

**Statistical Analysis for Business Decision Making**

Course-End Project Problem Statement





**Course-End Project: Employee Satisfaction Survey**

**Problem Scenario:** An auto parts manufacturing company, Xio Ltd., conducts regular training and development programs for its 1,000 employees. As a result, it wants to know how satisfied the employees are with the training program since satisfaction leads to tangible results, which boosts the profitability and success of the organization.  As a result, the company conducted the survey and gathered feedback from its 1,000 employees' pre- and post-training programs.

**Problem Objective:** As a chief data scientist in a consultant firm hired by Xio Ltd., you are asked to examine how training impacts employee satisfaction using the R tool and a given dataset. To analyze the dataset, you need to compute and analyze descriptive measures, as well as perform statistical interference analyses through confidence interval estimation and hypothesis testing.

**Dataset**: Data.xlsx

**Dataset Description**:

**Important:** Since individuals are biased towards the upside or downside when they give their responses, and therefore the **pre and post variables** are scaled to eliminate bias. This bias correction converts the original data (responses in the integer form) into fractional form.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Employee\_id | Unique identifier of each employee |
| Pre | Pre represents survey responses on a scale of 1-8 before the training and development program. |
| Post | Post represents survey responses on a scale of 1-8 after the training and development program. |
| Cold-Drink | Cold-drink preferences of the employees |
| Status | It represents if the employee is a member (less than 5 years of experience) or an observer (more than and equal to 5 years of experience). |
| Rating | Rating is given by the management to an employee’s overall performance in training on a scale of AAA to B. |
| Outlook | It is a categorical level as stable, negative, or positive given by the bank as per their account activity. |
| Salary | It represents the month-end salary account credit balances of the employees. |

**Note:** Download **Data.xlsx**dataset from here

**Steps to perform:**

1. Import the data to check its class and structure and display the head and tail of the data
2. Calculate the:
3. Difference in the means of the pre and post variables
4. Values that divide the pre and post variable data into equal halves
5. Mode for the pre variable
6. First and third quantile for the pre and post variables
7. Range of the pre and post variables
8. Variance and standard deviation for the pre and post variables
9. Coefficient of variation and mean absolute deviation for the pre and post variables
10. Interquartile range of the pre and post variables
11. Measure the skewness for pre and post variables and apply the Agostino test to check the skewness
12. Identify the nature of distribution through kurtosis for both pre and post variables and confirm the result through the Anscombe test
13. Plot a graph to check the skewness and peakedness in the distribution of pre and post variables
14. Compute the frequency and relative frequency for each brand of cold drink
15. Create a pie chart and bar chart to show the preferences of the cold drinks available and provide the necessary labels
16. Plot a density graph on the cold-drink frequency and comment on the skewness and kurtosis
17. Convert the ‘Status’, ‘Rating’, and ‘Outlook’ variables into factor types and summarize them
18. Calculate the difference in the average pre-training satisfaction ratings of member and observer status and for the post-training member and observer status
19. Compute the average pre-satisfaction and post-satisfaction ratings of employees with a ‘Stable’ Outlook
20. Construct a confidence interval at a 2.5%, 5%, and 1% level of significance for the salary variable
21. Construct a 99%, 95%, and 90% confidence interval estimate for the pre and post variables
22. Considering the Data.xlsx as a population:
23. Take a sample of 50 observations from the pre and post dataset (without replacement)
24. Construct a null hypothesis to examine whether the sample (50 observations) mean score of pre and post variables is significantly different from the population (1000 observations)
25. Compute corresponding Z values for pre and post variables in the sample
26. Using the p-value method, determine whether the sample mean for the pre and post variables differs significantly from the population mean at the 10% significance level
27. Calculate the critical Z value for the 10% level of significance and the decision rule using the critical value approach
28. Compute the T-statistics value for the pre and post variables
29. Calculate the p-value and the decision using the p-value approach for pre and post variables at a 10% level of significance
30. Calculate the critical T value for the level of significance of 10% and the decision rule using the critical value approach