**University of Mumbai**

**Practical Journal of**

**Research in Computing**

**&**

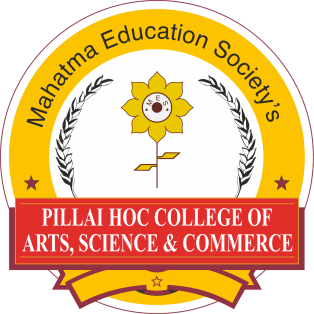
**Data Science**

**M.Sc. (Information Technology) Part-I**

**Submitted by**

**PATIL ASHUTOSH SHASHANK**

**Seat No: 3254080**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**PILLAI HOC COLLEGE OF ARTS, SCIENCE &COMMERCE, RASAYANI**

***(Affiliated to Mumbai University)* RASAYANI, 410207 MAHARASHTRA**

**2022-2023**

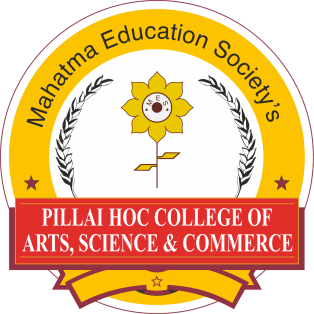
**Mahatma Education Society’s**

**Pillai Hoc College of Arts, Science & Commerce, Rasayani**

***(Affiliated to Mumbai University)***

**RASAYANI – MAHARASHTRA - 410207**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

****

**CERTIFICATE**

This is to certify that the experiment work entered in this journal is as per the syllabus in **M.Sc. (Information Technology) Part-I, Semester-I**; class prescribed by University of Mumbai for the subject **Research in Computing** was done in computer lab of Mahatma Education Society’s Pillai HOC College of Arts, Science & Commerce, Rasayani by **ASHUTOSH PATIL** during Academic year 2022-2023.

**Exam Seat No: 3254080**

**Subject In-Charge Coordinator**

**External Examiner Principal**

**Date: College Seal**

RESEARCH IN COMPUTING

|  |  |  |
| --- | --- | --- |
| **Practical No.** | **Title** | **Page No.** |
| **1** | 1. **Write a program for obtaining descriptive statistics of data.** 2. **Import data from different data sources (from Excel, csv, mysql, sql server, oracle to R/Python/Excel).** | **05** |
| **2** | 1. **Design a survey form for a given case study, collect the primary data and analyze it.** 2. **Perform analysis of given secondary data.** | **10** |
| **3** | 1. **Perform testing of hypothesis using one sample t-test.** 2. **Write a program for t-test comparing two means for independent samples.** 3. **Perform testing of hypothesis using paired t-test.** | **15** |
| **4** | 1. **Perform testing of hypothesis using chi-squared goodness- of-fit test.** 2. **Perform testing of hypothesis using chi-squared test of independence.** | **21** |
| **5** | 1. **Perform testing of hypothesis using Z-test.** 2. **Two-sample Z-test.** | **25** |
| **6** | 1. **Perform testing of hypothesis using One-way ANOVA.** 2. **Perform testing of hypothesis using Two-way ANOVA.** 3. **Perform testing of hypothesis using MANOVA.** | **27** |
| **7** | 1. **Perform the Random sampling for the given data and analyze it.** | **35** |
| **8** | 1. **Write a program for computing different correlation.** | **37** |

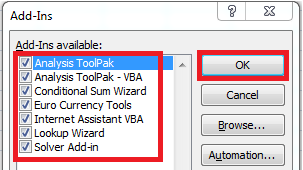
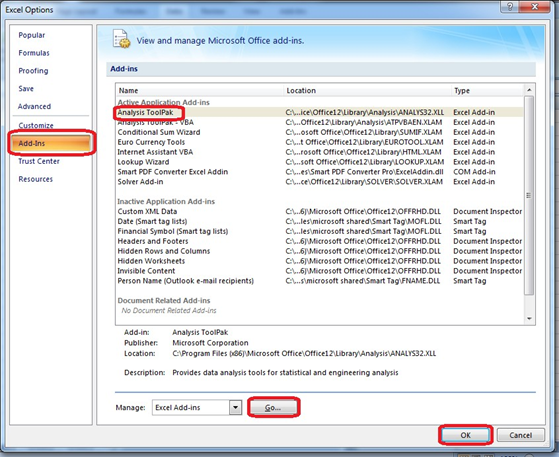
**INDEX**

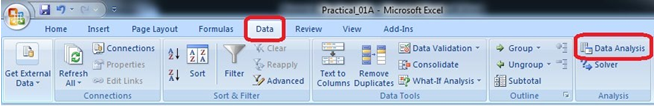
**Practical No. 01**

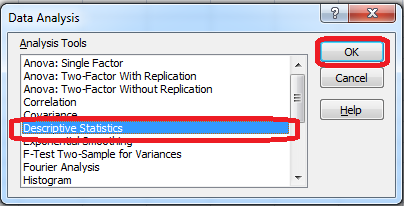
**Aim: A) Write a program for obtaining descriptive statistics of data.**

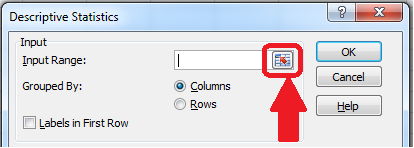
Using Excel

Go to File Menu  Options  Add-Ins Select Analysis ToolPak  Press OK

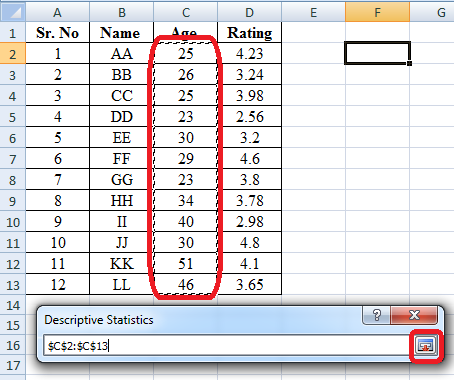


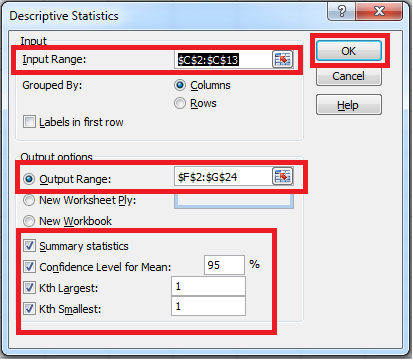




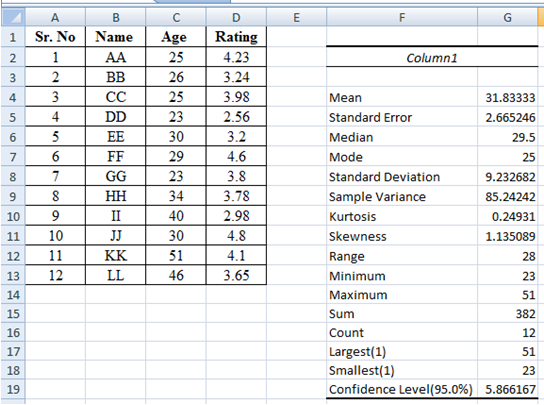
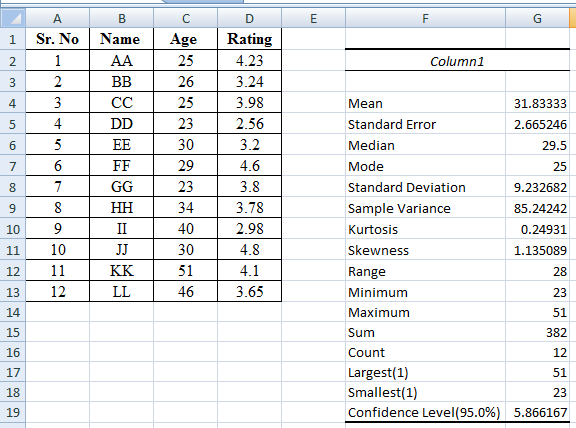


Select the data range from the excel worksheet.





