point estimators include the sample mean for estimating the population mean and the sample proportion for estimating the population proportion.

3.2 Interval Estimation

Interval estimation provides a range of values within which the parameter is expected to lie, with a certain level of confidence. This range is known as a confidence interval.

4. Confidence Intervals Method

4.1 Constructing Confidence Intervals

A confidence interval for a parameter is calculated using the sample statistic, the standard error, and a critical value from the relevant probability distribution (e.g., t-distribution for small samples).

$$\mathrm{CI} = \bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

4.2 Confidence Level and Estimation Error

The confidence level represents the proportion of times that the confidence interval would contain the true parameter if we were to repeat the sampling process an infinite number of times. It is denoted as $1-\alpha$ where α is the significance level. The confidence level is usually expressed as a percentage, such as 90%, 95%, or 99%.

Critical Values and Confidence Levels:

The critical value is a point on the distribution that corresponds to the desired confidence level. For a 95% confidence interval, the critical value for the normal distribution (assuming the population variance is known or the sample size is large) is typically denoted as $z\alpha/2$ For a 95% confidence level, $\alpha\ is 0.05,\ \alpha/2$ is 0.025. The critical value $z0.025$ is 1.96, meaning that 95% of the data lies within 1.96 standard deviations from the mean in a standard normal distribution.

The formula for the confidence interval for a population mean $\mu\backslash mu$μ when the population standard deviation $\sigma\backslash sigma$σ is known is:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Where:

- X^{-} is the sample mean,
- $z\alpha/2$ is the critical value from the standard normal distribution,
- σ is the population standard deviation,
- n is the sample size.

Estimation Error:

The estimation error, also known as the margin of error, is the range within which we expect the true parameter to lie. It is calculated as

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The margin of error decreases with larger sample sizes, providing more precise estimates.

For example, if you have a sample mean (X) of 50, a population standard deviation (σ) of 10, and a sample size (n) of 100, the 95% confidence interval would be calculated as follows:

1. Determine the critical value:

$$z_{0.025} = 1.96$$

2. Calculate the margin of error:

$$1.96 \times \frac{10}{\sqrt{100}} = 1.96 \times 1 = 1.96$$

3. Construct the confidence interval:

$$50 \pm 1.96, \text{ which gives } [48.04, 51.96]$$

Thus, we are 95% confident that the true population mean lies between 48.04 and 51.96.

In summary, the confidence level and the associated critical value are fundamental in constructing confidence intervals, which provide a range of values that likely contain the true population parameter. The margin of error quantifies the uncertainty associated with the estimate and is influenced by the sample size and the variability of the data.

4.3 Practical Examples with MATLAB Implementation

See attachment ex1.m

In this code, we calculate the 95% confidence interval for the sample mean. The function `tinv` is used to find the critical value from the t-distribution.

5. Significance Tests

5.1 Null and Alternative Hypotheses

In hypothesis testing, the null hypothesis $(H0)$ rep resents the default assumption, while the alternative hypothesis (Ha)represents the outcome we seek evidence for.

For example, $H0$: $\mu = 7$ and Ha : $\mu \neq 7$

5.2 General Procedure for Hypothesis Testing

The steps for hypothesis testing are:

1. State the null and alternative hypotheses.
2. Choose a significance level (α).
3. Calculate the test statistic.
4. Determine the p-value.
5. Compare the p-value to and make a decision.

5.3 Practical Examples with MATLAB Implementation

See attachment ex2.m

This example demonstrates a t-test for the population mean, where we test if the sample mean significantly differs from 7.

6. Types of Data

6.1 Qualitative and Quantitative Data

Qualitative Data

Definition: Qualitative data (or categorical data) describe characteristics or qualities that cannot be measured numerically. They are used to classify or categorize items.

Types of Qualitative Data:

Nominal: Categories without a specific order (e.g., gender, marital status).
Ordinal: Categories with a specific order but no significant distance between them (e.g., satisfaction levels: low, medium, high).

Example: In a survey on patient satisfaction at a hospital:

- Gender (male, female, other).
- Marital status (single, married, divorced, widowed).
- Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied).

MATLAB Implementation:

See attachment qualitative.m

Quantitative Data

Definition: Quantitative data are those that can be measured and expressed numerically. They represent quantities.

Types of Quantitative Data:

Discrete: Countable values, integers (e.g., number of children, number of hospital visits).
Continuous: Values in a continuous range (e.g., weight, height, blood glucose level).

Example: Measuring blood glucose levels in a sample of patients.

MATLAB Implementation:

See attachment quantitative.m

Example 1: Estimation of Mean Blood Glucose Level

Context: Analyzing blood glucose levels in a group of people. The aim is to estimate the mean blood glucose level using a sample of data from a medical database.

MATLAB Implementation:

See attachment glucose.m

7. Data Selection and Application in Databases

Sample Selection: Selecting representative samples is crucial in data analysis. Common methods include:

- **Simple Random Sampling**: Randomly selecting individuals.
- **Stratified Sampling**: Dividing the population into strata and selecting samples from each stratum.

MATLAB Implementation:

Simple Random Sampling:

See attachment randomSampling.m

Stratified Sampling:

See attachment stratifiedSampling.m

8. Estimation Methods

8.1 Method of Moments

Definition: The Method of Moments is a technique for estimating the parameters of a probability distribution by equating sample moments to theoretical moments.

Formulas:

First Moment (Mean): $\mu = E[X]$

Second Moment (Variance): $\sigma 2 = E[X2] - (E[X])2\sigma 2 = E[X2] - (E[X])2$

Steps:

Calculate the sample moments from the data.
Equate the sample moments to the theoretical moments to solve for the parameters.

Example:

To estimate the parameter θ of a distribution, we set the sample mean equal to the population mean:

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

8.2 Maximum Likelihood Method

Definition: Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a probability distribution by maximizing the likelihood function.

Formula: The likelihood function L(θ) is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(\,Xi \mid \theta\,)$$

where θ is the parameter vector and $f(xi;\theta)$ is the probability density function.

Steps:

Write down the likelihood function for the given data and distribution.
Take the logarithm of the likelihood function to obtain the log-likelihood function.
Differentiate the log-likelihood function with respect to the parameters.
Set the derivatives equal to zero and solve for the parameters.

9. MATLAB Implementation

9.1 Estimation using the Method of Moments

See attachment ex3.m

9.2 Estimation using the Maximum Likelihood Method

See attachment ex4.m

These codes show how to implement the method of moments and maximum likelihood estimation in MATLAB, providing practical insights into these estimation methods.

10. Conclusions

10.1 Summary of Key Concepts

We have explored the fundamental concepts of estimation theory, different types of estimations, and the methods to construct confidence intervals and perform significance tests. Additionally, we discussed two primary estimation methods: the method of moments and the maximum likelihood method, with practical MATLAB implementations.

10.2 Importance of Estimation Techniques in Statistics

Estimation techniques are vital for making inferences about population parameters based on sample data. These techniques are widely used in various fields to inform decision-making processes.

10.3 Final Reflections on MATLAB Implementation

Implementing estimation methods in MATLAB provides a practical approach to understanding and applying statistical concepts, reinforcing theoretical knowledge with hands-on experience.

11. References

- Casella, G., & Berger, R. L. (2002). Statistical Inference (2nd ed.). Duxbury.
- Rice, J. A. (2006). Mathematical Statistics and Data Analysis (3rd ed.). Duxbury.
- MATLAB Documentation. MathWorks.