Introduction to Data Science — Assignment 1

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TA(s): Omid Panakari

Deadline: Tuesday, Esfand 29th, 11:59 PM

Introduction

In this assignment, we are going to get acquainted with and implement some tools for statistical analysis. These tools could come to your help in your future research and projects.

Task

Monte Carlo Simulation

In this section, we will learn more about Monte Carlo Simulation and understand how it can help us to model and analyze complex systems, where analytical or closed-form solutions are difficult.

Pi Calculation

First, you are going to estimate the value of Pi using Monte Carlo simulation. The idea is to randomly generate points within a square and determine how many fall in a circle inscribed by the square. By comparing the points inside the circle to the total number of points you can approximate the Pi number. Repeat your simulation with a different number of points and analyze their results.

Mensch Game

Mensch is a very old German game, which is also popular in our country. You can learn more about Mensch and its rules in this link.
We are going to analyze the simpler version of this game in which every player only has one piece in the game. So basically, every player only rolls dice in his turns and moves his piece. So, everything is purely based on chance. We want to calculate the probability of winning for each of the 1st, 2nd, 3rd, and 4th players in this game. Perform the Monte Carlo Simulation over the specified game to calculate these probabilities.

Central Limit Theorem(CLT)

The objective of this section is to provide you with a hands-on opportunity to observe and understand the Central Limit Theorem in action. The CLT is a fundamental result that

supports many statistical techniques and methods. It provides a theoretical basis for making inferences about population parameters based on sample statistics.
First of all, select three different probability distributions. These distributions will serve as the population distributions from which you will take out samples.
Now for each of the distributions, perform the following steps:

1. Generate a large number of random samples with a specific sample size from the chosen distribution.
2. Calculate the mean of each sample.
3. Plot the histogram of the sample means and overlay it with the expected normal distribution based on the Central Limit Theorem.
4. Repeat steps a to c for increasing sample sizes and observe how the distribution of sample means changes as we increase the sample size.

Document your observations and insights from each experiment. Compare the distribution of sample means for each sample size and discuss how they align with the principles of the Central Limit Theorem.

Hypothesis Testing

Hypothesis testing is an essential tool in statistics and scientific research that allows us to make informed decisions and come to a conclusion about population parameters based on sample data. In this section, you will understand how hypothesis testing can help you to analyze data and make a decision based on them in different situations.

Unfair Coin

First, you need to simulate an unfair coin that is biased toward landing on one face more often than the other one (about 10% more probable). Your task will be to perform a hypothesis testing on this coin to determine whether it is fair or not. Use both confidence interval and p-value approach for your test.
Conduct these tests using different sample sizes: 30, 100, 1000. Report and analyze the result of the hypothesis testing for different sample sizes. Include the calculated z-scores, p-values, the decision made regarding the null hypothesis, and the justification for the decision. Also, discuss the impact of sample size on the hypothesis testing result.

T-Test

A t-test is a statistical test used to determine if there is a significant difference between the means of two groups or samples. In this test, we first calculate the **t-statistic** for the two groups or samples and then calculate the probability of having this t-statistic using the **t-distribution**. In the end, we only need to compare the p-value with our significance level(α) and make our decision. You can learn more about t-test, t-statistics, degrees of freedom, and student's t-distribution in this link.

Job Placement

There is a common belief that working alongside studying has a negative impact on their grades. You are given a job placement dataset that contains information about students studying in USA various universities alongside their job status. Perform a hypothesis test to test whether this belief is true or not. In your test, suppose that variance is unknown but equal for both of these groups.

1. Split students into two groups based on their job placement status.
2. Calculate the t-statistic and degrees of freedom for these two groups. Do not use any library for implementing this part. Write the formula for each of these values in your report.
3. Now, determine the p-value for the calculated t-statistic and degrees of freedom with the help of t-distribution. You can use the SciPy library for this purpose.
4. Finally, report the result of your test and make your decision.
5. Repeat this test one more time. This time, only use the SciPy library for performing the test. Compare the results of the tests with each other.

Questions

1. Read a little bit about the applications of Monte Carlo Simulation in real life. What are some of these applications?
2. How does the sample size affect your plots in part 2(CLT)? What can you understand from these plots?
3. How does increasing the sample size affect your coin test?
4. What are t-statistic, degrees of freedom, and t-distribution in t-test? How can they help us to compare two data sets?
5. What are the preliminary conditions for using t-test on our data?
6. Read about some other types of tests that are used in scientific research. Write a line about each of them.

Notes

- Upload your work as a zip file in this format on the website: DS_CA1_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.