**Output:**



**B) Import data from different data sources (from Excel, csv, mysql, sql server, oracle to R/Python/Excel)**

SQLite:

######################################### #######################

# -\*- coding: utf-8 -\*- ################################################################

import sqlite3 as sq import pandas as pd

################################################################ Base='C:/VKHCG'

sDatabaseName=Base + '/01-Vermeulen/00-RawData/SQLite/vermeulen.db' conn = sq.connect(sDatabaseName) ################################################################

sFileName='C:/VKHCG/01-Vermeulen/01-Retrieve/01-EDS/02-Python/Retrieve\_IP\_DATA.csv' print('Loading :',sFileName) IP\_DATA\_ALL\_FIX=pd.read\_csv(sFileName,header=0,low\_memory=False) IP\_DATA\_ALL\_FIX.index.names = ['RowIDCSV']

sTable='IP\_DATA\_ALL'

print('Storing :',sDatabaseName,' Table:',sTable) IP\_DATA\_ALL\_FIX.to\_sql(sTable, conn, if\_exists="replace") print('Loading :',sDatabaseName,' Table:',sTable) TestData=pd.read\_sql\_query("select \* from IP\_DATA\_ALL;", conn) print('################')

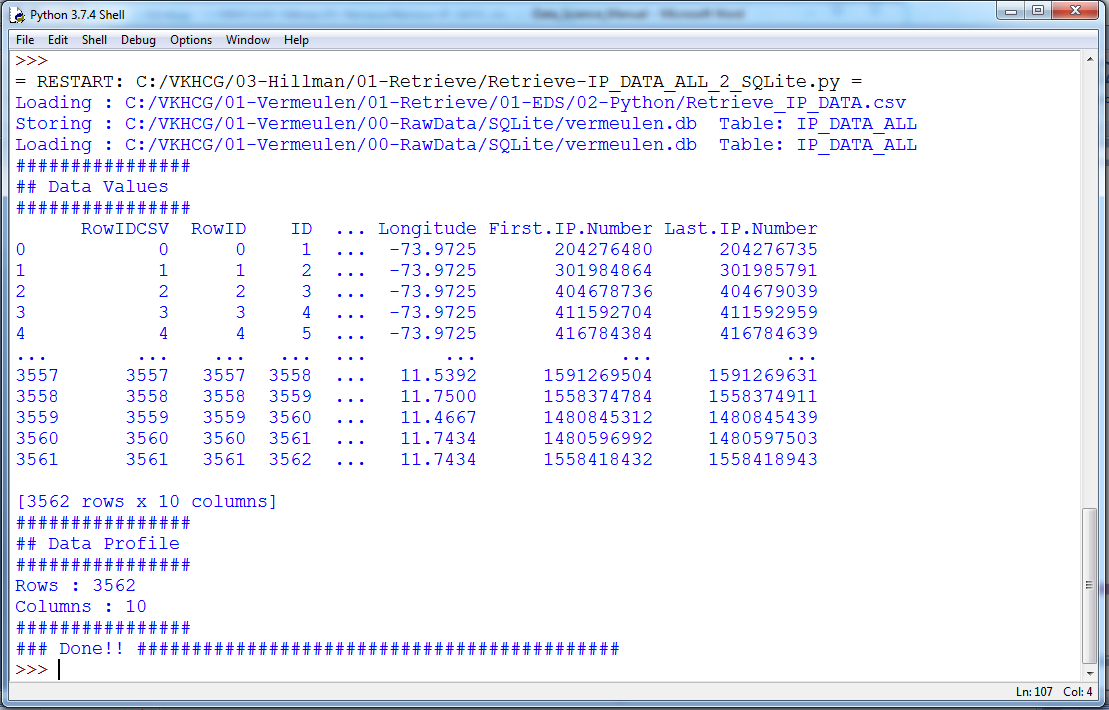
print('## Data Values') print('################')

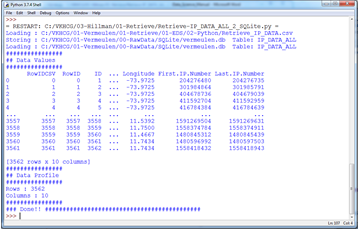
print(TestData) print('################')

print('## Data Profile') print('################')

print('Rows :',TestData.shape[0]) print('Columns :',TestData.shape[1]) print('################')

print('### Done!! ############################################')

**Output:**

****

**Microsoft Excel**

##################Retrieve-Country-Currency.py ################################################################

# -\*- coding: utf-8 -\*- ################################################################

importos

import pandas as pd ################################################################ Base='C:/VKHCG' ################################################################

sFileDir=Base + '/01-Vermeulen/01-Retrieve/01-EDS/02-Python' #if not os.path.exists(sFileDir):

#os.makedirs(sFileDir) ################################################################

CurrencyRawData = pd.read\_excel('C:/VKHCG/01-Vermeulen/00-RawData/Country\_Currency.xlsx') sColumns = ['Country or territory', 'Currency', 'ISO-4217']

CurrencyData = CurrencyRawData[sColumns] CurrencyData.rename(columns={'Country or territory': 'Country', 'ISO-4217': 'CurrencyCode'}, inplace=True) CurrencyData.dropna(subset=['Currency'],inplace=True) CurrencyData['Country'] = CurrencyData['Country'].map(lambda x: x.strip()) CurrencyData['Currency'] = CurrencyData['Currency'].map(lambda x: x.strip())

