Module 2

Data Analytics with Python - Statistics

Section: Advance statistics

Lab: Statistical Tests

Topic: Par Inc. Golf balls new scheme

Lab brief

Par Inc. is a major manufacturer of golf equipment. Management believes that Par's market share could be increased with the introduction of a cut-resistant, longer-lasting golf ball. Therefore, the research group at Par has been investigating a new golf ball coating designed to resist cuts and provide a more durable ball. The tests with the coating have been promising. One of the researchers voiced concern about the effect of the new coating on driving distances. Par would like the new cut-resistant ball to offer driving distances comparable to those of the current-model golf ball. To compare the driving distances for the two balls, 40 balls of both the new and current models were subjected to distance tests. The testing was performed with a mechanical hitting machine so that any difference between the mean distances for the two models could be attributed to a difference in the design.

Objective

- Use hypothesis testing technique to analyze effect of action.
- Understand how to use hypothesis testing techniques to solve business problem.

Problem Statement

Check whether there is any effect on driving distances due to the new coating on golf balls?

Dataset - Golf.csv

https://github.com/bluedataconsulting/AIMasteryProgram/blob/main/Projects/Module-2/Golf.csv

https://sourceforge.net/p/aimasteryprogram/code/ci/main/tree/Projects/Module-2/Golf.csv?format=raw

The dataset has 40 records in total and 2 columns.

The columns in datasets are-

- **Current:** contains driving distances of golf balls without coating.
- **New:** contains driving distances of golf balls with new coating.

Hint: (Use t-test or z-test to do so.)

Task:

1. Create Histograms and Box Plots for both columns in the dataset.
2. Formulate a Hypothesis to achieve the objective - Check whether there is any effect on driving distances due to the new coating on golf balls?
3. Test the Hypothesis using statistical tests.

Topic: IBM Attrition Dependency

Lab brief

Attrition is a problem that impacts all businesses, irrespective of geography, industry, and size of the company. It is a major problem to an organization, and predicting turnover is at the forefront of the needs of Human Resources (HR) in many organizations. Organizations face huge costs resulting from employee turnover. You are given with an Attrition Dataset IBM HR Analytics Employee Attrition & Performance and are supposed to answer if attrition is related with education and job satisfaction.

Objective

- Understand use of hypothesis testing technique to perform root cause analysis
- Using hypothesis testing techniques for suitable problems related to employee attrition problem.
- Implement hypothesis testing using python.

Problem Statement

Dataset – IBM Attrition.csv

https://github.com/bluedataconsulting/AIMasteryProgram/blob/main/Projects/Module-2/IBM%20Attrition.csv

https://sourceforge.net/p/aimasteryprogram/code/ci/main/tree/Projects/Module-2/IBM%20Attrition.csv?format=raw

The dataset consists of 1470 records and 35 columns. The columns in datasets are as follows-

- **Age -** Age of the employee.
- **Attrition -** Did Attrition occurred? (Yes, No) – whether employee left or not
- **Business Travel -** ['Travel_Rarely', 'Travel_Frequently', 'Non-Travel']
- **DailyRate –** Daily Rate.
- **Department -** ['Sales', 'Research & Development', 'Human Resources']
- **DistanceFromHome -** Distance from home.

- **EducationField -** ['Life Sciences', 'Other', 'Medical', 'Marketing', 'Technical Degree', 'Human Resources']
- **EmployeeCount -** A static field
- **EmployeeNumber –** Unique Employee Number.
- **Gender -** ['Female', 'Male']
- **HourlyRate -** Hourly Rate.
- **JobLevel -** [1,2,3,4,5]
- **JobRole -** ['Sales Executive', 'Research Scientist', 'Laboratory Technician', 'Manufacturing Director', 'Healthcare Representative', 'Manager', 'Sales Representative', 'Research Director', 'Human Resources']
- **MaritalStatus -** ['Single', 'Married', 'Divorced']
- **MonthlyIncome -** Monthly Income of an employee.
- **MonthlyRate -** Monthly Rate.
- **NumCompaniesWorked -** Number of Companies worked at.
- **Over18 -** Is the employee over 18? ['Y']
- **OverTime –** Did the employee do overtime? ['Yes', 'No']
- **PercentSalaryHike -** Percentage Hike in Salary.
- **StandardHours -** Standard Hours [80]
- **StockOptionLevel -** Stock Option Level [0,1,2,3]
- **TotalWorkingYears -** Total working years.

- **TrainingTimesLastYear –** Training times last year.
- **YearsAtCompany -** Years at the company.
- **YearsInCurrentRole -** Years spent in current role.
- **YearsSinceLastPromotion -** Years since last promotion.
- **YearsWithCurrManager -** Years worked with current manager.

The columns below are interpreted as follows:

- **Education**
  1 'Below College'
  2 'College'
  3 'Bachelor'
  4 'Master'
  5 'Doctor'

- **EnvironmentSatisfaction**
  1 'Low'
  2 'Medium'
  3 'High'
  4 'Very High'

- **JobInvolvement**
  1 'Low'
  2 'Medium'
  3 'High'
  4 'Very High'

- **JobSatisfaction**
  1 'Low'
  2 'Medium'
  3 'High'
  4 'Very High'

- **PerformanceRating**
  1 'Low'
  2 'Good'

3 'Excellent'
4 'Outstanding'

- **RelationshipSatisfaction**
  1 'Low'
  2 'Medium'
  3 'High'
  4 'Very High'

- **WorkLifeBalance**
  1 'Bad'
  2 'Good'
  3 'Better'
  4 'Best'

Task:

- Perform descriptive statistics on the dataset.
- Formulate a Hypothesis to achieve below objective-
  - Whether age of employees has information related to their decision to leave the company or not
  - whether job satisfaction and education of the employees relates to their decision to stay or leave