JSS MAHAVIDYAPEETHA

JSS SCIENCE AND TECHNOLOGY UNIVERSITY, MYSURU

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

VI Semester **20IS610**

USER MANUAL FOR STATISTICAL METHODS IN INFORMATION PROCESSING

20IS67L STATISTICAL METHODS LABORATORY

Course Title: Statistical Methods Laboratory	***Course Code: 20IS67L***
Credits: 1.5	***Total Contact Hours (L:T:P): 0:0:39***
Type of Course: Laboratory	***Category: Professional Core Course***
CIE Marks: 50	***SEE Marks: 50***

Prerequisite: R / Python Programming.

Course Objective: To implement and analyze data using various statistical methods.

Course Outcomes: After completing this course, students should be able to:

CO1	Perform Correlation and Regression Analysis using statistical tools.
CO2	Analyze data using different statistical methods.
CO3	Analyze results and draw inferences.

SL No.	Program List	No. of Hours
	PART A	
1	Extract a dataset of your choice and compute the correlation between any two variables and visualize the relationship using scatter plot using JMP Pro tool. Interpret the results.	3
2	Apply the Pearson correlation test on a dataset, show the normality of variables using Q-Q plot and interpret the results using JMP Pro Tool. Interpret the results.	3
3	Select any dataset from JMP Pro tool and perform ANOVA test and Non-Parametric tests (The Mann Whitney test and The Kruskal-Wallis test). Interpret the results and draw inferences.	3
	PART B	
4	You have been provided with a dataset containing information about the heights of students in a college. Perform univariate analysis on this dataset using Python and calculate the mean, median, mode, and standard deviation of the heights. Also, create a histogram to visualize the distribution of heights. Write a Python program to implement this analysis. heights = [165, 170, 168, 172, 175, 169, 180, 160, 165, 172, 168, 176, 170, 173, 165]	4
5	You are given a dataset containing the prices of houses in a neighborhood. Perform univariate analysis on this dataset using Python and calculate the mean, median, mode, and standard deviation of the house prices. Additionally, plot a	4

	box plot to visualize the distribution of prices. Write a Python program to implement this analysis. house_prices = [300000, 350000, 320000, 280000, 400000, 380000, 330000, 310000, 290000, 270000, 350000, 380000, 370000]	
6	Perform univariate analysis on a dataset containing information about the performance of students in a school. The dataset includes variables such as student ID, test scores in different subjects, attendance, and socio-economic background. Use Python to analyze the distribution of test scores in each subject separately. Calculate the mean, median, and standard deviation for each subject and visualize the distribution using box plots.	5
7	Consider the scores of ten students in SMIP and DBMS and Compute the Spearman rank correlation and Interpret the results using Python/R programming.	4

SMIP	70	46	94	34	20	86	18	12	56	64	42
DBMS	60	66	90	46	16	98	24	08	32	54	62

8	Analyze a dataset containing information about the sales revenue and advertising expenditure of a company over a period of time. Calculate the Karl Pearson correlation coefficient between sales revenue and advertising expenditure using Python. Interpret the correlation coefficient and discuss the strength and direction of the relationship between advertising and sales.							5
9	Develop a Python/R code to build a simple Linear Regression model to predict sales units based on the advertising budget spent on TV. Display the statistical summary of the model.							4
	Sales	2	4	6	9	12	34	45
	TV	1	2	4	7	9	11	15
10	Consider Australian Drug Sales dataset and develop a Python/R code to perform Time Series Analysis and visualize using plots.							4

Evaluation for 50 Marks

One program from Part A (20 marks):

Write-up (Steps)	05 Marks
Program Execution	05 Marks
Interpretation	05 Marks
Output	05 Marks

One program from Part B (20 Marks):

Write-up (Python Program)	05 Marks
Program Execution	10 Marks
Output	05 Marks

Viva Voce : 10 Marks

PART A Solution:

1) **Extract a dataset of your choice and compute the correlation between any two variables and visualize the relationship using scatter plot using JMP Pro tool.**