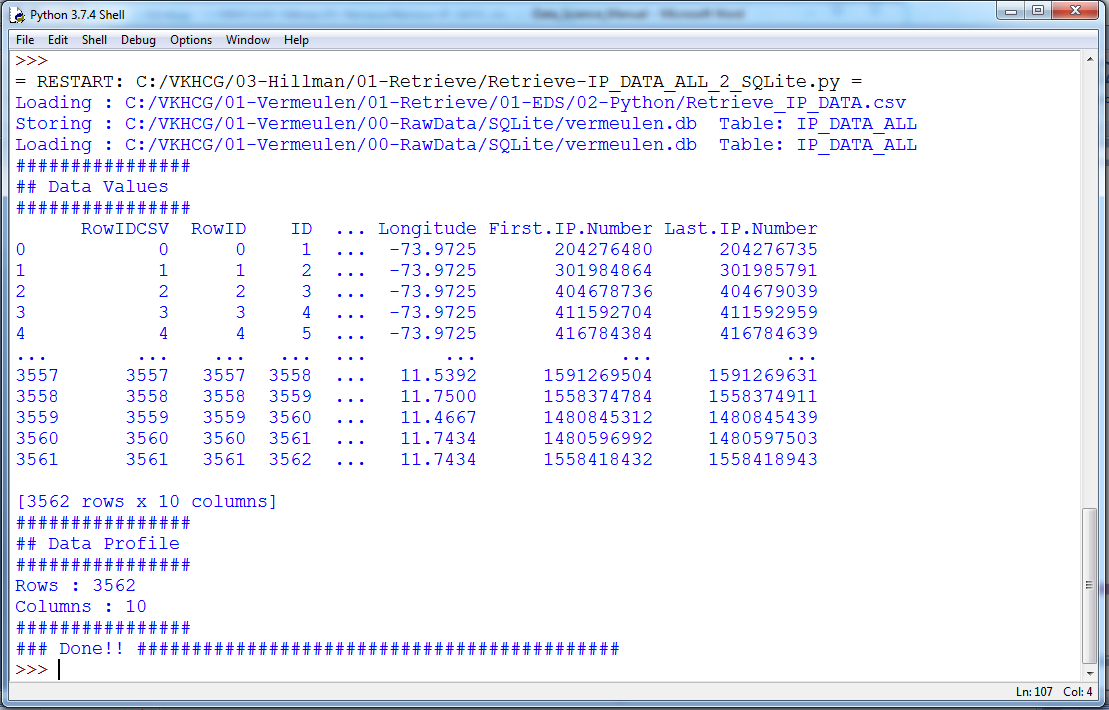
CurrencyData['CurrencyCode'] = CurrencyData['CurrencyCode'].map(lambda x: x.strip())

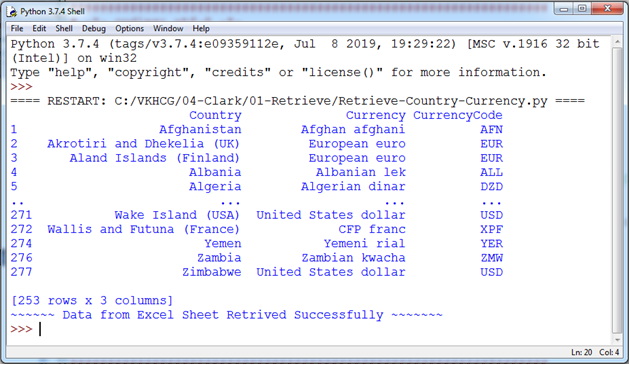
print(CurrencyData)

print('~~~~~~ Data from Excel Sheet Retrived Successfully ~~~~~~~ ') ################################################################

sFileName=sFileDir + '/Retrieve-Country-Currency.csv'

CurrencyData.to\_csv(sFileName, index = False) ################################################################

**Output:**



**Practical No. 02**

**Aim: A) Design a survey form for a given case study, collect the primary data and analyze it**

Case 1:

A researcher wants to conduct a Survey in colleges on Use of ICT in higher education from Mumbai, Thane and Navi Mumbai. The survey focuses on access to and use of ICT in teaching and learning, as well as on attitudes towards the use of ICT in teaching and learning.

Design questionnaire addressed to teachers seeks information about the target class, his experience using ICT for teaching, access to ICT infrastructure, support available, ICT based activities and material used, obstacles to the use of ICT in teaching, learning activities with the target class, your skills and attitudes to ICT, and some personal background information.

Arrange question in following groups:

1. Information about the target class you teach

2. Experience with ICT for teaching

3. ICT access for teaching

4. Support to teachers for ICT use

5. ICT based activities and material used for teaching

6. Obstacles to using ICT in teaching and learning

7. Learning activities with the target class

8. Teacher skills

9. Teacher opinions and attitudes

10. Personal background information

Case 2:

A research agency wants to study the perception about App based taxi service in Mumbai, Thane and Navi Mumbai. The survey focuses on customers attitude towards app base taxi service as well as on attitudes towards regular taxi cab.

Design questionnaire seeks information about the target taxi service, his experience using taxi services, access, support available, obstacles and some personal background information, with the following objectives:

1. To find out the customer satisfaction towards the App based-taxi services.

2. To find the level of convenience and comfort with App based -taxi services.

3. To know their opinion about the tariff system and promptness of service.

4. To ascertain the customer view towards the driver behaviour and courtesy.

5. To provide inputs to enhance the services to delight the customers.

6. To examine relationship between service quality factors and taxi passenger satisfaction.

7. To suggest better regulations for transportation authorities regarding customer protection

and effective monitoring of taxi services.

Case 3:

A popular electronic store want to conduct a survey to develop awareness of branded laptop baseline estimates and determine popularity of different company’s laptop. It suggests steps to be initiated or strengthened in the field of demand in a region. The key indicators are among the general population, demand branded laptop and the problem users.

The objectives of this particular study are:-

1. To know the preferences of different types of branded laptops by students and professionals.

2. To study which factor influence for choosing different types of branded laptops.

3. To know about the level of satisfaction towards different types of branded laptops.

4. To identify the perception of consumers towards the laptop positioning strategy.

5. To know the consumer preference towards laptop in the present era

Use the collected data for analysis.

**B) Perform analysis of given secondary data.**

Steps in Secondary Data Analysis

1. Determine your research question – Knowing exactly what you are looking for.

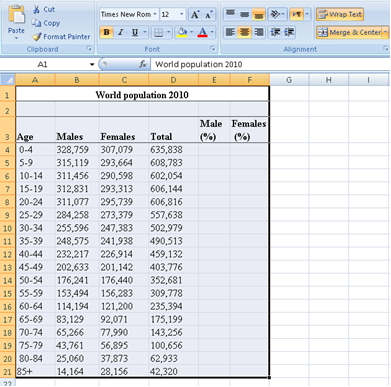
2. Locating data– Knowing what is out there and whether you can gain access to it. A quick Internet search, possibly with the help of a librarian, will reveal a wealth of options.

3. Evaluating relevance of the data – Considering things like the data’s original purpose, when it was collected, population, sampling strategy/sample, data collection protocols, operationalization of concepts, questions asked, and form/shape of the data.

4. Assessing credibility of the data – Establishing the credentials of the original researchers, searching for full explication of methods including any problems encountered, determining how consistent the data is with data from other sources, and discovering whether the data has been used in any credible published research.

5. Analysis – This will generally involve a range of statistical processes.

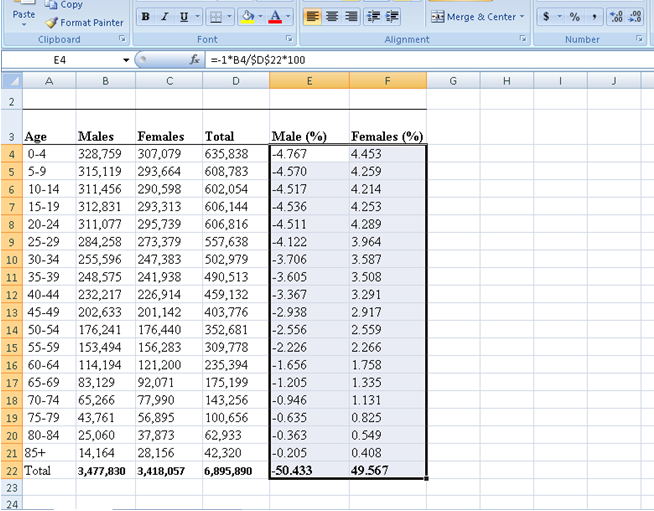
Example: Analyze the given Population Census Data for Planning and Decision Making by using the size and composition of populations.



