Introduction to Data Science
CA1-Report
1403/01/05

Amirreza Akbari |810899045
Mohammad Yaser Azad |810800003
Reza Baghestani |810899046
Hannah Jamali |810899053

Monte Carlo Simulation

Pi calculation

The simulation was performed for 1000, 10000, 100000, and 1000000 iterations, and to improve the results and prevent errors, 5 simulations were conducted for each number. The average of these simulations was reported as the result. It seems that as the number of simulations increases, the resulting value gets closer to pi.

Mensch

We simulated the game according to the rules mentioned before, running 10,000 simulation showed that the chances of winning for all players are equal and approximately 25%, which seems reasonable.

Question 1) Application of Monte Carlo simulation int real life

1. **Finance**: It is widely used in financial modeling to assess the risk and uncertainty associated with investment portfolios, pricing options, and analyzing market trends.

2. **Engineering**: In engineering, It is used to analyze the reliability and performance of complex systems, such as bridges, buildings, and aircraft, by considering various input parameters and their uncertainties.

3. **Healthcare**: It is applied in healthcare for predicting patient outcomes, optimizing treatment plans, and evaluating the effectiveness of medical interventions.

4. **Energy**: In the energy sector, it is used to assess risks in energy trading, forecast energy demand, and optimize resource allocation in power generation.

5. **Project Management**: It helps project managers in estimating project timelines, costs, and resource allocation by considering uncertainties in project variables.

6. **Insurance**: Insurance companies use it to model and analyze potential risks related to insurance policies, claims, and pricing strategies.

7. **Supply Chain Management**: It is utilized in supply chain management to optimize inventory levels, forecast demand, and evaluate the impact of disruptions on the supply chain.

8. **Environmental Science**: It is used in environmental modeling to assess pollution levels, predict climate change scenarios, and analyze the impact of environmental policies.

Introduction to Data Science
CA1-Report
1403/01/05

Amirreza Akbari |810899045
Mohammad Yaser Azad |810800003
Reza Baghestani |810899046
Hannah Jamali |810899053

Central Limit Theory

Explain code:

1. **Poisson Distribution**: Generate 1000 random samples with a mean rate of 5. Calculate sample means and plot a histogram.
2. **Binomial Distribution**: Generate samples with 20 trials and a success probability of 0.3. Calculate sample means and plot a histogram.
3. **Uniform Distribution**: Generate uniform samples between 0 and 10. Calculate sample means and plot a histogram.
4. **Visualization**:Create a figure with three subplots (one for each distribution). Observe how sample means converge to a normal distribution with increasing sample size.
5. **Printed Sample Means**: Display the means for each distribution.

Question 2) How does the sample size affect your plots

1. **Poisson Distribution**:

As the sample size increases, the histogram of sample means becomes more bell-shaped (resembling a normal distribution). Larger sample sizes lead to better alignment with the expected normal distribution. This aligns with the CLT, which states that the distribution of sample means approaches normality regardless of the original distribution.

2. **Binomial Distribution**:

Similar to the Poisson case, increasing the sample size results in a more normal-looking histogram of sample means. The binomial distribution, which may be skewed for small samples, becomes more symmetric as the sample size grows. Again, this behavior is consistent with the CLT.

3. **Uniform Distribution**:

Initially, the uniform distribution's sample means exhibit a flat histogram. However, as the sample size increases, the histogram becomes more bell-shaped. The uniform distribution converges toward normality due to the CLT.

Overall Insights:

Regardless of the original distribution (Poisson, Binomial, or Uniform), larger sample sizes lead to more normally distributed sample means. The CLT ensures that even if the population distribution is not normal, the distribution of sample means tends toward normality with sufficient sample size. Practically, this allows us to make statistical inferences about population parameters using sample statistics. Remember that the CLT assumes random sampling and sufficiently large sample sizes for accurate approximations.