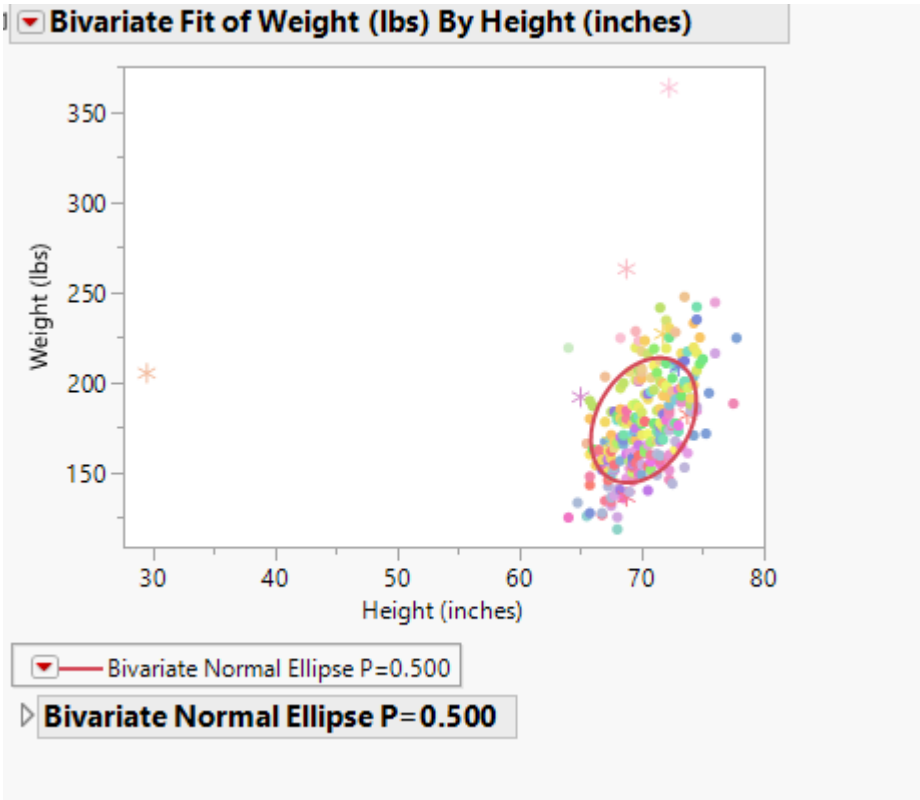
The tools used are JMP Pro

The correlation is a measure on the linear association between two variables

The correlation between two variables;

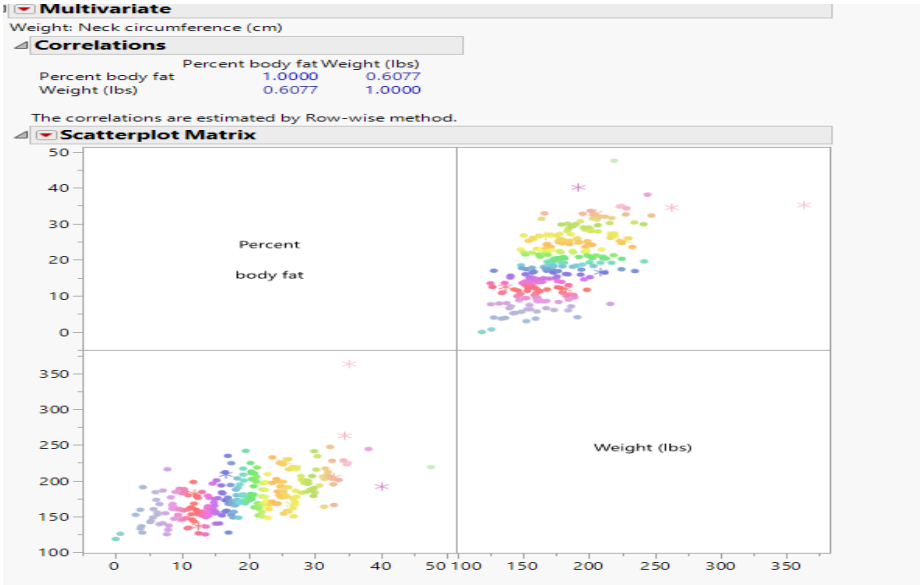
The steps required to perform scatter plot in JMP pro are:

1) Open JMP Pro data table, select Analyze, fit Y by X.
2) Click on a continuous variable from select columns and click Y response(continuous variables have blue triangles)
3) Click on a second continuous variable and click X, and click OK to generate a scatter plot.
4) To display the correlation, click on the red triangle and select the density eclipse> 0.55



Correlation between multiple pair of variables

1) Select analyze , and click on multivariate methods and atlast click on multivariate
2) Click on two or more continuous variables from select columns with density ellipse and table of correlation
3) Click OK to produce a scatter plot matrix with density ellipse and table of correlations



2) Apply Pearson correlation to test on a dataset .Show the normality on variables using Q-Q Plot and interpret the results using JMP Pro

Steps to produce a normal Quantile plot

1) Select help> sample.data folder and open big class JMP
2) Select analyse> Fit Y by X
3) Select height and click Y, Response
4) Select another other attributes eg> age from the attributes and display the Q-Q plot
5) We get the histogram and now click on the lower triangle and click on normal quantile plot

3) Select any dataset from JMP Pro tool and perform ANOVA test and Non- Parametric tests (The Mann Whitney test and The Kruskal-Wallis test).