Put the cursor in cell B22 and click on the AutoSum and then click Enter. This will calculate the total population. Then copy the formula in cell D22 across the row 22.

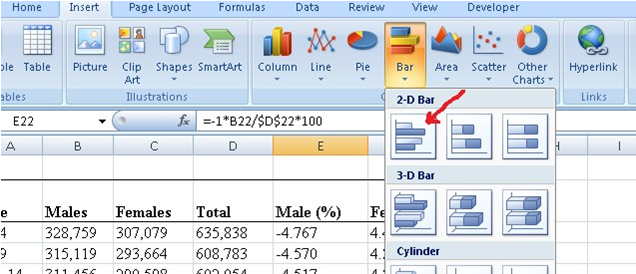
To calculate the percent of males in cell E4, enter the formula =-1\*100\*B4/$D$22 . And copy the formula in cell E4 down to cell E21.

To calculate the percent of females in cell F4, enter the formula =100\*C4/$D$22. Copy the formula in cell F4 down to cell F21.



To build the population pyramid, we need to choose a horizontal bar chart with two series of data (% male and % female) and the age labels in column A as the Category X-axis labels. Highlight the range A3:A21, hold down the CTRL key and highlight the range E3:F21

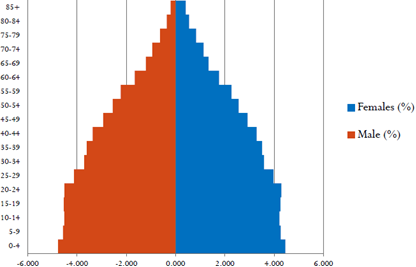
Under inset tab, under horizontal bar charts select clustered bar chart



Put the tip of your mouse arrow on the Y-axis (vertical axis) so it says “Category Axis”, right click and chose Format Axis

Choose Axis options tab and set the major and minor tick mark type to None, Axis labels to Low, and click OK.

Click on any of the bars in your pyramid, click right and select “format data series”. Set the Overlap to 100 and Gap Width to 0. Click OK.



**Practical No. 03**

**Aim:** **A)** **Perform testing of hypothesis using one sample t-test.**

One sample t-test : The One Sample t Test determines whether the sample mean is statistically different from a known or hypothesised population mean. The One Sample t Test is a parametric test.

Program Code:

fromscipy.stats import ttest\_1samp   
importnumpy as np

ages = np.genfromtxt('ages.csv')

print(ages)

ages\_mean = np.mean(ages)

print(ages\_mean)

tset, pval = ttest\_1samp(ages, 30)

print('p-values - ',pval)

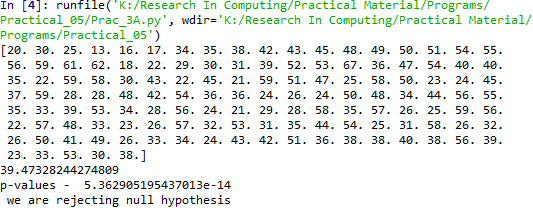
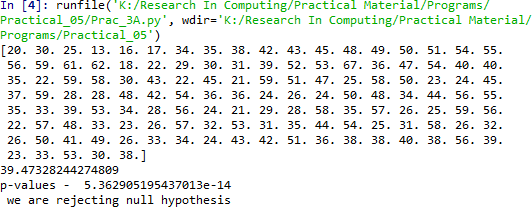
if pval< 0.05: # alpha value is 0.05

print(" we are rejecting null hypothesis")

else:

print("we are accepting null hypothesis")

**Output:**



**B) Write a program for t-test comparing two means for independent samples.**

The t distribution provides a good way to perform one sample tests on the mean when the population variance is not known provided the population is normal or the sample is sufficiently large so that the Central Limit Theorem applies.

Two Sample t Test

Example: A college Princiapal informed classroom teachers that some of their students showedunusual potential for intellectual gains. One months later the students identified to teachers ashaving potentional for unusual intellectual gains showed significiantly greater gains performanceon a test said to measure IQ than did students who were not so identified. Below are the data forthe students:

|  |  |  |
| --- | --- | --- |
| **Experimental** | **Comparison** |  |
| 35 | 2 |
| 40 | 27 |
| 12 | 38 |
| 15 | 31 |
| 21 | 1 |
| 14 | 19 |
| 46 | 1 |
| 10 | 34 |
| 28 | 3 |
| 48 | 1 |
| 16 | 2 |
| 30 | 3 |
| 32 | 2 |
| 48 | 1 |
| 31 | 2 |
| 22 | 1 |
| 12 | 3 |
| 39 | 29 |
| 19 | 37 |
| 25 | 2 |
| 27.15 | 11.95 | Mean |
| 12.51 | 14.61 | Sd |

Experimental Data

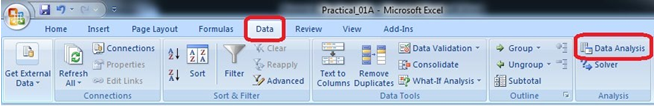
To calculate Standard Mean go to cell A22 and type =SUM(A2:A21)/20

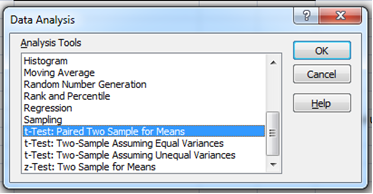
To calculate Standard Deviation go to cell A23 and type =STDEV(A2:A21)

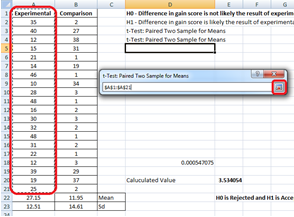
Comparison Data

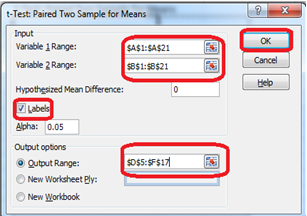
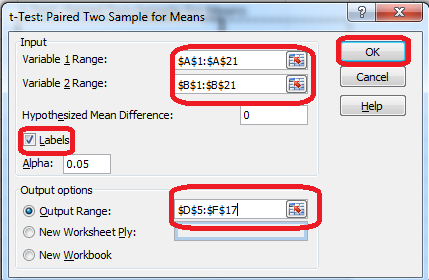
To calculate Standard Mean go to cell B22 and type =SUM(B2:B21)/20

To calculate Standard Deviation go to cell B23 and type =STDEV(B2:B21) To find T-Test Statistics go to data Data Analysis









To caluculate the T-Test square value go to cell E20 and type

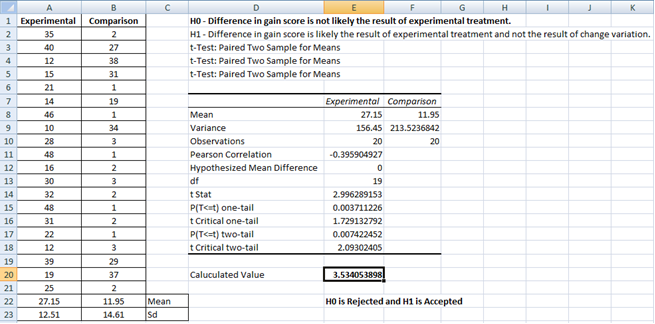
=(A22-B22)/SQRT((A23\*A23)/COUNT(A2:A21)+(B23\*B23)/COUNT(A2:A21))

Now go to cell E20 and type

=IF(E20<E12,"H0 is Accepted", "H0 is Rejected and H1 is Accepted")

Our calculated value is larger than the tabled value at alpha = .01, so we reject the null hypothesisand accept the alternative hypothesis, namely, that the difference in gain scores is likely the resultof the experimental treatment and not the result of chance variation.

