1. Read the enclosed file, bmi\_region.csv into your python system as follows:

bmi\_df = pd.read\_csv(bmi\_region.csv', sep = ',', header = 0)

Using the vector, bmi\_df.bmi do the following:

1. Use bmi\_df.groupby() get region-wise measures such as: -
2. Mean and median **– 2 marks**
3. Standard deviation, coefficient of variation, and Z scores - **– 4 marks**
4. Are there any outliers? You can use Z scores to detect outliers. **– 2 marks**
5. Draw a boxplot and check if there are any outliers. **– 2 marks**
6. Draw a histogram. Are there data skewed? If so, how? **– 2 marks**
7. Based on the results of a) through e), what conclusions can you reach concerning bmi in regions in India. **- 3 marks**
8. For this problem, we are using a built-in Boston housing dataset which is a famous dataset from the 1970s. It contains 506 observations on housing prices around Boston.

Convert the data to pandas data frame as follows:

from sklearn.datasets import load\_boston

boston\_data = load\_boston()

print(boston\_data.data.shape)

print(type(boston\_data))

df\_boston = pd.DataFrame(boston\_data.data,columns=boston\_data.feature\_names)

df\_boston['target'] = pd.Series(boston\_data.target)

**The variables are described below:**

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. Target - Median value of owner-occupied homes in $1000's
15. Construct a scatter plot to show the relationship between the variables df\_boston.LSTAT and df\_boston.target **– 2 marks**

Hint:

import matplotlib.pyplot as plt

Use plt.scatter()

1. Measure the relative strength of a linear relationship between the df\_boston.LSTAT and df\_boston.target by calculating the sample correlation coefficient. **– 2 marks**

Hint:

Import numpy as np

Use np.corrcoef() to calculate the sample correlation coefficient.

1. Give your inference on the relationship between these variables. **– 1 marks**
2. The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Use the following python code to populate data.

from sklearn import datasets

import pandas as pd

import matplotlib.pyplot as plt

iris = datasets.load\_iris()

print(iris.target\_names)

print(iris.feature\_names)

df\_iris = pd.DataFrame(iris.data,columns = iris.feature\_names)

a. List the five number summary for 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)' **– 4 marks**

b. Construct the boxplot **– 2 marks**

c. Compare the distance between the median and highest value with lowest value and median for each of the above four variables to describe its shape for each of the numerical variables.

**– 4 marks**

Choose one of the following for each variable:

i. data distribution is skewed to the right

ii. data distribution is skewed to the left

1. data distribution is symmetrical
2. We have a liver disease L and a test T, used to detect the disease. The test is not 100% precise, indicating a person as having the disease when they don't have the disease (false positive) and not detecting the disease in people that are having (false negatives).

Let us assume that 4% of the population has the disease. Our test detects the disease in a patient with liver problem correctly in 95% of the cases.

In patients without liver problems, the test detects correctly in 90% of the cases.

What is the probability of having the liver disease, given a positive test? – **5 marks**

Hint: Use Bayes' theorem.

1. The length of similar products produced by a company follow a normal distribution model with a mean of 10 cm and a standard deviation of 0.05 cm.

If a product is chosen at random,

a) what is the probability that the length of this component is below 10 cm?  **2 marks**

b) what is the probability that the length of this component is between 9 and 11 cm? **3 marks**

Hint:

Use the following code to calculate the Cumulative Density Function

from scipy.stats import norm

norm.cdf(x)

where x is the Z value.

1. A group of students want to improve their marks in NEET examination and attended extra coaching classes conducted by their university. Marks scored by these students – before and after are given in the following table to be obtained by running the following python code:

marks\_dict = {'before':[86, 86, 75, 85, 87, 95, 85, 75, 84, 73, 84, 97, 84, 75,

73, 89, 80, 81, 82, 76],

'after' : [97, 95, 87, 95, 76, 99, 86, 88, 82, 74, 80, 95, 90, 77,

75, 76, 75, 80, 84, 75]}

marks\_df = pd.DataFrame(marks\_dict)

print(marks\_df)

a. State your null and alternative hypothesis - 2 marks

b. Test the hypothesis and Calculate the p-value. - 2 marks

c. At the 5% level of significance, is there evidence that the marks scored after attending the coaching classes is higher than before attending the coaching classes? - 2 marks

d. Plot the histogram for the differences in marks. Give your inference on the distribution of difference in marks? Is the assumption of paired t test violated? - 4 marks

Hint: Use the following code to conduct paired t test

from scipy import stats

stats.ttest\_rel()

1. A company an experiment 10 times, each with an estimated probability of success of 0.2. All ten experiments fail. What is the probability of that happening? – **2 marks**

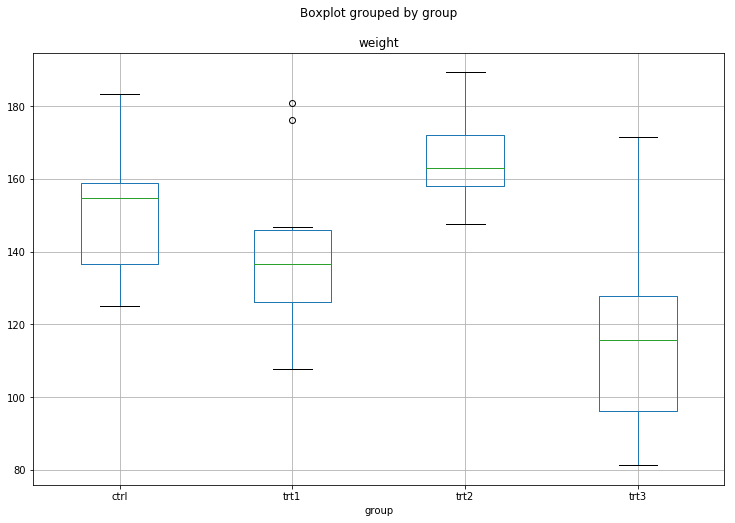
Hint: Let’s do 10,000 trials of the model and count the number that generate zero positive results.

np.random.binomial() function.

8) The file BodyWeight.csv contains results from an experiment to compare weights in kgs (as measured by weight of over-weight participants after undergoing a weight reduction program) obtained under a control and three different treatment conditions.

We have a data of 40 cases on 2 variables, weight, numeric and group, a factor variable with four levels ctrl, trt1, trt2 and trt3.

a. Judge by the Boxplot and state if there are differences in the weight of the participants for the three treatments. 2 marks



b. Perform a one-way ANOVA and draw your inferences about if there is any significant difference in weight due to the groups. 2 marks

Hint:

import pandas as pd

import statsmodels.api as sm

from statsmodels.formula.api import ols

datafile = "BodyWeight.csv"

data = pd.read\_csv(datafile)

Use sm.stats.anova\_lm() to perform a one-way ANOVA.

c. Write your null and alternative hypothesis. 2 marks

d. Write your conclusion based on the p-value. 2 mark

e. What are the assumptions of a ANOVA model? – 2 marks