One- way annova

Steps:

1) Select analyze> fit y by x
2) Click on a continuous variable from select columns and click y response
3) Click on a categorical variable and click x vector(variable have red or green bars)
4) Click ok
5) Click on red triangle
6) The null hypotheses is here and between the population means
7) Prob> F is p-value for whole model test since prob> F is less than 0.05, reject hypothesis on equal means

Two-way annova

1) Select analyze> fir model

2) Click on a continuous variable from select columns and click y, response
3) Click on two categorical variables from select columns and click macros, categorical(red or green)
4) Alternative hypothesis doesn't have an affect

Kruskals test

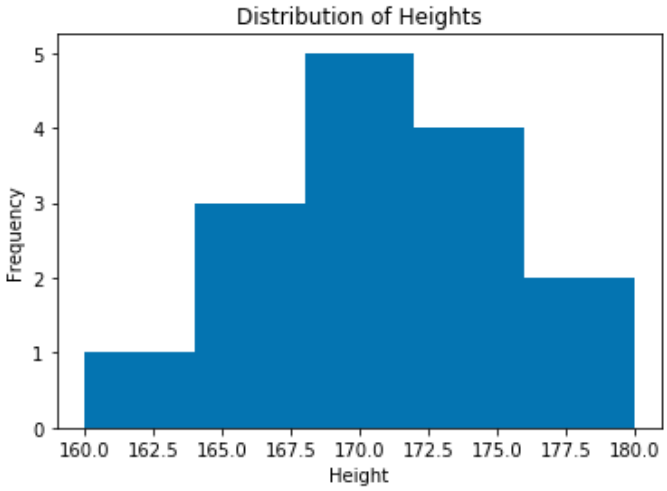
1) Select analyse> fit y by x
2) Continuous variable and categorical variable
3) Click ok, JMP will display the dropbox
4) Click on triangle go to the non-parametric test
5) Select x variable has 3 variables then we perform kruskal Wallis test

PART B Solution:

4) **You have been provided with a dataset containing information about the heights of students in a college. Perform univariate analysis on this dataset using Python and calculate the mean, median, mode, and standard deviation of the heights. Also, create a histogram to visualize the distribution of heights. Write a Python program to implement this analysis.**
heights = [165, 170, 168, 172, 175, 169, 180, 160, 165, 172, 168, 176, 170, 173, 165]

```
import numpy as np
import matplotlib.pyplot as plt
# Heights dataset
heights = [165, 170, 168, 172, 175, 169, 180, 160, 165, 172, 168, 176, 170, 173, 165]
# Calculate mean
mean_height = np.mean(heights)
print("Mean height:", mean_height)
# Calculate median
median_height = np.median(heights)
print("Median height:", median_height)
# Calculate mode
mode_height = np.argmax(np.bincount(heights))
print("Mode height:", mode_height)
# Calculate standard deviation
std_height = np.std(heights)
print("Standard deviation:", std_height)
# Create histogram
plt.hist(heights, bins=5)
plt.xlabel("Height")
plt.ylabel("Frequency")
plt.title("Distribution of Heights")
plt.show()
```

```
Mean height: 169.86666666666667
Median height: 170.0
Mode height: 165
Standard deviation: 4.910759162854106
```



5) You are given a dataset containing the prices of houses in a neighborhood. Perform univariate analysis on this dataset using Python and calculate the mean, median, mode, and standard deviation of the house prices. Additionally, plot a box plot to visualize the distribution of prices. Write a Python program to implement this analysis.
house_prices = [300000, 350000, 320000, 280000, 400000, 380000, 330000, 310000, 290000, 270000, 350000, 380000, 370000]

```
import numpy as np
import matplotlib.pyplot as plt
# House prices dataset
house_prices = [300000, 350000, 320000, 280000, 400000, 380000, 330000, 310000, 290000,
270000, 350000, 380000, 370000]
# Calculate mean
mean_price = np.mean(house_prices)
print("Mean price:", mean_price)
# Calculate median
median_price = np.median(house_prices)
print("Median price:", median_price)
# Calculate mode
mode_price = np.argmax(np.bincount(house_prices))
print("Mode price:", mode_price)
# Calculate standard deviation
std_price = np.std(house_prices)
print("Standard deviation:", std_price)
# Create box plot
plt.boxplot(house_prices)
plt.ylabel("Price")
plt.title("Distribution of House Prices")
plt.show()
```

```
Mean price: 333076.92307692306
Median price: 330000.0
Mode price: 350000
Standard deviation: 40455.980899399096
```