**Output:**

****

**C) Perform testing of hypothesis using paired t-test.**

The paired sample t-test is also called dependent sample t-test. It’s an univariate test that tests for a significant difference between 2 related variables. An example of this is if you where to collect the blood pressure for an individual before and after some treatment, condition, or time point. The data set contains blood pressure readings before and after an intervention. These are variables “bp\_before” and “bp\_after”.

The hypothesis being test is:

• H0 - The mean difference between sample 1 and sample 2 is equal to 0.

• H0 - The mean difference between sample 1 and sample 2 is not equal to 0

Program Code:

from scipy import stats

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv("blood\_pressure.csv")

print(df[['bp\_before','bp\_after']].describe())

df[['bp\_before', 'bp\_after']].plot(kind='box')

plt.savefig('boxplot\_outliers.png')

df['bp\_difference'] = df['bp\_before'] - df['bp\_after']

df['bp\_difference'].plot(kind='hist', title= 'Blood Pressure Difference Histogram')

plt.savefig('blood pressure difference histogram.png')

stats.probplot(df['bp\_difference'], plot= plt)

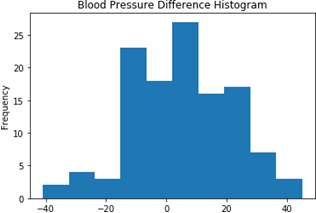
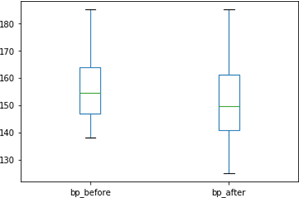
plt.title('Blood pressure Difference Q-Q Plot')

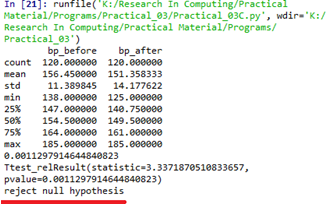
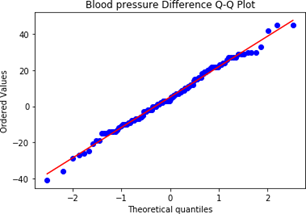
plt.savefig('blood pressure difference qq plot.png')

stats.shapiro(df['bp\_difference'])

stats.ttest\_rel(df['bp\_before'], df['bp\_after'])

**Output:**





A paired sample t-test was used to analyze the blood pressure before and after the intervention to test if the intervention had a significant affect on the blood pressure. The blood pressure before the intervention was higher (156.45 ± 11.39 units) compared to the blood pressure post intervention (151.36 ± 14.18 units); there was a statistically significant decrease in blood pressure (t(119)=3.34, p= 0.0011) of 5.09 units.

**Practical No. 04**

**Aim: A) Perform testing of hypothesis using chi-squared goodness- of-fit test.**

**Problem**

Ansystem administrator needs to upgrade the computers for his division. He wants to know what sort of computer system his workers prefer. He gives three choices: Windows, Mac, or Linux. Test the hypothesis or theory that an equal percentage of the population prefers each type of computer system .

|  |  |  |  |
| --- | --- | --- | --- |
| **System** | **O** | **Ei** |  |
| **Windows** | 20 | **33.33%** |  |
| **Mac** | 60 | **33.33%** |  |
| **Linux** | 20 | **33.33%** |  |

H0 : The population distribution of the variable is the same as the proposed distribution HA : The distributions are different

To calculate the Chi –Squred value for Windows go to cell D2 and type =((B2- C2)\*(B2-C2))/C2

To calculate the Chi –Squred value for Mac go to cell D3 and type =((B3-C3)\*(B3- C3))/C3

To calculate the Chi –Squred value for Mac go to cell D3 and type =((B4-C4)\*(B4- C4))/C4

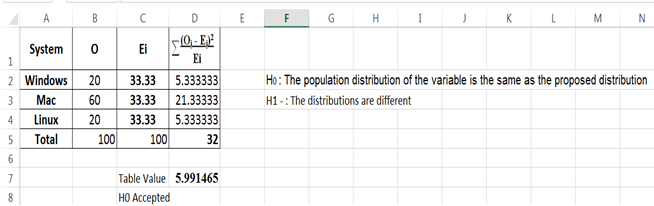
Go to Cell D5 for and type=SUM(D2:D4)

To get the table value for Chi-Square for α = 0.05 and dof = 2, go to cell D7 and type

=CHIINV(0.05,2)

At cell D8 type =IF(D5>D7, "H0 Accepted","H0 Rejected")

**Output:**



**B) Perform testing of hypothesis using chi-squared test of independence.**

In a study to understatnd the permormacne of M. Sc. IT Part -1 class, a college selects a random sample of 100 students. Each student was asked his grade obtained in B. Sc. IT. The sample is as given below

