IE6400 Foundations Data Analytics Engineering

Fall Semester 2024

Quiz 6: Lecture 8: Data and Sampling Distributions

Objective: To understand the concept of data and sampling distributions and apply Python programming skills to analyze and visualize these distributions.

Instructions: Use Python to solve the following problems. Ensure that you include necessary comments in your code for clarity. Submit your Python code along with the outputs.

Questions:

1. **Central Limit Theorem (CLT) Simulation:**
   - Generate a population of 10,000 random numbers between 1 and 100.
   - Randomly sample 30 numbers from this population and calculate the sample mean. Repeat this process 1,000 times.
   - Plot the distribution of the 1,000 sample means. What do you observe?
2. **Sampling from a Non-Normal Distribution:**
   - Generate a population of 10,000 numbers from an exponential distribution with a scale parameter of 2.
   - Randomly sample 50 numbers from this population and calculate the sample mean. Repeat this process 500 times.
   - Plot the distribution of the 500 sample means. How does the CLT apply here?
3. **Confidence Intervals:**
   - Calculate the 95% confidence interval for the sample means from the previous question. What does this interval tell you about the population mean?

4. **Sample Size and Sampling Distribution:**
    - Using the population from question 1, randomly sample 10 numbers and calculate the sample mean. Repeat this process 1,000 times.
    - Plot the distribution of the 1,000 sample means for the sample size of 10 and compare it with the distribution from question 1 (sample size of 30).
    - What do you observe about the spread of the distributions as the sample size changes?

Hints:

- Use Python libraries like numpy for numerical operations and matplotlib for plotting.
- For generating random numbers, you can use numpy.random.rand() or numpy.random.exponential().
- To calculate confidence intervals, you can use the formula: mean ± (1.96 x standard error) where standard error = standard deviation / sqrt(sample size).

IE6400 Foundations Data Analytics Engineering

Fall Semester 2024

Quiz 6: Lecture 9:

Statistical Experiments and Significance Testing

Objective: To understand the concepts of statistical experiments and significance testing and to apply Python programming skills to solve real-world statistical problems.

Problem 1: T-test

Objective: To determine if there is a significant difference between the means of two independent samples.

Problem Statement: You have been given the scores of students from two different classes for a mathematics test. Determine if there is a significant difference in the mean scores of the two classes using a t-test.

```
class_A_scores = [85, 90, 78, 92, 88, 76, 95, 87, 79, 91]
class_B_scores = [80, 82, 88, 85, 83, 87, 84, 86, 89, 81]
```

Hint: Use the scipy.stats.ttest_ind function.

Problem 2: Chi-Squared Test

Objective: To determine if there is a significant association between two categorical variables.

Problem Statement: You have been given the observed frequencies of students preferring different types of beverages in two different grades. Using a chi-squared test, determine if there is a significant association between grade level and beverage preference.

```
# Observed frequencies
# Columns: ['Tea', 'Coffee', 'Juice']
# Rows: ['Grade 10', 'Grade 11']
observed_frequencies = [[30, 10, 10], [10, 30, 10]]
```

Hint: Use the scipy.stats.chi2_contingency function.

Problem 3: One-way ANOVA

Objective: To determine if there are any statistically significant differences between the means of three or more independent groups.

Problem Statement: You have been given the scores of students from three different teaching methods for a science test. Determine if there is a significant difference in the mean scores of the students taught by the various techniques using one-way ANOVA.

```
method_1_scores = [85, 87, 88, 86, 84, 85, 87]
method_2_scores = [80, 82, 81, 83, 82, 80, 81]
method_3_scores = [90, 91, 92, 90, 91, 92, 93]
```

Hint: Use the scipy.stats.f_oneway function.

Problem 4: One-sample Z-Test

Objective: To determine if there's a significant difference between the sample mean and the population mean.

Problem Statement: You have been given the scores of a sample of students from a class for an English test. The population mean score for this test is known to be 75, with a standard deviation of 10. Determine if there's a significant difference between the sample mean and the population mean using a one-sample Z-test.

```
sample_scores = [78, 76, 74, 75, 77, 76, 78, 74, 79, 75]
population_mean = 75
population_std = 10
```

Hint: You can use the formula for the Z-test or use a Python library that provides this functionality.

Problem 5: Two-sample Z-Test

Objective: To determine if there's a significant difference between the means of two independent samples when population variances are known.

Problem Statement: You have been given the scores of students from two different classes for a history test. The population variances for these two classes are known. Using a two-sample Z-test, determine if there's a significant difference between the mean scores of the two classes.

```
class_X_scores = [85, 87, 88, 86, 84, 85, 87]
class_Y_scores = [80, 82, 81, 83, 82, 80, 81]
population_variance_X = 15
population_variance_Y = 20
```

Hint: You can use the formula for the two-sample Z-test or use a Python library that provides this functionality.

Problem 6:

You are provided with a dataset containing the scores of two groups of students: Group A and Group B. These students were subjected to two different teaching methods, and their scores reflect their understanding of the material taught.

Your task is to determine if there is a significant difference in the scores between the two groups using appropriate statistical tests. Additionally, visualize the data to gain insights and interpret your findings.

Dataset:

Student ID	Group	Score
1	A	85
2	A	87
3	A	82
4	A	90
5	A	88
6	B	78
7	B	80
8	B	79
9	B	83
10	B	81

Tasks:

1. Visualize the distribution of scores for both Group A and Group B. What observations can you make from the visualization?
2. Calculate the mean and standard deviation of scores for both groups. How do they compare?
3. Perform a t-test to determine if there is a significant difference in the scores between Group A and Group B. What is the p-value, and what does it indicate?
4. If the p-value is less than 0.05, what conclusion can you draw about the teaching methods?
5. Based on your analysis, which teaching methods are more effective, and why?

Problem 7:

A pharmaceutical company has developed a new drug intended to increase sleep duration. To test its effectiveness, it conducted an experiment with two groups: a control group that did not receive the drug and a treatment group that did. Both groups were asked to record their sleep duration in hours for a week.

Using appropriate statistical tests, you will determine if the drug significantly affects sleep duration. Additionally, you will visualize the data to gain insights and interpret your findings.

Dataset:

Day	Control Group (Hours)	Treatment Group (Hours)
1	7.5	8.2
2	7.8	8.4
3	7.6	8.1
4	7.7	8.3
5	7.9	8.5
6	7.8	8.2
7	7.6	8.4

Tasks:

1. Visualize the average sleep duration for the control and treatment groups over the week. What observations can you make from the visualization?
2. Calculate the overall mean sleep duration for both groups. How do they compare?
3. Perform a paired t-test to determine if there is a significant difference in the sleep duration between the control and treatment groups over the week. What is the p-value, and what does it indicate?
4. If the p-value is less than 0.05, what conclusion can you draw about the drug's effectiveness?
5. Based on your analysis, do you believe the drug is effective in increasing the duration of sleep? Provide reasons for your answer.