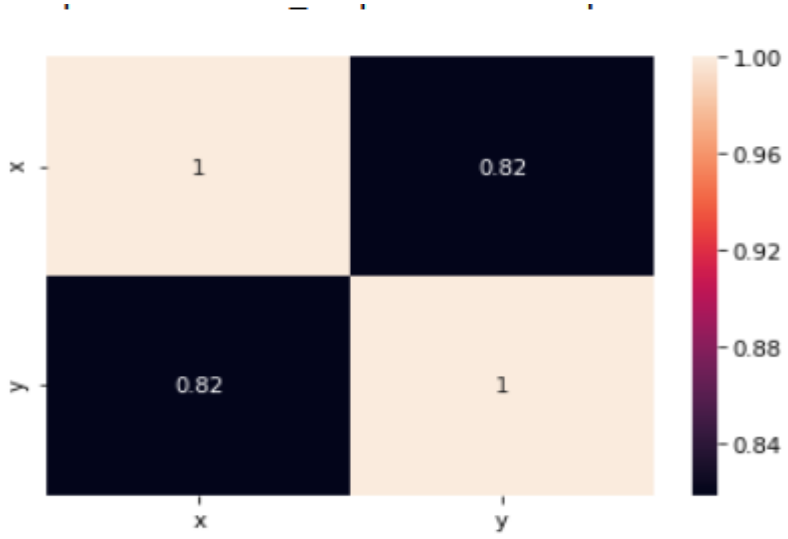
6) Perform univariate analysis on a dataset containing information about the performance of students in a school. The dataset includes variables such as student ID, test scores in different subjects, attendance, and socio-economic background. Use Python to analyze the distribution of test scores in each subject separately. Calculate the mean, median, and standard deviation for each subject and visualize the distribution using box plots.

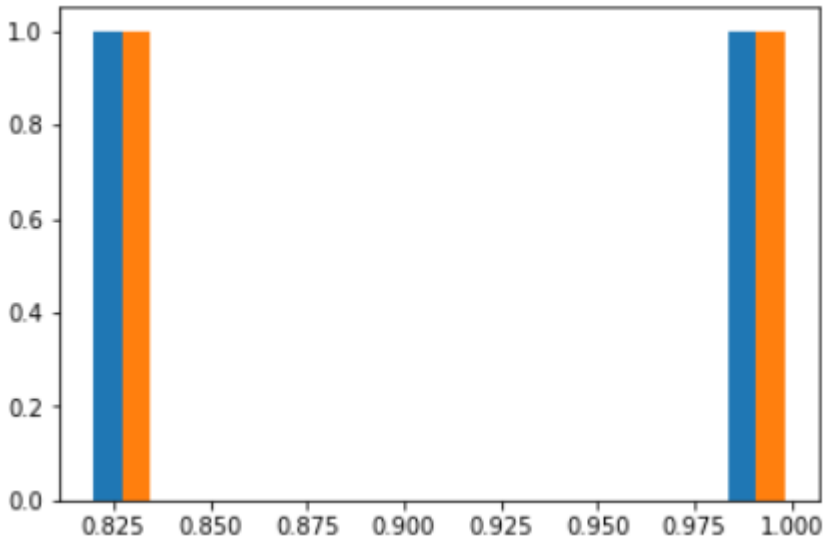
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# Read the dataset
data = pd.read_csv("student_performance_dataset.csv")  # Replace
"student_performance_dataset.csv" with the actual file name
# Analyze test scores in different subjects
subjects = ['math', 'science', 'english']  # Add more subjects if needed
for subject in subjects:
    subject_scores = data[subject]
    # Calculate descriptive statistics
    mean_score = np.mean(subject_scores)
    median_score = np.median(subject_scores)
    std_score = np.std(subject_scores)

    # Visualize the distribution using a box plot
    plt.boxplot(subject_scores)
    plt.xlabel(subject.capitalize() + " Score")
    plt.ylabel("Score")
    plt.title("Distribution of " + subject.capitalize() + " Scores")
    plt.show()