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr.**  **No** | **Roll No** | **Student's Name** | **Gen** | **Grade** |  | **Sr. No** | **Roll No** | **Student's Name** | **Gen** | **Grade** |
| **1** | 1 | Gaborone | m | O | **62** | 3 | Maun | f | O |
| **2** | 2 | Francistown | m | O | **63** | 7 | Tete | f | O |
| **3** | 5 | Niamey | m | O | **64** | 9 | Chimoio | f | O |
| **4** | 13 | Maxixe | m | O | **65** | 11 | Pemba | f | O |
| **5** | 16 | Tema | m | O | **66** | 14 | Chibuto | f | O |
| **6** | 17 | Kumasi | m | O | **67** | 25 | Mampong | f | O |
| **7** | 34 | Blida | m | O | **68** | 36 | Tlemcen | f | O |
| **8** | 35 | Oran | m | O | **69** | 40 | Adrar | f | O |
| **9** | 38 | Saefda | m | O | **70** | 41 | Tindouf | f | O |
| **10** | 42 | Constantine | m | O | **71** | 46 | Skikda | f | O |
| **11** | 43 | Annaba | m | O | **72** | 47 | Ouargla | f | O |
| **12** | 45 | Bejaefa | m | O | **73** | 10 | Matola | f | D |
| **13** | 48 | Medea | m | O | **74** | 20 | Legon | f | D |
| **14** | 49 | Djelfa | m | O | **75** | 21 | Sunyani | f | D |
| **15** | 50 | Tipaza | m | O | **76** | 72 | Teenas | f | D |
| **16** | 51 | Bechar | m | O | **77** | 73 | Kouba | f | D |
| **17** | 54 | Mostaganem | m | O | **78** | 75 | HussenDey | f | D |
| **18** | 55 | Tiaret | m | O | **79** | 77 | Khenchela | f | D |
| **19** | 56 | Bouira | m | O | **80** | 82 | HassiBahbah | f | D |
| **20** | 59 | Tebessa | m | O | **81** | 84 | Baraki | f | D |
| **21** | 61 | El Harrach | m | O | **82** | 91 | Boudouaou | f | D |
| **22** | 62 | Mila | m | O | **83** | 95 | Tadjenanet | f | D |
| **23** | 65 | Fouka | m | O | **84** | 4 | Molepolole | f | C |
| **24** | 66 | El Eulma | m | O | **85** | 8 | Quelimane | f | C |
| **25** | 68 | SidiBel Abbes | m | O | **86** | 23 | Bolgatanga | f | C |
| **26** | 69 | Jijel | m | O | **87** | 58 | Mohammadia | f | C |
| **27** | 70 | Guelma | m | O | **88** | 83 | Merouana | f | C |
| **28** | 85 | Khemis El Khechna | m | O | **89** | 24 | Ashaiman | f | B |
| **29** | 87 | Bordj El Kiffan | m | O | **90** | 76 | N'gaous | f | B |
| **30** | 88 | Lakhdaria | m | O | **91** | 90 | Bab El Oued | f | B |
| **31** | 6 | Maputo | m | D | **92** | 92 | BordjMenael | f | B |
| **32** | 12 | Lichinga | m | D | **93** | 93 | Ksar El Boukhari | f | B |
| **33** | 15 | Ressano Garcia | m | D | **94** | 74 | Reghaa | f | A |
| **34** | 19 | Accra | m | D | **95** | 78 | Cheria | f | A |
| **35** | 27 | Wa | m | D | **96** | 79 | Mouzaa | f | A |
| **36** | 28 | Navrongo | m | D | **97** | 80 | Meskiana | f | A |
| **37** | 37 | Mascara | m | D | **98** | 81 | Miliana | f | A |
| **38** | 44 | Batna | m | D | **99** | 94 | Sig | f | A |
| **39** | 57 | El Biar | m | D | **100** | 99 | Kadiria | f | A |
| **40** | 60 | Boufarik | m | D |  | | | | | |
| **41** | 63 | OuedRhiou | m | D |
| **42** | 64 | Souk Ahras | m | D |
| **43** | 71 | Dar El Befda | m | D |
| **44** | 86 | Birtouta | m | D |
| **45** | 18 | Takoradi | m | C |
| **46** | 22 | Cape Coast | m | C |
| **47** | 29 | Kwabeng | m | C |
| **48** | 30 | Algiers | m | C |
| **49** | 31 | Laghouat | m | C |
| **50** | 39 | Relizane | m | C |
| **51** | 52 | Setif | m | C |
| **52** | 53 | Biskra | m | C |
| **53** | 67 | Kolea | m | C |
| **54** | 100 | AefnFakroun | m | C |
| **55** | 26 | Nima | m | B |
| **56** | 32 | TiziOuzou | m | B |
| **57** | 33 | Chlef | m | B |

Null Hypothesis - H0 : The performance of girls students is same as boys students. Alternate Hypothesis - H1 : The performance of boys and girls students are different. Open Excel Workbook

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **O** | **A** | **B** | **C** | **D** | **Total** |  |
| **Girls** | 11 | 7 | 5 | 5 | 11 | **39** | 6.075 |
| **Boys** | 30 | 4 | 3 | 10 | 14 | **61** | 6.075 |
| **Total** | 41 | 11 | 8 | 15 | 25 | **100** | **12.150** |
| **Ei** | **20.5** | **5.5** | **4** | **7.5** | **12.5** | **50** |  |

Prepare a contingency table as shown above. To calculate Girls Students with ‘O’ Grade

Go to Cell N6 and type =COUNTIF($J$2:$K$40,"O")

To calculate Girls Students with ‘A’ Grade

Go to Cell O6 and type =COUNTIF($J$2:$K$40,"A")

To calculate Girls Students with ‘B’ Grade

Go to Cell P6 and type =COUNTIF($J$2:$K$40,"B")

To calculate Girls Students with ‘C’ Grade

Go to Cell Q6 and type =COUNTIF($J$2:$K$40,"C")

To calculate Girls Students with ‘D’ Grade

Go to Cell R6 and type =COUNTIF($J$2:$K$40,"D")

To calculate Boys Students with ‘O’ Grade

Go to Cell N7 and type =COUNTIF($D$2:$E$62,"O")

To calculate Boys Students with ‘A’ Grade

Go to Cell O7 and type =COUNTIF($D$2:$E$62,"A")

To calculate Boys Students with ‘B’ Grade

Go to Cell P7 and type =COUNTIF($D$2:$E$62,"B")

To calculate Boys Students with ‘C’ Grade

Go to Cell Q7 and type =COUNTIF($D$2:$E$62,"C")

To calculate Boys Students with ‘D’ Grade

Go to Cell R7 and type =COUNTIF($D$2:$E$62,"D")

To calculated the expected value Ei

Go to Cell N9 and type =N8/2 Go to Cell O9 and type =O8/2 Go to Cell P9 and type =P8/2 Go to Cell Q9 and type =Q8/2 Go to Cell R9 and type =R8/2

Go to Cell S6 and calculate total girl students = SUM(N6:R6) Go to Cell S7 and calculate total girl students = SUM(N7:R7)

Now Calculate

Go to cell T6 and type

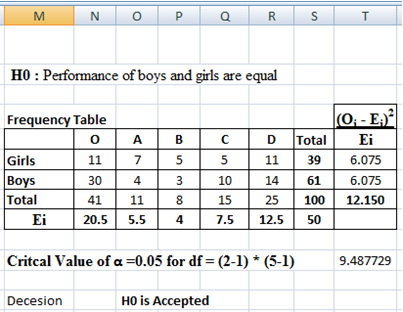
=SUM((N6-$N$9)^2/$N$9,(O6-$O$9)^2/$O$9,(P6-$P$9)^2/$P$9,(Q6-Q$9)^2/$Q$9, (R6-$R$9)^2/$R$9)

Go to cell T7 and type

=SUM((N7-$N$9)^2/$N$9,(O7-$O$9)^2/$O$9,(P7-$P$9)^2/$P$9,(Q7-Q$9)^2/$Q$9, (R7-$R$9)^2/$R$9)

To get the table value go to cell T11 and type =CHIINV(0.05,4)

Go to cell O13 and type =IF(T8>=T11," H0 is Accepted", "H0 is Rejected")



**Practical No. 05**

**Aim: Perform testing of hypothesis using Z-test.**

Use a Z test if:

• Your sample size is greater than 30. Otherwise, use a t test.

• Data points should be independent from each other. In other words, one data point isn’t related or doesn’t affect another data point.

• Your data should be normally distributed. However, for large sample sizes (over 30) this doesn’t always matter.

• Your data should be randomly selected from a population, where each item has an equal chance of being selected.

• Sample sizes should be equal if at all possible.

**Ho - Blood pressure has a mean of 156 units**

Program Code for one-sample Z test.

from statsmodels.stats import weightstats as stests

import pandas as pd

from scipy import stats

df = pd.read\_csv("blood\_pressure.csv")

df[['bp\_before','bp\_after']].describe()

print(df)

ztest ,pval = stests.ztest(df['bp\_before'], x2=None, value=156)

print(float(pval))

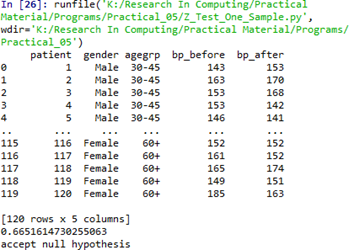
if pval<0.05:

print("reject null hypothesis")

else:

print("accept null hypothesis")

**Output:**



**Two-sample Z test-** In two sample z-test , similar to t-test here we are checking two independent data groups and deciding whether sample mean of two group is equal or not.