Introduction to Data Science
CA1-Report
1403/01/05

Amirreza Akbari |810899045
Mohammad Yaser Azad |810800003
Reza Baghestani |810899046
Hannah Jamali |810899053

Hypothesis Testing

Unfair Coin

- I've used the `proportions_ztest` function from `statsmodels` to calculate the z-score and p-value for hypothesis testing.
- I've also used the `proportion_confint` function to calculate the confidence interval.

Question 3) How does increasing the sample size affect your coin test?

Increasing the sample size is important for two reasons. One is the concept of probability, which is more accurately represented with a larger sample size (as shown by simulation results). The other is the standard error, as the denominator of the standard error formula is the square root of the sample size, and the standard error itself is in the denominator of the z-score. With an increase in sample size, the z-value increases, leading to a decrease in the p-value. Which means it's less likely to consider null hypothesis true.

Question 4) t-statistic

In a t-test, the t-statistic is a measure of the difference between the means of two data sets relative to the variability within the data sets. It is calculated as the difference between the sample means divided by the standard error of the difference. The t-statistic follows a t-distribution.

Degrees of freedom (df) in a t-test represent the number of independent observations in a sample. In a two-sample t-test, the degrees of freedom are calculated as the sum of the sample sizes minus 2.

The t-distribution is a probability distribution that is symmetric and bell-shaped, similar to the normal distribution. The shape of the t-distribution is determined by the degrees of freedom. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution.

These components help us compare two data sets by providing a statistical framework to assess whether the difference between the means of the two data sets is statistically significant. By calculating the t-statistic and comparing it to the critical values from the t-distribution based on the degrees of freedom, we can determine if the difference between the means is likely due to a real difference or just due to random variation.

Question 5) Preliminary conditions for using t-test

1. **Normality**: The data should be approximately normally distributed. This assumption is important for the validity of the t-test results. Normality can be assessed through visual inspection of histograms or using statistical tests such as the Shapiro-Wilk test.

2. **Homogeneity of Variance**: The variances of the two data sets being compared should be approximately equal. This assumption is known as homogeneity of variance or homoscedasticity. It can be checked using statistical tests such as Levene's test or by inspecting plots of the data.

Amirreza Akbari |810899045
Mohammad Yaser Azad |810800003
Reza Baghestani |810899046
Hannah Jamali |810899053

3. **Independence**: The observations within each data set should be independent of each other. This means that the values in one group should not be influenced by the values in another group.

4. **Random Sampling**: The data should be collected through a random sampling process to ensure that the sample is representative of the population.

5. **Continuous Data**: The t-test is suitable for continuous data. If the data is categorical, other statistical tests such as the chi-square test should be used.

Job Placement

We analyzed the relationship between job placement status and GPA using a two-sample t-test in Python by using these formula.

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{S_p(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

This operation has been performed both manually and with a library, and very similar results have been obtained. The t-statistic and p-value are equal in both cases, and both indicate that working alongside studying has a significant impact on grades.

Question 6) Other types of tests that are used in scientific research

1. **Analysis of Variance (ANOVA)**: ANOVA is a statistical test used to compare the means of three or more groups to determine if there are statistically significant differences between them. It helps to understand if there is a significant variation in the means of multiple groups.

2. **Chi-Square Test**: The Chi-Square test is used to assess the association between categorical variables by comparing the observed frequencies to the expected frequencies. It is commonly used to test for independence between two categorical variables.

3. **Pearson's Correlation Coefficient**: Pearson's correlation coefficient is a measure of the strength and direction of a linear relationship between two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

4. **Wilcoxon Signed-Rank Test**: The Wilcoxon Signed-Rank test is a non-parametric test used to compare two related samples or repeated measures data when the assumptions of a paired t-test are not met.

5. **Mann-Whitney U Test**:The Mann-Whitney U test is a non-parametric test used to compare the medians of two independent samples when the data do not follow a normal distribution.

Amirreza Akbari |810899045
Mohammad Yaser Azad |810800003
Reza Baghestani |810899046
Hannah Jamali |810899053

6. **Kruskal-Wallis Test**: The Kruskal-Wallis test is a non-parametric test used to compare the medians of three or more independent samples to determine if there are significant differences between them.