    # Print the results
    print(subject.capitalize() + " Scores:")
    print("Mean:", mean_score)
    print("Median:", median_score)
    print("Standard Deviation:", std_score)
```

print()

7) Consider the scores of ten students in SMIP and DBMS and Compute the Spearman rank correlation and Interpret the results using Python/R

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats

x=  [70, 46, 54, 34, 20, 86, 18, 12, 56, 64, 42]
y=[60,66,90,46,16,48,74,8,32,54,62]

df= pd.DataFrame((‘ SMIP’: X, ‘DBMS’:Y))
sns.heatmap(corr, annot=true)
print(corr)
plt.show()
plt.hist(corr)
plt.title(‘histogram’)
plt.show()

corr, pval=st.spearman(x,y)
print(“correlation co-efficient:”, corr)
print(“p-value”,pval)
```





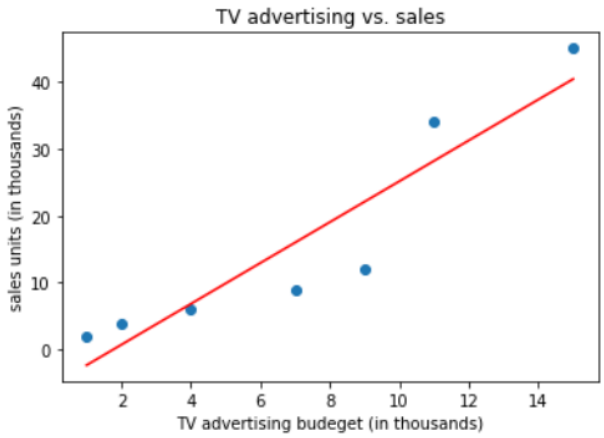
8) **Analyze a dataset containing information about the sales revenue and advertising expenditure of a company over a period of time. Calculate the Karl Pearson correlation coefficient between sales revenue and advertising expenditure using Python. Interpret the correlation coefficient and discuss the strength and direction of the relationship between advertising and sales.**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

# Read the dataset
data = pd.read_csv("sales_dataset.csv")  # Replace "sales_dataset.csv" with the actual file name

# Extract sales revenue and advertising expenditure columns
sales_revenue = data['sales_revenue']
advertising_expenditure = data['advertising_expenditure']

# Calculate the Karl Pearson correlation coefficient
correlation_coefficient, p_value = pearsonr(sales_revenue, advertising_expenditure)

# Interpret the correlation coefficient
if correlation_coefficient > 0:
    correlation_direction = "positive"
elif correlation_coefficient < 0:
    correlation_direction = "negative"
else:
    correlation_direction = "no"

correlation_strength = abs(correlation_coefficient)

# Print the results
print("Correlation Coefficient:", correlation_coefficient)
print("Correlation Direction:", correlation_direction)
print("Correlation Strength:", correlation_strength)

# Visualize the relationship using a scatter plot
```

```
plt.scatter(advertising_expenditure, sales_revenue)
plt.xlabel("Advertising Expenditure")
plt.ylabel("Sales Revenue")
plt.title("Relationship between Advertising Expenditure and Sales Revenue")
plt.show()
```

9) **Develop a Python/R code to build a simple Linear Regression model to predict sales units based on the advertising budget spent on TV. Display the statistical summary of the model.**

```
import numpy as np
import matplotlib.pyplot as plt
From sklearn.linear_model.import linear regression

sales=np.array([2,4,6,9,12,34,45])
tv_budgets=np.array([1,2,4,7,9,11,15])

x=tv_budgets.reshape(-1,1)
y=sales

model=linear.regression()
model.list(x,y)

print(“coeff”, model_coeff)

print(“intercept:”,model.intercept)

y_prod=model.predict(x)

plt.scatter(x,y,color=’b’, label=”actual.sales”)
plt.plot(x,y_prod,color=’r’,label=’linear regression”)
plt.xlabel(“tv budget”)
plt.ylabel(“sales”)
plt.legend()
plt.show()
```



10) Consider Australian Drug Sales dataset and develop a Python/R code to perform Time Series Analysis and visualize using plots.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df=pd.read_csv('ausdrug.csv')
df=['ds']=pd.to_datetime(df["ds"])
sns.set(style="whitegrid",color_codes=true)

plt.figure(figsize=(16,6))
plt.plot(df["ds"],df["y"])
plt.xlabel("time")
plt.ylabel("$millions")
plt.title("antibiotic drug")
plt.show()
```