**H0 : mean of two group is 0**

**H1 : mean of two group is not 0**

import pandas as pd

from statsmodels.stats import weightstats as stests

df = pd.read\_csv("blood\_pressure.csv")

df[['bp\_before','bp\_after']].describe()

print(df)

ztest ,pval = stests.ztest(df['bp\_before'], x2=df['bp\_after'], value=0,alternative='two- sided')

print(float(pval))

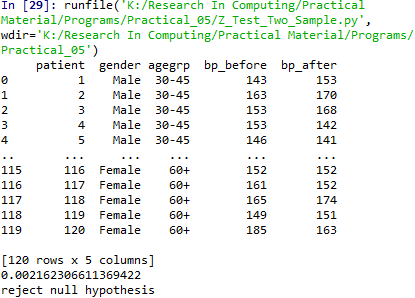
if pval<0.05:

print("reject null hypothesis")

else:

print("accept null hypothesis")

**Output:**



**Practical No. 06**

**Aim: Perform testing of hypothesis using One-way ANOVA.**

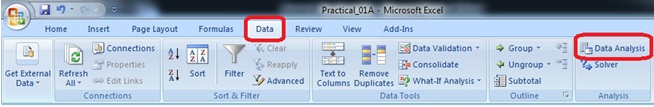
**Using Excel**

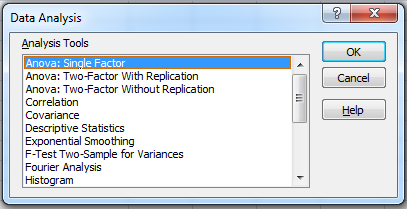
**H0 - There are no significant differences between the Subject’s mean SAT scores.**

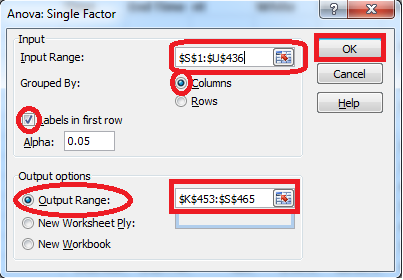
**µ1 = µ2 = µ3 = µ4 = µ5**

**H1 - There is a significant difference between the Subject's mean SAT scores.**

To perform ANOVA go to data Data Analysis







**Input Range:** $S$1:$U$436 (Select columns to be analyzed in group)

**Output Range:** $K$453:$S$465 (Can be any Range)

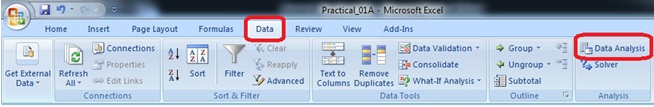
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Average Score (SAT Math) | 375 | 162354 | 432.944 | 5177.144 |  |  |
| Average Score (SAT Reading) | 375 | 159189 | 424.504 | 3829.267 |  |  |
| Average Score (SAT Writing) | 375 | 156922 | 418.4587 | 4166.522 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 39700.57 | 2 | 19850.28 | 4.520698 | 0.01108 | 3.003745 |
| Within Groups | 4926677 | 1122 | 4390.977 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 4966377 | 1124 |  |  |  |  |

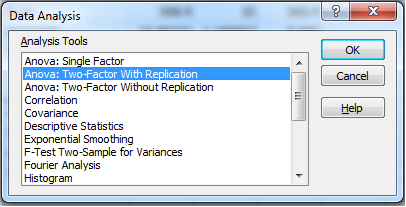
Since theresulting pvalueis less than 0.05. The null hypothesis (H0) is rejected and conclude that there is a significant difference between the SAT scores for each subject.

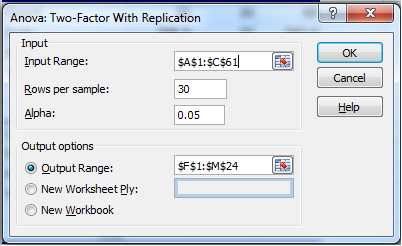
**B) Perform testing of hypothesis using Two-way ANOVA.**

**Using Excel:**

Go to Data tab Data Analysis







Rows Per Sample – 30 (Beacause 30 Patients are given each dose)

Alpha – 0.05

Output Range - $F$1:$M$24

**Output:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Anova: Two-Factor With Replication** | | |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY | len | dose | Total |  |  |  |
| *1* |  |  |  |  |  |  |
| Count | 30 | 30 | 60 |  |  |  |
| Sum | 508.9 | 35 | 543.9 |  |  |  |
| Average | 16.96333 | 1.166667 | 9.065 |  |  |  |
| Variance | 68.32723 | 0.402299 | 97.22333 |  |  |  |
|  |  |  |  |  |  |  |
| *31* |  |  |  |  |  |  |
| Count | 30 | 30 | 60 |  |  |  |
| Sum | 619.9 | 35 | 654.9 |  |  |  |
| Average | 20.66333 | 1.166667 | 10.915 |  |  |  |
| Variance | 43.63344 | 0.402299 | 118.2854 |  |  |  |
|  |  |  |  |  |  |  |
| *Total* |  |  |  |  |  |  |
| Count | 60 | 60 |  |  |  |  |
| Sum | 1128.8 | 70 |  |  |  |  |
| Average | 18.81333 | 1.166667 |  |  |  |  |
| Variance | 58.51202 | 0.39548 |  |  |  |  |
| **ANOVA** |  |  |  |  |  |  |
| *Source of*  *Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Sample | 102.675 | 1 | 102.675 | 3.642079 | 0.058808 | 3.922879 |
| Columns | 9342.145 | 1 | 9342.145 | 331.3838 | 8.55E-36 | 3.922879 |
| Interaction | 102.675 | 1 | 102.675 | 3.642079 | 0.058808 | 3.922879 |
| Within | 3270.193 | 116 | 28.19132 |  |  |  |
| Total | 12817.69 | 119 |  |  |  |  |

P-value = 0.0588079 column in the ANOVA Source of Variation table at the bottom of the output. Because the p-values for both medicin dose and interaction are less than our significance level, these factors are statistically significant. On the other hand, the interaction effect is not significant because its p-value (0.0588) is greater than our significance level. Because the interaction effect is not significant, we can focus on only the main effects and not consider the interaction effect of the dose.

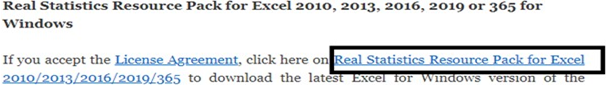
**C) Perform testing of hypothesis using MANOVA.**

**Excel:**

Go to http://www.real-statistics.com/free-download/

1. Download Real Statistics Resource Pack



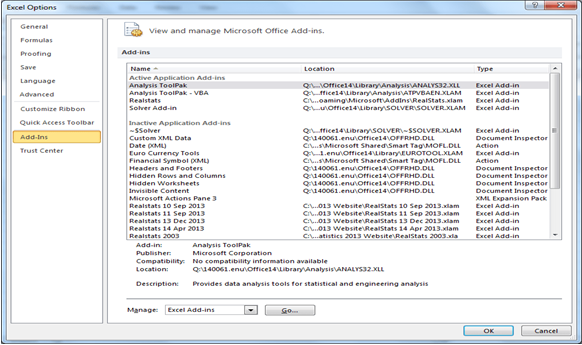


OR

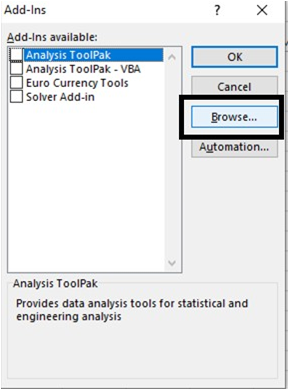
http://www.real-statistics.com/wp-content/uploads/2019/11/XRealStats.xlam

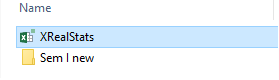
Install Add-in in excel. Select File > Help Options > Add-Ins and click on the Go button at the bottom of the window (see Figure 1).

Add-ins -> Analysis Pack -> Go

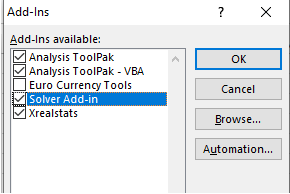


Click on browse and select XrealStats file (previously downloaded).



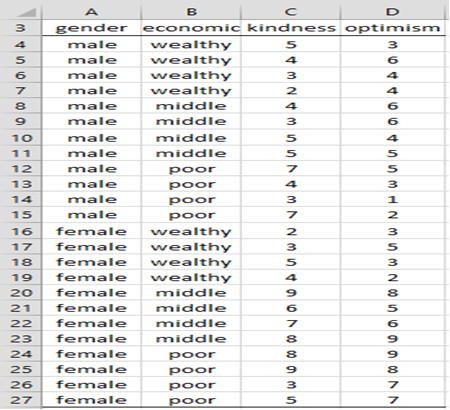


Select the following Add-Ins. Click OK.

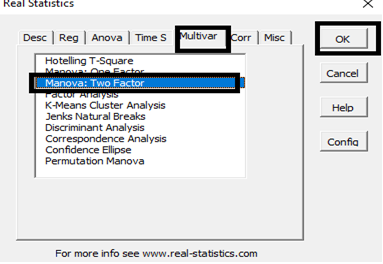


Now create an excel sheet with following data.

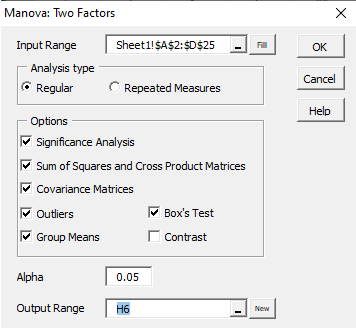
A study was conducted to see the impact of social-economic class (rich, middle, poor) and gender (male, female) on kindness and optimism using on a sample of 24 people based on the data in Figure 1.



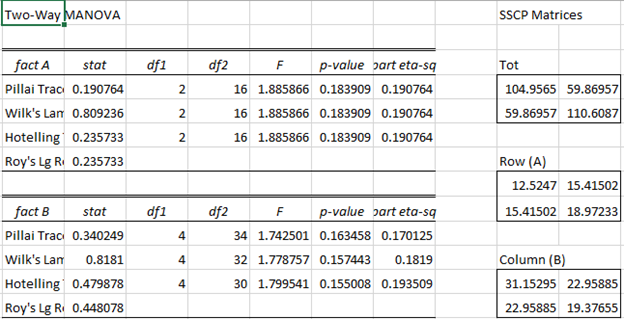
Press ctrl-m to open Real Statistics menu.



Select the data excluding column names. Select a cell for output.



**Output:**



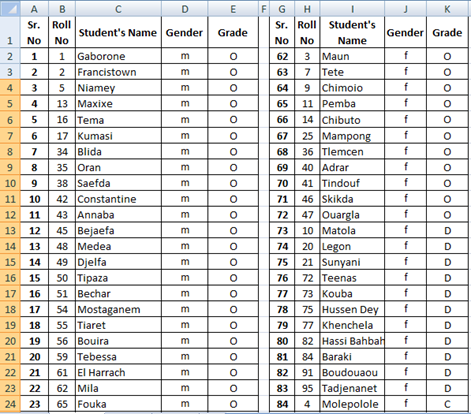
**Practical No. 07**

**Aim: Perform the Random sampling for the given data and analyze it.**

Example 1: From a population of 10 women and 10 men as given in the table in Figure 1 on the left below, create a random sample of 6 people for Group 1 and a periodic sample consisting of every 3rd woman for Group 2.

You need to run the sampling data analysis tool twice, once to create Group 1 and again to create Group 2. For Group 1 you select all 20 population cells as the Input Range and Random as the Sampling Method with 6 for the Random Number of Samples. For Group 2 you select the 10 cells in the Women column as Input Range and Periodic with Period 3.

Open existing excel sheet with population data Sample Sheet looks as given below:



Set Cell O1 = Male and Cell O2 = Female

To generate a random sample for male students from given population go to Cell O1 and type

=INDEX(E$2:E$62,RANK(B2,B$2:B$62))

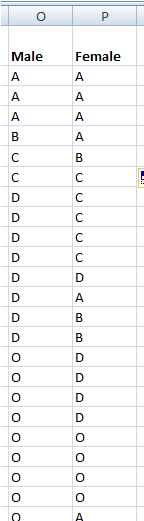
Drag the formula to the desired no of cell to select random sample.

Now, to generate a random sample for female students go to cell P1 and type

=INDEX(K$2:K$40,RANK(H2,H$2:H$40))

Drag the formula to the desired no of cell to select random sample.

**Output:**



**Practical No. 08**

**Aim: Write a program for computing different correlation.**

1. **Positive Correlation:**

Let’s take a look at a positive correlation. Numpy implements a corrcoef() function that returns a matrix of correlations of x with x, x with y, y with x and y with y. We’re interested in the values of correlation of x with y (so position (1, 0) or (0, 1)).

**Code:**

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

# 1000 random integers between 0 and 50

x = np.random.randint(0, 50, 1000)

# Positive Correlation with some noise

y = x + np.random.normal(0, 10, 1000)

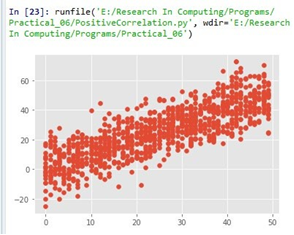
np.corrcoef(x, y)

matplotlib.style.use('ggplot')

plt.scatter(x, y)

plt.show()

**Output:**



1. **Negative Correlation**

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

#1000 random integers between 0 and 50

x = np.random.randint(0, 50, 1000)

# Negative Correlation with some noise

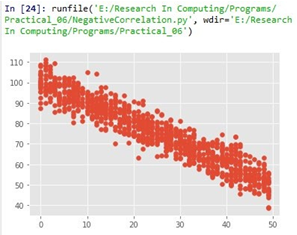
y = 100 - x + np.random.normal(0, 5, 1000)

np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()

**Output:**



1. **No/Weak Correlation:**

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(1)

x = np.random.randint(0, 50, 1000)

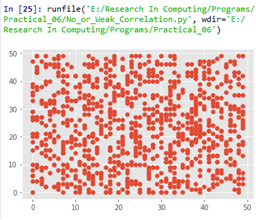
y = np.random.randint(0, 50, 1000)

np.corrcoef(x, y)

plt.scatter(x, y)

plt.show()

**Output:**

****